

STATA para Ciencias Sociales y Gestión Pública

Sesión 1: Introducción y procesamiento de datos

César Mora Ruiz

QLAB - PUCP

Junio de 2022

Temas a abordar

Esta sesión brindará una rápida introducción al software STATA, y presentará una serie de comandos de uso general, así como la implementación de procesos para el procesamiento de datos.

- Presentación de STATA como software para análisis de datos e investigación
- Navegando entre las ventanas
- El uso de Do-files
- Help files
- Configuración
- Importación de bases de datos
- Plataforma Nacional de datos abiertos
- Exploración de datos
- Distribución de los datos
- Vista de observaciones
- Ordenando información
- Creación de variables
- Resúmenes con tablas
- Guardado, preservación y restauración de bases

¿Qué es STATA?

- Es un paquete muy amigable, pero poderoso, para realizar análisis de datos con fuerte potencial para:
 - Análisis estadístico
 - Manipulación y gestión de datos
 - Visualización de datos al detalle
- STATA nos ofrece muchas herramientas que implementan métodos analíticos y econométricos de uso estándar, así como métodos nuevos y avanzados que se van incorporando como parte de los nuevos lanzamientos del software.

STATA - Ventajas

- La sintaxis de los comandos es muy sencilla, corta e intuitiva
- Dicha sintaxis es muy consistente cuando se utilizan diversos comandos al mismo tiempo, por lo que es más sencilla de comprender y aprender
- Un paquete muy competitivo ya que cuenta con diversidad de métodos
- Amplia documentación existente a nivel de libros y en la web
- Cuenta con componentes ideales para realizar análisis estadístico, econométrico y de encuestas de diseño complejo.

STATA - Desventajas

- Solo puede cargar una base de datos en cada sesión
- Para trabajar con más bases de datos al mismo tiempo es necesario abrir otras sesiones de Stata
- Menor número de funcionalidades escritas por los mismos usuarios

Navegando entre las ventanas de STATA

Ventana de comandos

- En esta ventana escribimos directamente los comandos .
- Escribe: `webuse auto, clear` , y fíjate qué sucede



Ventana de variables

- Cuando tenemos datos cargados, esta ventana nos muestra el listado de variables que contiene la base con sus respectivas etiquetas.
- Al seleccionar una variable, aparecerá su información asociada en la **ventana de propiedades**.
- Doble click en la variable causa que esta aparezca en la ventana de comandos

The screenshot displays the Stata software interface. The main window shows the Stata logo and version information (15.1). The Command window at the bottom contains the following text:

```
(R)
sis
15.1 Copyright 1985-2017 StataCorp LLC
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)

tial license:
301506219756
Andy Lin
IDRE UCLA

ported; see help unicode_advice.
ailable; type -update all-
```

The Variables window (top right) lists the following variables and their labels:

Name	Label
make	Make and Model
price	Price
mpg	Mileage (mpg)
rep78	Repair Record 1978
headroom	Headroom (in.)
trunk	Trunk space (cu. ft.)
weight	Weight (lbs.)
length	Length (in.)
turn	Turn Circle (ft.)
displacement	Displacement (cu....
gear_ratio	Gear Ratio
foreign	Car type

The Properties window (bottom right) shows the details for the selected variable 'make':

Variables	
Name	make
Label	Make and Model
Type	str18
Format	%-18s
Value label	
Notes	

The Data window (bottom right) shows the following information:

Data	
Filename	auto.dta
Label	1978 Automobile Data
Notes	
Variables	12
Observations	74
Size	3.11K
Memory	64M
Sorted by	foreign

A red circle highlights the Variables window, and a red arrow points from the 'make' variable in the Variables list to the Command window.

Ventana de propiedades

- Brinda información sobre cada una de las variables (cuando son seleccionadas de la lista), así como de toda la base de datos.
- Describe las características de la base de datos cargada

The screenshot shows the Stata Properties window with two panes: Variables and Properties. The Variables pane lists variables like make, price, mpg, etc. The Properties pane shows details for the selected variable 'make', including its name, label, type, format, and value label. Below this, the Data section provides summary statistics for the dataset, such as the number of observations (74) and the size of the file (3.11K). A red circle highlights the Properties pane, and a red arrow points from the main Stata window to it.

(R)

15.1 Copyright 1985-2017 StataCorp LLC
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC <http://www.stata.com>
979-696-4600 stata@stata.com
979-696-4601 (fax)

trial license:
301506219756
Andy Lin
IDRE UCLA

ported; see [help unicode_advice](#).
ailable; type `-update all-`

Variables

Filter variables here

Name	Label
make	Make and Model
price	Price
mpg	Mileage (mpg)
rep78	Repair Record 1978
headroom	Headroom (in.)
trunk	Trunk space (cu. ft.)
weight	Weight (lbs.)
length	Length (in.)
turn	Turn Circle (ft.)
displacement	Displacement (cu....
gear_ratio	Gear Ratio
foreign	Car type

Properties

Variables

Name	make
Label	Make and Model
Type	str18
Format	%-18s
Value label	
Notes	

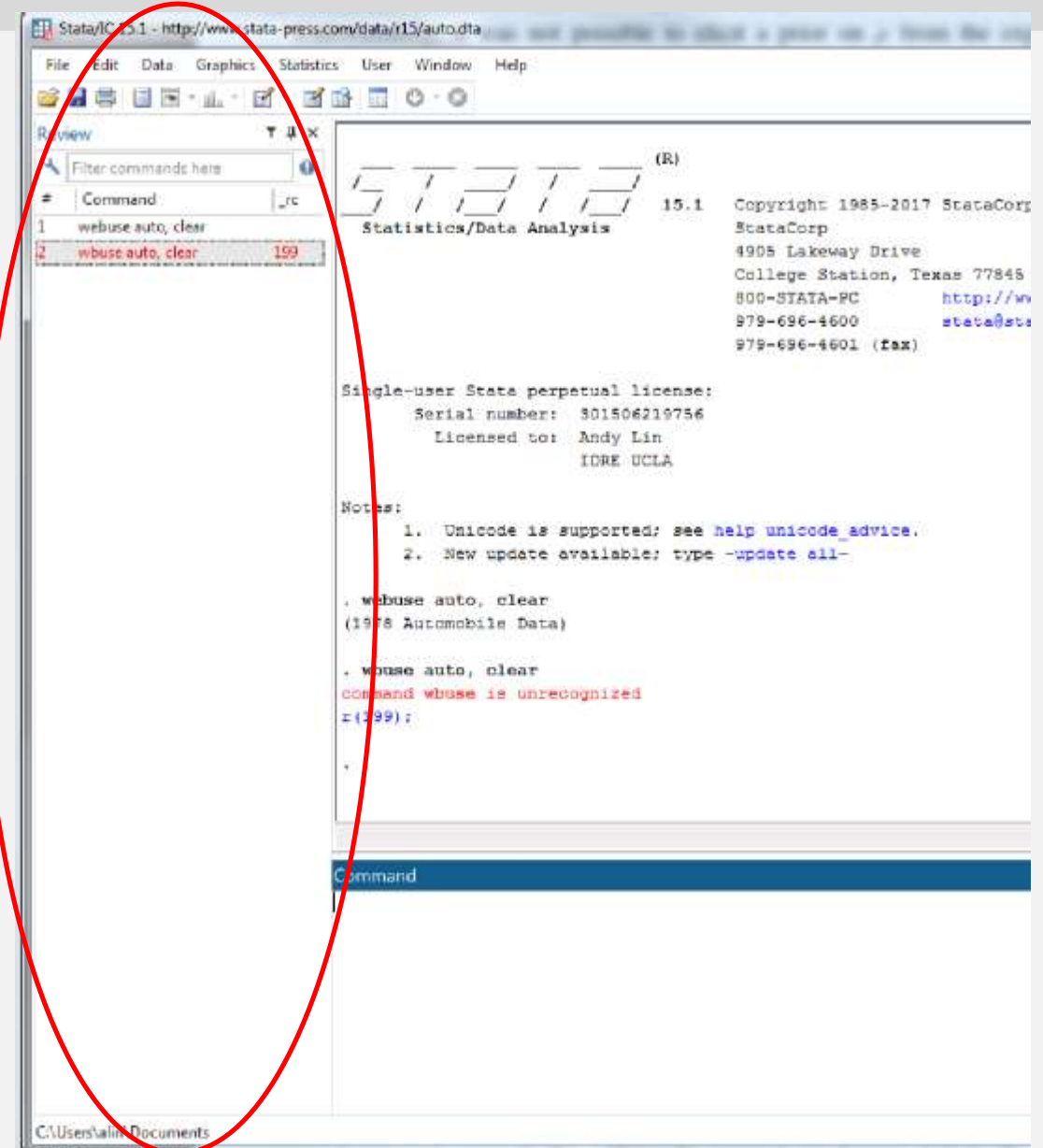
Data

Filename	auto.dta
Label	1978 Automobile Data
Notes	
Variables	12
Observations	74
Size	3.11K
Memory	64M
Sorted by	foreign

CAP NUM OVR

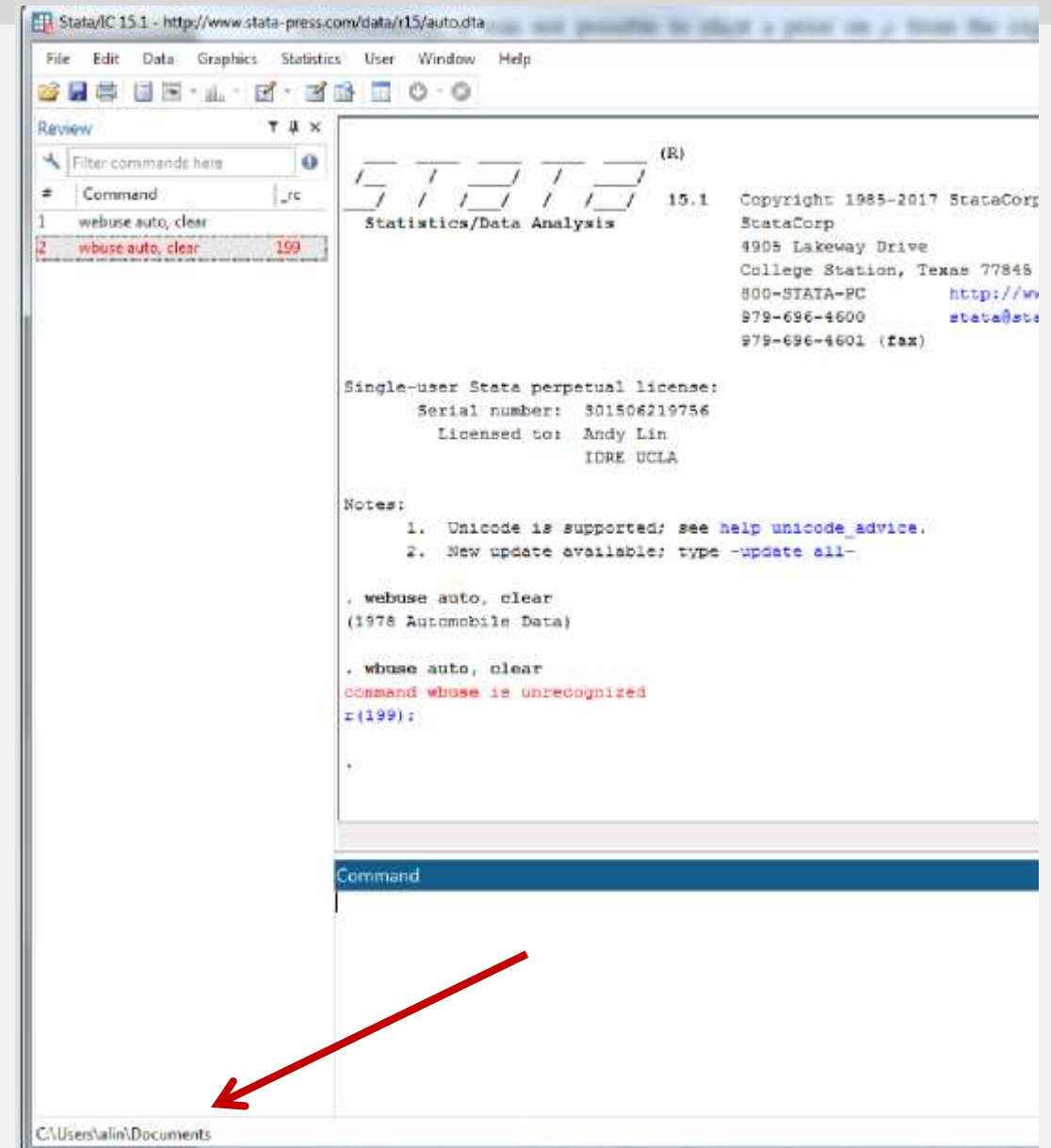
Ventana de revisión

- Esta ventana registra todas las líneas de comando ejecutadas desde la ventana de comandos, así como aquellas que son ejecutadas desde un “Do-file” (concepto a revisar más adelante)
- Los errores serán marcados en color rojo
- Doble click en la línea de comando para que aparezca nuevamente en la ventana de comando
- Si presiono la tecla “Repag” del teclado, aparecerá el último comando ejecutado



Directorio de trabajo

- El directorio de trabajo se muestra en la parte inferior izquierda de la ventana principal
- Los archivos serán cargados y guardados a partir de este directorio, a menos que se le especifique otro al programa
- Escribe “cd” (abreviatura de “current directory”) en la barra de comandos y observa el resultado
- Usando este comando se puede cambiar el directorio actual por otro



The screenshot shows the Stata/IC 15.1 interface. The top menu bar includes File, Edit, Data, Graphics, Statistics, User, Window, and Help. Below the menu bar is a toolbar with various icons. The main window is divided into two panes. The left pane, titled 'Review', contains a list of commands: 1. webuse auto, clear and 2. wbuse auto, clear. The right pane displays the Stata startup screen, which includes the Stata logo, version 15.1, copyright information (1985-2017 StataCorp), and a single-user perpetual license for Andy Lin at IDRE UCLA. Below the license information, there are notes about Unicode support and a new update available. The command window at the bottom shows the command '. webuse auto, clear' and the output '(1978 Automobile Data)'. A red arrow points to the status bar at the bottom left, which displays the current directory: C:\Users\alim\Documents.

```
Stata/IC 15.1 - http://www.stata-press.com/data/r15/auto.dta
File Edit Data Graphics Statistics User Window Help
Review
Filter commands here
# Command _rc
1 webuse auto, clear
2 wbuse auto, clear 199

(R)
STATA 15.1 Copyright 1985-2017 StataCorp
Statistics/Data Analysis StataCorp
4905 Lakeway Drive
College Station, Texas 77845
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)

Single-user Stata perpetual license:
Serial number: 301506219756
Licensed to: Andy Lin
IDRE UCLA

Notes:
1. Unicode is supported; see help unicode_advice.
2. New update available; type -update all-

. webuse auto, clear
(1978 Automobile Data)

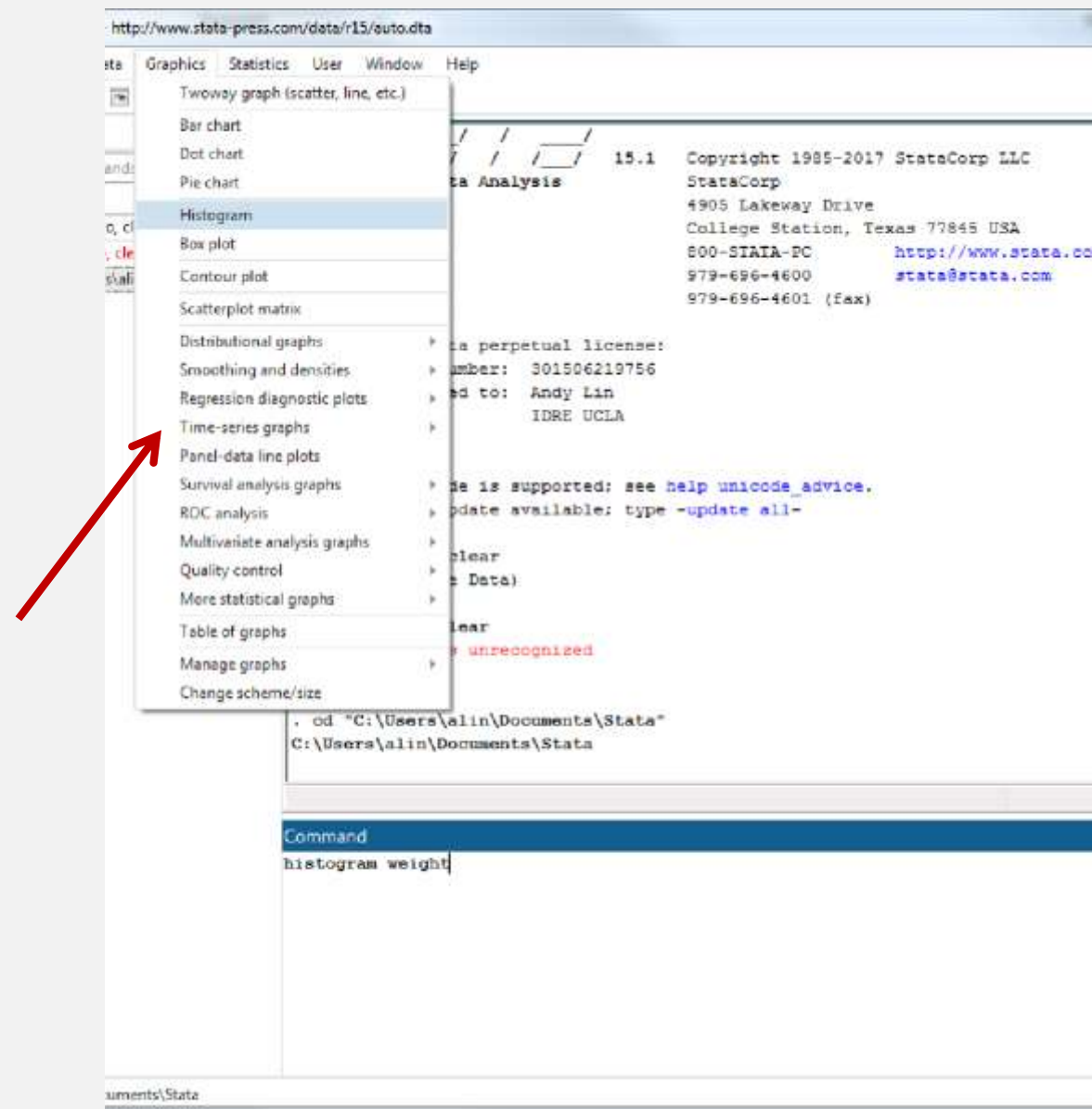
. wbuse auto, clear
command wbuse is unrecognized
r(199);

Command

C:\Users\alim\Documents
```

Trabajando con Menú

- STATA también puede ser utilizado a través de ventanas y menús para ejecutar funciones.
- No obstante, la gran mayoría de usuarios prefiere ejecutar las funciones a través de líneas de comando o de Do-files (scripts de comandos)



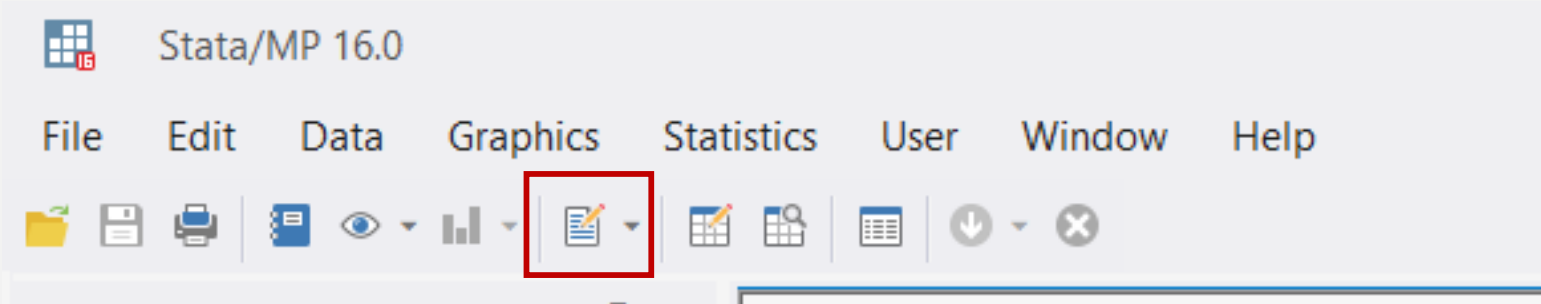
El uso de Do-files

Los Do-files

- Los Do-files son archivos de texto que almacenan una serie de comandos para ser reutilizados en el futuro, sin necesidad de tener que volver a escribirlos en la ventana de comandos.
- Brindan la ventaja de que son reproducibles, fáciles de ajustar y cambiar de acuerdo a nuestras necesidades
- Es altamente recomendable usar Do-files en vez de escribir solamente en la barra de comandos.
- Su extensión de archivo es **.do**

Abriendo el editor de Do-files:

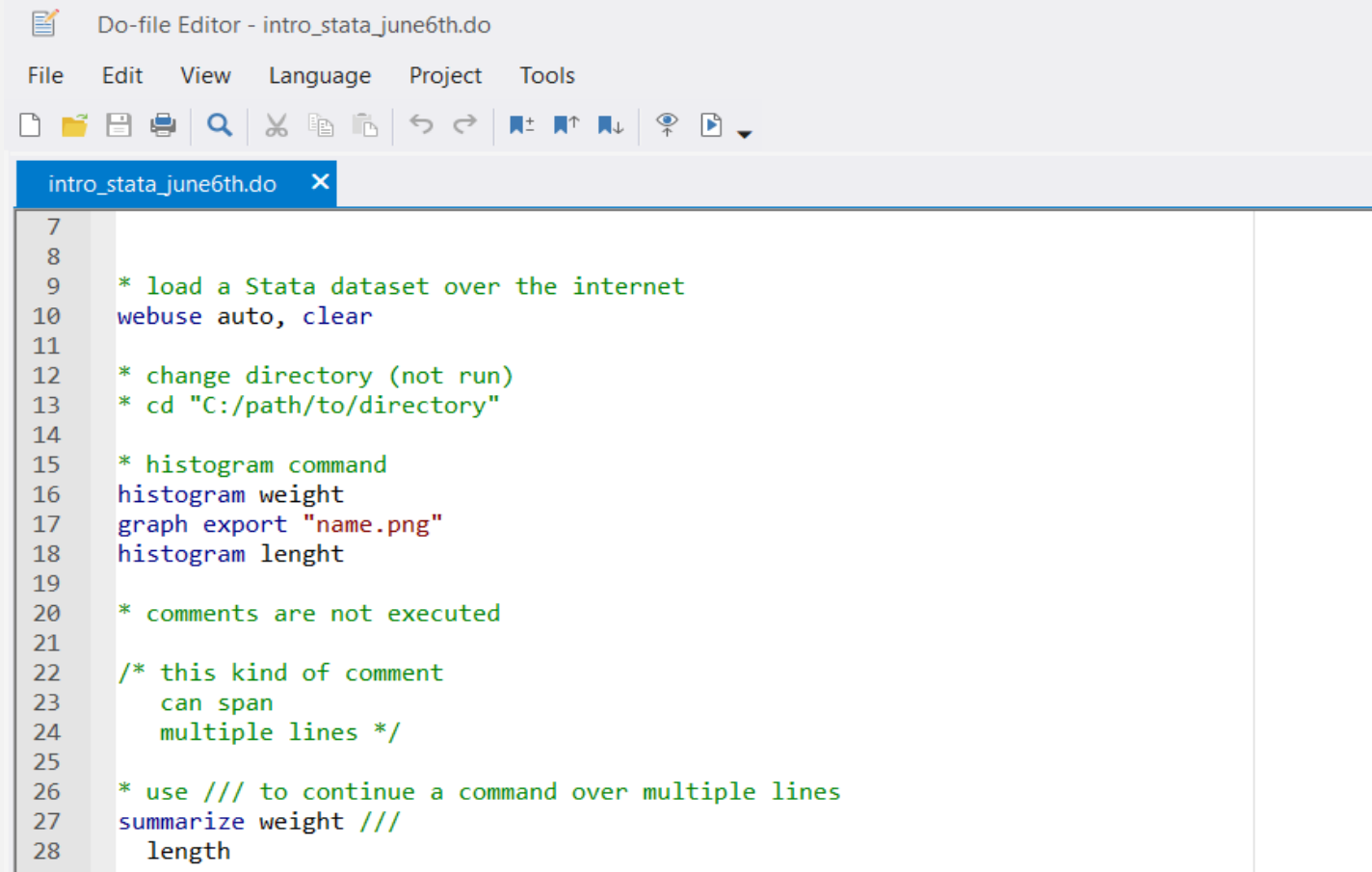
- Escribir el comando **doedit** en la barra de comandos, o hacer click en el icono de un lápiz y papel en la ventana principal de Stata:



Los Do-files

Colores de sintaxis:

- Comandos de Stata: azul
- Comentarios: verde, y deben ser precedidos por *
- Palabras en comillas (como nombres de archivos, valores strings) son de color rojo oscuro
- Stata 16 presenta la función de autocompletar comandos.



The screenshot shows the 'Do-file Editor - intro_stata_june6th.do' window. The menu bar includes File, Edit, View, Language, Project, and Tools. The toolbar contains icons for opening files, saving, printing, searching, and other editing functions. The editor window displays the following code with syntax highlighting: line 9 is a comment, line 10 is a command, line 12 is a comment, line 13 is a command with a string in quotes, line 15 is a comment, line 16 is a command, line 17 is a command with a string in quotes, line 18 is a command, line 20 is a comment, line 22 is a multi-line comment, line 24 is a comment, and line 27 is a command with a multi-line continuation using ///.

```
7
8
9  * load a Stata dataset over the internet
10 webuse auto, clear
11
12 * change directory (not run)
13 * cd "C:/path/to/directory"
14
15 * histogram command
16 histogram weight
17 graph export "name.png"
18 histogram length
19
20 * comments are not executed
21
22 /* this kind of comment
23    can span
24    multiple lines */
25
26 * use /// to continue a command over multiple lines
27 summarize weight ///
28 length
```

Los Do-files

Líneas extensas de comandos:

- Stata asume que una serie de comandos termina en cada línea de Do-file. No obstante, en algunas ocasiones requerimos escribir líneas de comando más extensas.
- Para escribir una serie de comandos que ocupe más de una línea, debemos usar el acompañamiento de múltiples líneas oblicuas: `///` al final de cada línea escrita
- Para ejecutar este conjunto extenso de comandos, es necesario seleccionar todas las líneas que lo componen:

* Se puede usar `///` al final de cada línea para que un comando pueda ser escrito en varias líneas y ser ejecutado sin problemas:

```
summarize weight length mpg price trunk  
summarize weight length ///  
    mpg price ///  
    trunk
```


Help files

Help files

- En la barra de comandos precede un nombre de comando por la palabra **help** y aparecerán recursos que brindan información sobre el uso de dicho comando y sus opciones, incluyendo ejemplos.
- Prueba: `help summarize`

```
Viewer - help summarize
File Edit History Help
help summarize
help summarize x
+ Dialog - Also see - Jump to -

[R] summarize — Summary statistics
      (View complete PDF manual entry)

Syntax

summarize [varlist] [if] [in] [weight] [, options]

options      Description
-----
Main
  detail      display additional statistics
  meanonly    suppress the display; calculate only the mean; programmer's option
  format      use variable's display format
  separator(#) draw separator line after every # variables; default is separator(5)
  display_options control spacing, line width, and base and empty cells

varlist may contain factor variables; see fvvarlist.
varlist may contain time-series operators; see tsvarlist.
by, rolling, and statsby are allowed; see prefix.

aweights, fweights, and iweights are allowed. However, iweights may not be used with the detail
option; see weight.
```

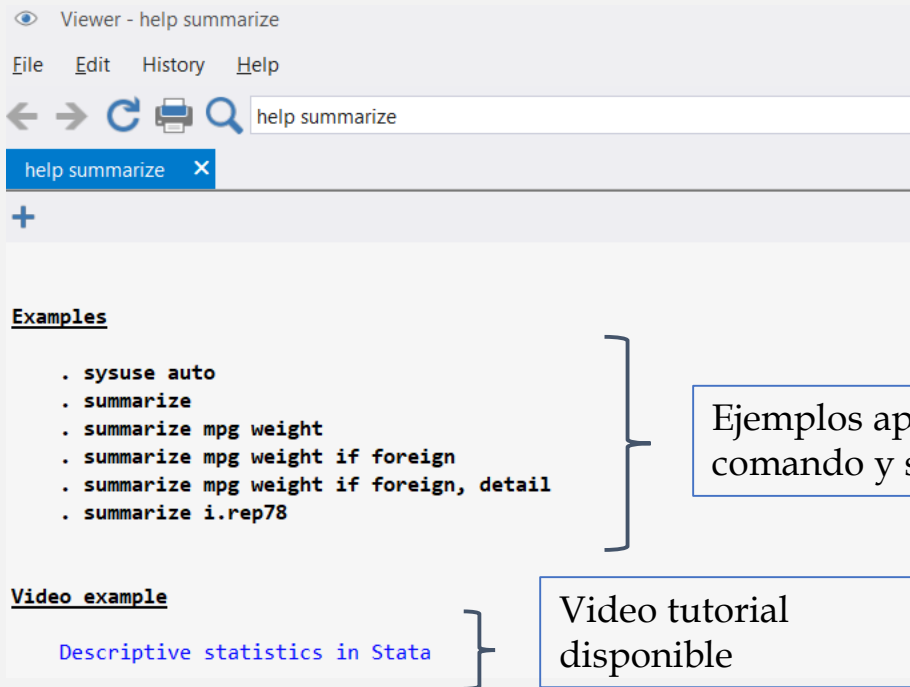
Información general
y sintaxis

Opciones y detalles
adicionales

Help files

Ejemplos:

- Los help files también brindan ejemplos aplicados con bases destinadas para aprendizaje
- En algunas ocasiones, incluso existen tutoriales en video para el comando requerido.



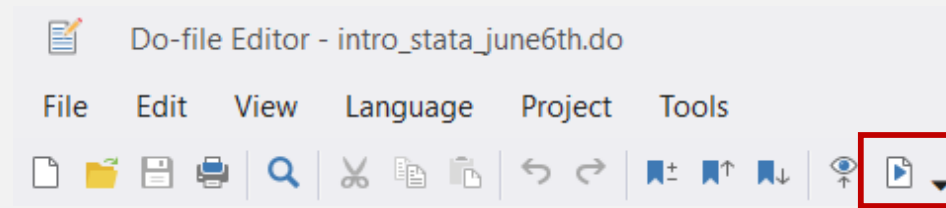
Se puede obtener más videotutoriales en el canal oficial de StataCorp en Youtube:

<https://www.youtube.com/user/statacorp>

Los Do-files

Correr Do-files:

- Seleccionar la línea (o líneas) que se desea correr, y presionar las teclas CTRL+D (SHIFT+CMD+D en Mac), o el ícono de ejecución en la ventana del editor de Do-files:



Añadir comentarios:

- Los comentarios no son ejecutados, y para insertarlos deben ser precedidos por un asterisco (*)
- Si se quiere añadir un comentario o texto en varias líneas, el texto completo debe estar encerrado entre `/* ... */`
- Cuando el texto aparezca en verde, entonces Stata no los ejecutará como comando.

```
/* Comentarios como este pueden ser escritos
en varias líneas
sin necesidad de colocar asteriscos en cada línea */

* Se puede usar /// al final de cada línea para que un comando pueda ser escrito en varias líneas y
ser ejecutado sin problemas:

summarize weight length mpg price trunk
summarize weight length ///
    mpg price ///
    trunk
```

Configuración

Configuración

Comandos a utilizar:

- **pwd:** inspecciona el directorio de trabajo
- **cd "...":** indica al programa la ruta del directorio de trabajo deseado
- **dir:** revisa los archivos en el directorio de trabajo
- **log using "****.log":** crea un archivo de texto que guarda los resultados obtenidos en la pantalla
- **log close:** cierra el archivo de texto
- **cls:** limpia el espacio de trabajo
- **ssc install *comando*:** instala comandos nuevos

Importación de bases de datos

Importación de datos

Archivos en formato dta

- El formato de bases de datos en Stata es **.dta**
- A diferencia de otros softwares estadísticos, Stata solo puede tener abierta una base de datos en simultáneo, aunque se pueden añadir nuevas bases a la cargada sin necesidad de abrirlas.

El comando use:

- Puede abrir archivos almacenados en el disco duro, pero también disponibles en Internet (a través de su respectiva dirección web)

```
/*-----  
-- Importación de bases de datos --  
-----*/  
  
* Cargando base en formato Stata desde el disco duro:  
use "hs0stata"
```


Importación de datos

Limpieza de memoria:

- Ya que Stata solo puede cargar una base de datos al mismo tiempo, es necesario limpiar la memoria cada vez que se requiera cargar una nueva base.
- Para ello, se debe usar el comando **clear** que remueve la base en uso de la memoria.

```
* Cargando base en formato Stata desde el disco duro:  
use "hs0stata"  
  
* Limpiando la memoria del programa (cuando ya se tiene cargada una base de datos)  
clear
```

Importación de datos

Importando archivos de Excel:

- Para importar archivos de formato Excel, se utiliza el comando **import excel using**
- Es necesario indicar el “path” del archivo Excel (y en general de cualquier otro formato) para abrir aquellas bases que no se encuentran en el actual directorio de trabajo.
- La opción **sheet()**, permite indicar si queremos trabajar con una hoja particular del libro de Excel.
- Además, usar la opción **firstrow** para ordenar a Stata que considere a la primera fila de la base de datos como los nombres de las columna.
- Incluso se puede indicar el rango de datos que se quiere abrir usando la opción **cellrange**

```
* Cargando una tabla de excel:
import excel using "hs0_excel.xlsx",sheet("Hoja2") firstrow clear

* Colocación de la opción "firstrow" para colocar encabezados:
import excel using "hs0_excel.xlsx",sheet("Hoja2") firstrow clear

* Opción para cargar solo una subtabla:
import excel using "hs0_excel.xlsx",sheet("Hoja1") cellrange(A1:F51) firstrow clear
```

Importación de datos

Importando archivos .csv

- Archivos Csv (comma-separated values) pueden ser abiertos usando el comando **import delimited using**
- La sintaxis es muy semejante a la usada para abrir archivos de Excel, pero en esta ocasión no se especifica una hoja de trabajo, porque los archivos .csv no las tienen.
- También existe la opción de cargar una submuestra de la tabla usando las opciones **colrange()** y **rowrange()**

```
* Cargando un csv:  
import delimited using "hs0.csv", clear  
  
* Opción para cargar solo una subtabla:  
import delimited using "hs0.csv", colrange(1:5) rowrange(1:20) clear
```

Importando archivos de SPSS:

- Se logra con el comando **import spss**

La Plataforma Nacional de Datos Abiertos



gob.pe

Plataforma Nacional de Datos Abiertos

Datos Abiertos

Marco de Gobernanza de Datos del Estado Peruano está constituido por instrumentos técnicos y normativos que establecen los requisitos mínimos que las entidades de la Administración Pública deben implementar conforme a su contexto legal, tecnológico y estratégico para asegurar un nivel básico y aceptable para la recopilación, procesamiento, publicación, almacenamiento y apertura de los datos que administre.

COVID-19

Tipos de contenido

 Recurso (4600)

 Dataset (2360)

 Entidades (78)

 Harvest Source (28)

 Página (3)

Categorías

 Economía y Finanzas (832)

7069 Distribución de Datos

Search

Ordenar por

Fecha cambiada

Pedido

Descendente

Consultar

Reiniciar



Situación de Naves

 Gobernabilidad

 Transporte

 Economía y Finanzas

Estado actual de las naves que hacen eso de los puertos administrado por la Empresa Nacional de Puertos.

Plataforma Nacional de datos abiertos

- Usaremos una base de datos de hospitalizados por COVID-19 disponible en dicha plataforma:
<https://www.datosabiertos.gob.pe/dataset/hospitalizados-vacunados-y-fallecidos-por-covid-19>

Datos Abiertos

Marco de Gobernanza de Datos del Estado Peruano está constituido por instrumentos técnicos y normativos que establecen los requisitos mínimos que las entidades de la Administración Pública deben implementar conforme a su contexto legal, tecnológico y estratégico para asegurar un nivel básico y aceptable para la recopilación, procesamiento, publicación, almacenamiento y apertura de los datos que administre.

COVID-19

[Home](#) / [Datasets](#) / Hospitalizados, vacunados y fallecidos por COVID-19

[Ver](#) [Revisiones](#)

Ministerio de Salud

Ministerio de Salud

Licencia

[Open Data Commons Attribution License](#)

OPEN DATA

Hospitalizados, vacunados y fallecidos por COVID-19

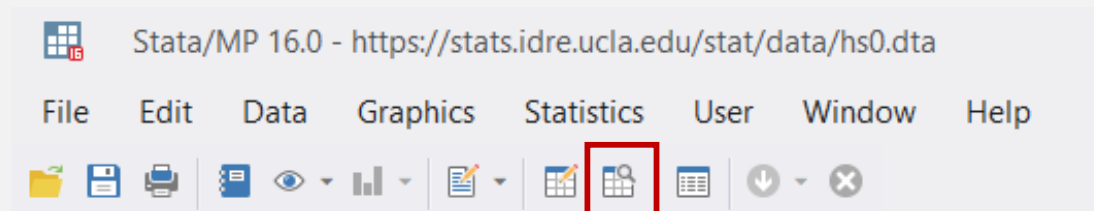
[COVID-19](#) [Salud](#)

Esta tabla toma como referencia el universo de hospitalizados de la f500 (en base al último registro de la fecha de hospitalización), vinculando información de dosis de vacunas y fallecimiento por COVID (obtenida por fallecimientos informados por el CDC).

Exploración de los datos

Exploración de datos

- Una vez que los datos han sido cargados, pueden ser visualizados como una tabla u hoja de cálculo usando el comando **browse**, o usando el correspondiente icono.



Data Editor (Browse) - [hs0.dta]

File Edit View Data Tools

gender[1] 1

	gender	id	race	ses	schtyp	prgtype	read	write	math	science	socst
1	1	70	white	low	1	general	57	52	41	47	57
2	2	121	white	middle	1	vocati	68	59	53	63	61
3	1	86	white	high	1	general	44	33	54	58	31
4	1	141	white	high	1	vocati	63	44	47	53	56
5	1	172	white	middle	1	academic	47	52	57	53	61
6	1	113	white	middle	1	academic	44	52	51	63	61
7	1	50	african-amer	middle	1	general	50	59	42	53	61
8	1	11	hispanic	middle	1	academic	34	46	45	39	36
9	1	84	white	middle	1	general	63	57	54	.	51

- Columnas negras: datos numéricos
- Columnas rojas: datos tipo "string" (texto)
- Columnas azules: datos numéricos pero etiquetados.

Exploración de datos

- El comando **ds** nos brinda el listado de variables en la tabla cargada.
- Con la opción **has(type...)** podemos indicar qué tipo de variables queremos visualizar

Table 2.1. Stata's numeric storage types

Storage type	Bytes	Minimum	Maximum
byte	1	-127	100
int	2	-32,767	32,740
long	4	-2,147,483,647	2,147,483,620
float	4	$-1.70141173319 \times 10^{38}$	$1.70141173319 \times 10^{38}$
double	8	$-8.9984656743 \times 10^{307}$	$8.9984656743 \times 10^{307}$

variable name	storage type	display format
ieess_renaes	int	%8.0g
eess_diresa	str17	%17s
eess_red	str26	%26s
eess_nombre	str80	%80s
id_eess	int	%8.0g
id_persona	long	%12.0g
edad	int	%8.0g
sexo	str1	%9s
fecha_ingreso	str10	%10s

```
.      * Listado de variables (solo string):
.      ds, has(type string)
eess_diresa  sexo      fecha_ingr~n  fecha_dosis1  fabricante~2  cdc_fecha_~d  dist_domic~o
eess_red     fecha_ingr~p  fecha_segu~o  fabricante~1  fecha_dosis3  dep_domici~o
eess_nombre  fecha_ingr~i  evolucion_~o  fecha_dosis2  fabricante~3  prov_domic~o

.      * Listado de variables (solo integer):
.      ds, has(type int)
ieess_renaes  id_eess      edad
```

Exploración de datos

Otros comandos útiles son:

- **describe:** presenta el listado de variables y sus características con más detalle
- **lookfor:** permite buscar palabras en la información cargada (en forma de nombre o etiqueta)
- **isid:** analiza si cada caso de una variable identifica a una observación o a más.

Distribución de los datos

Distribución de datos

El commando `codebook`:

- Brinda un resumen de las variables, incluyendo: número de valores únicos y missings, rangos, quintiles, media, desviación estándar para el caso de variables numéricas.
- Si las variables son de tipo “string” nos brindará información sobre frecuencias

eess_diresa

(unlabeled)

```
type: string (str17)
unique values: 29          missing "": 0/143,399
examples: "CALLAO"
           "JUNIN"
           "LIMA DIRIS ESTE"
           "LIMA DIRIS SUR"
warning: variable has embedded blanks
```

edad

(unlabeled)

```
type: numeric (int)
range: [0,111]          units: 1
unique values: 110      missing .: 9,763/143,399
mean: 51.7484
std. dev: 20.6939
percentiles:    10%    25%    50%    75%    90%
                24     36     53     67     79
```

Distribución de datos

Resumen estadístico con summarize:

- Información sobre el número de observaciones no missing, media, desviación estándar, mínimo y máximo

Summarize con opción detail:

- Agrega más detalles como valores en percentiles, valores más altos y bajos, varianza, skewness y kurtosis

```
. summ edad flag_uci
```

Variable	Obs	Mean	Std. Dev.	Min	Max
edad	133,636	51.74844	20.69394	0	111
flag_uci	143,399	.1331739	.3397637	0	1

```
.  
end of do-file
```

```
. do "C:\Users\cesar\AppData\Local\Temp\STD3854_000000.tmp"
```

```
.      sum edad,detail
```

edad				
			Percentiles	Smallest
1%	4	0		
5%	18	0		
10%	24	0		
25%	36	0		
			Obs	133,636
			Sum of Wgt.	133,636
50%	53		Mean	51.74844
			Std. Dev.	20.69394
		Largest		
75%	67	107		
90%	79	107	Variance	428.2391
95%	84	109	Skewness	-.1736396
99%	92	111	Kurtosis	2.388225

Inspección con un histograma básico

- Para ello usamos el comando **inspect** indicando la variable deseada

```
. * Inspección con un histograma básico:  
. inspect edad
```

edad:		Number of Observations		
		Total	Integers	Nonintegers
#	Negative	-	-	-
#	Zero	56	56	-
# #	Positive	133,580	133,580	-
# # #				
# # #	Total	133,636	133,636	-
# # # # .	Missing	9,763		
0		143,399		
111				
(More than 99 unique values)				

```
. inspect flag_vacuna
```

flag_vacuna:		Number of Observations		
		Total	Integers	Nonintegers
#	Negative	-	-	-
#	Zero	53,865	53,865	-
#	Positive	89,534	89,534	-
#				
#	Total	143,399	143,399	-
# . # #	Missing	-		
0		143,399		
3				
(4 unique values)				

Vista de observaciones

Vista de observaciones

- **display var [**]**: muestra el valor de la variable en la posición señalada en []
- **levelsof var**: muestra la lista de valores únicos para una variable

```
. * Muestra de valores únicos para una variable:
. levelsof edad
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 3
> 8 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
> 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 1
> 05 106 107 109 111

. levelsof eess_diresa
`"AMAZONAS"' ` "ANCASH"' ` "APURIMAC"' ` "AREQUIPA"' ` "AYACUCHO"' ` "CAJAMARCA"' ` "CALLAO"' ` "CUSCO"' ` "HUANC
> AVELICA"' ` "HUANUCO"' ` "ICA"' ` "JUNIN"' ` "LA LIBERTAD"' ` "LAMBAYEQUE"' ` "LIMA DIRIS CENTRO"' ` "LIMA DIR
> IS ESTE"' ` "LIMA DIRIS NORTE"' ` "LIMA DIRIS SUR"' ` "LIMA PROVINCIAS"' ` "LORETO"' ` "MADRE DE DIOS"' ` "MO
> QUEGUA"' ` "PASCO"' ` "PIURA"' ` "PUNO"' ` "SAN MARTIN"' ` "TACNA"' ` "TUMBES"' ` "UCAYALI"'
```


Vista de observaciones

- El comando **list** nos permite visualizar los datos en la pantalla principal de Stata
- Es recomendable especificar el nombre de las variables de interés luego del comando para que la pantalla no se llene de mucha información

	LIMA	LIMA	EL AGUSTINO	.	
	empresa2 .	empresa3 .	_Icon_~1 1	_Iflag~1 0	_IconX~1 0

5275.	ieess_~s 753	eess_diresa JUNIN	eess_red NO PERTENECE A NINGUNA RED				
eess_nombre HOSPITAL REGIONAL DOCENTE CLINICO QUIRURGICO DANIEL ALCIDES CARRION				id_eess 753			
id_per~a 36224391	edad 90	sexo M	fecha_in~p 21/02/2021	flag_uci 0	fecha_in~i	fecha_in~n	con_ox~o 1
con_ve~n 0	fecha_se~o 24/02/2021	evolucion_hos~o defuncion		flag_v~a 0	fecha_do~1	fabricant~1	fecha_do~2
fabricant~2	fecha_do~3	fabricant~3	cdc_po~d 1	cdc_fech~d 23/02/2021	cdc_fa~d 1	ubigeo~o 120101	
dep_domicilio JUNIN	prov_domicilio HUANCAYO		dist_domicilio HUANCAYO			empresa1 .	
empresa2 .	empresa3 .	_Icon_~1 1	_Iflag~1 0	_IconX~1 0			

5276.	ieess_~s 23159	eess_diresa LIMA DIRIS SUR	eess_red NO PERTENECE A NINGUNA RED		
-------	-------------------	-------------------------------	--	--	--

Vista muy cargada y desordenada

Vista de observaciones

Opción in:

- La opción **in** selecciona la fila de observaciones que queremos visualizar, indicando la primera y última observación de interés (ver ejemplo)
- También se puede usar la opción en negativo, y la especificación “L”, particularmente cuando queremos visualizar la información de las últimas filas.
- **Ejercicio:** visualiza las últimas 10 observaciones para tres variables de tu elección.

```
. list id_persona edad sexo flag_uci flag_vacuna in 1/5
```

	id_persona	edad	sexo	flag_uci	flag_vacuna
1.	23822739	57	F	0	3
2.	38730191	.		0	0
3.	21982368	67	F	0	0
4.	36217805	.		0	0
5.	36217873	49	F	1	0

```
. * Lista de edad para las últimas 5 observaciones:  
. li edad in -5/L
```

	edad
143395.	20
143396.	18
143397.	20
143398.	80
143399.	45

Vista de observaciones

Opción if:

- La opción **if** permite seleccionar un subconjunto de información de acuerdo a una serie de condiciones.
- Esta opción se especifica luego del comando, pero antes de la coma.

Veamos más opciones de comandos relacionales:

- ✓ Igual a: ==
- ✓ Mayor que: >
- ✓ Mayor o igual que: >=
- ✓ Menor que: <
- ✓ Menor o igual que: <=
- ✓ No: !
- ✓ No es igual que: !=
- ✓ Y: &
- ✓ O: |

```
. list sexo edad cdc_fallecido if edad>65 in 1/50
```

	sexo	edad	cdc_fa~d
2.		.	0
3.	F	67	1
4.		.	0
6.	M	85	1
7.	M	.	1
8.	M	72	1
9.	M	75	1
11.		.	0
12.	M	88	0
14.		.	0
16.	F	91	0
17.	M	84	1
18.		.	0
19.	F	.	0
21.	F	83	0
22.	M	83	1
23.		.	0
24.		.	0
25.		.	0
27.	M	.	1

Vista de observaciones

Observaciones duplicadas:

- El comando **duplicates report** nos brinda un reporte de las observaciones duplicadas en la tabla cargada, teniendo en cuenta valores repetidos en todas las variables.
- Para eliminar dichos casos, simplemente aplicamos **duplicates drop**

```
. * Informe de observaciones duplicadas:  
. duplicates report
```

Duplicates in terms of all variables

copies	observations	surplus
1	142460	0
2	930	465
3	9	6

```
. display 465+6  
471
```

```
.  
. * Eliminando los duplicados:  
. duplicates drop
```

Duplicates in terms of all variables

(471 observations deleted)

Ordenando información

Ordenando información

- La mayoría de las veces, la información en las tablas cargadas no se encuentra ordenada por ningún criterio particular
- STATA cuenta con comandos para ordenar los datos de manera ascendente (**sort**) o descendente (**gsort**)

Ascendente por id_persona

```
. sort id_persona edad  
  
. list id_persona edad sexo in 1/10
```

	id_persona	edad	sexo
1.	11	58	F
2.	100	64	F
3.	172	85	M
4.	221	62	M
5.	242	68	F
6.	297	63	F
7.	341	85	F
8.	408	84	M
9.	811	83	M
10.	864	68	M

Descendente por edad

```
gsort -edad  
  
list id_persona edad sexo in 1/10
```

	id_persona	edad	sexo
1.	5928246	111	F
2.	5261527	109	F
3.	12070722	107	F
4.	20480188	107	F
5.	12559013	106	M
6.	15973464	106	F
7.	2524751	105	M
8.	34876844	104	M
9.	131803	103	F
10.	20050511	103	F

Creación de variables

Creación de variables

- A menudo es necesario crear nuevas variables para obtener la información que necesitamos.
- Usamos el comando **generate (gen)** para realizar transformaciones y operaciones entre variables.
- Existen algunos atajos para crear variables binarias, índices e indicadores de número de observaciones con información y/o por grupo (ver ejemplos en Do-file)

Creación de variables

Los missing values:

- Los missing values de una variable numérica son representados por “.”
- En el caso de variables string, los missing values son “” (comillas vacías)
- Los valores missing son obviados de las operaciones y los análisis numéricos de manera automática

```
. list edad sexo flag_vacuna in -10/L if edad==.
```

	edad	sexo	flag_v~a
142919.	.	M	0
142920.	.		0
142921.	.		0
142922.	.		0
142923.	.		0
142924.	.		0
142925.	.		0
142926.	.		0
142927.	.		0
142928.	.		0

Creación de variables

Creación de variables extendida

- Con el comando **egen** se puede crear variables haciendo uso de un abanico más extenso de funciones incluyendo cálculos estadísticos, estandarizaciones, entre otras de una manera más rápida y eficiente.
- Además, si alguna variable tiene valor missing, no es tomada en cuenta para hacer el cálculo requerido (ver ejemplo en Do-file)

```
. codebook suma_prueba suma_prueba2

suma_prueba (unlabeled)

  type: numeric (float)
  range: [0,112]
  unique values: 110
  mean: 53.5474
  std. dev: 20.4655
  missing: 9,744/142,928
  percentiles: 10% 27 25% 38 50% 55 75% 69 90% 80

suma_prueba2 (unlabeled)

  type: numeric (float)
  range: [0,112]
  unique values: 110
  mean: 49.9008
  std. dev: 23.9177
  missing: 0/142,928
  percentiles: 10% 14 25% 34 50% 53 75% 68 90% 79
```

Creada con suma simple
(tiene missings)

Creada con rowtotal
(sin missings)

Resúmenes con tablas

Resúmenes con tablas

Tabulaciones con tabulate (tab):

- Muestra información sobre frecuencias de valores.
- Particularmente útil para variables categóricas.
- En el caso de variables etiquetadas, dichas etiquetas se remueven con la opción nolabel

```
. tabulate eess_diresa
```

eess_diresa	Freq.	Percent	Cum.
AMAZONAS	1,838	1.29	1.29
ANCASH	12,895	9.02	10.31
APURIMAC	2,742	1.92	12.23
AREQUIPA	27	0.02	12.25
AYACUCHO	183	0.13	12.37
CAJAMARCA	9,251	6.47	18.85
CALLAO	3,811	2.67	21.51
CUSCO	11,798	8.25	29.77
HUANCAVELICA	1,833	1.28	31.05

Tabulaciones de doble entrada:

- Muestra frecuencias en tablas cruzadas. Solo hace falta indicar dos variables luego de tabulate
- Se puede incorporar información porcentual de filas y columnas usando las opciones **row** y **col** respectivamente.

```
. tab evolucion_hosp sexo
```

evolucion_hosp_ultimo	sexo		Total
	F	M	
alta	47,221	43,657	90,878
alta_voluntaria	1,102	1,223	2,325
defuncion	12,622	22,602	35,224
desfavorable	193	324	517
estacionario	195	274	469
favorable	62	60	122
referido	1,327	1,632	2,959
Total	62,722	69,772	132,494

```
. tab evolucion_hosp sexo, col row
```

Key			
frequency			
row percentage			
column percentage			
evolucion_hosp_ultimo	sexo		Total
	F	M	
alta	47,221	43,657	90,878
	51.96	48.04	100.00
	75.29	62.57	68.59
alta_voluntaria	1,102	1,223	2,325
	47.40	52.60	100.00
	1.76	1.75	1.75

Resúmenes con tablas

Comand tabstat:

- Este commando permite incluir indicadores estadísticos tales como promedio, desviación estándar, etc. para variables seleccionadas.

```
. tabstat flag_vacuna edad, by(evolucion_hosp) stat(mean sd)
```

Summary statistics: mean, sd

by categories of: evolucion_hosp_ultimo

evolucion_hosp_ultimo	flag_v~a	edad
alta	2.202343	46.36416
	1.155325	20.10264
alta_voluntaria	1.895859	53.17558
	1.270698	21.93738
defuncion	.1593579	65.73577
	.5643583	15.01809

Tabulaciones flexibles (table):

- Las tablas son más flexibles, permitiendo la elaboración de una tabla más amplia, con indicadores asociados a terceras variables.
- Es posible obtener datos agregados a nivel de columna y/o fila con las opciones **col** y **row** respectivamente

```
. table flag_uci sexo, content(mean edad) col row
```

flag_uci	sexo		
	F	M	Total
0	47.5565	55.2859	51.4885
1	52.5749	54.5916	53.8778
Total	48.0499	55.1739	51.8026

Resúmenes con tablas

Tablas multinivel

- El commando **table** tiene opciones que permiten obtener subcategorías en filas y columnas para mostrar las respectivas frecuencias.
- Para ello se deben indicar las filas/columnas entre paréntesis separados (ver ejemplos en Do-file)
- Se puede incluir, además, información sobre una tercera variable en la misma table cruzada, como por ejemplo el promedio de edad, u otra variable continua.

```
. table (evolucion_hosp) (flag_uci sexo), content(mean edad) format(%9.2f)
```

evolucion_hosp_ultimo	sexo and flag_uci			
	F		M	
	0	1	0	1
alta	42.65	47.72	50.04	49.96
alta_voluntaria	51.85	48.96	54.84	52.14
defuncion	67.99	60.95	66.65	60.16
desfavorable	60.78	57.09	60.74	52.87
estacionario	55.42	57.37	57.89	58.84
favorable	53.88	42.50	55.58	51.60
referido	45.86	49.97	55.89	52.89

Resúmenes con tablas

Colapso de tablas para ser guardadas como datasets en diferentes formatos:

- El commando **collapse** con sus diversas opciones permite obtener tablas resumen con información estadísticas (conteo de observaciones, promedio, desviación estándar, entre otros), separando dichos indicadores por categorías (ver ejemplo en Do-file)
- La table resultante, puede ser guardada en diversos formatos

eess_diresa[1]		AMAZONAS		
eess_diresa	id_persona	edad	flag_vacuna	
AMAZONAS	1838	53.5479	1.78237	
ANCASH	12895	53.5706	1.78666	
APURIMAC	2742	55.2139	1.89752	
AREQUIPA	27	57.4444	1.66667	
AYACUCHO	183	53.0769	2.30601	
CAJAMARCA	9251	52.911	1.77332	
CALLAO	3811	58.8069	.994227	
CUSCO	11798	51.707	1.9082	
HUANCAVELICA	1833	50.3213	2.02946	
HUANUCO	1646	54.3134	2.1695	
ICA	6914	56.363	1.63205	
JUNIN	8198	54.3734	1.74128	
LA LIBERTAD	1	68	3	

Guardado, preservación y restauración de bases

Guardado, preservación y restauración de bases

- El comando **save** guarda las bases de datos trabajadas en el formato **dta**, el cual es el más eficiente para trabajar en Stata.
- Si se quiere sobrescribir una base existente, y que ha sido modificada, usar la opción **replace**
- Tener en cuenta que en la línea de comandos, la extensión **dta** para guardar los archivos puede ser omitida, ya que Stata la entiende de manera predeterminada.
- Los archivos se guardan en el “path” de la sesión (¡identifícalo!)
- Existe la posibilidad de guardar las bases como **Excel** o **csv** usando las opciones correspondientes:

Guardado de la tabla como dataset en diversos formatos:

* En formato Stata:

```
save "data_trabajada",replace
```

* En formato Excel:

```
export excel "data_trabajada.xlsx", firstrow (variables) replace
```

```
export excel "data_trabajada_.xlsx", firstrow (variables) sheet(mi_hoja) replace
```

* En formato csv:

```
export delimited "data_trabajada.csv", delimiter (",") replace
```

Guardado, preservación y restauración de bases

- En muchas ocasiones, hemos realizado diversos cambios a las bases de datos con las que venimos trabajando, y podemos cometer errores. En dichos casos, **no podemos deshacer** los cambios hechos y debemos volver a cargar la base original y trabajarla, lo que resulta ineficiente
- Algo similar ocurre cuando aplicamos un **collapse** u otras transformaciones que reducen el número de observaciones o las unidades de análisis.
- Para realizar cambios, guardarlos en una nueva base de datos, y finalmente regresar a la base original con la que veníamos trabajando, podemos utilizar las opciones **preserve/restore**.

```
* Conservamos una base con solo 9 variables, y solo los casos que tengan información sobre edad:
keep eess_diresa id_persona edad sexo flag_uci con_oxigeno con_ventilacion evolucion_hosp flag_vacuna
drop if edad==.
browse

** Procederemos a obtener una tabla resumen, guardarla, y luego reestablecer la base con la que
estabamos trabajando

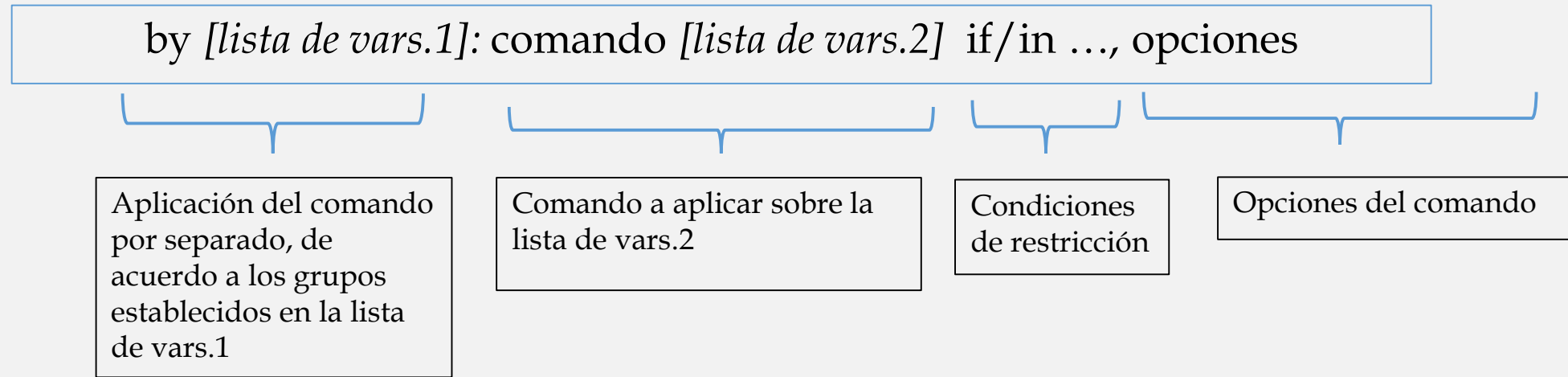
preserve
collapse (mean) edad flag_vacuna,by(eess_diresa sexo)
save "resumen2",replace
restore

browse

/* Confirmamos que la base resumen fue creada y guardada, y en STATA continua
cargada la base con la que estabamos trabajando antes de hacer el collapse */
```

Entre “preserve” y “restore” se escriben las operaciones que llevaremos a cabo, las cuales no serán tomadas en cuenta luego de la restauración.

Sintaxis general



Ejercicio aplicado

Ejercicio aplicado

Use la base de datos de instituciones educativas que reciben el Programa Qali Warma, cuyo link de acceso es:
<https://www.datosabiertos.gob.pe/sites/default/files/ListadoInstitucionesEducativasPublicas-2022-04-22.csv>

- Determina cuántos valores únicos tiene la variable “provincia”
- Contabiliza y elimina las observaciones duplicadas
- Obtener un tabulado de registros de instituciones educativas por departamento
- ¿Cuál es el promedio de usuarios (nrousuarios) en las escuelas de Cusco?
- Obtener una base de datos resumen que contenga una variable que indique el nivel de usuarios a nivel de departamento y provincia. Pista: usar el comando collapse, con la opción (*sum*) y el agrupamiento *by(departamento provincia)*
- Exportar la base de datos resumen obtenida en formato Excel, en una hoja llamada “Usuarios”, en la que claramente la primera fila indique el nombre de las variables.