Guía práctica para usar los datos de la Encuesta Longitudinal de Protección Social (ELPS)

Joaquín Prieto, 2015

OBSERVATORIO DE SEGURIDAD SOCIAL
DE AMERICA LATINA Y EL CARIBE

IDB
Inter-American
Development Bank

GUÍA PRÁCTICA PARA USAR LOS DATOS DE LA ENCUESTA LONGITUDINAL DE PROTECCIÓN SOCIAL (ELPS)

Joaquín Prieto[1][2]

17 de junio de 2015

Resumen: Este documento consiste en una guía práctica para trabajar datos longitudinales sobre seguridad social. Mediante el uso del programa estadístico Stata y bases de datos ficticias de cuatro rondas se describen los procedimientos para: i) preparar las bases de datos panel previo a cualquier análisis, ii) realizar las primeras inspecciones de las variables usando tablas y gráficos, iii) incluir los factores de expansión transversales y longitudinales en determinados análisis, y iv) realizar una serie de ejercicios que muestran análisis estáticos (corte transversal) y análisis dinámicos (longitudinales) en temas de seguridad social.

[1] Ingeniero Civil Industrial y Magíster en Economía Ambiental de la Universidad Católica de Chile. Candidato a Doctor en Políticas Sociales en la London School of Economics UK y profesor asociado al Departamento de Sociología de la universidad Alberto Hurtado. Actualmente es profesor visitante en el Programa Internacional de Derechos Humanos, Boston College US. jjprieto@uahurtado.cl; +1 617 775 7690; 205 Walden Street 4R, Cambridge, MA, 2140, USA.

[2] Se agradecen los comentarios recibidos por Lucía Madrigal en la elaboración del presente documento.

Contenidos

1. Introducción

Una encuesta longitudinal permite describir fenómenos que cambian en el tiempo, entender de mejor manera procesos socioeconómicos y comportamientos de las personas y, como resultado de esta información diseñar mejores instrumentos de política social Jenkins (2011). Entre las ventajas de los datos generados con este tipo de encuestas, comparado con los que se obtienen de las encuestas de corte transversal, están: i) al seguir a las mismas unidades en el tiempo, es posible medir los cambios de los individuos, y así estudiar las transiciones entre estados; ii) permite analizar el efecto de determinadas intervenciones de políticas sociales; iii) implementar modelos de comportamientos a través de técnicas de historias de vida, y iv) controlar el efecto de las variables no observadas a través del tiempo y entre las propias unidades (Rose, 2000).

En particular, una Encuesta Longitudinal de Protección Social (ELPS) es una herramienta de diagnóstico confiable que provee información estadística certera para evaluar el impacto de los programas de previsión y seguridad social. La ELPS permite construir una base de datos que recoge a lo largo del tiempo, y en un mismo instrumento, la historia laboral y previsional de los encuestados profundizando en áreas como educación, salud, seguridad social, capacitación laboral, patrimonio y activos, historia familiar e información del hogar.

No obstante lo anterior, las encuestas tipo panel presentan complicaciones adicionales a las de corte transversal. Entre ellas figuran la no respuesta o atrición entre ola y ola, que puede afectar la representatividad de la muestra, el condicionamiento del panel, es decir que las respuestas del encuestado estén influenciadas por la experiencia de la entrevista anterior, y una alta demanda de trabajo en el diseño, planificación e implementación en cada ronda (Lynn, 2009). Por otra parte, la incorporación en las bases de datos panel de la dimensión temporal y de varios factores de expansión para representar a la población en el tiempo (pesos longitudianles) y en cada ronda (pesos transversales) puede generar algunas dificultades iniciales para los usuarios de bases de datos que no han trabajado datos panel.

El presente documento consiste en una guía práctica para aquellos usuarios que desean trabajar por primera vez en datos longitudinales, en particular en temas de previsión y seguridad social. Mediante el uso del programa estadístico Stata y una base de datos longitudinal ficticia se ilustran, en dos capítulos, ejemplos y ejercicios que dan cuenta de las características de los datos longitudinales y su potencial analítico. En el capítulo 2 se entregan los procedimientos para preparar las bases de datos panel previo a cualquier análisis junto a las primeras inspecciones de variables usando tablas y gráficos. Además se explica conceptualmente la lógica de los distintos factores de expansión, y su uso en los análisis descriptivos. En el capítulo 3, se desarrollan una serie de ejercicios que tienen como objetivo mostrar algunos análisis estáticos (corte transversal) y principalmente profundizar en los análisis dinámicos (longitudinales) en temas de previsión y seguridad social.

2. Preparación de los datos en las encuestas longitudinales

2.1 UNA PRIMERA MIRADA DE DATOS PANEL DE PERSONAS USANDO STATA

El objetivo de este capítulo es doble: entender la lógica de los datos longitudinales a través de una inspección inicial de una base de datos panel y mostrar algunos comandos en Stata[3] que faciliten esa exploración. Esta primera mirada de los datos tiene tres etapas. La primer etapa consiste en disponer y revisar la documentación que acompaña a las bases de datos que serán utilizadas. La segunda etapa consiste en conocer qué variables están disponibles en la base datos, cómo han sido codificadas esas variables y el número de datos faltantes en cada una de ellas. Finalmente, la tercera etapa consiste en estudiar la distribución de frecuencia de las variables de interés junto algunos análisis descriptivos de sus medias y desviaciones estándar. Ahora bien, previo a las etapas mencionada y a realizar cualquier tipo de análisis, es necesario entender los formatos que presentan las bases de datos longitudinales y como se preparan los datos cuando éstos provienen de más de un archivo como usualmente ocurre con las encuestas panel.

Para el uso práctico de esta guía se recomienda bajar los datos que están en el siguiente enlace:

https://www.dropbox.com/sh/vo38muo8xpx3or4/AACbwbAeyog3W3vA4aZNRWOMa?dl=0

2.1.1 Los formatos "long" y "wide" en las bases de datos panel

Los datos longitudinales presentan una complejidad adicional comparados con los datos de corte transversal puesto que incorporan en su estructura la dimensión temporal. En las bases de datos de las encuestas de corte transversal cada fila representa una unidad de análisis (en el caso de la ELPS la unidad de análisis son individuos). El número de columnas está determinado por el total de variables que se midieron en la encuesta. Por lo tanto, este tipo de base datos está compuesta por dos dimensiones: las unidades y las variables. En cambio, los datos longitudinales contienen repetidas observaciones en el tiempo de las mismas unidades (personas). Así, los datos longitudinales pueden ser visto como datos que incluyen una tercera dimensión: las observaciones a lo largo del tiempo. La inclusión de la dimensión temporal incide en la disposición de los datos panel para ser utilizados en Stata. Tal como lo muestra la Figura 1[4], existen dos formas de organizar los datos panel: el formato "wide" y el formato "long".

[3] Todos los ejercicios de este documento están diseñados en Stata 12 (http://www.stata.com/). El Banco Interamericano de Desarrollo no provee este paquete estadístico. La adquisición de la licencia para usar Stata es responsabilidad de cada usuario y/o institución.

[4] En la Figura 1 la variable id se refiere al identificador de la persona. Las variables que tienen el prefijo IngTrab hacen referencia al ingreso líquido mensual declarado por cada persona en las rondas realizadas los años 2002, 2004, 2006 y 2009.

Figura 1: Organización datos panel en formatos "wide" y "long"

a) Formato "wide"

id	IngTrab2002 Ronda 1	IngTrab2004 Ronda 2	IngTrab2006 Ronda 3	IngTrab2009 Ronda 4
1001	200.000	200.000	260.000	300.000
1002	315.000	Sin Dato	420.000	450.000
1003	210.000	320.000	650.000	700.000
1004	150.000	300.000	.	400.000

b) Formato "long"

id	Ronda / Año	IngTrab
1001	1 / 2002	200.000
1001	2 / 2004	200.000
1001	3 / 2006	260.000
1001	4 / 2009	300.000
1002	1 / 2002	315.000
1002	3 / 2006	420.000
1002	4 / 2009	450.000
1003	1 / 2002	210.000
1003	2 / 2004	320.000
1003	3 / 2006	650.000
1004	4 / 2009	700.000
1004	1 / 2002	150.000
1004	2 / 2004	300.000
1004	4 / 2009	400.000

En el formato "wide" cada unidad de análisis ocupa una fila. Entonces, las variables que entregan información de cada una de las mediciones (la primera ronda más la ronda dos, tres y cuatro) aparecen en diferentes columnas y se identifican con un sufijo. En la Figura 1 los sufijos serían 2002, 2004, 2006 y 2009 correspondientes a los años de ejecución de la encuesta. Por lo tanto, mientras más rondas tenga la encuesta mayor será el número de variables que contendrá la base de datos en este formato. Las unidades que no tienen respuesta en alguna medición se dejan como celdas vacías en la base de dato en formato "wide" y aparecen como punto (.). Aunque este formato es más claro para el investigador porque las filas de las bases de datos son independientes entre sí y el número de filas corresponde al total de unidades se utiliza menos que el formato "long". La razón es simple. Los paquetes estadísticos prefieren el formato "long" para los análisis de regresiones en vez del "wide".

En el formato "long" cada medición ocupa una fila. Este formato entrega las observaciones de cada individuo en múltiples filas. En otras palabras, las observaciones de una variable para diferentes periodos de tiempo son puestas en una fila adicional. De esta manera es posible reconocer la matriz de observaciones de un individuo en el tiempo a través de dos identificadores: uno para el individuo (en la base de datos de la Figura 1 sería id), y el otro para el tiempo (ronda o año). Se sugiere utilizar el formato "long" cuando los datos cubren varias rondas. Solo las unidades que entregan información en cada ronda son incluidas en las bases de datos en formato "long", es decir, no hay celdas vacías para las unidades que no responden la encuesta en alguna ronda. Adicionalmente, Stata necesita el formato "long" para utilizar los comandos que analizan datos longitudinales.

Con el fin de aprender a transformar los archivos de formato "wide" a formato "long" y viceversa en Stata, se utilizará una base en formato "wide" llamada PanelMuestraR1234Wide.dta. Pero antes es necesario conocer algunos consejos prácticos en Stata que también serán útiles para los ejemplos y ejercicios que se verán en el presente documento[5].

[5] En los anexos se entrega la programación completa del Do-File que se ejecuta durante el capítulo 2. También se incluye el contenido de los Do-Files de los demás capítulos. Para quienes recién están comenzando a usar Stata o quienes no han tenido la oportunidad de trabajar con este programa se recomienda el libro *A Gentle Introduction to Stata, Fourth Edition* de Alan Acock (2014).

2.1.2 Directorio de trabajo en Stata y carga de base de datos

Se sugiere utilizar un directorio de trabajo en Stata cuando el usuario tiene distintas bases de datos o cuando se quiere automatizar los análisis[6]. La ubicación actual del directorio se puede encontrar utilizando el comando pwd.

```
. pwd
C:\Stata\Datos
```

En este ejemplo Stata está trabajando con el directorio C:\Stata\Datos. Ahora se deben descargar los archivos de la web[7] y copiarlos en una carpeta llamada GuiaUsoDatosELPS. Esta carpeta debe ser creada en un directorio del computador cuya ubicación sea conocida por el usuario (RutaEnTuComputador). Para cambiar de la carpeta Datos a la nueva capeta de trabajo GuiaUsoDatosELPS se debe usar el comando cd.

```
. cd C:\RutaEnTuComputador\GuiaUsoDatosELPS
```

Si además se quiere obtener una descripción del directorio de trabajo es necesario escribir el comando dir:

```
. dir
```

En el directorio se encuentra la base de datos PanelMuestraR1234Wide.dta que pesa 176K y el Do-File que lleva el nombre GuiaPracticaUsoDatosELPS.do. Además se visualizan las bases de datos y Do-Files que se utilizará más adelante y en los siguientes capítulos.

Una vez que se ha establecido el directorio donde se trabajará en Stata se debe cargar la base de datos PanelMuestraR1234Wide.dta utilizando el comando use. Con la misma lógica, los Do-Files se obtienen escribiendo doedit más el nombre del archivo que se desea obtener.

```
. use PanelMuestraR1234Wide.dta
. doedit GuiaPracticaUsoDatosELPS
```

Una vez que los datos se encuentran en Stata, el primer objetivo será reproducir el formato "wide" de la figura 1 agregando la variable Sexo del entrevistado para ver cómo se visualiza una característica de la persona que no cambia en el tiempo junto a otra que si lo hace como la variable IngresoTrababajador para los años 2002, 2004, 2006 y 2009. Para eso se utiliza el comando keep que genera una base de datos que solo dispone de las variables id Sexo IngresoTrabajador2002 IngresoTrabajador2004 IngresoTrabajador2006 IngresoTrabajador2009. Por su lado, el comando list entrega en pantalla la información de esas variables para las personas que están identificadas con los números 1001, 1002, 1003 y 1004.

```
. keep    id Sexo IngresoTrabajador2002 IngresoTrabajador2004 IngresoTrabajador2006
          IngresoTrabajador2009

. list    id Sexo IngresoTrabajador2002 IngresoTrabajador2004 IngresoTrabajador2006
          IngresoTrabajador2009 if id <= 1004
```

[6] Para estudiar con mayor profundidad el manejo de bases de datos y automatización de procesos en Stata se sugiere revisar el *libro The Workflow of Data Analysis Using Stata* de Scott Long (2009).

[7] Descargar archivos en https://www.dropbox.com/sh/vo38muo8xpx3or4/AACbwbAeyog3W3vA4aZNRWOMa?dl=0

	id	Sexo	Ing~2002	Ing~2004	Ing~2006	Ing~2009
1.	1001	Hombre	200000	200000	260000	300000
2.	1002	Hombre	315000	Sin Dato	420000	450000
3.	1003	Mujer	210000	320000	650000	700000
4.	1004	Mujer	150000	300000	Sin Dato	400000

Se aprecia que las variables id y Sexo son constantes en el tiempo mientras las variables que dan cuenta el ingreso líquido mensual promedio del entrevistado pueden cambiar en cada medición. Utilizando el comando describe es posible saber que la base de datos en formato "wide" contiene información de 6 variables para 400 individuos.

. describe

```
Contains data from PanelMuestraR1234Wide.dta
  obs:           400
 vars:             6                          25 Apr 2015 12:54
 size:         8,400
```

2.1.3 Transformación de archivos de formato "*wide*" a formato "*long*" y viceversa.

El siguiente paso será utilizar la misma base de datos para aprender a usar el comando reshape que transforma los datos desde formato "wide" a "long" y viceversa (Baum & Cox, 2007). La información que necesita reconocer Stata para utilizar reshape son tres: i) el identificador de las unidades de análisis (en este caso es la variable id), ii) las variables que representan características de la unidad de observación, y ii) el momento en que fue realizada cada observación. La manera que se ingresa esa información es la siguiente:

. reshape long IngresoTrabajador, i(id) j(Anho)

```
Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                      400   ->    1600
Number of variables                   6   ->       4
j variable (4 values)                     ->   Anho
xij variables:
IngresoTrabajador2002 IngresoTrabajador2004 ... IngresoTrabajador2009->IngresoTrabajador
-----------------------------------------------------------------------------
```

Stata provee información sobre la nueva base de datos en formato "long". Ésta tiene 4 variables en vez de 6 y 1600 observaciones en lugar de 400. Es importante tener en cuenta que siguen habiendo 400 individuos, pero ahora con 4 observaciones cada uno. A continuación se muestran los datos en este nuevo formato para los identificadores menores o igual a 1004.

. list id Anho Sexo IngresoTrabajador if id <= 1004

	id	Anho	Sexo	Ingres~r
1.	1001	2002	Hombre	200000
2.	1001	2004	Hombre	200000
3.	1001	2006	Hombre	260000
4.	1001	2009	Hombre	300000
5.	1002	2002	Hombre	315000
6.	1002	2004	Hombre	Sin Dato
7.	1002	2006	Hombre	420000
8.	1002	2009	Hombre	450000
9.	1003	2002	Mujer	210000
10.	1003	2004	Mujer	320000
11.	1003	2006	Mujer	650000
12.	1003	2009	Mujer	700000
13.	1004	2002	Mujer	150000
14.	1004	2004	Mujer	300000
15.	1004	2006	Mujer	Sin Dato
16.	1004	2009	Mujer	400000

Una vez utilizado el comando reshape es fácil transformar los archivos de formato "wide" a formato "long" y viceversa.

```
. reshape wide
. reshape long
```

2.1.4 Agrupar archivos para construir bases de datos panel: comandos merge y append

En general, las bases de datos longitudinales de individuos se entregan para cada ronda en archivos separados. Es trabajo del usuario combinar los datos de una manera longitudinal ya sea en formato "wide" o "long". Es posible que existan algunas excepciones en este punto. Por ejemplo, puede ocurrir que el organismo encargado de la encuesta en un país decida unir los datos a nivel individual antes de liberar las bases de datos para su uso público. Sin embargo, es posible que incluso en ese caso, ciertos módulos del cuestionario como la historia laboral o la información de los otros miembros del hogar vengan en archivos independientes que necesitan ser combinados. Los comandos merge y append son de gran utilidad a la hora de unir archivos cuyas columnas (variables) o filas (observaciones) están relacionadas.

Para el propósito de este documento se han creado bases de datos ficticias para ser usadas en los ejercicios. Se cuentan con base de datos que tienen la información de los entrevistados para cada año de medición. Las rondas corresponden a los años 2002, 2004, 2006 y 2009 (PanelMuestraR1.dta, PanelMuestraR2.dta, PanelMuestraR3.dta y PanelMuestraR4.dta). Además se dispone de las bases de datos de los miembros del hogar que vivían con el entrevistado. Cada ronda tiene su propia base de datos y se llaman HogarR1.dta, HogarR2.dta, HogarR3.dta y HogarR4.dta.

Con el fin de construir la base de datos panel de todos los individuos que participaron en las cuatro entrevistas se utiliza el comando merge. El primer paso es cargar en Stata la base de datos de la primera medición realizada el año 2002 llamada PanelMuestraR1.dta. El comando list permite ver tres variables de la base de datos: el número identificador del encuestado (id), su sexo y el ingreso promedio líquido mensual el 2002 (IngTrab2002).

```
. use  PanelMuestraR1.dta
. list id Sexo IngresoTrabajador* if id <= 1004
```

	id	Sexo	Ingres~2
1.	1001	Hombre	200000
2.	1002	Hombre	315000
3.	1003	Mujer	210000
4.	1004	Mujer	150000

Luego se combinan las rondas dos, tres y cuatro sucesivamente utilizando el comando merge tal como se muestra a continuación.

```
. merge 1:1 id using "PanelMuestraR2", keepusing () update nogenerate
. merge 1:1 id using "PanelMuestraR3", keepusing () update nogenerate
. merge 1:1 id using "PanelMuestraR4", keepusing () update nogenerate
. list id Sexo IngresoTrabajador* if id <= 1004
```

	id	Sexo	Ingres~2	Ingres~4	Ingres~6	Ingres~9
1.	1001	Hombre	200000	200000	260000	300000
2.	1002	Hombre	315000	Sin Dato	420000	450000
3.	1003	Mujer	210000	320000	650000	700000
4.	1004	Mujer	150000	300000	Sin Dato	400000

Se puede apreciar para los cuatro primeros casos una base idéntica a la que se utilizó para aprender el uso del formato "wide". Colocando asterisco (*) al final de la variable IngresoTrabajador no es necesario escribir las variables[8] IngesoTrabajador2002, IngresoTrabajador2004, IngresoTrabajador2006 e IngresoTrabajador2009.

En el caso que se quiera unir dos bases de datos que tienen las mismas variables se utiliza el comando append. Por ejemplo, si el usuario está interesado en construir el ingreso per cápita promedio del entrevistado, necesita utilizar los ingresos que aportan todos los individuos al hogar y dividirlo por el número de personas del hogar durante el año de medición. Para construir esta variable es necesario unir la base de datos del entrevistado con la base de datos del hogar para cada ronda. La secuencia de instrucciones para realizar esta combinación de bases de datos se encuentra en el Do-File de esta sección.

2.2 INSPECCIONANDO ALGUNAS VARIABLES DE LA PANEL FICTICIA

Hasta el momento se ha revisado la manera de construir una base de datos panel en formato "wide" y, en caso que sea necesario, transformarla a un formato "long". Paralelamente se han revisado algunos comandos básicos en Stata como use, list, describe y otros diseñados especialmente para

[8] El símbolo * en Stata es usado como marcador de caracteres no especificados. Por lo tanto, de existir en la base de datos otras variables que comiencen con IngresoTrabajador también serán listadas. Para esas situaciones se recomienda escribir cada una de las variables de interés con el fin de ejecutar sólo esas variables en el comando.

manejar datos longitudinales como el reshape y varias bases de datos como son los comandos merge y append.

A continuación el foco estará puesto en las variables más que en la estructura de las bases de datos panel. El objetivo es realizar una primera inspección de los datos, aprender a renombrar o recodificar las variables de manera conveniente para varios tipo de análisis y conocer algunos comandos en Stata que permiten graficar variable continuas y discretas. Se comenzará entonces cargando en la memoria una de las bases de datos ficticias (primera ronda) y seleccionando las variables que serán usadas en los siguientes ejercicios.

```
. use PanelMuestraR1.dta
. keep id Sexo Edad2002 StatusLaboral2002 IngresoTrabajador2002 CotizaActual2002
```

2.2.1 El uso de los comando summarize y tabulate en la exploración de datos

El comando summarize entrega las medias, desviaciones estándares, mínimos, máximos y el número de observaciones de cada variable. Para la base de datos de 400 casos se aprecia entre otras cosas que el promedio de edad de la muestra es 44 años y que hay personas entre los 18 y 83 años. El ingreso líquido mensual de los 254 trabajadores de la base de datos ficticia tiene un máximo de Ch$1,750,000, y un valor mínimo de 7777 correspondiente al código "No aplica". En la variable CotizaActual2002 el valor máximo 9999 corresponde a los que declaran no saber si están cotizando en el momento de la entrevista. Como se explicará más adelante, el análisis de variables que entregan información sobre los datos faltantes de la entrevista debe tomar en cuenta algunas consideraciones para no cometer errores al calcular sus medias y desviaciones estándares.

```
. summarize, sep (6)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	400	1200.5	115.6143	1001	1400
Sexo	400	1.4825	.5003194	1	2
Edad2002	400	44.0775	14.15831	18	83
StatusL~2002	400	1.78	1.241149	1	4
Ingreso~2002	400	122051.7	177962.7	7777	1750000
CotizaA~2002	400	2833.852	3790.108	1	9999

Como era de esperar en la variable Sexo el comando summarize informa que el valor mínimo es 1 y el valor máximo es 2. Para la construcción de la base de datos ficticia la convención adoptada fue otorgar valor 1 para indicar que se entrevistó a un hombre y 2 si la entrevistada fue una mujer. Sin embargo, al tabular la variable Sexo, Stata despliega "Hombre" y "Mujer" en vez de los valores 1 y 2. La explicación está dada porque se han etiquetado los valores numéricos para entender lo que significan cada uno de ellos de una manera más directa y clara. En este caso Hombre corresponde a la etiqueta del valor 1 y Mujer es la etiqueta del valor 2.

```
. tabulate Sexo
```

Sexo	Freq.	Percent	Cum.
Hombre	207	51.75	51.75
Mujer	193	48.25	100.00
Total	400	100.00	

La relación entre los valores y las etiquetas de las variables puede ser mostrada de varias maneras en Stata. Por ejemplo, si a la variable StatusLaboral2002 se le asigna el término nolabel en el comando tabulate es posible visualizar solamente los valores de la variable, es decir, la tabla no muestra las etiquetas.

```
. tabulate StatusLaboral2002
. tabulate StatusLaboral2002, nolabel
```

2002 StatusLaboral	Freq.	Percent	Cum.
Trabajando	270	67.50	67.50
Cesante	39	9.75	77.25
Inactivo	91	22.75	100.00
Total	400	100.00	

2002 StatusLaboral	Freq.	Percent	Cum.
1	270	67.50	67.50
2	39	9.75	77.25
4	91	22.75	100.00
Total	400	100.00	

Un comando que permite agregar el valor numérico a la etiqueta es numlabel _all, add. Por ejemplo para la variable Sexo este comando de Stata cambia las etiquetas a: 1. Trabajando, 2. Cesante, y 4. Inactivo. Para volver al formato anterior solo hay que escribir numlabel _all, remove.

```
. numlabel _all, add
. tabulate StatusLaboral2002
```

2002 StatusLaboral	Freq.	Percent	Cum.
1. Trabajando	270	67.50	67.50
2. Cesante	39	9.75	77.25
4. Inactivo	91	22.75	100.00
Total	400	100.00	

2.2.2 Importancia de recodificar los datos faltantes en los análisis descriptivos

La variable CotizacionActual2002 da cuenta si la persona en el momento de la entrevista realizada el año 2002 estaba cotizando o no. Al tabular esta variable se observa que la categoría "No sabe"

corresponde al valor 9999 y el valor 8888 indica que la persona no quiso contestar esa pregunta ("No responde")[9]. La falta de información de una pregunta no solo se explica porque el encuestador no sabe la respuesta o se niega a entregarla sino también porque el flujo del cuestionario no permite hacer la pregunta. Por ejemplo, si alguien está desempleado o inactivo es razonable no preguntarle si está cotizando. En este caso el código 7777 indica que preguntar por la cotización actual del trabajador "No aplica". La manera de identificar los valores perdidos o *missing values* puede variar de una base de datos a otra dependiendo de la agencia responsable de la codificación de las variables. Independiente de cuales sean los códigos utilizados en cada país, lo importante es disponer de esa información en la documentación para estudiar la no respuesta de esas preguntas y entender sus orígenes.

```
. tabulate CotizaActual2002
```

2002 CotizaActual	Freq.	Percent	Cum.
1. Si	191	47.75	47.75
2. No	65	16.25	64.00
7777. No aplica	137	34.25	98.25
8888. No responde	2	0.50	98.75
9999. No sabe	5	1.25	100.00
Total	400	100.00	

En algunos análisis es necesario recodificar los valores que explican la no respuesta de la pregunta por un código que el paquete estadístico utilizado identifique como valor faltante. Stata considera el punto (.) como *missing value* o valor faltante. La razón principal para recodificar los motivos que explican la falta de respuesta es que valores como 9999 u 8888 pueden ser incluidos erróneamente en cualquier cálculo de medida de tendencia y dispersión. De esta forma se aconseja recodificar esos valores en punto (.) ocupando el comando recode como se describe a continuación.

```
. recode CotizaActual2002 (9999 = .) (8888 = .) (7777 = .)
. tabulate CotizaActual2002, missing
```

2002 CotizaActual	Freq.	Percent	Cum.
1. Si	191	47.75	47.75
2. No	65	16.25	64.00
.	144	36.00	100.00
Total	400	100.00	

La recodificación de la variable CotizaActual2002 permite mirar la proporción de personas que cotizan/no cotizan según hombres/mujeres usando el comando tabulate. Con el mismo comando y agregando r chi2 es posible testear si existe alguna relación entre ambas variables (test de Chi-

[9] La codificación de los datos faltantes puede variar entre las agencias que generan datos oficiales en cada país. Por ejemplo, en algunas encuestas la categoría "No aplica" tiene un valor -4, "No responde" se clasifica como 8, y "No sabe" es designado con un 9.

Cuadrado[10]). Para la base de datos ficticia que está siendo usada en este ejercicio, los resultados indican que no hay una relación estadísticamente significativa entre la situación de cotización y el género del entrevistado el año 2002 (Chi-Cuadrado con un grado de libertad = 0,2106 y p=0,646).

```
. tabulate CotizaActual2002 Sexo
. tabulate CotizaActual2002 Sexo, r chi 2
```

2002 CotizaActual	Sexo: 1. Hombre	2. Mujer	Total
1. Si	**112**	**79**	**191**
2. No	**36**	**29**	**65**
Total	**148**	**108**	**256**

2002 CotizaActual	Sexo: 1. Hombre	2. Mujer	Total
1. Si	**112**	**79**	**191**
	58.64	**41.36**	**100.00**
2. No	**36**	**29**	**65**
	55.38	**44.62**	**100.00**
Total	**148**	**108**	**256**
	57.81	**42.19**	**100.00**

Pearson chi2(**1**) = **0.2106** Pr = **0.646**

2.2.3 Renombrar y etiquetar variables

En algunos casos, el usuario prefiere trabajar sus análisis con variables cuyo nombre no exceda los 12 caracteres. De esta manera Stata visualiza el nombre completo de la variable sin necesitad de truncarlo con el uso del tilde (~) cuando el nombre excede el máximo de caracteres. Por ejemplo, en las tabulaciones de más arriba la variable IngresoTrabajador2002 el programa Stata la entrega como Ingreso~2002. Para algunos usuarios es más cómodo usar el nombre de IngTrab02 en lugar de IngresoTrabajador2002 (nombre actual de la variable en la base de datos ficticia). Stata utiliza el comando rename para cambiar el nombre de la variable seleccionada.

```
. rename IngresoTrabajador2002 IngTrab02
```

[10] El estadístico Chi-Cuadrado se obtiene de la comparación entre las frecuencias de la muestra observada y los valores esperados en una situación donde no existe relación alguna entre las variables. Su valor calculado se compara con el valor teórico considerando los grados de libertad y una posibilidad de error en la estimación de un 5 por ciento. Para una explicación detallada del test de Chi-Cuadrado se recomienda el capítulo 2.4 del libro de Agresti (2007).

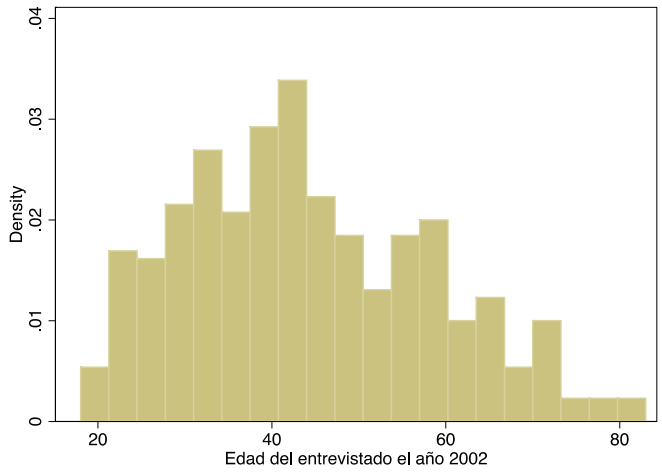
Las etiquetas de las variables son textos que tienen un máximo de 80 caracteres y están asociados a variables. Estas etiquetas aparecen en las tablas y gráficos documentando las variables analizadas como se verá más adelante. Con el fin de ilustrar el uso del comando label var se etiqueta la variable recién renombrada (IngTrab02)y también la variable Edad2002.

```
. label var IngTrab02 "Ingreso Mensual líquido del trabajador"
. label var Edad2002  "Edad del entrevistado el año 2002"
```

2.2.4 Graficar variables continuas

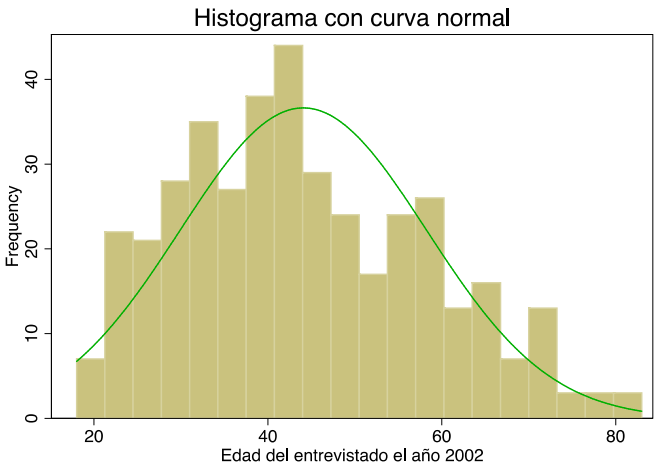
Las variables que conforman una base de datos son variables discretas o continuas. En las variables discretas no se pueden encontrar valores intermedios entre dos valores dados, en las variables continuas sí. La variable Edad2002 es una variable continua que en la base de datos ficticia tiene valores entre 18 y 83 años. Stata permite graficar la distribución de frecuencia de este tipo de variables a través de histogramas. Los histogramas dividen los valores de la variable en una serie de intervalos y representan cada uno de éstos con un área proporcional a su tamaño. Solo dos palabras generan un histograma en Stata: el comando histogram y el nombre de la variable.

```
. histogram Edad2002
```



Una manera de mejorar la información que entrega el histograma es mostrar las frecuencias en vez de los porcentajes, y sobrescribir la distribución normal sobre la distribución empírica. Además se puede agregar un título para describir el contenido del histograma.

```
. histogram Edad2002, frecuency normal title (Histograma con curva normal)
```



Una variable clave en cualquier estudio que aborda el tema laboral y de seguridad social es la remuneración del trabajo principal del entrevistado. En el caso de la base de datos ficticia esa variable es la lngTrab02. Se utilizará esta variable continua para mostrar la implementación de gráficos de densidad en Stata. Los gráficos de densidad también dividen el rango de la variable en una serie de intervalos, pero en lugar de asignar una probabilidad constante, le asigna a cada valor un peso el cual explica la probabilidad final. De esta manera se suaviza la frecuencia de la variable estudiada.

Tal como se discutió en la sección anterior antes de graficar la variable lngTrab02 es necesario revisar si tiene códigos que no guardan relación con la remuneración declarada por el trabajador. Específicamente hay que chequear si las razones de no respuesta de esa pregunta tienen algún código que erróneamente puede ser incluido al visualizar los datos de la variable. El comando sort ordena de menor a mayor los valores de la variable lngTrab02. Al tabular los primeros 200 casos ordenados por el comando sort se observa que a 159 personas no se les preguntó su ingreso líquido mensual. Los dos motivos para que la pregunta no aplique (código 7777) son que el entrevistado declaró previamente que se encuentra cesante o inactivo (146 casos). Los 13 trabajadores que no quisieron entregar su salario mensual el año 2002 aparecen con el código 8888 en la base de datos.

```
. sort lngTrab02
. tabulate lngTrab02 in 1/200
```

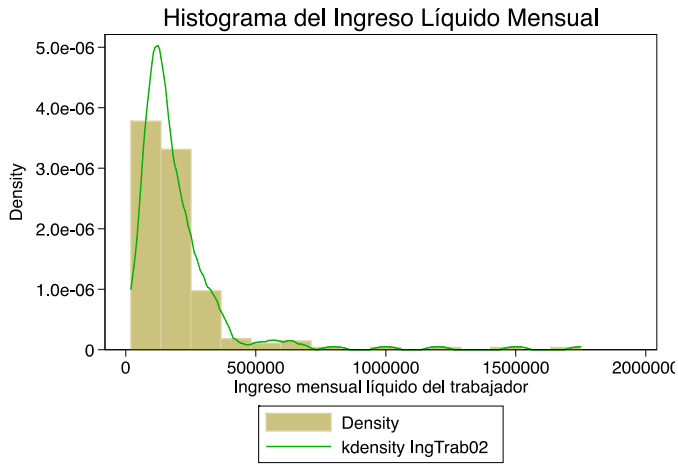
Ingreso mensual líquido del trabajador	Freq.	Percent	Cum.
7777. No aplica	146	73.00	73.00
8888. No responde	13	6.50	79.50
20000	2	1.00	80.50
22000	1	0.50	81.00
24000	1	0.50	81.50
30000	2	1.00	82.50
32000	2	1.00	83.50
35000	1	0.50	84.00
40000	3	1.50	85.50
50000	3	1.50	87.00
53000	1	0.50	87.50
55000	1	0.50	88.00
60000	5	2.50	90.50
70000	4	2.00	92.50
79125	1	0.50	93.00
80000	5	2.50	95.50
82500	1	0.50	96.00
83000	1	0.50	96.50
85000	5	2.50	99.00
85151	1	0.50	99.50
90000	1	0.50	100.00
Total	200	100.00	

Se recodifican entonces las razones de no respuesta de la variable lngTrab02 como punto (.) Valor que Stata reconoce como *missing value* o valor faltante.

```
. recode lngTrab02 (8888 = .)
```

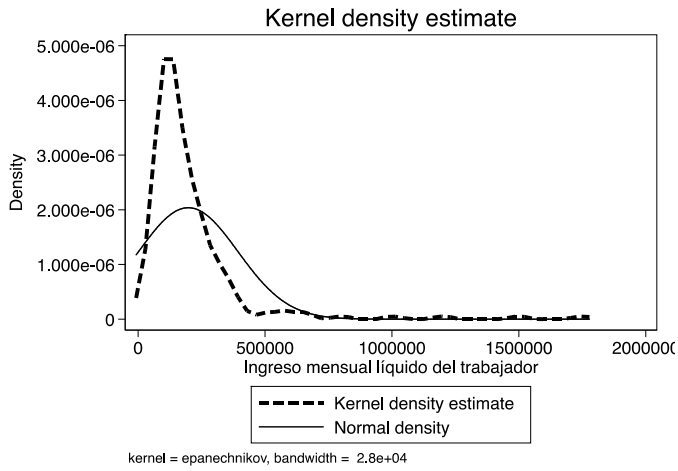
Recién ahora es posible ajustar a la densidad estimada de la variable lngTrab02 en forma no paramétrica (gráfico de densidad) a través de la siguiente programación que utiliza el comando histogram. Se incluye la leyenda del gráfico con la opción legend y se entrega cinco valores de referencia dispuestos con un ángulo de cero grado en el eje de la densidad con la opción ylabel.

```
. histogram     lngTrab02, kdensity ///
                title (Histograma del Ingreso Líquido Mensual) ///
                legend (on) ylabel(#5,angle(0))
```



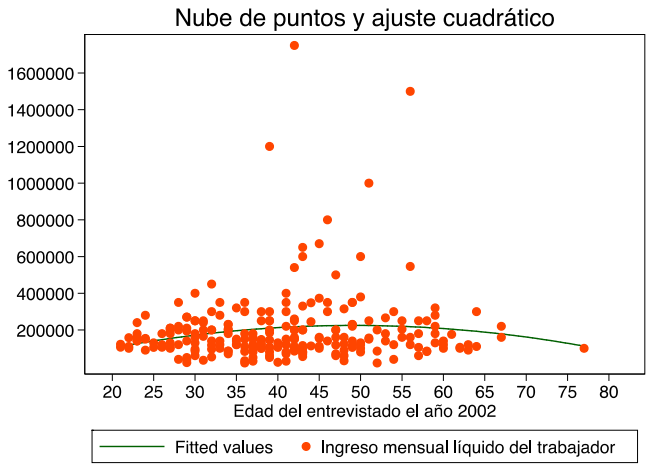
Stata además permite mostrar la representación de densidad de la variable junto a la densidad normal. De esta manera se puede tener una idea de la distancia entre ambas densidades. Al mismo tiempo ofrece la posibilidad de cambiar el formato de las líneas para mejorar la presentación del gráfico en situaciones donde no se puede imprimir el gráfico en colores.

```
. kdensity lngTrab02, lcolor (black) lwidth (thick) lpattern (dash) ///
                      normal normopts (lcolor (black) lpattern (solid)) ///
                    ylabel(#5,angle(0))
```



El comando graph twoway permite dibujar en dos ejes varios gráficos al mismo tiempo. Solo es necesario expresar los distintos gráficos entre paréntesis. En el siguiente ejercicio se grafica la nube de puntos que representa el ingreso líquido mensual del trabajador y los años de vida que el trabajador tenía el año 2002. Además se dibuja en el mismo gráfico el ajuste cuadrático que modela la relación entre las variables lngTrab02 y Edad2002.

```
. graph twoway (qfit lngTrab02 Edad2002)(scatter lngTrab02 Edad2002), ///
            title (Nube de puntos y ajuste cuadrático) ///
            ylabel(200000 (200000) 1600000, angle(0)) xlabel(20 (5) 80)
```



El usuario también puede utilizar los comandos aprendidos en Stata para graficar variables continuas en datos longitudinales. Primero necesita cargar en la memoria datos panel. Hasta el momento se ha trabajado con una base de datos con información de la primera medición, es decir, con información de corte transversal. Para tener datos panel es necesario tener por lo menos dos rondas de medición. Para unir las bases de datos de las rondas uno y dos se puede usar el comando merge de la siguiente forma.

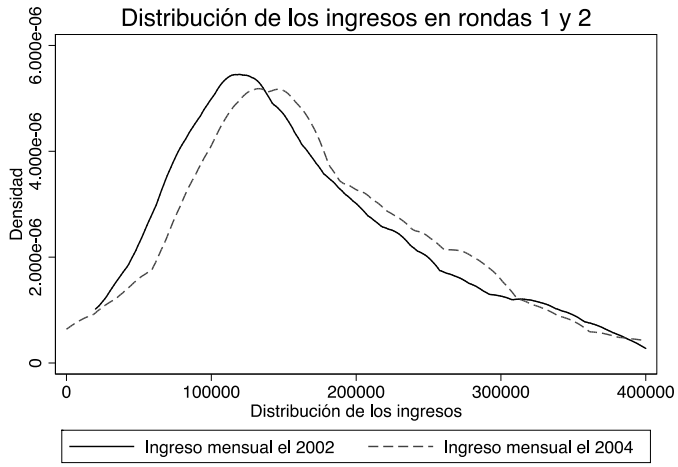
```
. sort id
. merge id using PanelMuestraR2.dta
. drop _merge
```

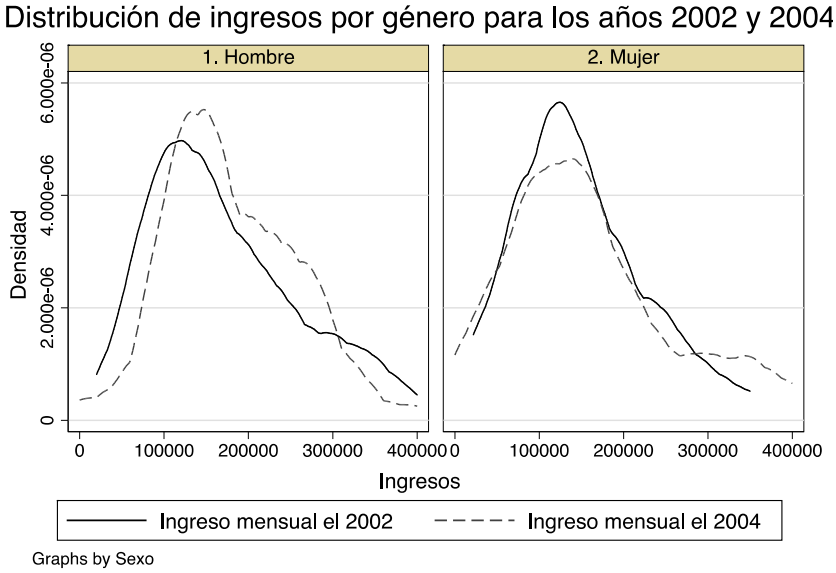
La siguiente sintaxis que usa el comando twoway muestra la distribución del ingreso para las dos rondas de medición en la base de datos ficticia.

```
. rename IngresoTrabajador2004 IngTrab04
. graph twoway (kdensity IngTrab02, lcolor(black)) ///
        (kdensity IngTrab04, lcolor(gs4) lp(dash)) ///
        if IngTrab02<=400000 & IngTrab04<=400000, ///
        legend (label(1 "Ingreso mensual el 2002") ///
                label(2 "Ingreso mensual el 2004")) ///
        title (Distribución de los ingresos en rondas 1 y 2) ///
        ytitle (Densidad) xtitle (Distribución de los ingresos)
```

Agregando la siguiente línea a las instrucciones anteriores se puede obtener la distribución de ingresos (líquida y mensual) de los trabajador para los años 2002 y 2004.

```
by (Sexo, title(Distribución de ingresos por género para los años 2002 y 2004))
```





2.2.5 Crear variables y etiquetar categorías usando *loops*

Para realizar algunos análisis es necesario generar nuevas variables. Por ejemplo, la relación entre remuneración y edad del trabajador puede ser distinta entre hombres y mujeres. Por lo tanto es necesario crear las variables IngTrabHom02 y IngTrabMuj02 para graficar los ingresos de los hombres y mujeres por separado (el sufijo 02 indica el año 2002 en ambas variables). El comando generate es usado habitualmente en Stata para crear nuevas variables.

Al trabajar con datos longitudinales la programación para generar variables puede ser extensa si no se ocupan circuitos recursivos o *loops*. En muchos casos, la estructura de la sintaxis de los comandos se va repitiendo y solo cambia la nominación de las variables que identifica cada ronda del panel. Stata tiene distintas posibilidades para diseñar circuitos recursivos que son de suma utilidad para algunas tareas. Por ejemplo, crear y etiquetar variables en más de una ronda de la base de datos panel.

A continuación se muestra el comando foreach que junto al comando generate crean y etiquetan las variables que contienen las remuneraciones de los hombres y mujeres para los años 2002 y 2004 (primera y segunda ronda del panel ficticio).

```
. foreach i in 02 04  {
            generate IngTrabHom`i'=.
            replace IngTrabHom`i'=IngTrab`i' if Sexo==1
            label var IngTrabHom`i' "Ing. Liq. Men."
            generate IngTrabMuj`i'=.
            replace IngTrabMuj`i'=IngTrab`i' if Sexo==2
            label var IngTrabMuj`i' "Ing. Liq. Men."
                                  }
. summarize, sep (8)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	400	1200.5	115.6143	1001	1400
Sexo	400	1.4825	.5003194	1	2
Edad2002	400	44.0775	14.15831	18	83
StatusL~2002	400	1.78	1.241149	1	4
IngTrab02	241	197384.5	195745.2	20000	1750000
CotizaA~2002	256	1.253906	.4360972	1	2
Edad2004	400	46.45	14.12307	20	86
StatusL~2004	400	1.79	1.220502	1	4
IngTrab04	254	212800.6	210600.4	2000	2000000
CotizaA~2004	377	1.477454	.5001552	1	2
IngTrabHom02	139	222157.2	224081.1	20000	1750000
IngTrabMuj02	102	163625.7	142989.1	20000	1200000
IngTrabHom04	151	245684.5	250664.7	2000	2000000
IngTrabMuj04	103	164592.2	116888.3	20000	600000

El comando summarize informa que se agregaron cuatro nuevas variables a la base de datos y que el año 2002 el promedio de ingresos de los hombres de la muestra es Ch$222,157 y el de las mujeres es un 26 por ciento menor que el de los hombres (Ch$163,626). El año 2004 el ingreso de los hombres de la muestra panel aumentó en un 11 por ciento (Ch$245,685), y las remuneraciones de las mujeres solo un 0,6 por ciento (Ch$164,592).

Una variante del comando generate en Stata es egen que permite crear variables con las medias, medianas, desviaciones estándares, suma total, etc. de las observaciones. Por ejemplo, se puede calcular el promedio del ingreso líquido mensual de hombres y mujeres en una sola línea de programación. El comando bysort indica que se realiza la operación para las dos categorías por separado y se crean las variable MediaIngreso02 y MediaIngreso04. Ambas variables asigna el promedio de ingreso mensual de los hombres a cada observación donde el individuo es hombre. Lo mismo hace para el caso de las mujeres. El comando list muestra un extracto de la base de datos donde se comprueba que los valores calculados son los mismos del ejercicio anterior. A continuación se describen todos los comandos señalados a través de un *loop* para las dos rondas.

```
. foreach i in 02 04  {
          bysort Sexo: egen MediaIngreso`i' = mean(IngTrab`i')
          list id Sexo IngTrab`i' MediaIngreso`i' in 205/210 if IngTrab`i' < .
                                  }
```

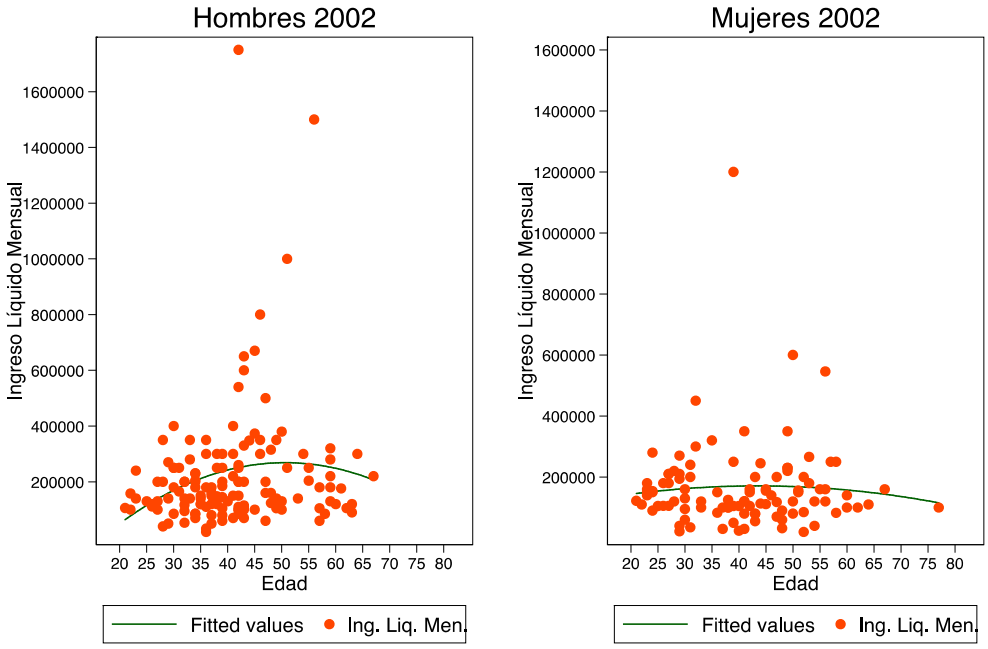
	id	Sexo	IngT~b02	Media~02
206.	1018	1. Hombre	260000	222157.2
207.	1365	1. Hombre	185000	222157.2
208.	1184	2. Mujer	220000	163625.7

	id	Sexo	IngT~b04	Media~04
206.	1018	1. Hombre	300000	245684.5
207.	1365	1. Hombre	250000	245684.5
208.	1184	2. Mujer	300000	164592.2
209.	1127	2. Mujer	124000	164592.2

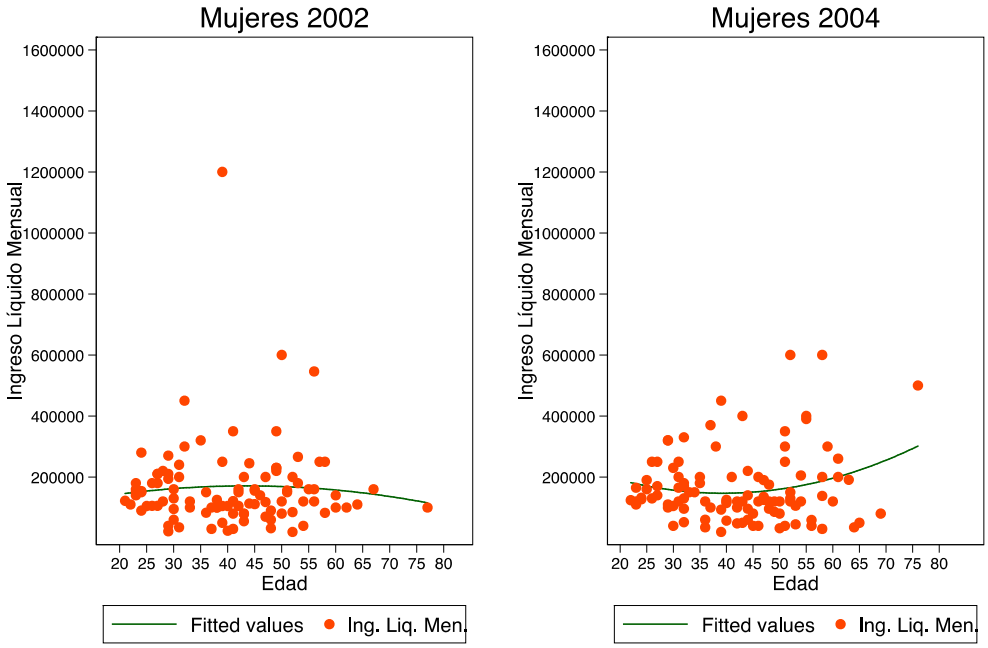
El uso de *loops* es eficiente también para construir más de un gráfico y después colocarlos en una sola visualización a través del comando combine. A continuación se entregan las líneas en Stata para generar cuatro gráficos con dos *loops* usando datos de las rondas 1 y 2. Posteriormente los cuatro gráficos son utilizados para comparar los ingresos de hombres y mujeres el año 2002 y el de las mujeres entre los años 2002 y 2004.

```
. foreach i in 02 04  {
        twoway (qfit lngTrabHom`i' Edad20`i')(scatter lngTrabHom`i' Edad20`i'), ///
        ytitle (Ingreso Líquido Mensual) xtitle (Edad) title (Hombres 20`i') ///
        ylabel(200000 (200000) 1600000, angle(0) labsize(small)) ///
        xlabel(20 (5) 80,labsize(small)) ///
        name (IngresosHombres`i')

        twoway (qfit lngTrabMuj`i' Edad20`i')(scatter lngTrabMuj`i' Edad20`i'), ///
        ytitle (Ingreso Líquido Mensual) xtitle (Edad) title (Mujeres 20`i') ///
        ylabel(200000 (200000) 1600000, angle(0)labsize(small)) ///
        xlabel(20 (5) 80,labsize(small)) ///
        name (IngresosMujeres`i')
                        }


. graph combine IngresosHombres02 IngresosMujeres02, ///
         title (Nube de Puntos y Ajuste Cuadrático por Género (Ronda 1))
. graph combine IngresosMujeres02 IngresosMujeres04, ///
         title (Nube de Puntos y Ajuste Cuadrático Mujeres (Rondas 1 y 2))
```

Nube de Puntos y Ajuste Cuadrático por Género (Ronda 1)



Nube de Puntos y Ajuste Cuadrático Mujeres (Rondas 1 y 2)



Existen otras situaciones que el usuario de la base de datos desea transformar una variable continua en una discreta. Por ejemplo, hay análisis que requieren utilizar solo tres categorías que clasifiquen los ingresos líquidos mensuales. Las personas que ganan menos de Ch$200,000, las que reciben entre Ch$200,000 y Ch$400,0000, y las que tienen un sueldo líquido mayor a Ch$400,000. Stata permite hacer esta transformación utilizando el comando generate y replace.

Siguiendo la lógica de los ejercicios anteriores se entregan las sentencias en un *loop* usando el comando foreach. El circuito recursivo es llamado tres veces porque ahora se trabaja con la tercera ronda del panel además de la primera y la segunda. A continuación las sentencias en Stata que realizan lo señalado.

```
. sort id
. merge id using PanelMuestraR3.dta
. drop _merge
. rename IngresoTrabajador2006 IngTrab06


. foreach i in 02 04 06 {
        generate IngTrab_BMA = 1 if IngTrab`i' <= 200000
        replace  IngTrab_BMA = 2 if IngTrab`i' >  200000 & IngTrab`i' <= 400000
        replace  IngTrab_BMA = 3 if IngTrab`i' >  400000 & IngTrab`i' < .
                                }
```

Al tabular dos variables discretas se pueden observar los cambios que presenta la muestra panel entre una medición y otra. Por ejemplo, si se tabulan las nuevas variables de remuneración de los trabajadores en la segunda y tercera ronda se obtiene una matriz de transición de seis celdas. La primera celda (fila y columna uno) entrega información de dos momentos en el tiempo. En el ejercicio realizado, 126 casos de la muestra panel tuvieron ingresos líquidos mensuales menores a Ch$200,000 el año 2004 y también el 2006.

```
. tabulate IngTrab_BMA04 IngTrab_BMA06
```

IngTrab_BMA04	IngTrab_BMA06			
	1	2	3	Total
1	**126**	**13**	**0**	**139**
2	**9**	**31**	**9**	**49**
3	**3**	**3**	**13**	**19**
Total	**138**	**47**	**22**	**207**

Etiquetar las variables permite una mejor lectura de las tabulaciones. En la sección anterior se señaló de qué manera Stata etiqueta las variables. Ahora se verá como se etiquetan los códigos de las nuevas variables. Siguiendo con el mismo ejercicio se muestra la forma que Stata etiqueta las categorías de las variables IngTrab_BMA02, IngTrab_BMA04 y IngTrab_BMA06. A continuación se entrega el *loop* con las sentencias para etiquetar las categorías.

```
. foreach i in 02 04 06 {
        label define IngTrab_BMA`i'_labels 1 "Bajo" 2 "Medio" 3 "Alto"
        label value IngTrab_BMA`i' IngTrab_BMA`i'_labels
                                }
```

El siguiente paso es etiquetar las nuevas variables y tabularlas para analizar las transiciones entre las rondas 1 y 2, y entre las rondas 2 y 3. Una opción útil que entrega el comando tabulate es row cell

nofreq. Esta especificación entrega los valores de la matriz de transición en porcentajes absolutos y relativos a nivel de filas.

```
. label var IngTrab_BMA02 "Ingreso mensual 2002"
. label var IngTrab_BMA04 "Ingreso mensual 2004"
. label var IngTrab_BMA06 "Ingreso mensual 2006"

. tabulate IngTrab_BMA02 IngTrab_BMA04, row cell nofreq
. tabulate IngTrab_BMA04 IngTrab_BMA06, row cell nofreq
```

Ingreso mensual 2002	Ingreso mensual 2004			
	Bajo	Medio	Alto	Total
Bajo	**82.48**	**16.06**	**1.46**	**100.00**
	56.22	**10.95**	**1.00**	**68.16**
Medio	**32.08**	**58.49**	**9.43**	**100.00**
	8.46	**15.42**	**2.49**	**26.37**
Alto	**0.00**	**0.00**	**100.00**	**100.00**
	0.00	**0.00**	**5.47**	**5.47**
Total	**64.68**	**26.37**	**8.96**	**100.00**
	64.68	**26.37**	**8.96**	**100.00**

Ingreso mensual 2004	Ingreso mensual 2006			
	Bajo	Medio	Alto	Total
Bajo	**90.65**	**9.35**	**0.00**	**100.00**
	60.87	**6.28**	**0.00**	**67.15**
Medio	**18.37**	**63.27**	**18.37**	**100.00**
	4.35	**14.98**	**4.35**	**23.67**
Alto	**15.79**	**15.79**	**68.42**	**100.00**
	1.45	**1.45**	**6.28**	**9.18**
Total	**66.67**	**22.71**	**10.63**	**100.00**
	66.67	**22.71**	**10.63**	**100.00**

Para cerrar esta sección sobre transformación de variables continuas a discretas se mostrará el uso del comando xtile en Stata. En algunas ocasiones el usuario de las bases de datos panel necesita de una variable que identifique la posición del trabajador en la distribución de ingreso para cada ronda. Se denomina quintil cuando la clasificación corresponde al 20 por ciento de la distribución y decil cuando es el diez por ciento. Por ejemplo, si el usuario está interesado en trabajar con quintiles de ingresos el comando xtile transforma el ingreso líquido mensual del trabajador en una variable discreta en tres etapas. Primero, ordena la variable (lngTrab) de menor a mayor. Luego, identifica sus quintiles dividiendo la distribución en 5 grupos de similar tamaño (nq(5)) y asigna el valor 1 a todas las personas del primer quintil (el 20 por ciento más bajo de la distribución). Finalmente, asigna los valores 2, 3, 4 y 5 a los siguientes grupos. A continuación se entrega el *loop* para las construcción de los quintiles en las cuatro rondas de la panel ficticia.

```
. sort id
. merge id using PanelMuestraR3.dta
. drop _merge
. rename IngresoTrabajador2009 lngTrab09

. foreach i in 02 04 06 09 {
        xtile ingresoquintil`i' = lngTrab`i' if lngTrab`i' < ., nq(5)
                                }
```

Al tabular las rondas 3 y 4 utilizando la opción row cell nofreq se observan las permanencias y transiciones de la muestra panel entre los años 2006 y 2009. Por ejemplo, entre las celdas de la diagonal de la matriz, la celda del quintil 5 presenta el porcentaje más alto. Esto significa que el 67.5 por ciento de las personas que estaban en el quintil con mayores ingresos el año 2006 se mantuvieron en el mismo quintil tres años después.

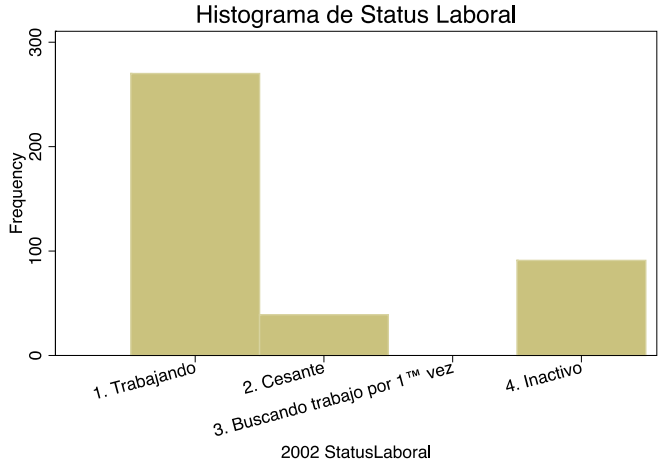
```
. tabulate ingresoquintil06 ingresoquintil09, row cell nofreq
```

5 quantiles of IngTrab06	5 quantiles of IngTrab09					
	1	2	3	4	5	Total
1	**52.94**	**32.35**	**5.88**	**5.88**	**2.94**	**100.00**
	9.14	**5.58**	**1.02**	**1.02**	**0.51**	**17.26**
2	**20.51**	**48.72**	**17.95**	**10.26**	**2.56**	**100.00**
	4.06	**9.64**	**3.55**	**2.03**	**0.51**	**19.80**
3	**19.23**	**26.92**	**38.46**	**13.46**	**1.92**	**100.00**
	5.08	**7.11**	**10.15**	**3.55**	**0.51**	**26.40**
4	**3.12**	**0.00**	**15.62**	**59.38**	**21.88**	**100.00**
	0.51	**0.00**	**2.54**	**9.64**	**3.55**	**16.24**
5	**5.00**	**0.00**	**0.00**	**27.50**	**67.50**	**100.00**
	1.02	**0.00**	**0.00**	**5.58**	**13.71**	**20.30**
Total	**19.80**	**22.34**	**17.26**	**21.83**	**18.78**	**100.00**
	19.80	**22.34**	**17.26**	**21.83**	**18.78**	**100.00**

2.2.6 Graficar variables discretas

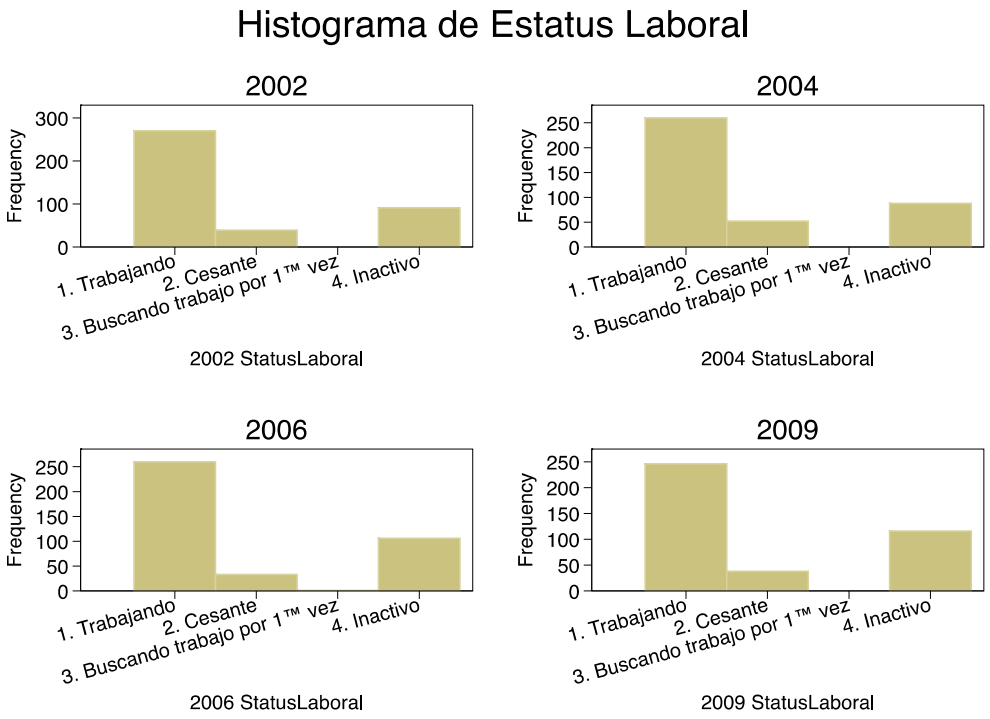
Para el caso de las variables discretas uno de los gráficos más útiles son los histogramas. Por ejemplo, la variable estatus laboral (StatusLaboral2002) puede graficarse usando el comando histogram y haciendo la especificación que es una variable discreta (discrete). La siguiente programación entrega la frecuencia de cada categoría y también incluye las etiquetas de la variable y el título del gráfico.

```
. histogram StatusLaboral2002, ///
            discrete freq xlabel (1(1)4, valuelabel angle(15)) ///
            ylabel(, angle(0)) ///
            title (Histograma de Status Laboral)
```



Utilizando los comandos foreach y combine explicados en la sección 2.2.5 es posible replicar la última sintaxis pero ahora para las cuatro rondas de la panel ficticia. De esta manera es posible mostrar los gráficos de la variable estatus laboral en los años 2002, 2004, 2006 y 2009 en un solo gráfico. A continuación se entregan las instrucciones descritas.

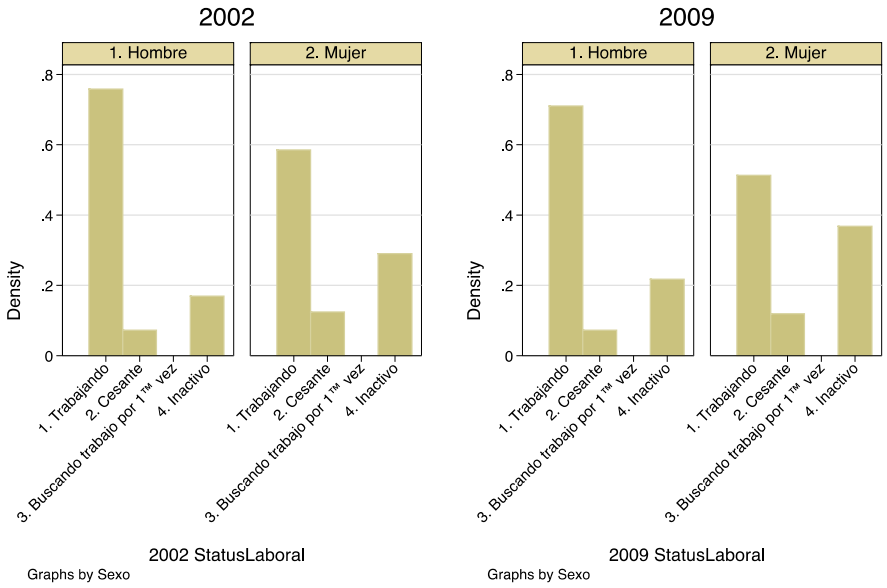
```
. foreach i in 02 04 06 09 {
        histogram StatusLaboral20`i', ///
            discrete freq xlabel (1(1)4, valuelabel angle(15)) ///
            ylabel(, angle(0)) ///
            name (EstatusLaboral`i') ///
            title (20`i')
                                    }

. graph combine EstatusLaboral02 EstatusLaboral04 ///
                                EstatusLaboral06 EstatusLaboral09, ///
                                title (Histograma de Estatus Laboral en Cuatro Rondas)
```



El histograma de una variable discreta puede cruzarse con otra variable dicotómica como el género de los entrevistados (Sexo). Por ejemplo el usuario podría estar interesado en hacerlo para las siguientes variables discretas: estatus laboral de las personas que componen la muestra (StatusLaboral), el ingreso líquido mensual por nivel (lngTrab_BMA), y quintiles de ingreso (ingresoquintil). A continuación se entregan las líneas de programación para generar los gráficos entre hombres y mujeres y comparar los cambios entre rondas. Los valores del eje y de los gráficos se entregan en términos de densidad y no de frecuencia.

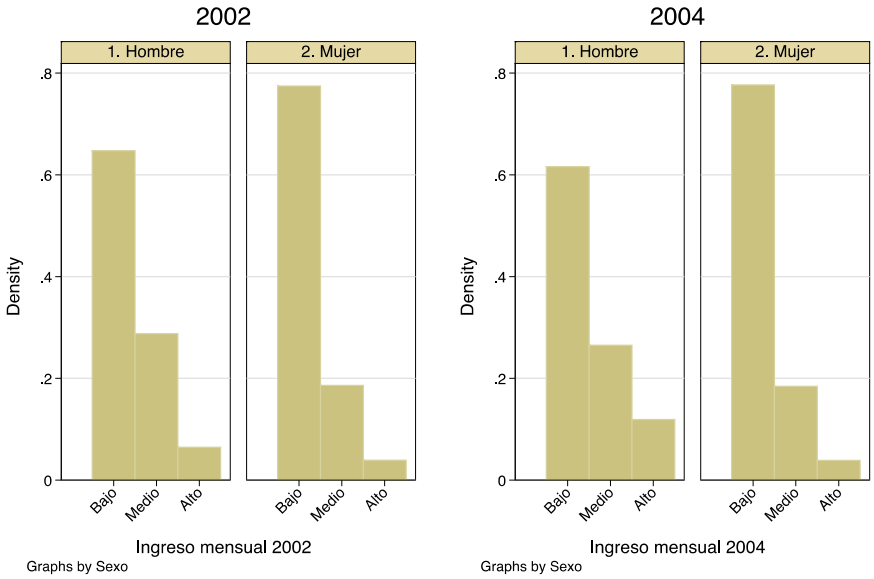
```
. foreach i in 02 04 06 09 {
        histogram StatusLaboral20`i', ///
            discrete xlabel (1(1)4, valuelabel angle(45)) ///
            ylabel(, angle(0)) ///
            by (Sex, title (20`i')) ///
            name (EstLabGen`i')
                }

. graph combine EstLabGen02 EstLabGen09, ///
        title (Histograma de Estatus Laboral y Género (2002-2009))
```

Histograma de Estatus Laboral y Género (2002 - 2009)

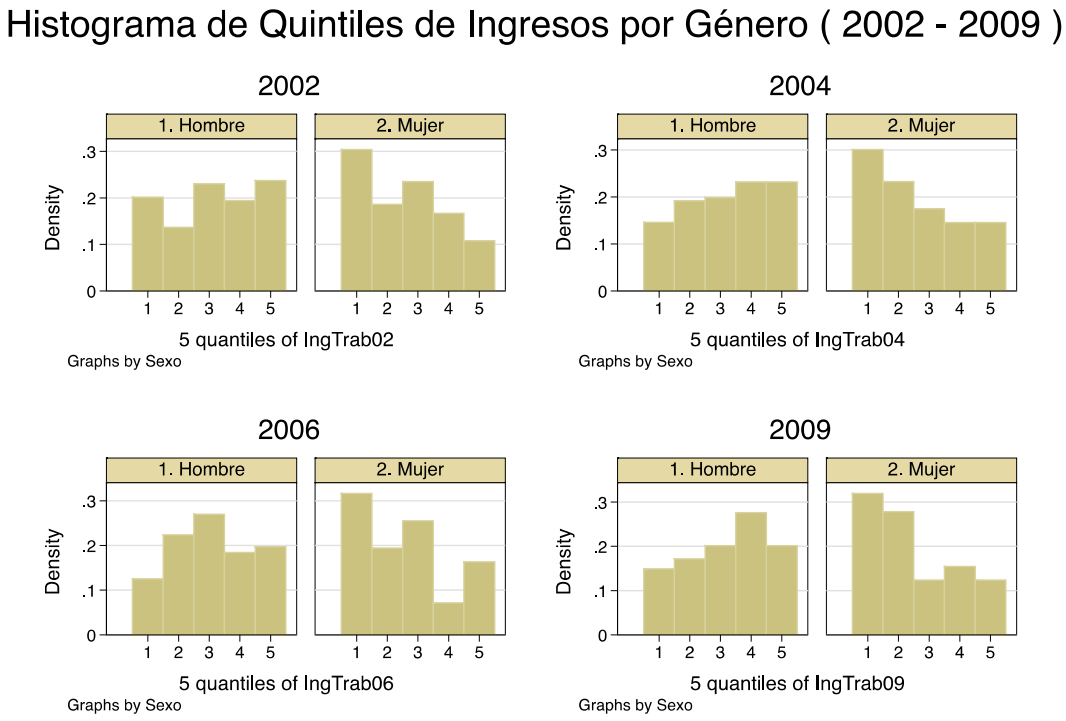



```
. foreach i in 02 04 {
        histogram lngTrab_BMA`i', ///
            discrete xlabel (1(1)3, valuelabel angle(45)) ///
            ylabel(, angle(0)) ///
            by (Sex, title (20`i')) ///
            name (IT_BMAGen`i')
                }

. graph combine IT_BMAGen02 IT_BMAGen04, ///
            title (Histograma de Rango de Ingresos por Género (2002-2004))
```

Histograma de Rango de Ingresos por Género (2002 - 2004)



```
. foreach i in 02 04 06 09 {
        histogram ingresoquintil`i', ///
            discrete xlabel (1(1)5, valuelabel angle(0)) ///
            ylabel(, angle(0)) ///
            by (Sex, title (20`i')) ///
            name (ingresoquintil`i')
                }

. graph combine ingresoquintil02 ingresoquintil04 ///
                ingresoquintil06 ingresoquintil09, ///
            title (Histograma de Quintiles de Ingresos por Género (2002-2009))
```



2.3 EL USO DE FACTORES DE EXPANSIÓN TRANSVERSALES Y LONGITUDINALES

En los ejemplos de la sección anterior se han usado los datos sin factores de expansión. Los factores de expansión o pesos estadísticos se utilizan para compensar el hecho que la probabilidad de selección de una unidad de la muestra es diferente de otra unidad. Por ejemplo, al no usar factores de expansión en los ejercicios anteriores se asume que el factor de expansión es 1, es decir, todas las unidades tienen la misma influencia en los análisis. En general, no siempre la probabilidad de selección de todas las unidades de la muestra es la misma. Además hay situaciones que es necesario ajustar los factores de expansión para balancear la influencia que las unidades van adquiriendo durante las rondas de un panel. Por lo tanto, los factores de expansión difieren entre las distintas unidades de la muestra permitiendo entregar resultados que dan cuenta de la población representada en cada medición (pesos transversales) y entre distintas rondas (pesos longitudinales).

Hay dos razones que explican por qué los factores de expansión son distintos de 1 y diferentes entre sí. Una razón está dada por la estrategia de diseño muestral utilizada en el estudio. La decisión del

diseño muestral guarda estrecha relación con el cálculo del factor de expansión de corte transversal o el peso estadístico de la unidad de la muestra en el año de medición. La otra razón está dada por las unidades (individuos) que van dejando de participar en las rondas posteriores y que tienen características diferentes de aquellas unidades que se mantienen en el panel. Los pesos estadísticos de las unidades entrevistadas en cada ronda que consideran el efecto de las unidades que abandonan la muestra se denominan factores de expansión longitudinales. A continuación se explica con mayor detalle cada uno de ellos y como trabajar como ambos factores de expansión en Stata.

2.3.1 Influencia del diseño muestral en los factores de expansión transversales

La elección del diseño muestral puede definir que los pesos estadísticos de las unidades muestrales sean distintos entre sí. Son muy escasas las encuestas que presentan un diseño muestral aleatorio simple (MAS) que permite pesos iguales para cada unidad de la muestra. El MAS consiste en una selección aleatoria, donde todas las unidades tuvieron igual probabilidad de selección y la selección de una unidad no afecta la probabilidad de selección de las otras unidades. Dado que el factor de expansión se calcula como el inverso de la probabilidad de selección de cada unidad, los pesos estadísticos de un MAS de n unidades de una población de N unidades es 1/n/N. Por ejemplo, si la población a estudiar es de 4,000,000 de personas (N) y el tamaño de la muestra es de 400 casos (n) el factor de expansión de cada persona de la muestra es 10,000 (1/400/4,000,000). De esta manera cada persona entrevistada representa a 10,000 personas de la población.

El ajuste de los presupuestos (costo del trabajo de campo) a los objetivos de los estudios explica en gran medida por qué los responsables de las encuestas sociales de cobertura nacional prefieren los diseños muestrales multietápicos (selección de la muestra en varias etapas) y uso de conglomerados[11] en vez de diseños MAS. En el caso de las encuestas de hogares es común seleccionar la muestra con un diseño de dos etapas. En la primera etapa se seleccionan las unidades primarias de muestreo con probabilidad proporcional al tamaño de su población y en la segunda se seleccionan las viviendas con igual probabilidad. Para este tipo de diseño los pesos siguen siendo los mismos para cada hogar. Si se considera una tercera etapa de selección en el diseño de la muestra, como es el caso de los estudios que buscan representatividad de una población determinada, los factores de expansión para cada miembro de la muestra ya no serán idénticos.

En el caso de las ELPS el diseño muestral es trietápico[12]. En la tercera etapa se selecciona a una persona del hogar entre todas las personas que viven en la vivienda. Como se mencionó, la probabilidad de selección de la vivienda es la misma para todas las viviendas pero no así para la persona seleccionada en cada vivienda[13]. Por ejemplo, si una vivienda seleccionada tiene un hogar compuesto por solo una persona, la probabilidad de selección de esa persona es 1 pero si en otra

[11] En general los diseños muestrales de encuestas de hogares seleccionan conglomerados de viviendas en la segunda etapa de selección, es decir, seleccionan viviendas que se encuentran cercas entre sí. La razón que justifica esta decisión es minimizar costos de traslado de encuestadores a las viviendas de la muestra. El problema del muestreo por conglomerado es que introduce un efecto diseño en los resultados de las estimaciones porque supone una pérdida de heterogeneidad de la muestra. Esto se explica porque los individuos en un mismo conglomerado son más parecidos entre sí (homogéneos) que los individuos en conglomerados diferentes.

[12] Solo la ELPS Chile no posee este diseño muestral porque el marco muestral utilizado fueron los registros administrativos de los afiliados al sistema previsional de capitalización individual.

[13] La selección de la persona a entrevistar en cada vivienda se realiza de manera aleatoria utilizando la Tabla de Kish. Este método de selección utiliza una tabla pre-asignada de números aleatorios para encontrar a la persona a entrevistar (Kish, 1949). Este método no se usa en todas las ELPS, en Colombia la ELPS entrevistó a todos los individuos del hogar.

vivienda de la muestra viven tres personas elegibles la probabilidad de selección de ese entrevistado será 0,333 (1/3). Por lo tanto, los pesos estadísticos de las personas seleccionadas con un diseño muestral de tres etapas serán distintos entre sí dependiendo del número total de personas que viven en los hogares seleccionados en la etapa anterior del diseño muestral.

Algunos diseños muestrales incorporan en la construcción del factor de expansión de cada ronda o peso de corte transversal, una etapa de post estratificación. La post estratificación consiste en ponderar los pesos estadísticos por las proyecciones oficiales de algunas variables como edad y sexo en el año de aplicación de la encuesta. En el caso de encuestas paneles la post estratificación es una técnica muy usada y va adquiriendo importancia en la medida que las rondas avanzan y los países van generando estadísticas oficiales.

2.3.2 Paneles balanceados y los factores de expansión longitudinales

Las encuestas longitudinales presentan bases de datos balanceadas y no balanceadas. La no respuesta del total de una encuesta o incluso la no respuesta de alguna pregunta del cuestionario provocan que las bases de datos sean desbalanceadas. Esto quiere decir que una unidad de la muestra o una variable específica no es observada en alguna ronda de la entrevista. Por ejemplo, si tenemos un panel de 600 personas entrevistadas durante cuatro años, se debería contar con cuatro entrevistas por cada una de las 600 personas. Sin embargo, esto habitualmente no sucede por el desgaste de la muestra (también conocido como atrición) y la no respuesta. Por ejemplo, el panel de 600 personas puede llegar a tener al final de la cuarta ronda 400 casos que respondieron la encuesta en las cuatro mediciones.

Para poder realizar análisis descriptivos que relacionen dos o más mediciones de una encuesta panel es necesario usar factores de expansión de las unidades que presentan información en el tiempo. Es decir se necesita utilizar una base de datos balanceada donde cada unidad de la muestra disponga de datos en todas las ronda analizadas. Por ejemplo, si se desea analizar los cambios del nivel de cotización en un país entre la primera y segunda ronda se necesita de una base de datos balanceada para los dos periodos. Una base de datos no balanceada se transforma en una base de datos panel balanceada eliminando de la base de datos las personas que no fueron entrevistadas en los años que se desea analizar o botando los casos donde la variable estudiada no presenta información o algún tipo de imputación. Siguiendo con el ejemplo de la base de datos de 600 personas se deben eliminar 200 casos para disponer de una base de datos balanceada de 400 unidades[14]. Con esta base de datos balanceada se pueden conocer los cambios de los niveles de cotización de la población en el periodo estudiado (cuatro rondas).

El problema de eliminar datos no solo es generar estimaciones menos precisas por contar con una muestra más pequeña que la original. También las estimaciones pueden estar sesgadas si los individuos que no se pueden entrevistar difieren en algunas características de los casos que se mantienen en la base de datos balanceada. Por ejemplo, una encuesta puede tener una alta participación de las mujeres entre 50 y 60 años, y un bajo porcentaje de hombres entre 30 y 40 años que respondieron la encuesta. Por lo tanto, el factor de expansión de las unidades sobrerepresentadas (mujeres entre 50 y 60 años) debiera tener un peso menor que en la primera ronda y las unidades subrepresentadas tendrán un peso mayor (hombres entre 30 y 40 años). Los pesos que corrigen este sesgo se conocen como factores de longitudinales. En general, el cálculo de los factores longitudinales es el inverso de las probabilidades de responder usando un modelo de respuesta (Kalton & Brick, 2000). Para estimar los modelos de respuesta es necesario tener

[14] En los ejemplos de la sección 2.2 se ha usado una base de datos balanceadas.

información de los que responden todas las rondas analizadas como los que no lo hacen en alguna de ellas. Por lo tanto, en las encuestas panel la última información de los individuos que abandonaron la muestra es utilizada en los modelos de respuesta[15] (Paredes, Prieto, & Zubizarreta, 2006).

2.3.3 Uso de los factores transversales y longitudinales en Stata

El uso de los factores transversales permite hacer análisis descriptivos de la población para cada año de medición. Los factores de expansión de una muestra que representa a un corte transversal de la población objetivo deben estar incluidos en las bases de datos no balanceadas. Así, cada ronda del panel debiera tener su peso transversal. Por lo tanto, una encuesta panel de cuatro rondas debiera tener una base de datos no balanceada con cuatro factores de expansión de corte transversal.

A continuación se carga en la memoria la base de datos ficticia no balanceada (PanelMuestraR1234NoBal.dta) y se dejan solamente las variables id, Secuencia y FactorTransv* en la base de datos.

```
. use PanelMuestraR1234NoBal.dta, clear
. keep id Secuencia FactorTransv*
```

Con el objetivo de mostrar la estructura de la base de datos no balanceada se tabula la variable Secuencia. La variable Secuencia entrega información de la historia de entrevistas de cada entrevistado. Existen cuatro secuencias posibles para cada miembro de la muestra inicial. La persona pudo haber sido entrevistada cuatro veces (secuencia XXXX)[16]. Se entrevistó a la persona en las tres primeras rondas (XXXO). La persona respondió la encuesta en la primera y segunda ronda (XXOO) o solo tiene una medición al inicio del panel (XOOO).

```
tabulate Secuencia
```

Historia de Entrevistas	Freq.	Percent	Cum.
XOOO	**87**	**14.50**	**14.50**
XXOO	**64**	**10.67**	**25.17**
XXXO	**49**	**8.17**	**33.33**
XXXX	**400**	**66.67**	**100.00**
Total	**600**	**100.00**	

Los datos de la base de datos ficticia indican que entre la primera y segunda ronda 87 personas no respondieron la encuesta. Por lo tanto la tasa de no respuesta fue de 14.50 por ciento. Entre la ronda dos y tres la atrición o tasa de no respuesta fue de 10.67 por ciento. La atrición en la cuarta ronda en

[15] Las variables que generalmente se utilizan para comparar a los individuos que no respondieron la última ronda con los que sí lo hicieron son a nivel de hogar (tipo de vivienda, propiedad de la vivienda, características de la vivienda, hogar unipersonal, decil de ingreso per cápita), variables geográficas, y variables personales (edad, sexo, educación, salud, situación ocupacional y variables laborales).

[16] La base de datos ficticia que se ha utilizado hasta el momento es una base de datos balanceada de 400 casos, es decir, incluye solo los casos que han sido entrevistados en las cuatro rondas (secuencia XXXX).

relación a la tercera fue de 8.17 por ciento. Por lo tanto, la tasa de respuesta después de cuatro mediciones es de 66.67 por ciento quedando 400 casos de los 600 iniciales.

El comando codebook con la especificación compact entrega información similar al comando summarize pero agregando la etiqueta de la variable.

```
. codebook, compact
```

Variable	Obs	Unique	Mean	Min	Max	Label
id	**600**	**600**	**1300.5**	**1001**	**1600**	Identificador numerico entrevistado
Secuencia	**600**	**4**	**.**	**.**	**.**	Historia de Entrevistas
FactorT~2002	**600**	**86**	**4454.4**	**3010**	**18370**	Factor Transversal 2002
FactorT~2004	**513**	**86**	**4688.46**	**2900**	**24030**	Factor Transversal 2004
FactorT~2006	**449**	**298**	**4982.762**	**930**	**24720**	Factor Transversal 2006
FactorT~2009	**400**	**295**	**5210.45**	**360**	**36260**	Factor Transversal 2009

De esta manera se observa que el número de factores transversal para la primera ronda es 600, en la segunda medición es 513 y en la tercera y cuarta ronda es 449 y 400 respectivamente. Como era de esperar la media de los factores de expansión va aumentando en cada ronda desde 4,454.40 el año 2002 a 5,210.45 el año 2009. En la cuarta ronda el factor transversal tiene un rango mayor. El peso menor representa a 360 personas y el mayor a 36,260 personas de la población objetivo.

Ahora bien, cualquier estimación de la media o proporción de una variable que represente a la población bajo estudio deberá usar el factor de expansión transversal correspondiente. Hacerlo en Stata es bastante simple. Solo se agrega el nombre del factor de expansión al comando pweight después de escribir el análisis descriptivo[17]. Por ejemplo, calcular el nivel de cotización el año 2009 utilizando los datos de la cuarta ronda del panel ficticio necesita del comando proportion, la variable CotizaActual2009 y el factor de expansión transversal para ese año.

```
. use PanelMuestraR1234NoBal.dta, clear
. proportion CotizaActual2009 [pweight = FactorTransv2009]

Proportion estimation              Number of obs    =     357
```

	Proportion	Std. Err.	[95% Conf.	Interval]
CotizaActual2009				
SÌ	**.535268**	**.0296194**	**.477017**	**.5935191**
No	**.464732**	**.0296194**	**.4064809**	**.522983**

Con el fin de mostrar la importancia de usar los factores de expansión para los análisis descriptivos se realiza otro cálculo del nivel de cotización para el año 2009 con la misma variable pero sin los

[17] Para los pesos longitudinales la lógica del comando es la misma: instruccion_stata [pweight = peso_longitudinal].

pesos transversales. Al ocupar datos a nivel muestral de la cuarta ronda solo es necesario tabular la variable CotizaActual2009 para obtener el resultado.

```
. tabulate CotizaActual2009
```

2009 CotizaActual	Freq.	Percent	Cum.
SÌ	**179**	**50.14**	**50.14**
No	**178**	**49.86**	**100.00**
Total	**357**	**100.00**	

Al comparar ambas tablas se observan dos diferencias. La tasa de cotización estimada utilizando los factores de expansión es 53,5 por ciento y sin los pesos transversales el nivel de cotización el año 2009 es 50,1 por ciento. La segunda diferencia es que los valores de la muestra usando el comando tabulate no entregan un intervalo de confianza para el parámetro estimado. Sí lo hace el comando proportion al usar los factores de expansión transversal. De esta manera, se sabe que con un 95 por ciento de confianza las personas que declaran estar cotizando el 2009 se encuentran entre un 48 y 59 por ciento, y los que responden que no están cotizando en esos momentos pueden estar entre un 41 y un 52 por ciento. Cabe señalar que trabajar con tamaños de muestra más grande permite tener intervalos de confianza más estrechos.

Es importante mencionar que en Stata, algunos análisis que consideran la muestra expandida necesitan especificar pesos analíticos (aw) en lugar de pesos muestrales (pweight). Los pesos analíticos se usan para trabajar con las medias de las variables. Por ejemplo, al realizar una regresión lineal con factores de expansión se necesita utilizar la opción aw en vez de pweight. No se recomienda usar los pesos analíticos para el cálculo de la varianza de las estimaciones porque utiliza un método menos robusto y preciso que el usado con los pesos muestrales.

Hay comandos en Stata que entregan estimaciones de variables a nivel poblacional usando solo los pesos analíticos pero sin calcular los intervalos de confianza. Ese es el caso del comando tabulate que es de gran utilidad para los análisis de las matrices de transición con datos longitudinales como se verá en el capítulo 2. Para ilustrar este punto, se calcula el nivel de cotización el año 2009 usando el comando tabulate y los pesos analíticos aw. Se observa que el resultado de la estimación es el mismo que el entregado por el comando proportion y los pesos muestrales (pweight).

```
. tabulate CotizaActual2009 [aw = FactorTransv2009]
```

2009 CotizaActual	Freq.	Percent	Cum.
SÌ	**204.548563**	**57.30**	**57.30**
No	**152.451437**	**42.70**	**100.00**
Total	**357**	**100.00**	

2.3.4 Ejercicio 1: Determinar cuántas personas tienen algún tipo de seguridad social en la primera ronda de medición y cómo ha evolucionado en el tiempo

Se carga la base de datos no balanceadas y se mantiene la variable que identifica a cada persona (id), la historia de entrevistas (Secuencia), los factores transversales (FactorTransv*) y si está cotizando en el momento de la entrevista (CotizaActual*).

```
. use PanelMuestraR1234NoBal.dta, clear
. keep id Secuencia CotizaActual* FactorTransv*
```

Se revisa si la variable CotizaActual2002 contiene valores que explican la no respuesta.

```
. tabulate CotizaActual2002
```

2002 CotizaActual	Freq.	Percent	Cum.
Si	**271**	**45.17**	**45.17**
No	**329**	**54.83**	**100.00**
Total	**600**	**100.00**	

Se recodifican las categorías No aplica, No responde y No sabe con un punto (.) para que Stata reconozca que son datos faltantes. Luego se estima la proporción de personas que cotizan el año 2002 entre los que respondieron esa pregunta.

```
. recode CotizaActual2002 (7777 = .) (8888 = .) (9999 = .)
. proportion CotizaActual2002 [pweight = FactorTransv2002]
```

	Proportion	Std. Err.	[95% Conf.	Interval]
CotizaActual2002				
Si	**.7337786**	**.0238385**	**.6869015**	**.7806557**
No	**.2662214**	**.0238385**	**.2193443**	**.3130985**

Con un 95 por ciento de confianza las personas que declaran estar cotizando se encuentran entre un 69 y 78 por ciento, y los que responden que no están cotizando en esos momentos pueden estar entre un 22 y un 32 por ciento.

Bajo el supuesto que en los casos donde las categorías No aplica, No responde y No sabe corresponden a casos en que la persona no está cotizando, se recodifica la variable CotizaActual2002 y se estima la proporción de cotizante nuevamente.

```
. replace CotizaActual2002 = 2 if CotizaActual2002 == .
. proportion CotizaActual2002 [pweight = FactorTransv2002]
```

```
Proportion estimation              Number of obs    =     600
```

	Proportion	Std. Err.	[95% Conf.	Interval]
CotizaActual2002				
Si	**.451778**	**.0210796**	**.410379**	**.493177**
No	**.548222**	**.0210796**	**.506823**	**.589621**

Ahora utilizando información de los 600 casos, la proporción que cotiza se encuentra en un intervalo de confianza entre 41 y 49 por ciento, y los que no cotizan su parámetro se encuentra entre el 51 y 59 por ciento.

Repitiendo la misma estrategia de recodificación de la variable CotizaActual para las siguientes ronda es posible observar en cuánto aumentó la cotización de una medición a otra. Con el objetivo de ser eficiente en la programación se utiliza el comando foreach. De esta manera se diseña un *loop* para los años 2004, 2006 y 2009, tal como mostró en la sección 2.2.5.

```
. foreach i in 04 06 09 {
        replace CotizaActual20`i' = 2 if CotizaActual20`i' == .
        proportion CotizaActual20`i' [pweight = FactorTransv20`i']
                            }
```

	Proportion	Std. Err.	[95% Conf.	Interval]
CotizaActual2004				
SÌ	**.4686344**	**.0233123**	**.4228347**	**.514434**
No	**.5313656**	**.0233123**	**.485566**	**.5771653**

```
Proportion estimation              Number of obs    =     449
```

	Proportion	Std. Err.	[95% Conf.	Interval]
CotizaActual2006				
SÌ	**.5047737**	**.0264694**	**.452754**	**.5567934**
No	**.4952263**	**.0264694**	**.4432066**	**.547246**

```
Proportion estimation              Number of obs    =     400
```

	Proportion	Std. Err.	[95% Conf.	Interval]
CotizaActual2009				
SÌ	**.5155169**	**.0290695**	**.4583684**	**.5726653**
No	**.4844831**	**.0290695**	**.4273347**	**.5416316**

3. Análisis de datos longitudinales sobre seguridad social

En este capítulo se muestran algunos ejemplos de análisis dinámicos utilizando datos panel sobre seguridad social. En general, los comandos de Stata que se necesitan son los mismos del capítulo anterior. Los comandos nuevos serán explicados durante el desarrollo de los ejercicios. El objetivo sigue siendo conocer las funciones de Stata al mismo tiempo que aprender a realizar análisis en el tiempo usando datos longitudinales. Por ejemplo el comando collapse se usa para preparar la base de datos que estima la densidad de cotización de la población estudiada, o el comando tabulate con la opción aw (pesos analíticos) para calcular las matrices de transición, el comando regress al estimar el efecto de un programa social en sus beneficiarios.

3.1 ANÁLISIS DE LA DENSIDAD DE COTIZACIÓN

3.1.1 Ejercicio 2: Determinar cuál es la densidad de cotización de una población en el tiempo

Un set de preguntas importante en la ELPS son las relacionadas con el estatus laboral de la persona entre una entrevista y otra. Por ejemplo, en la base de datos que se utiliza a continuación se registran todos los meses que el entrevistado estuvo trabajando, cesante o inactivo entre los años 2006 y 2009 (rondas tres y cuatro). La información recogida en esas preguntas se guarda en una base de datos formato "long" que se llama historia laboral. En general, la base de datos de la historia laboral se construye separada de las bases de datos panel que se han visto en los ejemplos del capítulo 2. Por lo tanto, es necesario juntar ambas bases de datos cuando los análisis consideren usar algunas características de la historia laboral de las personas entrevistadas. Por ejemplo, comparar la densidad de cotización entre hombres y mujeres durante dos entrevistas necesita unir dos bases de datos: la que tiene la información del género de la persona con la que contiene su historia laboral. Pero antes de empezar a responder la pregunta del ejercicio 2 es importante entender como se estructura una base de datos de historia laboral.

Una vez cargada la base de datos de historia laboral de las rondas 2 y 3 de la panel ficticia, se describen las variables utilizando el comando codebook y se listan los primeros 15 casos de la base de datos.

```
. use HistoriaLaboralR34.dta, clear
. codebook, compact
. list id NumCambEstat MesInicio AnhoInicio EstatusLaboral ///
        MesTermino AnhoTermino in 1/15
```

Variable	Obs	Unique	Mean	Min	Max	Label
id	**661**	**400**	**1210.669**	**1001**	**1400**	Identificador numerico entrevistado
NumCambEstat	**661**	**12**	**2.018154**	**1**	**12**	Numero de cambios de estatus laboral ent...
MesInicio	**661**	**12**	**2.721634**	**1**	**12**	Mes de inicio del estatus laboral
AnhoInicio	**661**	**4**	**2006.702**	**2006**	**2009**	Ano de inicio del estatus laboral
EstatusLab~l	**661**	**4**	**1.987897**	**1**	**4**	Estatus laboral de la persona durante es...
MesTermino	**661**	**12**	**6.624811**	**1**	**12**	Mes de termino del estatus laboral
AnhoTermino	**661**	**4**	**2008.425**	**2006**	**2009**	Ano de termino del estatus laboral
CotizaMes	**355**	**2**	**1.264789**	**1**	**2**	Cotiza durante ese mes

Al ser una base de datos en formato "long" el número de observaciones es mayor al número de casos. El identificador numérico del entrevistado (id) indica que son 400 personas pero se contabilizan más de 221 cambios de estatus laboral en la base de datos (400+221=661 observaciones). Hay un máximo de 12 cambios de estatus laboral durante el periodo analizado (2006-2009). Los rangos de los meses (12 meses) y años (4) están dentro de los valores esperados y se dispone de información de situación de cotización para 355 condiciones o estatus laborales.

	id	NumCam~t	MesIni~o	AnhoIn~o	Estatu~l	MesTer~o	AnhoTe~o
1.	1001	1	enero	2006	trabajan	junio	2009
2.	1002	1	enero	2006	trabajan	noviembr	2008
3.	1002	2	diciembr	2008	trabajan	julio	2009
4.	1003	1	enero	2006	trabajan	agosto	2006
5.	1003	2	septiemb	2006	cesante	diciembr	2006
6.	1003	3	enero	2007	trabajan	agosto	2007
7.	1003	4	septiemb	2007	cesante	diciembr	2007
8.	1003	5	enero	2008	trabajan	agosto	2008
9.	1003	6	septiemb	2008	cesante	diciembr	2008
10.	1003	7	enero	2009	trabajan	marzo	2009
11.	1003	8	abril	2009	cesante	junio	2009
12.	1004	1	enero	2006	trabajan	julio	2009
13.	1005	1	enero	2006	trabajan	julio	2009
14.	1006	1	enero	2006	inactivo	junio	2009
15.	1007	1	enero	2006	inactivo	junio	2009

Al listar los primeros 15 casos de la base de datos de historia laboral se observa que la persona con el id 1001 estuvo trabajando sin interrupción entre enero de 2006 y junio de 2009. Esta base de datos ficticia considera que el mes de la entrevista de la tercera ronda fue en enero de 2006 y el mes de la medición de la cuarta ronda entre junio y agosto de 2009. La persona con el id 1002 estuvo trabajando entre enero de 2006 y julio de 2009 pero presenta un cambio de empleo en diciembre de 2008 (tercera fila del listado). El miembro de la muestra panel con el id 1003 tiene 8 cambios de condición laboral pasando de trabajar a estar cesante en cuatro oportunidades (filas 4 a 11). Las personas con los id 1003 y 1004 declararon que estuvieron trabajando entre enero de 2006 y julio de 2009 (55 meses fue la duración de ese estatus laboral). Las personas con id 1006 y 1007 presentan inactividad entre enero de 2006 y junio de 2009 (equivalente a 54 meses).

La información que reporta la base de datos de historia laboral permite construir la densidad de cotización para cada miembro de la muestra panel. La densidad de cotización es la razón entre el número de meses que el trabajador ha cotizado y el número de meses potenciales de cotización. Por lo tanto, para calcular la densidad de cotización en primer lugar se debe crear una variable que contenga el total de meses que dura el estatus laboral declarado por el entrevistado. Una manera rápida de construir en Stata la variable que tiene el total de meses de cada condición laboral (MesesEstLab)es usando el comando generate y la especificación ym para un año y mes dado. Stata genera una variable con el total de meses desde enero de 1960 hasta la fecha dada. Por ejemplo, para enero de 2006 Stata entrega 552 meses. De esta manera la variable MesesEstLab es la diferencia

entre el número de meses del término de la condición laboral declarada en la entrevista y el inicio de ese mismo estatus laboral más un mes adicional.

```
. generate MesesDesdeEnero1960_Inicio = ym(AnhoInicio, MesInicio)
. generate MesesDesdeEnero1960_Termino = ym(AnhoTermino, MesTermino)
. generate MesesEstLab = ///
        MesesDesdeEnero1960_Termino - MesesDesdeEnero1960_Inicio < + 1
```

Con la variable MesesEstLab se crean las variables que entregan la duración (en número de meses) de las tres categorías de estatus laboral (trabajando, cesante, inactivo) además del número de meses que la persona estuvo cotizando.

```
. generate MesesTrab = MesesEstLab if EstatusLaboral==1
. replace MesesTrab=0 if  MesesTrab==.
. generate MesesCesan = MesesEstLab if EstatusLaboral==2 | EstatusLaboral ==3
. replace MesesCesan =0 if  MesesCesan ==.
. generate MesesInact = MesesEstLab if EstatusLaboral==4
. replace MesesInact =0 if  MesesInact ==.
. generate MesesCotiz = MesesEstLab if CotizaMes==1
. replace MesesCotiz =0 if  MesesCotiz ==.
```

Con el comando collapse Stata condensa la información de cada identificador numérico (id) en una sola observación por persona. La especificación sum en las variables creadas indica que deben sumarse los casos correspondientes al mismo id de la persona.

```
. collapse (sum) MesesEstLab (sum) MesesTrab (sum) MesesCesan ///
        (sum) MesesInact (sum) MesesCotiz, by(id)
```

Se revisa la estructura de la nueva bases de datos con el comando summarize. Ahora la base de datos tiene 400 casos correspondiente a todas las personas que entregaron información en la cuarta ronda. El rango de meses entre la cuarta y tercera entrevista se encuentra entre 40 y 46 meses. A nivel muestral el promedio de meses trabajando y desempleado es 23 y 4,6 respectivamente. La media de los meses cotizando de la muestra de 400 casos es 17.

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	**400**	**1200.5**	**115.6143**	**1001**	**1400**
MesesEstLab	**400**	**42.2725**	**1.143027**	**40**	**46**
MesesTrab	**400**	**22.9825**	**19.75418**	**0**	**46**
MesesCesan	**400**	**4.5475**	**11.41769**	**0**	**44**
MesesInact	**400**	**14.7425**	**19.5283**	**0**	**46**
MesesCotiz	**400**	**17.105**	**19.55556**	**0**	**46**

Se junta la actual base de datos de historia laboral con la base de datos que tiene el factor de expansión longitudinal (PanelMuestraR1234Wide.dta)

```
. merge id using PanelMuestraR1234Wide.dta
. keep id MesesEstLab MesesTrab MesesCesan MesesInact MesesCotiz FactorLong
```

A partir de la definición del concepto de densidad de cotización se crea la variable DensCotiz. Usando los factores de expansión longitudinales se calcula la media de los meses cotizados sobre el

total de duración de la historia laboral analizada, es decir se estima la densidad de cotización para la población estudiada entre la ronda tres y cuatro.

```
. generate DensCotiz= MesesCotiz/MesesEstLab*100
. mean DensCotiz [pweight = FactorLong]

Mean estimation                   Number of obs    =       400
```

	Mean	Std. Err.	[95% Conf.	Interval]
DensCotiz	**39.67768**	**2.461032**	**34.83947**	**44.51589**

La densidad de cotización promedio entre los años 2006 y 2009 es 39,7 por ciento. En otras palabras, la población cotiza en el sistema de pensiones un 39,7 por ciento del tiempo. Es importante señalar que en estricto rigor el resultado indica que con un 95 por ciento de seguridad la densidad de cotización está comprendida entre 34,9 y 44,5 por ciento. El amplio rango se explica principalmente por el tamaño de la muestra de la base de datos ficticia.

3.2 ANÁLISIS DE MATRICES DE TRANSICIONES

En los siguientes dos ejercicios se calcula el porcentaje de personas que han cambiado su estatus laboral entre empleo y sin empleo durante los años 2002 y 2009, y las personas que se han movido desde un trabajo formal a otro informal (y viceversa) durante el mismo periodo. La población objetivo es la población longitudinal del año 2002, vale decir, que está presente en las cuatro mediciones de la encuesta panel. Se utilizan los datos de la panel ficticia y la definición de "sin empleo" corresponde a todos aquellas personas que están cesantes, inactivas o buscan empleo por primera vez. En el segundo ejercicio se considera como empleo informal aquel trabajo que no está relacionado con una cotización del trabajador en el sistema de pensiones.

Ambos casos ilustran el máximo potencial que ofrecen las bases de datos longitudinales. El análisis de las dinámicas laborales y las dinámicas del trabajo informal entregan información que permanece oculta en los datos de las encuestas de empleo de corte transversal. De esta manera, los datos paneles permiten estimar el riesgo que tiene una persona de estar desempleada en la siguiente ronda de medición o la probabilidad que tiene un trabajador informal de llegar a tener un empleo formal más adelante en el tiempo.

En general, los resultados de este tipo de análisis dinámico se visualizan en matrices de transición. Las matrices de transición son tabulaciones que muestran en las filas de la matriz la condición del fenómeno para el año x. En las columnas de la matriz se observan los cambios de la variable estudiada para el año $x+t$, donde t corresponde al número de años entre las dos mediciones del panel. Por lo tanto, una matriz de transición de dos rondas tiene cuatro celdas o combinaciones. Por ejemplo, si se representa con un 1 a las personas que están empleadas y con un 0 a las que no tienen trabajo, las cuatro combinaciones de la matriz de transición son: i) las personas que estuvieron empleadas en las dos mediciones (11), ii) las que pasaron de estar empleada a estar sin trabajo (10), iii) las que no tenían empleo y continuaron sin estar trabajando en la siguiente ronda (00), y iv) las personas que dejaron de estar sin empleo porque encontraron trabajo (01). Para el caso de las personas que estaban en un empleo informal (0) y pasaron a un empleo formal (1) las cuatro celdas de la matriz de transición se configuran de manera similar.

Como se ha explicado en la sección 2.3 del capítulo 2 los análisis dinámicos se realizan con datos de la población inicial que es encuestada a lo largo del tiempo. Por lo tanto, se consideran como parte de la base de datos balanceada solo las personas que fueron entrevistadas (en la panel ficticia) en las cuatro mediciones (2002, 2004, 2006 y 2009). De esta manera, los individuos que no participaron en la ronda dos o tres, pero sí en la primera y en la cuarta ronda del panel no son parte de la población longitudinal entre el 2002 y el 2009.

3.2.1 Ejercicio 4: Determinar cuál es el flujo de estar empleado a sin empleo (y viceversa) en un periodo determinado

Se carga en la memoria la base de datos balanceada para las cuatro rondas y se dejan las variables que dan cuenta del identificador de la persona, su estatus laboral y factor de expansión longitudinal.

```
. use PanelMuestraR1234Wide.dta, clear
. keep id StatusLaboral* CotizaActual* FactorLong
. codebook, compact
```

```
Variable      Obs Unique      Mean   Min   Max  Label
──────────────────────────────────────────────────────────────────────────────
id            400    400    1200.5  1001  1400  Identificador numerico entrevistado
StatusL~2002  400      3      1.78     1     4  2002 StatusLaboral
StatusL~2004  400      3      1.79     1     4  2004 StatusLaboral
StatusL~2006  400      4    1.8825     1     4  2006 StatusLaboral
StatusL~2009  400      3     1.965     1     4  2009 StatusLaboral
FactorLong    400    321  1221.487   458  3877  Peso longitudinal olas 02-04-06-09-12
──────────────────────────────────────────────────────────────────────────────
```

Se construyen las variables dicotómicas para las personas empleadas en cada ronda utilizando el comando foreach.

```
. foreach i in 02 04 06 09 {
        generate Empleados20`i'=.
                replace Empleados20`i' = 1 if StatusLaboral20`i'==1
                replace Empleados20`i' = 0 if StatusLaboral20`i'==2
                replace Empleados20`i' = 0 if StatusLaboral20`i'==3
                replace Empleados20`i' = 0 if StatusLaboral20`i'==4
                label define Empleados20`i'_labels 1 "Si" 0 "No"
                label value Empleados20`i' Empleados20`i'_labels
                        }

. label var Empleados2002 "Empleado el 2002"
. label var Empleados2004 "Empleado el 2004"
. label var Empleados2006 "Empleado el 2006"
. label var Empleados2009 "Empleado el 2009"
```

La matriz de transición entre las personas con empleo y sin empleo se calcula utilizando el comando tabulate y la especificación aw (pesos analíticos) para el factor de expansión longitudinal. A continuación se entrega la línea de programación en Stata para la matriz de transición entre la primera y segunda ronda del panel.

```
. tabulate Empleados2002 Empleados2004 [aw = FactorLong], row cell nofreq
```

Empleado el 2002	Empleado el 2004 No	Si	Total
No	**72.43**	**27.57**	**100.00**
	24.05	**9.16**	**33.21**
Si	**20.84**	**79.16**	**100.00**
	13.92	**52.87**	**66.79**
Total	**37.97**	**62.03**	**100.00**
	37.97	**62.03**	**100.00**

La matriz de transición muestra que de todos los individuos que no tenía empleo el 2002, el 72.4 por ciento continuó sin trabajo el año 2004, mientras que el 27,6 por ciento llegó a tener un empleo el 2004. De las personas que si tenían empleo el año 2002, el 20,8 por ciento no estaba trabajando dos años después, mientras que el 79,2 por ciento permaneció con trabajo.

Las matrices de transición del empleo de las personas para los años 2004 y 2006, como también para los años 2006 y 2009 se estiman a continuación y se interpretan en forma análoga.

```
. tabulate Empleados2004 Empleados2006 [aw = FactorLong], row cell nofreq
. tabulate Empleados2006 Empleados2009 [aw = FactorLong], row cell nofreq
```

Empleado el 2004	Empleado el 2006 No	Si	Total
No	**70.83**	**29.17**	**100.00**
	26.90	**11.08**	**37.97**
Si	**14.98**	**85.02**	**100.00**
	9.29	**52.74**	**62.03**
Total	**36.19**	**63.81**	**100.00**
	36.19	**63.81**	**100.00**

Empleado el 2006	Empleado el 2009 No	Si	Total
No	**78.94**	**21.06**	**100.00**
	28.57	**7.62**	**36.19**
Si	**17.51**	**82.49**	**100.00**
	11.17	**52.64**	**63.81**
Total	**39.74**	**60.26**	**100.00**
	39.74	**60.26**	**100.00**

3.2.2 Ejercicio 5: Determinar cuál es el flujo del trabajo informal al formal (y viceversa) en un periodo determinado

Se generan las variables tipo de empleo (formal/informal) para las cuatro rondas a partir de la información entregada por el estatus laboral y si la persona declara cotizar o no.

```
. foreach i in 02 04 06 09 {
        generate TipoEmpleo20`i'=.
        replace TipoEmpleo20`i' = 1 if Empleados20`i'==1 & CotizaActual20`i'==1
        replace TipoEmpleo20`i' = 0 if Empleados20`i'==1 & CotizaActual20`i'!=1
        label define TipoEmpleo20`i'_labels 1 "Formal" 0 "Informal"
        label value TipoEmpleo20`i' TipoEmpleo20`i'_labels
                                }
. label var TipoEmpleo2002 "Tipo de empleo el 2002"
. label var TipoEmpleo2004 "Tipo de empleo el 2004"
. label var TipoEmpleo2006 "Tipo de empleo el 2006"
. label var TipoEmpleo2009 "Tipo de empleo el 2009"
```

De la misma forma que en el ejercicio anterior se construye la matriz de transición entre el empleo formal e informal para la ronda tres (2006) y cuatro del panel (2009).

```
. tabulate TipoEmpleo2006 TipoEmpleo2009 [aw = FactorLong], row cell nofreq
```

Tipo de empleo el 2006	Tipo de empleo el 2009		
	Informal	Formal	Total
Informal	**68.33**	**31.67**	**100.00**
	17.14	**7.95**	**25.09**
Formal	**5.90**	**94.10**	**100.00**
	4.42	**70.49**	**74.91**
Total	**21.56**	**78.44**	**100.00**
	21.56	**78.44**	**100.00**

La matriz de transición nos indica que entre el año 2006 y 2009 hubo un 31,7 por ciento de personas que pasaron de un empleo informal a uno formal. El flujo inverso, es decir, pasar de un empleo formal a uno informal fue cinco veces más bajo (5,9 por ciento). A continuación se entregan las sentencias para calcular las matrices de transición entre los años 2004 y 2006 y los años 2002 y 2004.

```
. tabulate TipoEmpleo2004 TipoEmpleo2006 [aw = FactorLong], row cell nofreq
. tabulate TipoEmpleo2002 TipoEmpleo2004 [aw = FactorLong], row cell nofreq
```

Tipo de empleo el 2004	Tipo de empleo el 2006		
	Informal	Formal	Total
Informal	**67.78**	**32.22**	**100.00**
	17.15	**8.15**	**25.30**
Formal	**12.37**	**87.63**	**100.00**
	9.24	**65.46**	**74.70**
Total	**26.39**	**73.61**	**100.00**
	26.39	**73.61**	**100.00**

Tipo de empleo el 2002	Tipo de empleo el 2004		
	Informal	Formal	Total
Informal	**63.31**	**36.69**	**100.00**
	17.35	**10.06**	**27.41**
Formal	**10.66**	**89.34**	**100.00**
	7.74	**64.85**	**72.59**
Total	**25.09**	**74.91**	**100.00**
	25.09	**74.91**	**100.00**

3.3 ANÁLISIS PARA EVALUAR EL BENEFICIO DE UN PROGRAMA SOCIAL

En varios países de Latinoamérica se están diseñando o se han implementado una serie de programas en el ámbito de la previsión y seguridad social. En general estos programas se focalizan en determinados grupos de la población como son los jóvenes, mujeres o adultos mayores. Por ejemplo, en Chile además de la Reforma Previsional que entró en vigencia en julio de 2008, se han ejecutado en los últimos años programas de inserción laboral (Subsidio al Empleo Joven, Bono al Trabajo de la Mujer) y políticas sociales de emprendimiento (financiamiento para las microempresas informales y capacitación a los trabajadores independientes), junto a subsidios de cesantía y de ingreso ético familiar, entre otros. La amplia gama de programas sociales que pueden diseñarse e implementarse han abierto la pregunta sobre cuáles son los que tienen un mayor impacto entre los beneficiarios considerando que el presupuesto para las políticas sociales siempre es limitado. Realizar evaluaciones de impacto de los programas sociales permite compararlos entre sí y decidir mantenerlos o replicarlos cuando el efecto es alto o cerrarlos cuando no hay efecto o es muy bajo en relación a su costo.

Aunque la ELPS no tiene como objetivo principal evaluar el impacto de un programa en particular su naturaleza longitudinal permite comparar el efecto de algunas políticas sociales que se han implementado durante el mismo tiempo que se han realizado las rondas del panel. En general, los beneficios de estos programas pueden ser estimados a través de técnicas econométricas. Ejemplos de beneficios en los programas son los incrementos en las remuneraciones de los trabajadores, aumento en la probabilidad de estar empleado, aumento en la probabilidad de cotizar en el sistema de previsión social o aumento en los ingresos totales del hogar. La razón de usar técnicas

econométricas para estimar el efecto de los programas sociales es que permiten controlar los errores que surgen habitualmente al seleccionar a los individuos que serán beneficiados. Por ejemplo, un programa de emprendimiento puede seleccionar a trabajadores independientes que poseen ciertas habilidades que no todos los trabajadores independientes han adquirido. Al terminar el programa es posible que el aumento de las ventas de su emprendimiento (u otro beneficio) sea explicado por esas habilidades y no necesariamente se puede atribuir al programa el efecto positivo que se ha medido en el tiempo.

La idea básica detrás de esta técnica es analizar los efectos del programa comparando un grupo que es beneficiado por el programa (grupo tratado) con un grupo que tiene características similares a los beneficiarios pero no participan en el programa (grupo de control). De esta manera el grupo de control se acerca a un escenario hipotético donde el grupo de beneficiarios es comparado con ellos mismos pero sin ser tratados por el programa. A esto se conoce como la situación contrafactual porque dicho escenario no puede ser observado. Una persona es o no es tratada. No es posible saber qué hubiera pasado si la persona no hubiera sido parte del programa.

En términos prácticos, se recomienda que cada individuo esté debidamente identificado y no pertenezca al grupo de beneficiario y al grupo de control al mismo tiempo. En las bases de datos los estados se identifican con un "1" a las personas tratadas (beneficiarios) y con un "0" a los individuos no tratados (grupo de control). Además se debe identificar el tiempo antes de aplicar el programa y el tiempo después. En general se crea una variable dicotómica que entrega un "0" cuando el programa aún no ha comenzado (Antes) y un "1" cuando ya ha terminado (Después). De esta manera, se definen cuatro combinaciones para las observaciones de la base de datos : i) Grupo Tratado Antes del Programa (10), ii) Grupo de Control Antes del Programa (00), iii) Grupo Tratado Después del Programa (11), iv) Grupo de Control Después del Programa (01).

Al tener los promedios en los cuatro escenarios de la variable que medirá el beneficio del programa es posible usar tres estimadores para calcular el efecto del programa evaluado: i) el "Corte Transversal", el cual compara el resultado de la media de los que participaron en el programa y los que no lo hicieron después de terminado el programa (media(11) - media(01)), ii) el "Antes y Después", el cual compara el resultado de la media de los participantes al inicio y al final del programa (media(10) - media(11)), y iii) el "Diferencias en Diferencias"[18], el cual compara los cambios de los resultados de la variable de interés antes y después del programa, y entre los beneficiarios y el grupo de control ((media(00) - media(10)) - (media(01) - media(11))).

3.3.1 Ejercicio 6: Determinar cuáles son los beneficios de un programa que entrega un bono anual a las madres que trabajan (PBMT)

Para responder a la pregunta de este ejercicio se va a suponer que el año 2005 se implementó en el mismo país donde se ha realizado la panel ficticia un programa de inserción laboral femenino focalizado en las mujeres con hijos. Este programa consiste en que las madres que puedan acreditar que estuvieron trabajando durante el último año y tienen al menos un hijo o hija menor de 12 años viviendo con ella puede recibir un bono en dinero. En adelante, se hará referencia al programa como PBMT (Programa Bono Madre Trabajadora).

[18] Para una explicación más extensa del estimador Diferencias en Diferencias (DD) se recomienda leer el capítulo 6 del libro "La Evaluación de Impacto en la Práctica" de Paul Gertler y su equipo (2011).

La base de datos ficticia que se utiliza para estimar los efectos del PBMT se llama PanelMuestraPBMT.dta y contiene 1,000 observaciones que representan a la población longitudinal de mujeres que han sido entrevistadas el año 2002, 2004, 2006 y 2009. Las bases de datos que tienen información de los miembros del hogar de la mujer entrevistada en las cuatro rondas se denominan HogarR1PBMT.dta, HogarPBMTR2.dta, HogarPBMTR3.dta y HogarPBMTR4.dta. De estas bases de datos se obtiene la información si la mujer tiene un hijo o hija de 12 años o menos viviendo con ella en el momento de la entrevista. El grupo de mujeres que tienen un hijo o hija conformará el grupo potencial que puede ser beneficiado por PBMT. Las mujeres que no registran tener un hijo o hija en los datos panel serán parte del grupo de control. Se asigna con un "1" a las mujeres que son parte del grupo tratado y un "0" a las mujeres que conforman el grupo de comparación o de control.

A continuación se entrega la sintaxis que genera la variable Grupo con los códigos Tratado (1) y Control (0) en la base de datos PanelMuestraPBMT.dta.

```
. foreach i in 06 09  {
        use Hogar20`i'PBMT.dta, clear
                sort id orden
        generate Hijos20`i' = 0
                by id : replace Hijos20`i' = 1 if Edad20`i'<=12
        collapse (sum) Hijos20`i', by(id)
        save Hijos20`i'.dta, replace

        use PanelMuestraPBMT.dta, clear
                sort id
        merge id using Hijos20`i'.dta
                drop _merge
        generate Grupo20`i'= 0
                replace Grupo20`i'=1 if Hijos20`i'>=1
                label define Grupo20`i'_labels 1 "Tratado" 0 "Control"
                label value Grupo20`i' Grupo20`i'_labels
        save PanelMuestraPBMT.dta, replace
                                                }
```

Ahora se tabulan las variables creadas para los años 2006 y 2009 con el fin de revisar la manipulación de datos realizadas.

```
. tabulate Grupo2006 Grupo2009
```

	Grupo2009		
Grupo2006	Control	Tratado	Total
Control	**369**	**59**	**428**
Tratado	**110**	**462**	**572**
Total	**479**	**521**	**1,000**

De las 1,000 mujeres que conforman la base de datos panel ficticia, 462 mujeres pudieron haber accedido al PBMT los años 2006 y 2009, mientras que 110 mujeres no podían continuar con la opción del beneficio el año 2009 porque ya no tenían hijos o hijas menores de 12 años de edad en su hogar. Esos 110 casos serán considerados como parte del grupo de tratados. En relación al grupo de control hay 59 mujeres que tenían un hijo en la ronda del año 2009. Esos casos serán sacados del análisis para no sesgar los resultados al tener la posibilidad de estar en ambas situaciones una vez que el PBMT ya empezó a ejecutarse. A continuación se realizan las recodificaciones señaladas y se

tabula nuevamente las variables para chequear que no existan mujeres que estén en el tiempo en ambos grupos.

```
. replace Grupo2009=1 if Grupo2006==1
. drop if Grupo2006==0 & Grupo2009==1

. tabulate Grupo2006 Grupo2009
```

Grupo2006	Grupo2009 Control	Tratado	Total
Control	**369**	**0**	**369**
Tratado	**0**	**572**	**572**
Total	**369**	**572**	**941**

Por último, se botan todos las observaciones de mujeres que el año 2009 tenían un año de jubilación (61 años de edad) y se crea una sola variable para identificar a las mujeres que están en el grupo de tratamiento y en el grupo de control.

```
. drop if Edad2009>=61
. generate Grupo=.
. replace Grupo=1 if Grupo2006==1
. replace Grupo=1 if Grupo2009==1
. replace Grupo=0 if Grupo2006==0
. replace Grupo=0 if Grupo2009==0
. replace Grupo=1 if Grupo2006==1
. tab Grupo
```

Grupo	Freq.	Percent	Cum.
0	**248**	**33.56**	**33.56**
1	**491**	**66.44**	**100.00**
Total	**739**	**100.00**	

Se observa que para analizar los beneficios del PBMT - utilizando las cuatro rondas del panel ficticio - se tiene finalmente un grupo de tratamiento de 491 mujeres y un grupo de control de 248 mujeres.

Dado que el programa hipotético que se está presentando en este ejercicio se ejecutó el 2005, es necesario crear una variable que dé cuenta si las mediciones fueron realizadas antes de ese año o después que el PBMT empezó a implementarse. Para eso se crea la variable PBMT2005 que identifica con un "0" si la entrevista fue realizada antes de 2005 y un "1" si fue posterior a esa fecha. A continuación se escribe las líneas de programación para ejecutar lo señalado. Se utiliza el comando reshape para transformar la base de datos de formato "wide" a "long".

```
. keep id StatusLaboral2002 IngresoTrabajador2002 CotizaActual2002 ///
                    StatusLaboral2004 IngresoTrabajador2004 CotizaActual2004 ///
                    StatusLaboral2006 IngresoTrabajador2006 CotizaActual2006 ///
                    StatusLaboral2009 IngresoTrabajador2009 CotizaActual2009 Grupo

. reshape long StatusLaboral IngresoTrabajador CotizaActual, i(id) j(Anho)

. generate PBMT2005=.
```

```
. replace PBMT2005=0 if Anho==2002
. replace PBMT2005=0 if Anho==2004
. replace PBMT2005=1 if Anho==2006
. replace PBMT2005=1 if Anho==2009

. tabulate PBMT2005 Grupo
```

	Grupo		
PBMT2005	Control	Tratado	Total
Antes	**496**	**982**	**1,478**
Después	**496**	**982**	**1,478**
Total	**992**	**1,964**	**2,956**

La tabulación de las variables Grupo y PBMT2005 informa que se disponen de 2,956 observaciones para analizar. El grupo tratado tiene 982 casos después que el PBMT entró en vigencia. En el grupo de control hay 496 observaciones. El mismo número de observaciones se tiene en el grupo de tratamiento y de control para antes del año 2005. De esta manera, recién ahora se dispone de una base de datos que puede ser usada para estimar los beneficios del programa.

En la evaluación del PBMT analizan dos tipos de beneficios: la probabilidad de estar empleado y la probabilidad de estar cotizando en el sistema de previsión social. Por lo tanto, el siguiente paso es recodificar las variables StatusLaboral y CotizaActual de la base de datos longitudinal en variables binarias, es decir, "1" si la mujer está empleada y un "0" si no lo está en el caso de StatusLaboral. Para la variable CotizaActual se recodifica con un "1" si la mujer está cotizando y un "0" en caso contrario.

```
. generate Trabajando=.
        replace Trabajando= 1 if StatusLaboral==1
        replace Trabajando= 0 if StatusLaboral!=1
        label define Trabajando_labels 1 "Si" 0 "No"
        label value Trabajando Trabajando_labels

. generate Cotizando=.
        replace Cotizando = 1 if CotizaActual==1
        replace Cotizando = 0 if CotizaActual!=1
        label define Cotizando_labels 1 "Si" 0 "No"
        label value Cotizando Cotizando_labels

. tabulate Trabajando Cotizando, mis
```

	Cotizando		
Trabajando	No	Si	Total
No	**1,136**	**50**	**1,186**
Si	**495**	**1,275**	**1,770**
Total	**1,631**	**1,325**	**2,956**

La tabulación de las variables Trabajando y Cotizando muestra que 1,136 observaciones declararan no estar trabajando ni cotizando en el sistema de previsión social. Hay 495 observaciones que a pesar de indicar estar trabajando en el momento de la entrevista no estaban cotizando. La base de datos

contiene 1,275 observaciones de momentos en el tiempo donde la mujer entrevistada declara estar trabajando y cotizando. Solo 50 mujeres cotizan pese a no estar empleada.

La base que se dispone después de la preparación de los datos está en formato "long" e identifica si la observación corresponde a una mujer del grupo de tratamiento o que pertenece al grupo de control. Además se dispone de información del tiempo cuando fue realizada la observación. En este caso es antes del año 2005 o despúes de la implementación del PBTM. Por último, se cuenta con las variables que podrían ser afectadas por el programa: empleo femenino y nivel de cotización de las mujeres en el sistema de previsión social. Los beneficios o efectos positivos en esas variables se basan en la siguiente hipótesis. Se espera que el PBTM incentive a las madres a buscar trabajo, por lo tanto, al estar empleadas la probabilidad que las mujeres coticen en el sistema de previsión social aumenta.

A continuación se calculan las medias para las observaciones del grupo de tratamiento y del grupo de control el año 2009, es decir, 4 años después de la implementación del PBMT.

```
. mean Trabajando if Grupo==1 & PBMT2005==1
. mean Trabajando if Grupo==0 & PBMT2005==1
```

	Mean	Std. Err.	[95% Conf.	Interval]
Trabajando	.6496945	.0152315	.6198044	.6795846

	Mean	Std. Err.	[95% Conf.	Interval]
Trabajando	.5766129	.0222079	.5329794	.6202464

```
. display 0.6496945 - 0.5766129
0.0730816
```

La diferencia de ambos promedios entrega el beneficio del programa utilizando el estimador de Corte Transversal. El beneficio del programa es positivo en un 7,3 por ciento aunque al observar los intervalos de confianza de las medias estos se traslapan entre sí. Por lo tanto, no es posible decir que las medias son estadísticamente distintas.

El segundo estimador conocido como el Antes y Despúes compara las medias para las observaciones del grupo de beneficiarios del programa antes y después del año que comenzó el PBTM.

```
. mean Trabajando if Grupo==1 & PBMT2005==0
. mean Trabajando if Grupo==1 & PBMT2005==1
```

	Mean	Std. Err.	[95% Conf.	Interval]
Trabajando	.5600815	.0158481	.5289814	.5911815

	Mean	Std. Err.	[95% Conf.	Interval]
Trabajando	**.6496945**	**.0152315**	**.6198044**	**.6795846**

```
. display 0.6496945 - 0.560081
0.089613
```

Los resultados muestran que los intervalos de confianza de las medias no se traslapan y el beneficio del programa mejora en un 8,9 por ciento la participación laboral femenina.

El tercer estimador del impacto del PBMT en el empleo femenino es el de Diferencias en Diferencias (DD). El cálculo de este estimador es la diferencia entre: i) la diferencia del grupo de tratamiento (estimador Antes y Después) y ii) la diferencia del grupo de control (la situación antes y después del inicio del programa). El resultado de este estimador controla las diferencias no observadas en el tiempo entre los grupos de tratamiento y de comparación. De esta manera, el DD entrega el impacto del programa considerando cualquier diferencia que sea constante en el tiempo de las características de las mujeres beneficiadas que podrían estar explicando la diferencia en los resultados entre los dos grupos.

```
. mean Trabajando if Grupo==0 & PBMT2005==0
. mean Trabajando if Grupo==0 & PBMT2005==1
```

	Mean	Std. Err.	[95% Conf.	Interval]
Trabajando	**.5967742**	**.0220484**	**.5534543**	**.6400941**

	Mean	Std. Err.	[95% Conf.	Interval]
Trabajando	**.5766129**	**.0222079**	**.5329794**	**.6202464**

```
. display (0.6496945 - 0.5600815) - (0.5766129 - 0.5967742)
0.1097743
```

El estimador DD da cuenta de un impacto positivo del PBMT de un 10,9 por ciento en el empleo femenino.

Ahora, se correrá una regresión para calcular el estimador DD del efecto del PBMT. Como se verá, hacer la regresión lleva al mismo resultado anterior (0.1097743) pero usando mínimos cuadrados ordinarios[19]. La ecuasión de la regresión es la siguiente:

$$Trabajando = \beta_0 + \delta 0 \quad PBMT2005 + \beta_1 Grupo + \delta_1 (PBMT2005 \text{ x } Grupo) + \varepsilon$$

[19] Si bien existe una extensa literatura en econometría, se recomiendan las lecturas de (Gujarati & Porter, 2008) y (Greene, 2008) para un mejor entendimiento de la regresión lineal basado en el criterio de mínimos cuadrados.

Donde ε es el error de la perturbación aleatoria y δ_1 es el efecto del PBMT en el grupo de mujeres que pueden participar en el programa. Por lo tanto, el coeficiente de la regresión para la variable $PBMT2005 \times Grupo$ es el estimador DD. La instrucción en Stata que permite obtener los parámetros correspondientes a la recta obtenida con el criterio de mínimos cuadrados ordinarios es regress.

```
. generate PBMTxGrupo = PBMT2005*Grupo
. regress Trabajando Grupo PBMT2005 PBMTxGrupo
```

Source	SS	df	MS		
Model	4.26196548	3	1.42065516	Number of obs =	2956
Residual	705.89365	2952	.239123865	F(3, 2952) =	5.94
				Prob > F =	0.0005
Total	710.155616	2955	.240323389	R-squared =	0.0060
				Adj R-squared =	0.0050
				Root MSE =	.489

Trabajando	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Grupo	-.0366927	.0269372	-1.36	0.173	-.0895103	.0161248
PBMT2005	-.0201613	.0310517	-0.65	0.516	-.0810465	.0407239
PBMTxGrupo	.1097743	.0380949	2.88	0.004	.035079	.1844696
_cons	.5967742	.0219569	27.18	0.000	.5537218	.6398265

A continuación se calcula el estimador DD del impacto del PBMT en el nivel de cotización de las mujeres en el sistema de previsión social.

```
. regress Cotizando Grupo PBMT2005 PBMTxGrupo
```

Source	SS	df	MS		
Model	3.01459125	3	1.00486375	Number of obs =	2956
Residual	728.066261	2952	.246634912	F(3, 2952) =	4.07
				Prob > F =	0.0067
Total	731.080853	2955	.247404688	R-squared =	0.0041
				Adj R-squared =	0.0031
				Root MSE =	.49662

Cotizando	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Grupo	-.060451	.027357	-2.21	0.027	-.1140917	-.0068104
PBMT2005	.0060484	.0315356	0.19	0.848	-.0557857	.0678824
PBMTxGrupo	.0652347	.0386886	1.69	0.092	-.0106246	.1410941
_cons	.4637097	.0222991	20.80	0.000	.4199864	.5074329

El estimador DD da cuenta de un impacto positivo del PBMT de un 6,5 por ciento en el nivel de cotización femenino en el sistema de previsión social. Además los resultados de la regresión indican que ese valor es estadísticamente significativo.

4. Referencias

Agresti, A. (2007). *An introduction to categorical data analysis, 2nd Edition* (Vol. 135). John Wiley & Sons, Inc., Hoboken, New Jersey.

Baum, C. F., & Cox, N. J. (2007). Getting those data into shape. *The Stata Journal, 7*(2), 268-271.

Gertler, P. J., Martínez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2011). *La Evaluación de Impacto en la Práctica*: World Bank Publications.

Greene, W. H. (2008). *Econometric Analysis (6th ed.)*. Englewood Cliffs, New Jersey, Prentice Hall.

Gujarati, D., & Porter, D. C. (2008). *Basic Econometrics*. New York, McGraw-Hill Education.

Jenkins, S. P. (2011). *Changing fortunes: income mobility and poverty dynamics in Britain*: OUP Oxford.

Kalton, G., & Brick, M. (2000). Weighting in houselhold panel survey. In D. Rose (Ed.), *Researching Social and Economic Change. The use of Household Panel Studies* (pp. 96-112). London and New York: Routledege.

Kish, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association, 44*(247), 380-387.

Lynn, P. (2009). Methods for Longitudinal Surveys. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 1-19). Chichester, UK: John Wiley & Sons.

Paredes, R., Prieto, J., & Zubizarreta, J. R. (2006). Attrition in Longitudinal Data and Income Mobility in Chile. Mimeo, Observatorio Social, Universidad Alberto Hurtado.

Rose, D. (2000). Household panel studies In D. Rose (Ed.), *Researching Social and Economic Change: the uses of household panel studies* (pp. 3-35). London and New York: Routledge.