# Capítulo 2

# HERRAMIENTAS PARA EL ANALISIS DISTRIBUTIVO

Esta versión: 28 de septiembre, 2011 ♣

---

♣ Este documento es un borrador del capítulo 2 del libro "Pobreza y Desigualdad en América Latina. Conceptos, herramientas y aplicaciones" por Leonardo Gasparini, Martín Cicowiez y Walter Sosa Escudero. El libro se realiza en el marco del CEDLAS, el Centro de Estudios Distributivos, Laborales y Sociales de la Universidad Nacional de La Plata (cedlas.econo.unlp.edu.ar). Por favor, no citar sin permiso. Se agradecen los comentarios

# Índice del Capítulo 2



Pobreza y desigualdad en América Latina : conceptos, herramientas y aplicaciones / Leonardo Gasparini; Martín Cicowiez; Walter Sosa Escudero. - 1a ed. - La Plata : Universidad Nacional de La Plata, 2010.

CD-ROM.

ISBN 978-950-34-0667-0

1. Problemas Sociales. 2. Pobreza. 3. Enseñanza Universitaria. I. Cicowiez, Martín II. Sosa Escudero, Walter III. Título

CDD 362.5

Fecha de catalogación: 20/08/2010

## 2.1. Introducción

La pobreza y la desigualdad, los dos ejes centrales de este libro, son fenómenos intuitivamente claros, aunque complejos de definir con precisión. Todos tenemos una idea intuitiva del concepto de pobreza que asociamos a privaciones de distinto tipo, y de del concepto de desigualdad que vinculamos con diferencias, pero no resulta sencillo acordar definiciones estrictas. Esta dificultad es natural, dada la complejidad del fenómeno. La idea de pobreza, por ejemplo, está asociada a privaciones materiales concretas, como insuficiencia alimentaria, pero también a falta de oportunidades de progreso, vulnerabilidad ante shocks, marginalidad y estigmatización.

La manera de proceder ante un fenómeno complejo es analizarlo en su versión más simplificada y luego ir agregando complicaciones. Ese es el camino que vamos a seguir en el libro. Comencemos entonces asumiendo que todas las dimensiones en las que es relevante analizar privaciones o desigualdades pueden resumirse en una sola variable a la que denotamos con $x$, y a la que por comodidad llamamos *ingreso*. Existe un sinnúmero de cuestiones relacionadas a la elección de la variable sobre la cual se focaliza el análisis en la práctica. ¿Debemos usar el ingreso, el consumo u otra variable? ¿Debemos usar el ingreso per cápita o el ajustado por alguna escala de adultos equivalentes? Estas y muchas otras cuestiones de implementación práctica son derivadas al siguiente capítulo del libro. Mientras tanto supongamos que nuestra *proxy* de nivel de vida - el ingreso $x$ - está perfectamente definido, sin ambigüedades.[1]

Asumamos una comunidad de $N$ personas. A la lista que enumera los ingresos en esta población la llamamos "distribución empírica del ingreso", o directamente "distribución del ingreso". El término *distribución* de $x$ hace referencia a toda la colección de valores de $x$ en una circunstancia particular, es decir al vector $\{x_1, x_2, ..., x_N\}$, donde el subíndice indexa a los $N$ individuos de esta comunidad. Nótese que esta acepción es diferente a la usada coloquialmente, que asocia *distribución* a *reparto*, y por ende está vinculada al concepto de desigualdad. En contraste con ese uso coloquial, la literatura distributiva utiliza una acepción más amplia del término *distribución del ingreso* para hacer referencia a la lista completa de ingresos en una comunidad y no a alguna medida de disparidad de esos valores entre las personas.

¿Qué nos interesa de ese vector de valores de $x$ al que llamamos distribución de $x$? Por un lado nos preocupa el número y características de aquellas personas que no alcanzan un cierto nivel de $x$ considerado mínimo bajo algún criterio. Estas cuestiones están asociadas a uno de los temas centrales del libro: la pobreza. Por otro lado nos interesa conocer las discrepancias en los niveles de $x$ entre las personas. Este es un tema relacionado con el otro objetivo central del libro: la desigualdad.

La pobreza y la desigualdad son, entonces, dos *características* de la distribución del ingreso asociadas a la cantidad y ubicación de las observaciones debajo de un umbral, y

[1] Si el lector se siente incómodo con esta secuencia, puede estudiar primero el capítulo 3 para profundizar en temas conceptuales y prácticos sobre las variables de interés y luego volver a este capítulo.

a su nivel de dispersión, respectivamente.[2] Otras características de la distribución como la media o la mediana, que han ocupado tradicionalmente el centro de atención en Economía, tienen una relevancia menor en los estudios distributivos.

Vamos a destinar este capítulo a presentar un conjunto de herramientas gráficas y analíticas útiles para estudiar distribuciones, ejemplificándolas con casos concretos en varios países de América Latina. Una vez que desarrollemos el instrumental básico para presentar y estudiar distribuciones, será más sencillo analizar alguna de sus características, como la pobreza y la desigualdad, tarea que diferimos hasta el capítulo 4.

El análisis distributivo se complica (y se hace más interesante) cuando reconocemos que típicamente el investigador no puede observar toda la realidad, sino muestras imperfectas de la misma. A partir de información parcial un analista debe inferir resultados generalizables a toda la población. Esta consideración requiere detenerse en el análisis inferencial e introducir herramientas para estimar la significatividad estadística de los resultados, tareas que también abordamos en este capítulo.

El resto del capítulo está ordenado de la siguiente forma. La sección 2.2 presenta un conjunto de medidas resumen de la distribución y propone un primer examen de los microdatos de las encuestas de hogares latinoamericanas. La sección 2.3 introduce un conjunto de instrumentos gráficos que permiten ilustrar una distribución. La sección 2.4 extiende el análisis a funciones continuas que permiten un tratamiento más flexible y elegante. En la sección 2.5 se delinea el marco analítico general para el análisis inferencial necesario para desarrollar, en la sección 2.6, la idea de significatividad estadística de las mediciones distributivas. Finalmente, la sección 2.7 discute la aproximación de las distribuciones reales mediante formas paramétricas.

Como en el resto de los capítulos que componen el libro, este capítulo incluye un apéndice con explicaciones prácticas de cómo implementar en Stata los instrumentos y resultados presentados en el texto.

## 2.2. Medidas resumen

Una manera posible de presentar una distribución es a través de medidas resumen. Estas medidas sintetizan toda la distribución en uno o pocos valores, que representan alguna característica de la distribución subyacente. El proceso de resumir el vector de ingresos implica perder información para ganar en simplicidad analítica y comunicacional, y para permitir focalizar el análisis en alguna característica distributiva particular.

Comencemos el análisis con un ejemplo simple de una comunidad hipotética compuesta por 20 personas. La distribución empírica del ingreso de esta comunidad es un vector o lista que contiene los valores del ingreso de esas 20 personas. Supongamos que los

[2] Como veremos en el capítulo 4, hay visiones de la pobreza no necesariamente asociadas a la existencia de un umbral (pobreza relativa).

ingresos mensuales expresados en la moneda corriente del país (por comodidad, llamémosla *pesos*) ordenados de menor a mayor son:

{40, 65, 83, 101, 119, 137, 156, 176, 198, 223, 250, 279, 310, 350, 398, 456, 539, 651, 877, 1905}

Mientras que los primeros apartados de esta sección ilustran diversas medidas resumen en función de este ejemplo sencillo, en la sección 2.2.5 comenzamos a trabajar con microdatos de encuestas de hogares reales.

### 2.2.1. Tendencia central

Las medidas distributivas de uso más difundido en Economía son las de tendencia central, siendo el promedio o *media* el indicador más conocido. Analíticamente, la media aritmética de la distribución de $x$ es

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.1}$$

donde $i$ indexa a las personas y $N$ es el número de personas en la población o muestra disponible.[3] En el ejemplo, la media es 365.7: si bien ese número no se corresponde exactamente con ningún valor de la distribución de los ingresos, se ubica en una posición intermedia o "central".

La *mediana* es otra medida de tendencia central. Si ordenamos a los valores de $x$ de menor a mayor, como en el ejemplo, la mediana es aquel valor que deja por debajo (y por arriba) a la mitad de las observaciones. En nuestro ejemplo es fácil ver que dado que tenemos un número par de observaciones, todas distintas, cualquier valor entre 223 y 250 satisface este criterio. En estos casos usualmente la mediana se calcula como el promedio simple entre estos dos valores (236.5).

Si bien la media es una medida más popular que la mediana, esta última tiene una propiedad interesante: es considerablemente más robusta a la presencia de valores atípicos (*outliers*). Para ilustrar esta propiedad consideremos una distribución con cinco individuos con ingresos {1, 2, 4, 6, 7}. En este caso la media y la mediana coinciden (4) y ambas están en el "centro" de la distribución. Ahora bien, supongamos que por un error de tipeo al cargar los datos el último valor es registrado como 67 en vez de 7. Nótese que ante este error la media se cuadriplica a 16, mientras que la mediana se mantiene inalterada. Este caso simple ilustra la propiedad de robustez frente a valores atípicos que posee la mediana.

### 2.2.2. Cuantiles y proporciones

[3] Por ahora la distinción entre muestra y población no es importante. En la sección 2.5 de este capítulo esa distinción adquiere una relevancia fundamental.

Al trabajar con poblaciones con muchos individuos suele ser práctico ordenarlos de menor a mayor ingreso y dividirlos en grupos o segmentos contiguos iguales (con el mismo número de observaciones, dentro de lo posible). Por ejemplo, si dividimos a la población en diez grupos obtenemos *deciles*. El decil 1 de la distribución del ingreso hace referencia al grupo de personas pertenecientes al 10% de la población de menores ingresos, y el decil 10 al 10% más rico. En el ejemplo anterior el decil 1 está formado por las dos personas más pobres con ingresos 40 y 65. El ingreso promedio del decil inferior es 53, mientras que el ingreso promedio del decil superior es 1391. Los deciles surgen de dividir a la población en 10 segmentos contiguos iguales. Si, en cambio, la dividimos en 5 grupos obtenemos *quintiles*, si lo hacemos en 20 *ventiles* y en 100 *percentiles* o *centiles*. La denominación general de estos grupos es *cuantiles*.

Los términos introducidos en el párrafo anterior también son habitualmente usados para referirse a observaciones particulares y no a grupos de observaciones, lo cual puede generar confusiones. En esta acepción el $q$-ésimo cuantil de la distribución de los ingresos es un valor que deja por debajo una proporción $q$ de las observaciones, al ordenarlas de forma ascendente. En esta definición alternativa el decil 1 es el valor que deja por debajo al 10% de los ingresos y por arriba al 90%. El segundo decil se define en forma similar, dejando por debajo al 20% de los ingresos, y así sucesivamente hasta el noveno decil. Naturalmente, la mediana coincide con el quinto decil. En nuestro ejemplo hipotético el primer decil es cualquier valor entre 65 y 83, y el noveno decil cualquier valor entre 651 y 877.

De estas dos acepciones, la más usada en la literatura distributiva es la primera, donde *cuantil* hace referencia a un grupo de observaciones. Salvo cuando se indique lo contrario, esa será la alternativa utilizada en este libro.

Una característica de la distribución, que usaremos extensamente en los capítulos siguientes, es la proporción de observaciones cuyos ingresos son inferiores a algún valor arbitrario $x_m$. Formalmente,

$$M = \frac{1}{N}\sum_{i=1}^{N} 1(x_i < x_m) \tag{2.2}$$

donde 1(.) es una función *indicadora* que toma el valor 1 si la expresión entre paréntesis es verdadera y el valor 0 si es falsa. En la ecuación (2.2) la función indicadora vale 1 si el ingreso de la persona $i$ ($x_i$) es inferior al umbral $x_m$.

El indicador de pobreza más usado en la práctica y en gran parte de la literatura académica empírica -la tasa de incidencia- es simplemente la proporción de la población con ingresos inferiores a un umbral mínimo, conocido como línea de la pobreza, y en consecuencia se corresponde analíticamente con la ecuación (2.2).[4] Supongamos, siguiendo con el ejemplo anterior, que se identifica como pobres a todas aquellas

[4] La tasa de incidencia de la pobreza, o *headcount ratio*, es extensamente discutida en el capítulo 4, junto con otras medidas más sofisticadas de privaciones materiales.

personas con un ingreso inferior a 180 pesos. Es fácil calcular que bajo este criterio hay 8 personas pobres, de modo que la proporción de pobres es 0.4 (o 40%).

Otra característica distributiva a usar extensamente es la participación (o *share*) de un individuo o grupo en el ingreso total de la población. Analíticamente, la participación del grupo $J$ es

$$s_J = \frac{\sum_{i \in J} x_i}{\sum_{i=1}^{N} x_i} = \frac{\sum_{i=1}^{N} x_i 1[i \in J]}{N.\mu} \tag{2.3}$$

En nuestro ejemplo, el *share* del quintil superior en el ingreso total es 54.3%. Como veremos en el capítulo 6, la participación de algún cuantil extremo de la distribución en el ingreso total es a menudo utilizada como medida de desigualdad.

### 2.2.3. Dispersión

Las medidas de dispersión buscan resumir en un valor el grado de separación entre los valores de la distribución. El *rango de variación* - la diferencia entre el valor máximo y el mínimo - es una de esas medidas. Una versión menos extrema es el *rango intercuartílico*, es decir la diferencia entre el tercer y el primer cuartil, definidos como aquellos valores que, al ordenar a la población de forma ascendente según el ingreso, dejan por debajo al 75% y al 25% de las observaciones, respectivamente. Otra medida de separación usual es el cociente (o *ratio*) entre cuantiles. Si definimos los cuantiles en términos de grupos de observaciones, el ratio de ingresos medios entre el decil 10 y el decil 1 es 26.5, y el ratio entre los quintiles extremos es 13.7.

La varianza ($V$) es quizás la medida de dispersión más popular. Este indicador se define formalmente como

$$V = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \tag{2.4}$$

La varianza mide cuán lejos están en promedio las observaciones con respecto al centro de la distribución $\mu$. En nuestro ejemplo hipotético, $V$=168192.4. El desvío estándar $\sigma$, que es la raíz cuadrada positiva de la varianza, pone a ésta en unidades de medida similares a las utilizadas para construir la media. El coeficiente de variación ($CV$) expresa el desvío estándar como proporción de la media

$$CV = \frac{\sqrt{V}}{\mu} = \frac{\sigma}{\mu} \tag{2.5}$$

En nuestro ejemplo hipotético el desvío estándar es 410.1 y el coeficiente de variación 1.12. Nótese que a diferencia de la varianza, el valor del desvío pertenece al rango de las diferencias reales entre cualquier observación y la media. El coeficiente de variación en este ejemplo indica que el desvío estándar es un 12% superior a la media.

### 2.2.4. Asimetría

Intuitivamente, una distribución es simétrica en un punto $x$ si la frecuencia de observaciones es idéntica a ambos lados de $x$. En la práctica es relevante considerar el caso de distribuciones simétricas con respecto a alguna noción de tendencia central, como la media. Una forma simple de medir asimetría respecto de la media es el coeficiente de asimetría de Fisher, definido formalmente como[5]

$$A = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3} \tag{2.6}$$

A fines de entender la naturaleza de la asimetría es interesante explorar esta fórmula con cuidado. Consideremos el numerador, ya que el denominador es siempre positivo y cumple sólo un papel de normalización. El numerador de (2.6) busca medir la magnitud de las desviaciones con respecto a la media ($x_i$-$\mu$), comparando aquellas que ocurren a la derecha y a la izquierda. Nótese que si eleváramos la sumatoria de esas diferencias a la potencia 1 el resultado sería siempre cero, mientras que si lo hiciéramos a la potencia 2, siempre sería positivo. En cambio, al elevar a la potencia 3 (al cubo) la sumatoria puede ser positiva o negativa dependiendo de la magnitud de las diferencias entre $x_i$ y $\mu$ entre aquellos con mayor y menor ingreso que el valor promedio.

Nótese que si los ingresos fuesen simétricos en la media, la sumatoria del numerador de (2.6) debería dar cero, ya que los sumandos positivos (ingresos por arriba de la media) se cancelan con los negativos (ingresos por debajo de la media). En las distribuciones reales los ingresos de los ricos se encuentran muy por encima de la media, que se encuentra relativamente más cerca de los ingresos de los más pobres. Como las brechas relativas a la media son elevadas al cubo, los valores altamente positivos (la distancia de los ricos a la media) más que compensan los pequeños valores negativos (la distancia de los pobres a la media), produciendo un valor de $A$ positivo. En este caso se dice que la distribución es *asimétrica positiva* o asimétrica a la derecha. Todas las distribuciones del ingreso del mundo son asimétricas a la derecha, un fenómeno que documentaremos y analizaremos a lo largo del libro.

En general, tiende a pensarse que para distribuciones con asimetría positiva, la mediana está por debajo de la media. La intuición se deriva del análisis del párrafo anterior: los relativamente pocos valores muy altos tienen un efecto fuerte en la media y relativamente débil sobre la mediana, ya que esta última es más resistente a valores extremos. En la práctica, el hecho de que la media de los ingresos sea superior a la mediana es tomado como un síntoma natural de asimetría.[6]

---

[5] El coeficiente de Fisher es el tercer momento estándar. Otros indicadores de asimetría conocidos son el de Pearson y el de Bowley.

[6] Este resultado debe ser interpretado con cautela, ya que formalmente no es posible mostrar que la asimetría positiva induzca necesariamente un orden para la media y la mediana.

### 2.2.5. Un ejemplo: la distribución del ingreso en Brasil

Manos a la obra: trabajemos sobre una encuesta de hogares latinoamericana real; específicamente sobre la PNAD, la encuesta de hogares anual de Brasil, para el año 2007.[7] Esta encuesta tiene información de ingresos de 124794 hogares que reúnen a 394560 personas (cuadro 2.1). Esos individuos representan a cerca de 190 millones de brasileños que viven en una de las cinco grandes regiones geográficas en las que es posible dividir ese país: Norte, Nordeste, Sudeste, Sur y Centro-Oeste. Asumamos por ahora que la encuesta es una muestra perfectamente representativa de la población de Brasil.

Del cuadro 2.1 surge que el ingreso promedio per cápita mensual en Brasil es 574.3 reales (la moneda oficial en Brasil desde el año 1994). En este libro nos interesa ir más allá de los promedios y analizar toda la distribución del ingreso. Si las personas entrevistadas en la PNAD fueran toda la población brasileña, la distribución del ingreso en ese país sería una larga lista de 394560 números. Aun en este caso simplificado, trabajar con esa larga lista de números resulta impracticable, a menos que la logremos resumir de alguna forma. Comencemos por algunos estadísticos básicos como los del cuadro 2.1. Además de la media, se presentan los ingresos correspondientes a un conjunto de percentiles (definidos como observaciones, y no como grupos). En Brasil 2007, el 10% de la población tenía ingresos per cápita mensuales inferiores a 84 reales. La mitad de la población tenía un ingreso inferior a 330 reales: esa es la mediana de la distribución. Sólo el 1% de los brasileños representados en esta encuesta tenían en 2007 un ingreso per cápita igual o superior a 4400 reales mensuales. El rango intercuartílico es 621.5 – 165 = 456.5: el 50% central de las observaciones se encuentran agrupadas en un intervalo de esa magnitud.

**Cuadro 2.1**
**Resumen de la variable ingreso per cápita familiar**
**Brasil, 2007**

[7] El lector puede repetir el ejercicio con cualquiera de las bases de datos correspondientes a encuestas de hogares de los países de América Latina, disponibles en el sitio *web* del libro. Los comandos de Stata que generan los resultados siguientes están explicados con detalle al final del capítulo.

| | | Regiones | | | | |
|---|---|---|---|---|---|---|
| | Brasil | Norte | Nordeste | Sudeste | Sur | Centro-Oeste |
| Observaciones | | | | | | |
| Hogares | 124,794 | 15,619 | 38,156 | 37,197 | 19,826 | 13,996 |
| Individuos | 394,560 | 54,279 | 126,263 | 113,201 | 58,027 | 42,790 |
| Estadísticas de la distribución del ingreso per cápita familiar | | | | | | |
| Media | 574.3 | 391.0 | 344.7 | 693.7 | 710.7 | 685.5 |
| Percentiles | | | | | | |
| 1% | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 5% | 44.0 | 27.5 | 23.8 | 81.3 | 94.6 | 77.0 |
| 10% | 84.0 | 66.0 | 46.0 | 126.7 | 139.3 | 115.0 |
| 25% | 165.0 | 125.1 | 100.3 | 225.7 | 253.0 | 200.0 |
| 50% (mediana) | 330.0 | 224.4 | 192.3 | 418.0 | 450.0 | 361.7 |
| 75% | 621.5 | 425.0 | 373.2 | 757.2 | 788.3 | 666.7 |
| 90% | 1,200.0 | 815.4 | 665.4 | 1,433.7 | 1,430.0 | 1,422.7 |
| 95% | 1,870.0 | 1,223.8 | 1,085.3 | 2,200.0 | 2,163.3 | 2,350.0 |
| 99% | 4,400.0 | 2,757.9 | 2,909.7 | 4,895.0 | 4,669.5 | 5,720.0 |
| Mínimo | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Máximo | 66,000 | 49,592 | 30,120 | 66,000 | 45,650 | 55,000 |
| Coeficiente de Variación | 1.7 | 1.9 | 1.9 | 1.5 | 1.5 | 1.8 |
| Coeficiente de Asimetría - Fisher | 11.3 | 27.2 | 12.5 | 10.1 | 10.4 | 9.5 |
| Shares Deciles | | | | | | |
| Decil 1 | 0.7 | 0.7 | 0.6 | 1.0 | 1.1 | 0.9 |
| Decil 2 | 2.0 | 2.2 | 1.9 | 2.3 | 2.5 | 2.1 |
| Decil 3 | 2.9 | 3.2 | 2.9 | 3.3 | 3.6 | 2.9 |
| Decil 4 | 3.9 | 4.1 | 3.8 | 4.3 | 4.7 | 3.7 |
| Decil 5 | 5.1 | 5.2 | 4.9 | 5.5 | 5.8 | 4.6 |
| Decil 6 | 6.5 | 6.5 | 6.3 | 6.6 | 7.0 | 5.9 |
| Decil 7 | 8.2 | 8.4 | 8.0 | 8.4 | 8.6 | 7.4 |
| Decil 8 | 10.9 | 11.0 | 10.8 | 11.0 | 11.1 | 9.9 |
| Decil 9 | 16.1 | 16.0 | 15.1 | 16.1 | 15.8 | 15.4 |
| Decil 10 | 43.9 | 42.6 | 45.6 | 41.5 | 39.8 | 47.3 |
| Total | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Fuente: elaboración propia en base a microdatos de la PNAD

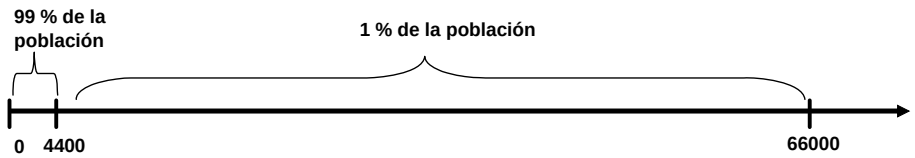
El cuadro indica también que el mínimo ingreso declarado es cero. De hecho, más del 1% de los encuestados en la PNAD declaran un ingreso mensual nulo. Por otro lado, el máximo ingreso declarado en la encuesta es 66000. De acuerdo a los datos de la PNAD 2007, el ingreso medio en las regiones Norte y Nordeste es considerablemente menor al ingreso en las regiones Sur y Sudeste, mientras que la distribución del ingreso en las primeras dos regiones es más dispersa, de acuerdo al coeficiente de variación.[8]

Es interesante notar que en todas las regiones, y en el agregado, la mediana del ingreso es claramente inferior a la media, lo cual es un signo de asimetría positiva de las distribuciones. De hecho, los coeficientes de asimetría resultan en todos los casos positivos y grandes. La inspección de los valores de los percentiles también revela la asimetría de la distribución. En el intervalo de ingresos que va de 0 a 330 están la mitad de las personas encuestadas. Si la distribución fuera simétrica, la mitad restante debería tener ingresos en el intervalo de 330 a 660. Según se desprende del cuadro 2.1 la realidad es muy diferente: el 20% más rico de la población brasileña tiene ingresos muy por encima de ese intervalo.

[8] El capítulo 6 discute el concepto de desigualdad, las bondades y defectos del coeficiente de variación como índice de desigualdad y otros indicadores alternativos.

Nótese la larga "cola" de la distribución. Mientras que el 99% de las personas encuestadas en la PNAD 2007 reportan ingresos en el intervalo [0, 4.400], el restante 1% superior reporta ingresos entre 4.400 y 66.000. El intervalo de ingresos del 1% más rico es 14 veces más grande que el intervalo donde se ubica el 99% restante de la población. La figura 2.1 ilustra estas diferencias. Esta larga cola superior no es una característica propia de la encuesta escogida en el ejemplo. De hecho, se trata de una característica de la mayoría de (quizás todas) las distribuciones del ingreso del mundo: un pequeño número de personas tienen ingresos desproporcionadamente altos respecto del resto de la población, y reúnen una alta proporción del ingreso total.[9]

**Figura 2.1**
**Ubicación de la población en la línea de ingresos per cápita familiar**
**Brasil, 2007**




Fuente: elaboración propia en base a microdatos de la PNAD.

El último panel del cuadro muestra los *shares* o participaciones de cada decil (interpretado como grupo de 10% de observaciones) en el ingreso total. El primer decil - el de menores ingresos - reúne apenas el 0.7% del ingreso total en Brasil. En el otro extremo, el 10% más rico de los brasileños tienen ingresos que representan el 43.9% del total. En virtud de estos *shares*, que examinaremos con más cuidado en el capítulo 6, la distribución del Sur de Brasil parece menos desigual que la del Noreste.

Un último ejercicio sencillo con la encuesta de Brasil. Supongamos que se fija la línea de pobreza en 130 reales mensuales.[10] Con esa línea, es posible deducir del cuadro 2.1 que la tasa de pobreza en Brasil (el porcentaje de personas con ingreso inferior a la línea) es superior al 10% e inferior al 25%. El porcentaje exacto es 18.2%. La pobreza así medida es 26.3% en la región Norte, 34.1% en la Nordeste, 10.4% en la Sudeste, 8.9% en la Sur y 12.3% en el Centro-Oeste.

El ejemplo nos ha permitido acercarnos a la distribución del ingreso real en un país concreto. Sin embargo, antes de entusiasmarnos con los números, es importante tratar algunas cuestiones conceptuales y aprender algunos instrumentos para graficar, resumir

[9] En la realidad, la cola superior es de hecho más larga que la ilustrada en la figura 2.1, dada la incapacidad de las encuestas de hogares (en Brasil y el resto del mundo) en captar a los grandes millonarios. El máximo ingreso en Brasil 2007 reportado en la encuesta (66000 reales) representaba unos US$35000 mensuales, un valor extraordinariamente alto comparado con el del resto de la población, pero seguramente inferior al de los grandes millonarios de ese país. El capítulo 3 y el Apéndice III discuten este punto.

[10] Esta, de hecho, es la línea internacional de 2.5 dólares por día por persona a paridad de poder adquisitivo para Brasil 2007, que discutiremos en el capítulo 4.

y comparar distribuciones y sus características. El resto de este capítulo trata esos temas.

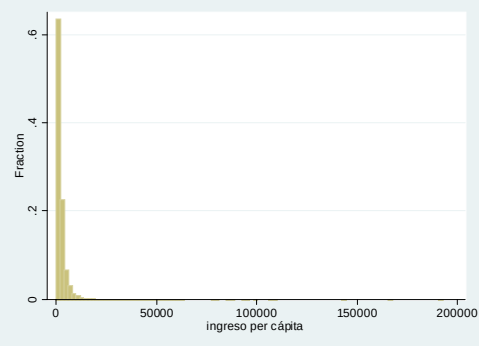
## 2.3. Gráficos

Las representaciones gráficas proporcionan una forma alternativa de ilustrar una distribución. Un gráfico es un modelo de la realidad en el que se presenta la información de una forma que nos resulta más fácil de aprehender que inspeccionando un largo vector de números. Adicionalmente, tienen la ventaja de representar un volumen de información mayor que las medidas resumen discutidas en la sección anterior y en consecuencia permiten visualizar conjuntamente varias características de una distribución.

### 2.3.1. Histograma

Una de las maneras más simples de representar una distribución es a través de un histograma. Para construirlo es necesario (i) dividir el rango de variabilidad de los ingresos (o *soporte*) en intervalos contiguos, preferentemente iguales, y (ii) graficar sobre el eje vertical la proporción de observaciones que caen dentro de cada intervalo (frecuencia relativa). Consecuentemente, las áreas de las barras que conforman el histograma suman 1. La figura 2.2 muestra el histograma de la distribución del ingreso per cápita familiar en México 2006, con 100 intervalos.

**Figura 2.2**
**Histograma del ingreso per cápita familiar**
**México, 2006**



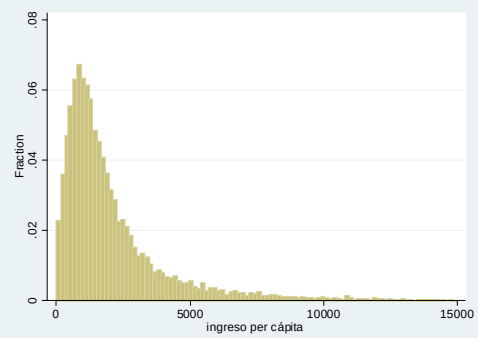

Fuente: elaboración propia en base a microdatos de la ENIGH
Nota: 100 intervalos.

El resultado es bastante frustrante. El ingreso máximo reportado en la encuesta de hogares de México en 2006 es casi 200000 pesos mexicanos mensuales, por lo que el eje horizontal debe llegar hasta ese valor. Al dividir el soporte de la distribución en 100 intervalos, el primero abarca desde 0 a 2000, pero resulta que en México 2006 ¡más del 60% de la población tiene ingresos en ese intervalo! Como consecuencia, el histograma

muestra una barra alta en el primer segmento, barras mucho más bajas en los cinco siguientes y luego barras imperceptibles. La larga "cola" derecha de la distribución en México vuelve al histograma poco útil en términos visuales.

Una posibilidad para aliviar este problema es restringir el soporte. Repitamos el histograma para ingresos inferiores a 15000, lo cual deja afuera al 1% más rico de los mexicanos captados en la encuesta. En este caso el histograma se vuelve más claro (figura 2.3). Nótese que pese al truncamiento de ingresos superiores, la forma de la distribución es claramente asimétrica, inclinada a la derecha y con una cola superior larga.

**Figura 2.3**
**Histograma del ingreso per cápita familiar**
**México, 2006**

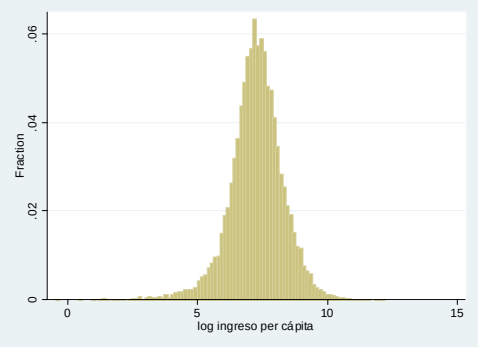



Fuente: elaboración propia en base a microdatos de la ENIGH.
Nota 1: ingresos restringidos a valores inferiores a 15000.
Nota 2: 100 intervalos.

Una práctica usual en el análisis distributivo es comprimir la escala de medición de los ingresos mediante alguna transformación que no altere el ordenamiento, típicamente la logarítmica. La figura 2.4 reproduce el histograma del logaritmo del ingreso per cápita familiar en México. Al comprimir la escala todas las observaciones pueden ser incluidas en el gráfico, sin que éste se degenere.[11] Una posible desventaja es que al aplicar la transformación logarítmica la asimetría positiva de la distribución ya no se visualiza en el gráfico.

**Figura 2.4**
**Histograma del logaritmo del ingreso per cápita familiar**
**México, 2006**

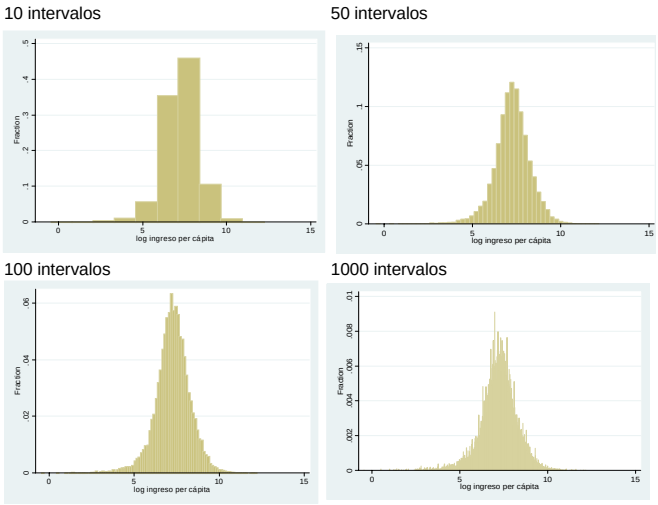
[11] Al tratarse de una escala logarítmica, el valor 5 en el eje horizontal corresponde a $148.4, mientras que el 10 corresponde a $22026.5.




Fuente: elaboración propia en base a microdatos de la ENIGH.
Nota: 100 intervalos.

La construcción de histogramas implica definir de antemano la cantidad de intervalos, o alternativamente el ancho de cada uno. La siguiente figura ilustra las complicaciones asociadas a esta elección. La misma presenta cuatro versiones del gráfico para un número variable de intervalos. Nótese que un intervalo excesivamente grande (es decir, un número pequeño de barras) provoca pocos saltos en el histograma, pero tiende a diferir notoriamente con respecto a la distribución verdadera, al agrupar en una misma barra a observaciones con valores muy diferentes. En el otro extremo, una elección de intervalos muy pequeños representa mejor a los verdaderos datos, pero al costo de un gráfico con muchos saltos. Se trata del *trade-off* entre precisión y volatilidad: cuanto menor es el intervalo, más precisa es la representación de los datos, pero a la vez menos útil, dado que se reproduce toda la variabilidad de la información original y la representación se vuelve confusa. El histograma se parece cada vez más a la distribución real, pero cumple cada vez menos con su función simplificadora.

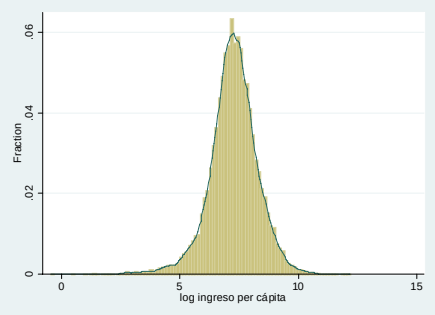
**Figura 2.5**
**Histograma del logaritmo del ingreso per cápita familiar**
**México, 2006**




Fuente: elaboración propia en base a microdatos de la ENIGH.

La figura 2.6 muestra, además del histograma, una versión "suavizada" del mismo en línea continua. Técnicamente, estos "histogramas suavizados" son estimaciones no paramétricas por el método de *kernels* de la función de densidad, en este caso del logaritmo del ingreso per cápita familiar. En la próxima sección presentaremos a las funciones de densidad y los métodos no paramétricos para estimarlas.

**Figura 2.6**
**Histograma del logaritmo del ingreso per cápita familiar y estimación de la función de densidad por kernels**
**México, 2006**



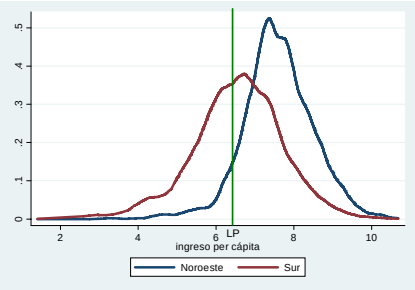

Fuente: elaboración propia en base a microdatos de la ENIGH.
Nota: 100 intervalos.

Una de las ventajas de estos "histogramas suavizados" es facilitar las comparaciones, ya que resulta incómodo superponer dos histogramas reales. La figura 2.7 ilustra los "histogramas suavizados" de la distribución del ingreso per cápita familiar (en logaritmos) en dos regiones de México: el Noroeste y el Sur. Las dos distribuciones son claramente diferentes. La distribución del Sur está desplazada a la izquierda, lo que sugiere que en general los individuos de esa región tienen menores ingresos que en el Noroeste. De hecho, el ingreso per cápita promedio en el Sur es menos de la mitad que en el Noroeste.

La línea vertical del gráfico marca la línea de pobreza internacional de USD 2.5 por día por persona para México (en logaritmos). Si recordamos que un histograma presenta frecuencias relativas, es intuitivamente claro que a la izquierda de la línea de pobreza hay más individuos, en proporción a la población de cada región, en el Sur que en el Noroeste.[12]

**Figura 2.7**
**Estimaciones por kernels de las funciones de densidad del logaritmo del ingreso per cápita familiar**
**Regiones Noroeste y Sur de México, 2006**

[12] La proporción de personas por debajo de la línea de USD 2.5 resulta ser 34.5% en el Sur y 9.3% en el Noroeste.

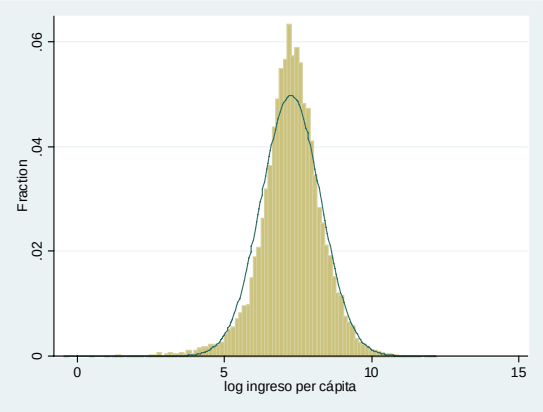



Fuente: elaboración propia en base a microdatos de la ENIGH.

El histograma suavizado del Sur está más "aplanado" que el del Noroeste, lo que es señal de mayor dispersión. En el Noroeste la mayor parte de las observaciones se concentran en un rango más estrecho de ingresos, lo que sugiere una menor dispersión en los datos, que va asociada a una menor desigualdad. Vamos a dedicar una gran parte de este libro a definir y medir pobreza y desigualdad, pero intuitivamente podemos inferir a partir de la figura 2.7 que el Sur de México es una región con más pobreza y más desigual que el Noroeste.

Algunos lectores habrán notado que los histogramas del logaritmo del ingreso se parecen al que resulta de una distribución normal (o Gaussiana). En la figura 2.8 repetimos el histograma resultante de tomar 100 intervalos, junto al gráfico de una distribución normal con media y varianza idénticas a la de los datos reales. La función normal se asemeja al histograma, pero no es igual. ¿Es posible asumir que el logaritmo del ingreso se ajusta a una distribución normal? Volveremos sobre este punto en la sección 2.7 de este capítulo.

**Figura 2.8**
**Histograma del logaritmo del ingreso per cápita familiar y distribución normal**
**México, 2006**

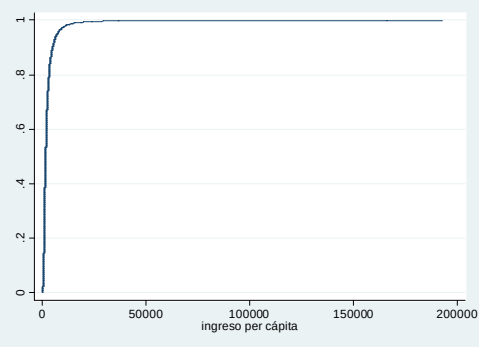



Fuente: elaboración propia en base a microdatos de la ENIGH.
Nota: 100 intervalos.

### 2.3.2. Función de distribución

Una manera alternativa de graficar una distribución es a través de su función de distribución acumulada (FDA), usualmente llamada simplemente función de distribución. La FDA grafica la proporción de personas con ingresos menores a cada valor del soporte de la distribución marcado en el eje horizontal. La FDA comienza en el origen de coordenadas. En el otro extremo, para todo valor mayor al ingreso más alto de la muestra la FDA es 1. La figura 2.9 muestra la FDA de México 2006.

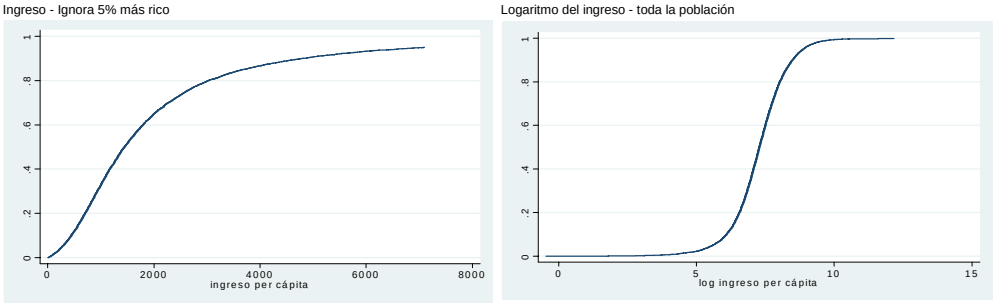
**Figura 2.9**
**Función de distribución del ingreso per cápita familiar**
**México, 2006**




Fuente: elaboración propia en base a microdatos de la ENIGH.

Nuevamente, la cola superior larga de la distribución vuelve al gráfico poco útil. Para aliviar este problema las alternativas son o bien truncar los valores superiores del ingreso, o trabajar en logaritmos. La figura 2.10 muestra ambas alternativas.

**Figura 2.10**
**Función de distribución del ingreso per cápita familiar**
**México, 2006**



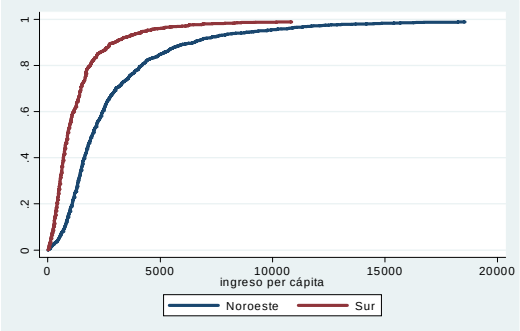

Fuente: elaboración propia en base a microdatos de la ENIGH.

La FDA es una función no decreciente en los ingresos, con saltos en cada punto donde observamos ingresos. De cualquier forma, dado el gran número de observaciones en una encuesta de hogares típica, gráficamente la función de distribución parece ser suave.

Es fácil ubicar los cuantiles en base a la FDA, marcando una proporción en el eje vertical e identificando el cuantil correspondiente en el horizontal. Por ejemplo, para ubicar la mediana debe marcarse el valor 0.5 en el eje vertical y leer el valor correspondiente implicado por la FDA en el eje horizontal (técnicamente, la preimagen de la FDA).

Las funciones de distribución son instrumentos muy útiles para evaluar pobreza. La figura 2.11 muestra las FDA del ingreso per cápita familiar en el Noroeste y el Sur de México. Nótese que la función de distribución del Sur está siempre por arriba de la del Noroeste. En la jerga estadística se dice que la FDA del Noroeste domina en sentido estocástico de primer orden a la FDA del Sur. Fijemos la línea de pobreza en cualquier valor arbitrario en el eje horizontal. Es sencillo ver que la proporción de personas con ingresos inferiores a ese nivel es siempre más grande en el Sur que en el Noroeste. El hecho que la FDA del Sur esté siempre por arriba garantiza que la tasa de pobreza es mayor en esa región, para cualquier línea de pobreza. Este es un resultado muy importante que examinaremos con más detalle en el capítulo 4.

**Figura 2.11**
**Función de distribución del ingreso per cápita familiar**
**Noroeste y Sur de México, 2006**



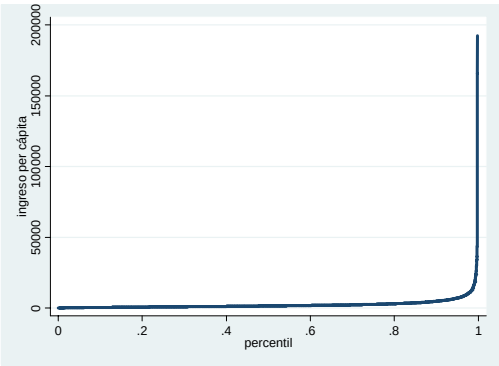

Fuente: elaboración propia en base a microdatos de la ENIGH.

### 2.3.3. El desfile de los enanos y unos pocos gigantes

Pueden seguir leyendo; el título no pertenece a otro libro. El "desfile de los enanos y unos pocos gigantes" es el nombre de un gráfico propuesto por Pen (1973) para visualizar distribuciones. La motivación de Pen para esta ilustración es la siguiente. Supongamos que ordenamos a toda la población de acuerdo a sus ingresos de forma ascendente - del más pobre al más rico - y hacemos que la altura de cada persona coincida con su ingreso. Ahora nos subimos a un estrado y hacemos desfilar a la población así ordenada. ¿Qué forma se va formando a medida que transcurre el desfile? La figura 2.12 muestra el desfile para el caso mexicano. Más concretamente, la curva de Pen muestra el ingreso correspondiente a cada cuantil de la distribución.

**Figura 2.12**
**Gráfico de Pen**

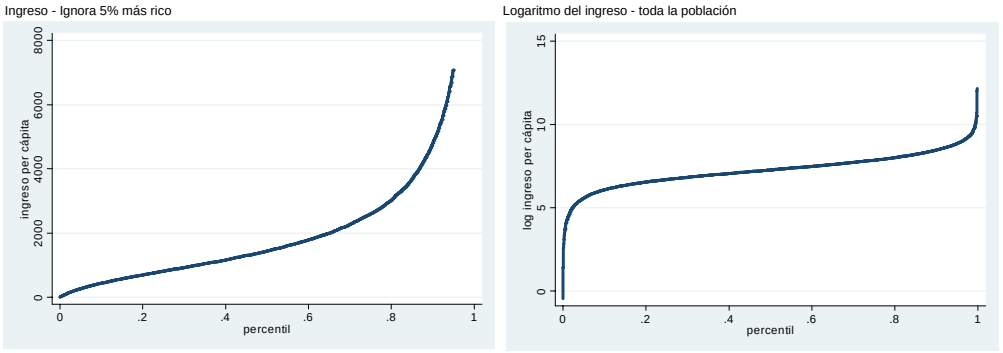
**México, 2006**




Fuente: elaboración propia en base a microdatos de la ENIGH.

El gráfico se mantiene visualmente casi horizontal hasta los últimos percentiles donde crece enormemente: es un desfile de enanos y unos pocos gigantes. La forma de este gráfico es consecuencia, una vez más, de la cola superior larga de las distribuciones. La figura 2.13 se vuelve más legible al eliminar al 5% más rico de la población, o al trabajar con el ingreso en logaritmos.

**Figura 2.13**
**Gráfico de Pen**
**México, 2006**




Fuente: elaboración propia en base a microdatos de la ENIGH.

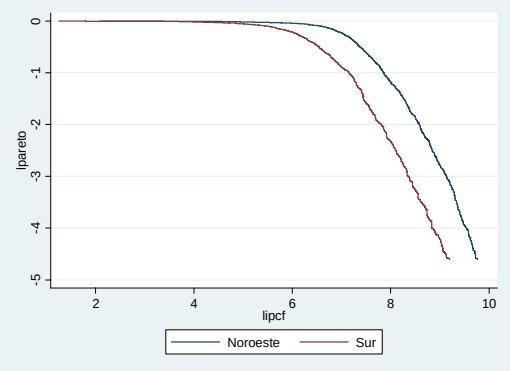
Nótese que pese a lo interesante de la motivación, la curva de Pen no agrega información respecto de la FDA. De hecho, se trata de la propia FDA pero graficada con los ejes invertidos.

### 2.3.4. El diagrama de Pareto

Este gráfico muestra para cada valor del ingreso $x$ el porcentaje de la población que recibe ingresos superiores a ese valor $x$, en una escala doble logarítmica. El cambio de escala genera una suerte de *zoom* óptico sobre los estratos de mayores ingresos, permitiendo un examen más detallado de esa parte de la distribución.

La figura 2.14 presenta el diagrama de Pareto para el Noroeste y Sur de México. El eje horizontal muestra el ingreso en logaritmos, mientras que el eje vertical mide la proporción de personas con ingreso superior a *x* en logaritmos. El valor 0 en ese eje corresponde al total de la población ya que ln(1)=0, mientras que el -4, por ejemplo, a menos del 2% más rico de la población, ya que ln(0.0184)=-4. En el ejemplo la proporción de personas con ingresos mayores a un determinado valor en la cola superior del soporte de la distribución es siempre más alta en el Noroeste que en el Sur de México.

**Figura 2.14**
**Diagrama de Pareto del ingreso per cápita familiar**
**Noroeste y Sur de México, 2006**




Fuente: elaboración propia en base a microdatos de la ENIGH.

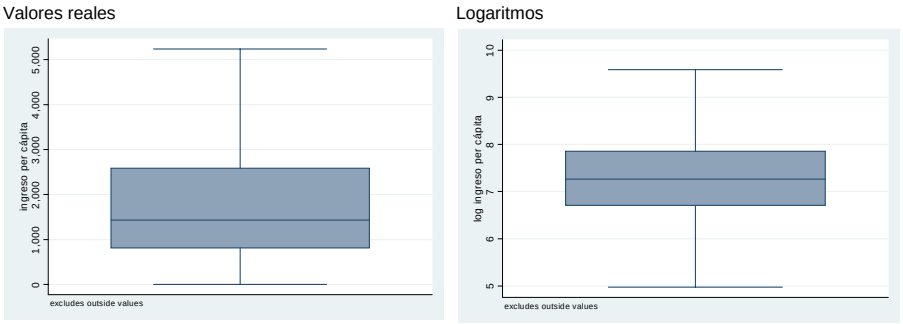
### 2.3.5. *Box-Plot*

Otro gráfico interesante para describir una distribución es el *box-plot* o diagrama de caja. El gráfico presenta una caja (*box*) cuyo lado inferior se corresponde con el primer cuartil y el superior con el tercer cuartil, de modo que la altura de la caja mide el rango intercuartílico. La línea horizontal dentro de la caja es la mediana. Del lado superior de la caja sale una línea vertical, cuyo extremo superior indica el valor máximo de la distribución. En forma análoga, la línea debajo de la caja tiene como punto extremo inferior al valor mínimo. El gráfico de *box-plot* suele construirse eliminando las observaciones extremas (*outliers*).[13] La figura 2.15 muestra el *box-plot* de la distribución del ingreso per cápita familiar de México 2006, tanto con los valores originales, como transformados en logaritmos.

**Figura 2.15**
**Box-plot**
**Distribución del ingreso per cápita familiar**
**México, 2006**

[13] Algunas versiones de este tipo de gráfico reemplazan los extremos inferiores y superiores del diagrama por cuantiles extremos (por ejemplo, 0.05 y 0.95).

Excluyendo valores extremos

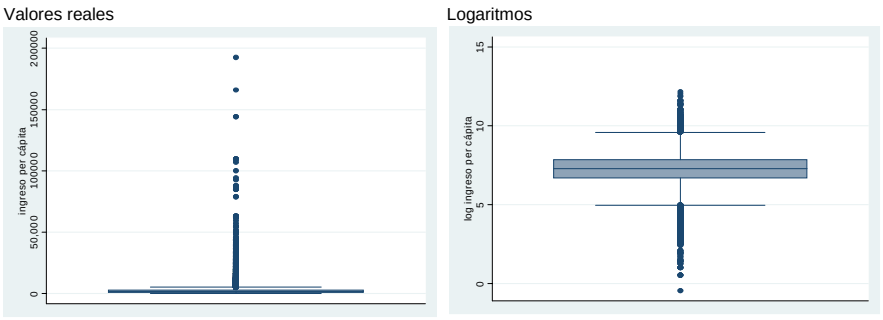



Fuente: elaboración propia en base a microdatos de la ENIGH.

La figura 2.16 incluye los valores extremos y los marca con puntos. Una vez más el gráfico en valores reales se hace difícil de leer, a diferencia del gráfico en logaritmos.

**Figura 2.16**
**Box-plot**
**Distribución del ingreso per cápita familiar**
**México, 2006**

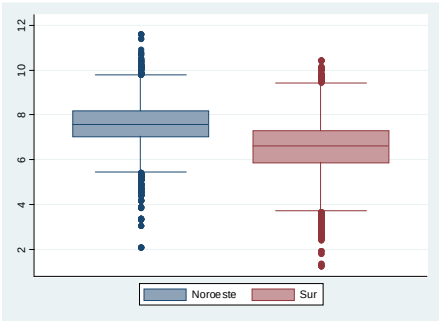
Incluyendo valores extremos




Fuente: elaboración propia en base a microdatos de la ENIGH.

El *box-plot* es una forma gráfica de resumir el rango de los ingresos, su tendencia central (medida por la mediana) y la dispersión, medida por el rango intercuartílico. De la figura 2.17 surge que en el Noroeste mexicano los ingresos son en general más altos y menos dispersos que en el Sur.

**Figura 2.17**
**Box-plot**
**Distribución del logaritmo del ingreso per cápita familiar**
**Noroeste y Sur de México, 2006**

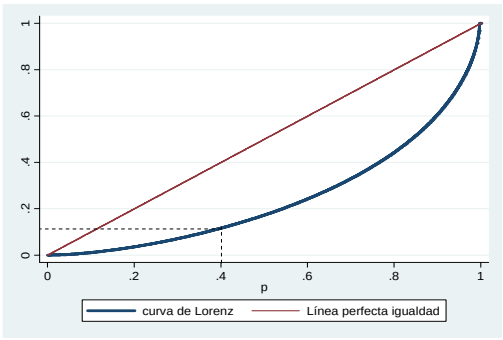



Fuente: elaboración propia en base a microdatos de la ENIGH.

### 2.3.6. Curva de Lorenz

Esta curva, introducida por Lorenz (1905), es una de las formas gráficas más utilizadas para estudiar desigualdad. La curva se grafica en una caja de dimensiones 1x1, donde el eje horizontal indica la proporción *p* de personas de menores ingresos en la población. Por ejemplo, un valor $p = 0.12$ hace referencia al 12% más pobre de la población. La curva de Lorenz grafica en el eje vertical el porcentaje acumulado del ingreso correspondiente al *p* por ciento más pobre de la población. La figura 2.18 ilustra la curva de Lorenz para México 2006. El gráfico indica, por ejemplo, que el 40% de la población con menores ingresos reúne poco más del 10% del ingreso nacional total.

**Figura 2.18**
**Curva de Lorenz**
**Distribución del ingreso per cápita familiar**
**México, 2006**



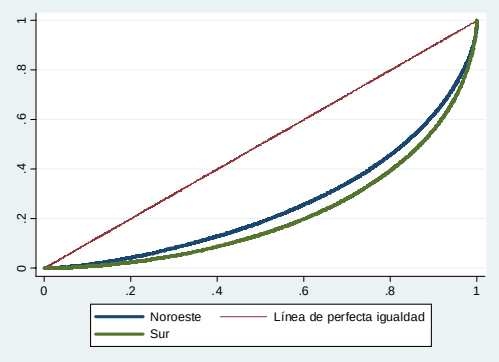

Fuente: elaboración propia en base a microdatos de la ENIGH.

Nótese que si todas las personas tuvieran exactamente el mismo ingreso, la curva de Lorenz coincidiría con la recta de 45º. Por esta razón, la diagonal de la caja recibe el nombre de *línea de perfecta igualdad* y proporciona una base útil para la comparación. En el otro extremo, si el ingreso fuera cero para toda la población, excepto para un individuo (que entonces sería quien concentra todo el ingreso), la curva de Lorenz coincidiría con los laterales inferior y derecho de la caja.

Es fácil observar las siguientes propiedades de la curva de Lorenz. Si se trata de magnitudes positivas (como el caso de los ingresos) la curva comienza en el punto (0,0) - el 0% de la población más pobre acumula el 0% de los ingresos -, es no-decreciente y termina en el punto (1,1) - el 100% de la población acumula todo el ingreso. La curva de Lorenz es homogénea de grado cero en los ingresos, implicando que si todos los ingresos se duplican (o se multiplican por cualquier otro escalar positivo) la curva permanece inalterada. Finalmente, la curva de Lorenz no puede estar por arriba de la línea de perfecta igualdad ni, naturalmente, por debajo de la curva de completa desigualdad.

Es fácil intuir que cuanto más alejada de la línea de perfecta igualdad esté la curva de Lorenz, más desigual resultará la distribución. La figura 2.19 muestra la curva de Lorenz de dos regiones en México, sugiriendo una distribución del ingreso más desigual en el Sur que en el Noroeste. El capítulo 6 trata la relación entre las curvas de Lorenz y la desigualdad con más detalle.

**Figura 2.19**
**Curva de Lorenz**
**Distribución del ingreso per cápita familiar**
**México, 2006**

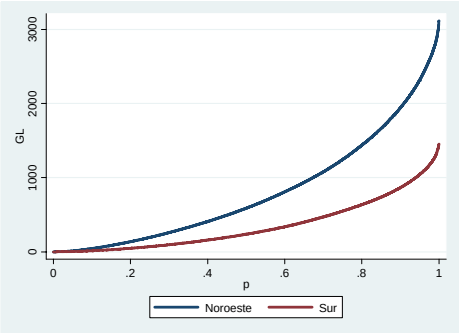



Fuente: elaboración propia en base a microdatos de la ENIGH.

## Curva generalizada de Lorenz

Esta generalización consiste en multiplicar la curva de Lorenz por la media de la distribución. Gráficamente se obtiene a través de una expansión $\mu$ veces de la curva de Lorenz. En consecuencia, la curva generalizada de Lorenz muestra el ingreso acumulado en el $p$% más pobre de la población, sobre el número de personas $N$. Esta curva parte del origen de coordenadas y llega hasta el punto (1, $\mu$). Como veremos en los capítulos 6 y 7, mientras que la curva de Lorenz se emplea para estudiar desigualdad, la generalizada de Lorenz es muy útil para analizar bienestar agregado. La figura 2.20 muestra que la curva del Noroeste de México está por encima de la del Sur, denotando un nivel de bienestar superior.

**Figura 2.20**
**Curva generalizada de Lorenz**
**Distribución del ingreso per cápita familiar**
**México, 2006**

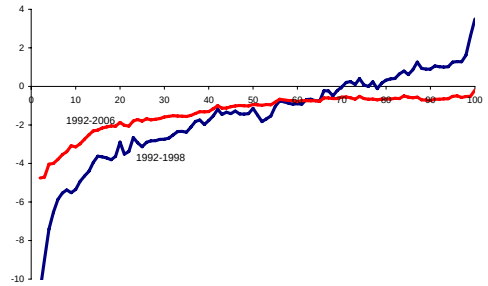



Fuente: elaboración propia en base a microdatos de la ENIGH.

### 2.3.7. Distribuciones en movimiento

Las distribuciones van cambiando en el tiempo, lo cual introduce una nueva dimensión en el análisis - la temporal -, volviéndolo a la vez más interesante y complicado. La dinámica distributiva será analizada en varios puntos del libro. En este apartado comenzamos por presentar algunos instrumentos gráficos. Uno de los más útiles y sencillos es la curva de incidencia del crecimiento (*growth-incidence curve*). Se trata simplemente de graficar en el eje vertical la tasa de crecimiento - o alternativamente el cambio proporcional - del ingreso real (es decir, a precios constantes) en un período de tiempo en cada uno de los cuantiles de la distribución.

**Figura 2.21**
**Curvas de incidencia del crecimiento del ingreso per cápita familiar**
**Argentina, 1992-1998 y 1992-2006**



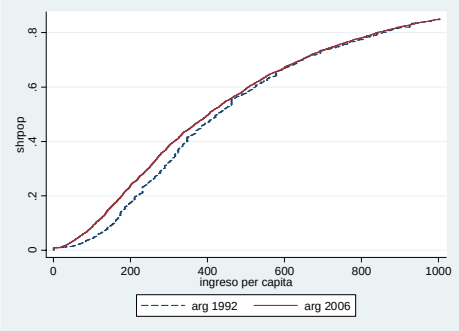

Fuente: elaboración propia en base a microdatos de la EPH.

Dejemos el ejemplo de México y tomemos el caso de Argentina para ilustrar cambios distributivos. La figura 2.21 muestra que la curva de incidencia del crecimiento de ese país del Cono Sur para el período 1992-2006 está completamente por debajo del eje horizontal, reflejando una caída en los ingresos de todos los percentiles de la distribución. Es claro que de acuerdo a este gráfico la pobreza de ingresos absoluta

aumentó en Argentina durante ese período, cualquiera sea la línea de pobreza escogida. Otra característica de las curvas de incidencia de la figura 2.21 es que son crecientes. Esta "pendiente" positiva implica caídas proporcionales del ingreso más grandes a medida que vamos descendiendo hacia estratos más pobres de la distribución. Es claro que la desigualdad de ingresos debe haber aumentado en Argentina, en particular entre 1992 y 1998.

Las tres figuras siguientes ilustran los cambios distributivos con gráficos conocidos. La 2.22 muestra la función de distribución y sugiere también caída de ingresos y aumento de la pobreza entre 1992 y 2006.

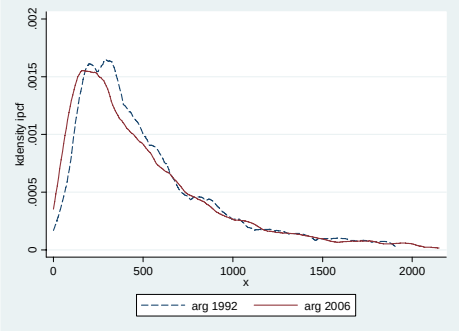
**Figura 2.22**
**Funciones de distribución del ingreso per cápita familiar**
**Argentina, 1992 y 2006**




Fuente: elaboración propia en base a microdatos de la EPH.

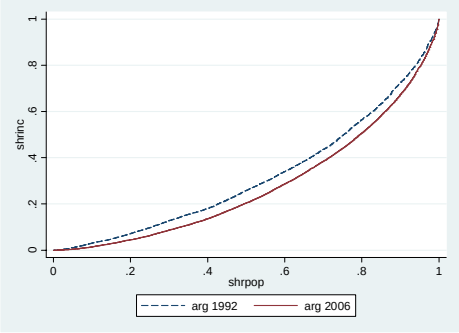
La figura 2.23 es clara al indicar el corrimiento horizontal hacia la izquierda de la función de densidad del ingreso, y por ende el aumento en la pobreza, mientras que las curvas de Lorenz de la figura 2.24 son sugerentes del aumento de la desigualdad.

**Figura 2.23**
**Funciones de densidad del ingreso per cápita familiar**
**Argentina, 1992 y 2006**




Fuente: elaboración propia en base a microdatos de la EPH.

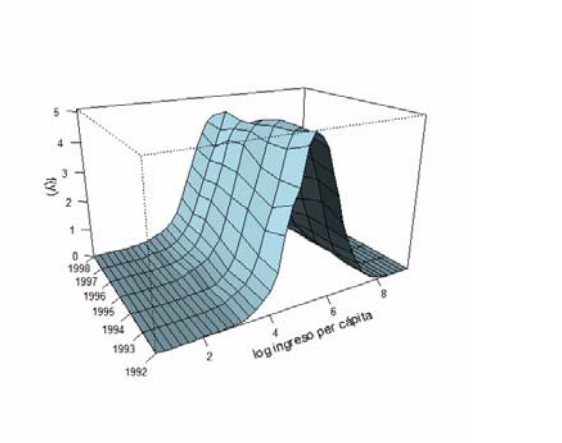
**Figura 2.24**
**Curvas de Lorenz del ingreso per cápita familiar**
**Argentina, 1992 y 2006**




Fuente: elaboración propia en base a microdatos de la EPH.

Es posible presentar varias funciones de densidad en un gráfico de tres dimensiones, aunque su lectura no siempre es sencilla. La figura 2.25 muestra las densidades anuales de la distribución del ingreso per cápita familiar en Argentina entre 1992 y 1998, sugiriendo un progresivo aumento de la dispersión de ingresos.

**Figura 2.25**
**Funciones de densidad del ingreso per cápita familiar**
**Argentina, 1992 a 1998**




Fuente: elaboración propia en base a microdatos de la EPH.

Las representaciones gráficas son útiles para visualizar una distribución, compararla con otras y evaluar sus cambios en el tiempo. Es altamente recomendable comenzar todo análisis distributivo desplegando un conjunto de ilustraciones como las presentadas en esta sección. En ocasiones, un gráfico es todo lo que necesitamos para acompañar un argumento. A menudo, sin embargo, pretendemos una evaluación más detallada de alguna característica de la distribución, o buscamos cuantificar diferencias con otras distribuciones o cambios temporales. Para estos casos es necesario ir más allá de una simple representación gráfica y trabajar una distribución en términos analíticos, para lo cual debemos pedir ayuda a las matemáticas. En el resto de este capítulo el enfoque

analítico ocupa un lugar central. El lector no especializado puede saltear las secciones siguientes, aunque es recomendable que haga el esfuerzo ahora para aprovechar plenamente luego todo el material del resto del libro.

## 2.4. Funciones continuas

Aunque en la realidad los datos disponibles son discretos, a menudo es útil trabajar con las versiones analíticas continuas de las funciones y gráficos presentados en la sección anterior.

### 2.4.1. Funciones

La versión suave del histograma es la *función de densidad* $f(x)$. Para un valor infinitesimal $dx$, $f(x)dx$ es la proporción de individuos cuyos ingresos pertenecen al intervalo $[x, x+dx]$. Consideremos los niveles de ingresos $x_1$ y $x_2$. El hecho que $f(x_1)$ sea mayor que $f(x_2)$ indica que la probabilidad de encontrar ingresos en un intervalo pequeño alrededor de $x_1$ es mayor que alrededor de $x_2$, es decir, hay relativamente más personas con ingresos similares a $x_1$ que a $x_2$.

Dado que en general se consideran sólo ingresos no negativos, la convención es trabajar en el soporte $[0, \infty)$. La función de densidad $f(x)$ de los ingresos tiene dos propiedades básicas

$$f(x) \geq 0\,; \qquad \int_0^\infty f(x)dx = 1 \tag{2.7}$$

A partir de la función de densidad es posible definir algunas de las medidas resumen discutidas anteriormente. Por ejemplo, la media es

$$\mu = \int_0^\infty xf(x)dx \tag{2.8}$$

y la varianza

$$V = \int_0^\infty (x-\mu)^2 f(x)dx \tag{2.9}$$

El ingreso acumulado entre dos valores $a$ y $b$, una magnitud a usar extensamente en el libro, es igual a

$$N\int_a^b xf(x)dx \tag{2.10}$$

donde $N$ es el total de la población. La función de distribución $F(x)$ o función de densidad acumulada (FDA), que indica la proporción de observaciones hasta un

determinado valor del ingreso $x$, es la integral de la función de densidad hasta ese valor $x$.

(2.11) $$F(x) = \int_0^x f(s)ds$$

En consecuencia,

(2.12) $$f(x) = \frac{dF(x)}{dx}$$

La función de distribución permite definir con facilidad los distintos cuantiles o percentiles. El percentil $p$ de la distribución es el valor del ingreso $x_p$ tal que [14]

(2.13) $$F(x_p) = p$$

Por ejemplo, la mediana es el valor del ingreso para el cual $F$ es igual a 0.5, y el primer decil el valor para el que $F$ es igual a 0.1.

La curva de Pen asociada a la distribución $F$ (recordar el "desfile de los enanos") puede escribirse como

(2.14) $$Q(F, p) = \min\{x / F(x) \geq p\}$$

es decir, el ingreso que le corresponde a la persona en la posición $p$ de la distribución.

La curva de Lorenz puede escribirse en términos continuos como

(2.15) $$L(p) = \int_0^y \frac{xf(x)dx}{\mu}, \text{ con } p=F(y)$$

Para interpretar esta ecuación nótese que $y$ es el valor tal que el $p$ por ciento de la población tiene ingresos menores a este valor. Ahora, por analogía con (2.10) nótese que

(2.16) $$N\int_0^y xf(x)dx$$

es el ingreso acumulado desde la persona más pobre hasta aquella con ingreso $y$. Luego $L(p)$ definido arriba resulta ser el porcentaje del ingreso total acumulado en el $p$ por ciento más pobre de la población.

De la definición de $L(p)$ es simple ver que $L(0) = 0$ y $L(1) = 1$. Derivando y asumiendo $f(y) > 0$ se llega a

(2.17) $$\frac{\partial L(p)}{\partial p} = \frac{\partial L(p)}{\partial y}\frac{\partial y}{\partial p} = \frac{yf(y)}{\mu}\frac{1}{f(y)} = \frac{y}{\mu} \geq 0$$

[14] Nótese que acá estamos aludiendo a los percentiles como observaciones singulares y no en la acepción alternativa de grupos de observaciones.

La pendiente de la curva de Lorenz es positiva (o cero para ingresos nulos). Derivando una vez más respecto de *p*,

$$\frac{\partial L^2(p)}{\partial p^2} = \frac{1}{\mu}\frac{1}{f(y)} \geq 0 \tag{2.18}$$

lo que indica que la curva de Lorenz es convexa. Dado que la curva parte del origen y llega al punto (1,1), y que es creciente y convexa, entonces se concluye que ningún punto de esa curva puede estar más allá de la recta de 45 grados en una caja de dimensiones 1x1. Nótese adicionalmente de (2.15) que la curva de Lorenz es homogénea de grado cero en los ingresos; un cambio en la escala de medición de los ingresos no modifica la ubicación de la curva.

Es posible obtener la función de distribución a partir de conocer su media μ y su curva de Lorenz $L(p)$. Denotando con $L'(p)$ a la pendiente de la curva de Lorenz y recordando que $p = F(y)$

$$L'(F(y)) = \frac{y}{\mu} \tag{2.19}$$

por lo que

$$F(y) = L'^{-1}\left(\frac{y}{\mu}\right) \tag{2.20}$$

donde la potencia -1 indica la inversa de la función. De (2.20), conociendo la media $\mu$ y la pendiente de la curva de Lorenz en cada punto, podemos rescatar la función de distribución de los ingresos original.

Recordemos que la curva generalizada de Lorenz indica el ingreso acumulado por el *p*% más pobre de la población dividido por el tamaño de la población *N*. Formalmente,

$$GL(p) = \int_0^y xf(x)dx \quad , F(y) = p \tag{2.21}$$

Nótese que si multiplicamos por *N* esta expresión el numerador indica el ingreso acumulado hasta el percentil *p* de la distribución. Si multiplicamos y dividimos (2.21) por la media de la distribución,

$$GL(p) = \mu\int_0^y \frac{xf(x)}{\mu}dx = \mu L(p) \tag{2.22}$$

La curva generalizada de Lorenz no es más que una expansión $\mu$ veces de la curva de Lorenz. Es fácil entonces ver que *GL* comienza en el punto (0, 0) y termina en (1, $\mu$) y que su pendiente es

$$\frac{\partial GL(p)}{\partial p} = \mu\frac{\partial L(p)}{\partial p} = \mu\frac{y}{\mu} = y \tag{2.23}$$

### 2.4.2. Gráficos

Mientras que los gráficos de $F(x)$, $Q(F, p)$, $L(p)$ o $GL(p)$ no ofrecen complicaciones y son una extensión natural de sus versiones discretas, la ilustración de $f(x)$ es, quizás sorprendentemente, complicada. Un histograma es ciertamente una forma de graficar la función de densidad $f(x)$, aunque rudimentaria, ya que supone una distribución uniforme dentro de cada intervalo, lo que genera saltos discretos en el gráfico. En lo que sigue discutiremos una estrategia para construir una representación más suave de la densidad, la cual adicionalmente permite ilustrar y aproximar con mayor precisión el problema de la elección del tamaño de los intervalos, mencionado en la sección anterior. Dicha representación no-paramétrica, denominada método de núcleos o *kernels*, puede ser apropiadamente vista como una generalización de la noción de histograma.

A partir de (2.12) la función de densidad en un punto $x_0$ es

$$f(x_0) = \left. \frac{dF(x)}{dx} \right|_{x_0} \tag{2.24}$$

Consecuentemente, recurriendo a la definición de derivada en un punto, vale la siguiente aproximación

$$f(x_0) \cong \frac{F(x_0 + h) - F(x_0 - h)}{2h} \tag{2.25}$$

donde $h > 0$. Naturalmente, esta aproximación tiende a ser exacta cuando $h$ tiende a 0. Ahora, nótese que $F(x_0+h) - F(x_0-h)$ es la proporción de observaciones con valores de ingreso entre $x_0-h$ y $x_0+h$. Ese valor dividido por $2h$ es una aproximación de $f(x_0)$. Lo que hemos realizado no difiere sustancialmente de un histograma. Gráficamente, comenzamos fijando un punto $x_0$, luego construimos un intervalo alrededor de este punto $(x_0-h,\ x_0+h)$ de ancho $2h$ y luego procedimos a calcular la proporción de observaciones que caen en este intervalo, normalizando por el ancho del mismo. A fines de construir un gráfico para toda la función de densidad podríamos repetir la estrategia anterior en una grilla de puntos (no necesariamente equiespaciada ni coincidente con los ingresos de nadie en la muestra).

El parámetro $h$, que cumple un rol fundamental en esta estrategia, es llamado "ancho de banda". La elección de este parámetro conlleva el mismo *trade-off* entre precisión y volatilidad comentado arriba para el caso del histograma. Cuanto menor es $h$, más precisa es la representación de los datos, pero vuelve el gráfico muy volátil y por consiguiente poco útil. La elección de un ancho de banda adecuado es, de hecho, el problema más delicado a resolver a la hora de utilizar este método. Existen varias estrategias a seguir para resolver este problema, pero ninguna de ellas ofrece una solución mecánica y confiable. Siguiendo a Deaton (2007), la recomendación práctica es explorar con varios anchos de banda, comenzando con uno muy pequeño y terminando con uno muy grande, a fines de ilustrar la ganancia (suavidad) y la pérdida (precisión).

El método de *kernels* nos ayuda a obtener estimaciones de $f(x)$ en cada punto. Para entender como funciona, en primer lugar nótese que si una observación $x_i$ cae en el intervalo entre $x_0$-$h$ y $x_0$+$h$ entonces

$$\left|\frac{x_i - x_0}{h}\right| < 1 \tag{2.26}$$

Entonces, un estimador de $f(x)$ puede ser reescrito de la siguiente forma

$$\hat{f}(x_0) = \frac{1}{N2h}\sum_{i=1}^{N} 1\left[\left|\frac{x_i - x_0}{h}\right| < 1\right] \tag{2.27}$$

La función 1[.] indica con 1 a todas las observaciones que caen dentro del intervalo y con 0 a aquellas que no. Consecuentemente, la sumatoria es igual a la cantidad de observaciones que caen dentro del intervalo. La fórmula anterior puede ser re-expresada de la siguiente forma

$$\hat{f}(x_0) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{h} K\left[\frac{x_i - x_0}{h}\right] \tag{2.28}$$

donde $K[(x_i\text{-}x_0)/h] = ½\ 1[|(x_i\text{-}x_0)/h| <1]$. La función $K(.)$ recibe el nombre de *kernel* y es interesante observar su papel. En cierto sentido, el kernel redefine cuán lejos está una observación $x_i$ de un valor $x_0$. El kernel utilizado en este caso, llamado kernel rectangular, lo hace de una forma peculiar y discontinua: al asignarle valor 1 a todas las observaciones que caen en el intervalo $x_0 \pm h$, sugiere una noción discontinua de distancia, donde "cerca" están todas las observaciones indicadas con 1 por el kernel (las que caen dentro del intervalo), y "lejos" todas las indicadas con 0 (las que caen fuera). El parámetro que controla esta noción de "cerca" o "lejos" es $h$: cuanto más grande es este valor, mayor es el intervalo alrededor del punto $x_0$ y consecuentemente más observaciones son consideradas como "cercanas" por el kernel.

Existen varias alternativas al kernel rectangular discutido anteriormente. Definamos $v=(x_i\text{-}x_0)/h$. El kernel gaussiano está dado por

$$K(v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2} \tag{2.29}$$

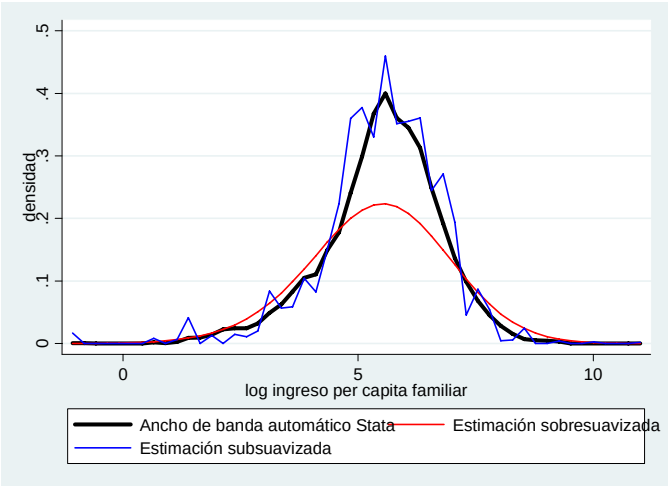
Notar que en este caso el kernel va otorgando importancia suavemente decreciente a las observaciones lejanas de $x_0$. Otros ejemplos son el kernel cuadrático o el triangular. Uno que recibe considerable atención en la práctica es el *kernel de Epanichnikov*

$$K(v) = \frac{3}{4}(1 - v^2),\ si\ |v| \le 1, \text{y } 0 \text{ en caso contrario.} \tag{2.30}$$

En la práctica, la elección del ancho de banda $h$ tiene mucho más impacto que la elección del kernel. La figura 2.26 ilustra el papel del ancho de banda, mostrando estimaciones no paramétricas alternativas de la densidad del logaritmo del ingreso per cápita familiar en Bolivia 2005, con kernels gaussianos. Se presentan tres estimaciones, para distintas elecciones de ancho de banda. Las estimaciones parecen sugerir la

simetría de la distribución de los logaritmos de los ingresos. La estimación en trazo grueso es aquella calculada con el ancho de banda escogido automáticamente por Stata.[15] La estimación con un ancho de banda pequeño es más errática, pero de cualquier manera tiende a sugerir la misma forma de la función de densidad que la producida por la estimación que surge al utilizar un ancho de banda intermedio. En el otro extremo, un ancho de banda exageradamente grande produce una estimación muy suave, que en ciertos tramos de ingresos difiere sistemáticamente de la intermedia. Nótese que, comparada con la intermedia, esta estimación demasiado suavizada sobreestima la densidad en los sectores de bajos y altos ingresos, y la subestima en el sector de ingresos medios.

**Figura 2.26**
**Estimaciones no paramétricas de la función de densidad**
**Logaritmo del ingreso per cápita, anchos de banda alternativos**
**Bolivia, 2005**




Fuente: elaboración propia sobre la base de datos de la ECH

## 2.5. El enfoque inferencial

Volvamos por un momento al ejemplo de Brasil del cuadro 2.1. Si el objetivo consistiese simplemente en caracterizar o resumir la información de ingresos de las 394560 personas relevadas por la encuesta, el enfoque descriptivo adoptado anteriormente alcanzaría. El análisis se torna más sofisticado (e interesante) al reconocer que estos datos son una muestra de una población más numerosa o general. El problema consiste ahora en aprender algo acerca de la población a través de la muestra. A modo de ejemplo, el ingreso promedio de los datos relevados por la encuesta PNAD es igual a 574.3 reales. La pregunta fundamental es cuán acertado es el valor 574.3 (la media muestral, observable) como estimación del ingreso medio de toda la población brasileña

[15] Este ancho de banda minimiza el error cuadrático medio integrado si la verdadera distribución fuera normal y se utilizara un kernel gaussiano.

(la media poblacional, inobservable). Este tipo de problema constituye la esencia del enfoque inferencial y, en general, de la Estadística: estudiar mecanismos que permitan aprender características poblacionales (su centro, dispersión, etc.) a partir de una muestra. Este enfoque requiere establecer un vínculo claro entre la población y la muestra, el cual es usualmente provisto en un marco probabilístico, que discutiremos brevemente a continuación.

El punto de partida es una variable aleatoria $X$, que en nuestro caso representa a alguna dimensión del bienestar individual y que por simplicidad pedagógica pensaremos nuevamente que es el ingreso. En este caso resulta conveniente representarlo a través de una variable aleatoria continua y positiva que toma valores en el intervalo real $[0, \infty)$. La *función de distribución acumulada* de dicha variable es $F(x)$: $[0, \infty) \rightarrow [0, 1]$ tal que

$$F(x) = Pr(X \leq x) \tag{2.31}$$

es decir, $F(x)$ indica la probabilidad de que la variable aleatoria $X$ tome valores menores o iguales a un valor del soporte $x$.

Una muestra aleatoria de tamaño $N$, independiente e idénticamente distribuida (*iid*) de la variable aleatoria $X$ consiste en una colección de $N$ variables aleatorias $X_1, X_2, \ldots, X_N$ todas ellas independientes entre sí, y cada una de ellas distribuidas de la misma manera que $X$, es decir con función de distribución acumulada $F(x)$. Las realizaciones de esta muestra aleatoria son los datos de ingreso. En este contexto, cada uno de los ingresos efectivamente captados por la PNAD de Brasil es visto como una realización de una variable aleatoria que representa al ingreso de cada persona.

En la práctica la variable aleatoria $X$ se intenta conocer a través de los datos de una muestra, típicamente en nuestro caso los microdatos de una encuesta de hogares. Definiremos a la función de distribución acumulada empírica $F_N(x)$ como la proporción de observaciones de ingresos en la muestra menores a $x$. El nexo entre $F(x)$ y su versión empírica $F_N(x)$ está dado por el Teorema Fundamental de la Estadística (Glivenko-Cantelli), que asegura que bajo condiciones generales, cuando el tamaño de la muestra crece indefinidamente

$$F_N(x) \rightarrow F(x) \tag{2.32}$$

donde $\rightarrow$ denota convergencia en probabilidad. Este resultado es muy importante, por lo que merece una explicación adicional. La función de distribución acumulada $F(x)$ contiene toda la información necesaria para caracterizar a la variable aleatoria $X$: conociendo $F(x)$ es posible realizar todo tipo de cálculo probabilístico acerca de $X$. En la práctica $F(x)$ no es conocida, pero sí lo es $F_N(x)$, ya que esta última se obtiene directamente de los datos de la muestra disponible. Este resultado, entonces, nos garantiza que para muestras grandes no hay mayor problema en reemplazar $F(x)$ (desconocida) por $F_N(x)$ (conocida), ya que en dicho caso ambas son prácticamente indistinguibles.[16] La distribución de ingresos muestral observable $F_N(x)$ constituye una estimación de la distribución poblacional $F(x)$ inobservable. Desde esta perspectiva,

[16] En la jerga estadística suele decirse que $F_N(x)$ es asintóticamente igual a $F(X)$.

nuestros dibujos de la función de distribución acumulada de la sección 2.3 son estimaciones de la "verdadera" función de distribución acumulada, que sólo podríamos dibujar si tuviésemos acceso a la información poblacional.[17]

Del mismo modo, las medidas resumen obtenidas de encuestas de hogares, es decir de muestras de una población, son en realidad estimaciones de conceptos poblacionales. De esta forma, la media o esperanza matemática de una variable aleatoria $X$, denotada con $E(X)$, puede ser estimada en base a una muestra aleatoria *iid*, a través de la media muestral, que denotamos $\bar{x}$. A su vez, la varianza de una variable aleatoria, definida como $V(X) = E(X - E(X))^2$ puede ser estimada mediante la varianza muestral

$$S^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1} \tag{2.33}$$

A menudo el interés recae en magnitudes simples como la proporción de ingresos debajo de un nivel determinado. Por ejemplo, la tasa de incidencia de la pobreza $H$ puede expresarse como la probabilidad de que una persona u hogar tenga ingresos por debajo de un umbral $z$ o línea de pobreza $H = Pr(X < z)$. Esta magnitud puede ser estimada en base a una muestra *iid* de ingresos simplemente como la proporción de individuos con ingresos inferiores a $z$.

**Censos, muestras, poblaciones y superpoblaciones**

En el uso coloquial de los términos "población" y "muestra" resulta cómodo pensar que el primero hace referencia a un conjunto de objetos y que el segundo es un subconjunto del mismo. En varios contextos esta caracterización provee una representación adecuada del fenómeno en cuestión. Sin embargo, para ciertos fines analíticos y prácticos hará falta una definición más certera y posiblemente sofisticada de las nociones de población y muestra.

Consideremos en primer lugar la definición más coloquial de estos conceptos, donde uno hace referencia a un subconjunto del otro. A modo de ejemplo, pensemos que es posible relevar los ingresos mensuales de todas las personas de Chile en un momento determinado, digamos, el 26 de marzo de 2007. Supongamos para facilitar el análisis que todas las personas relevadas ese día reportan su ingreso mensual (no hay no-respuestas) y que lo hacen correctamente (no hay errores de medición). Esta colección de ingresos será nuestra población de referencia. Si éste es nuestro interés (los ingresos mensuales de los individuos de Chile reportados a un censo relevado el 26 de marzo de 2007), todavía no hay ningún elemento aleatorio involucrado en el análisis; se trata de un simple evento administrativo o contable, que recoge ingresos en algún registro. Supongamos ahora que en lugar del censo se toma un subconjunto de observaciones de

[17] Si bien es clara la distinción conceptual entre $F(x)$ y $F_N(x)$, en la práctica, y por simplicidad, se suele llamar directamente como función de distribución a la versión empírica observable.

la población, o muestra. La aleatoriedad aparece en el análisis vinculada con la forma en la cual los elementos de la población fueron elegidos para integrar la muestra.

Alternativamente, consideremos la siguiente versión de los conceptos de población y muestra. Pensemos en un analista interesado en el ingreso de una familia cualquiera, de la cual todavía no dispone de ninguna información. Desde su punto de vista, el ingreso de esta familia puede ser representado a través de una variable aleatoria que denotaremos con $X$. Es decir, desde su punto de vista el ingreso que esta familia reporte será una realización de esta variable aleatoria $X$. La aleatoriedad en este caso es esencial a la naturaleza de los ingresos y debe ser entendida como una forma de modelar esta ignorancia *ex ante* por parte del analista. Supongamos ahora que el análisis se refiere a los ingresos de $N$ familias. En este caso la incertidumbre implícita en cada caso es la misma, es decir, los ingresos reportados por cada familia son vistos como $N$ realizaciones repetidas de la misma variable aleatoria. Más específicamente, los ingresos de las familias son vistos como $N$ variables aleatorias, $X_1, X_2, \ldots, X_N$, cada una con la misma distribución que la variable $X$ que representa el ingreso de cualquiera de ellas. En este contexto $X$ (la variable genérica) es la "población" y las variables $X_1, X_2, \ldots, X_N$, son la "muestra", entendidas como $N$ réplicas de la variable poblacional subyacente $X$.

Ciertamente, esta conceptualización provee una visión alternativa de los conceptos de población y muestra. La elección de cuál de ellas utilizar dependerá del objeto de interés. Si el interés recae en conocer detalles de la colección de objetos censales a partir de una subcolección de los mismos, claramente la primera de las visiones es la correcta. Pero en el análisis económico muchas veces el interés recae en el proceso causal del cual se desprenden los ingresos. Es relevante remarcar que se trata de objetivos distintos y ambas visiones no se contradicen. Por ejemplo, desde el punto de vista de la visión de "réplicas" discutida en segundo lugar, un censo en realidad debe interpretarse a su vez como una muestra de una superpoblación subyacente. Pensemos en la siguiente ficción, desde la perspectiva de uno de los hogares encuestados el 26 de marzo de 2007 en el censo (hipotético) de Chile. Supongamos que este hogar opera en un mercado informal y sus ingresos dependen de una enorme conjunción de factores, varios de ellos de naturaleza marcadamente fortuita. Entonces, lo que este hogar declare al censo es en realidad una magnitud sujeta a fuertes factores idiosincráticos, y es de esperar que los mismos hagan que la respuesta al censo varíe radicalmente si la misma pregunta es efectuada un mes antes o después. Una vez recolectados todos los ingresos del censo, la pregunta clave es si a través de una subcolección de ingresos el objeto de estudio es (i) la colección de ingresos en el censo, o (ii) el mecanismo subyacente del cual se desprenden cada uno de los ingresos.

Además de algunas consecuencias conceptuales y prácticas, esta distinción tiene consecuencias analíticas. Frecuentemente los métodos estadísticos son evaluados en un contexto de "muestra grande", es decir sus propiedades son estudiadas en el límite de un proceso que hace crecer el tamaño de la muestra indefinidamente. En la primera de las conceptualizaciones el límite superior de este proceso está dado por el tamaño de la

población, mientras que en la segunda esta restricción no opera, ya que la cantidad de potenciales réplicas de la noción poblacional se refiere no a individuos o períodos, sino a las distintas situaciones hipotéticas que podrían aparecer si el proceso generador de ingresos representado por la variable poblacional es replicado *ad-infinitum*.

## 2.6. Significatividad estadística

Supongamos que todos los habitantes de dos ciudades han sido censados, y que en la ciudad *A* el 8.3% de los habitantes declara ingresos por debajo de la línea de pobreza, mientras que en la ciudad *B* esa proporción asciende a 8.5%. Ciertamente es posible afirmar que al momento del censo la tasa de pobreza de ingresos en *B* es más alta que en A. Alternativamente, supongamos ahora que las cifras de pobreza de *A* y *B* son nuevamente 8.3% y 8.5%, pero que ambos valores provienen de muestras, en base a menos observaciones que el total de la población de cada ciudad. En este caso no es posible afirmar con certeza que la tasa de pobreza en *B* es mayor que en *A*, ya que la evidencia no está basada en la totalidad de la población.

### 2.6.1. La significatividad estadística de las estimaciones

El problema de la significatividad estadística es claramente un problema inferencial, propio del análisis estadístico, vinculado con la forma en la que la muestra se relaciona con la población. En el contexto inferencial números como 8.3% y 8.5% son estimaciones de las verdaderas (y no observables) tasas de pobreza poblacionales, que surgen de aplicar fórmulas (estimadores) sobre las observaciones de la muestra. Comencemos analizando la tasa de pobreza estimada para la ciudad *B* (8.5%). La misma se obtiene calculando la proporción de personas encuestadas que declaran ingresos por debajo de una línea de pobreza previamente establecida. Aun cuando resulte un tanto artificial, es conveniente pensar este número como proveniente del siguiente proceso. Existe un mecanismo aleatorio que primero "decide" qué personas de la población responden la encuesta y luego calcula la tasa de pobreza sólo para las personas a las cuales se ha encuestado. Desde este punto de vista, la regla "calcular la proporción de pobres para aquellos encuestados" es en realidad una variable aleatoria ya que el valor que efectivamente vaya a tomar depende de quiénes sean elegidos para integrar la muestra (un fenómeno claramente aleatorio). El valor 8.5% es entonces una realización de esta variable aleatoria, es decir una de las cifras que podrían haber resultado de las distintas muestras posibles.

El fenómeno de la variabilidad muestral está vinculado con la dispersión de valores que puede tomar la regla (que para ajustarnos a la terminología estadística, llamaremos *estimador*) en base a las distintas posibles muestras. Consideremos dos ejemplos extremos. En un caso supongamos que las muestras siempre tienen el mismo tamaño que la población. Ciertamente en este caso trivial la variabilidad muestral es nula: todas las muestras coinciden con la población, ergo, para cada "alteración" de la muestra obtendremos siempre la misma tasa de pobreza. En el otro extremo supongamos que la

muestra siempre tiene una sola persona elegida al azar de la población. En este caso la variabilidad muestral puede ser potencialmente muy alta ya que la tasa de pobreza "muestral" cambiará de 0 a 1 dependiendo de si la persona encuestada es pobre o no.

En síntesis, dado que los estimadores (entendidos como reglas de cálculo en base a datos muestrales) son variables aleatorias, es relevante dotar a las estimaciones de alguna medida de cuán grande es la variabilidad muestral. Una forma de computar esta medida es considerar la varianza del estimador (o su desvío estándar), que mide cuán dispares pueden ser las estimaciones en base a las potenciales muestras alternativas que pudiesen haber ocurrido. A modo de ejemplo, la media muestral $\bar{x}$ tiene varianza muestral estimada

$$\hat{V}(\bar{x}) = \frac{S^2}{N} \tag{2.34}$$

donde $S^2$ es la varianza de los datos y $N$ el tamaño de la muestra. Por su parte, la varianza de la tasa de pobreza muestral $\hat{H}$ (*i.e.* la proporción de personas con ingreso $x_i$ inferior a un umbral $z$) puede estimarse como

$$\hat{V}(\hat{H}) = \frac{\hat{H}(1-\hat{H})}{N} \tag{2.35}$$

$\hat{H}$ es en realidad una estimación de $H$, la probabilidad de que una variable binaria (0 ó 1) tome valor 1, que por ende tiene distribución *Bernoulli* con esperanza $H$ y varianza $H(1 - H)$.[18]

Volvamos sobre nuestro ejemplo de las ciudades $A$ y $B$, con tasas de pobreza de 8.3% y 8.5% respectivamente, en base a información muestral. Intuitivamente se trata de distinguir cuánto de la diferencia entre 8.3% y 8.5% se debe a diferencias entre las verdaderas (pero no observables) tasas de pobreza poblacionales y cuánto simplemente a variabilidad muestral. Podría suceder, por ejemplo, que las tasas de pobreza poblacionales de $A$ y $B$ sean idénticas y que las diferencias observadas se deban pura y exclusivamente a diferencias en las muestras tomadas. Una forma de aproximar este problema es verificando si los intervalos de confianza de estas dos estimaciones se solapan.[19] Si no lo hacen, podemos estar confiados en que la diferencia en las tasas de pobreza entre $A$ y $B$ es estadísticamente significativa.[20]

Siendo $\hat{H}$ asintóticamente normal, puede construirse para cada estimación de la pobreza (una en $A$ y la otra en $B$) un intervalo de confianza asintótico al, por ejemplo, 95% de la siguiente forma

---

[18] Es intuitivo pensar que la máxima variabilidad de la tasa de pobreza se corresponde cuando la mitad de las personas es pobre y la otra mitad es no pobre. Esto es fácil de chequear maximizando la varianza de una variable *Bernoulli* $p(1 - p)$ con respecto a $p$, lo cual arroja $p = 1/2$.

[19] Un "intervalo de confianza al 95%" es un intervalo tal que la probabilidad de que éste contenga al verdadero parámetro de interés es 95%.

[20] El hecho de que los intervalos no se solapen no es condición necesaria para la significatividad estadística. Puede existir cierto solapamiento y un test de hipótesis formal indicar que la diferencia en las estimaciones de pobreza es significativamente diferente de cero.

$$\hat{H} \pm c_{0.025} \sqrt{V\left(\hat{H}\right)} \tag{2.36}$$

donde $c_{0.025}$ es el percentil 0.975 de la distribución normal estándar.

Una forma dual de aproximar este problema es a través de un test de hipótesis. La hipótesis nula es que las tasas de pobreza poblacionales de *A* y *B* son idénticas y la hipótesis alternativa es que son distintas.

En algunos casos es relativamente sencillo calcular analíticamente la varianza o error estándar de un estadístico, a partir del cual realizar el análisis de significatividad. Desafortunadamente, esta tarea es muy engorrosa en casos donde el estadístico es una función compleja de las observaciones de la muestra. Como veremos a lo largo del libro, éste es de hecho el caso de muchos indicadores distributivos.

### 2.6.2. Bootstrap al rescate

Una estrategia alternativa es recurrir al principio de remuestreo o *bootstrap*.[21] Consideremos los siguientes pasos para producir una estimación de la varianza de la media muestral.

(i) Usar los *N* datos de la muestra original y tomar una muestra de tamaño *N*, con reemplazo. Nótese que es clave hacerlo con reemplazo, porque de lo contrario trivialmente siempre obtendríamos exactamente la muestra original. Al hacerlo con reemplazo estas pseudo-muestras pueden incluir una misma observación más de una vez.

(ii) Computar la media de esta pseudo-muestra.

(iii) Repetir el procedimiento anterior *B* veces (*B* es un número preferentemente grande).

(iv) Computar la varianza de las *B* medias computadas anteriormente. Esta es la estimación deseada.

Este método produce una estimación de la variabilidad de la media muestral a través de un esquema de remuestreo artificial conocido como *bootstrap*. Intuitivamente, hemos tomado a los datos de la muestra original como si fuesen ellos mismos la población y hemos remuestreado repetidas veces como si conociésemos esta población, a fines de producir *B* estimaciones alternativas de la misma media subyacente, y hemos aproximado la varianza de la media muestral a través de la varianza de estas medias *bootstrap* computadas en cada paso.

---

[21] La literatura sobre métodos de *bootstrap* aplicados a cuestiones distributivas es activa y creciente. En términos generales, el texto clásico de Efron y Tibshirani (1993) es una referencia muy accesible. Davison y Hinkley (1997) proveen un tratamiento más completo y avanzado. En cuanto a aplicaciones a problemas distributivos, Mills y Zandvakili (1997) y Sosa Escudero y Gasparini (2000) contienen aplicaciones al problema de la significatividad estadística de las medidas de desigualdad, éstos últimos para el caso argentino. Davidson y Flachaire (2007) presentan un tratamiento más definitivo y actual sobre los problemas de *bootstrap* aplicados a cuestiones de desigualdad y pobreza.

Si bien la intuición puede resultar convincente, la teoría que justifica el *bootstrap* es sorprendentemente más compleja. Nos limitaremos a señalar que cuando usamos a la muestra como si ella fuese la población, lo que hemos hecho es tomar una muestra de la distribución empírica (computable a través de los datos observados) en vez de hacerlo de la verdadera distribución "teórica" (no observable). El procedimiento detallado arriba será tan errado como grandes sean las diferencias entre la distribución empírica y la "teórica". Es justamente el Teorema Fundamental de la Estadística el que garantiza que estas diferencias son menores para tamaños de muestras lo suficientemente grandes.

En términos generales, si se busca computar la varianza para un estadístico genérico $\theta = g(.)$, análogamente los pasos a seguir son los siguientes: (i) usar los $N$ datos de la muestra original y tomar una muestra de tamaño $N$, con reemplazo, (ii) computar $g(.)$ para esta pseudo-muestra, (iii) repetir el procedimiento anterior $B$ veces (con $B$ grande) y (iv) calcular la varianza de las $B$ versiones de $g(.)$ computadas anteriormente. Esta es la estimación deseada. Por ejemplo, $g(.)$ podría ser la mediana de los datos. Calcular teóricamente la varianza de la mediana es una tarea sorprendentemente complicada. La alternativa de hacerlo por *bootstrap* implica repetir los pasos anteriores, computando en el paso (ii) la mediana de cada pseudo-muestra y finalmente la varianza muestral de las $B$ medianas obtenidas.

La estrategia de *bootstrap* puede ser extendida para calcular otros objetos estadísticos. Por ejemplo, podríamos usar el procedimiento anterior para construir un intervalo de confianza de nivel de significatividad $\alpha$. En el caso de la mediana el procedimiento comienza computando los primeros tres pasos y el último paso consiste en construir un intervalo tomando los cuantiles $\alpha/2$ y $(1 - \alpha/2)$ de la distribución empírica de las medianas obtenidas en los pasos anteriores. Es decir, una vez que obtenemos $B$ pseudo-estimaciones de la mediana, el intervalo de confianza es un intervalo que contiene a las $1 - \alpha$ observaciones centrales.

A modo de ejemplo, comparemos el desempeño del *bootstrap* con el de las aproximaciones asintóticas discutidas anteriormente con datos de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) de Ecuador para la región Amazónica, correspondiente a diciembre de 2009. La muestra incluye información del ingreso per cápita familiar de 3393 personas. Con una línea de pobreza de $46.32, la proporción de personas pobres es de 0.3457 (cerca del 35% de la población es pobre). En base a la fórmula (2.35) el error estándar (la raíz cuadrada de la varianza) para esta línea de pobreza es de 0.00816. Un intervalo de confianza al 95% está dado por (0.3297, 0.3617), utilizando la fórmula asintóticamente válida (2.36). El error estándar usando *bootstrap* con 500 replicaciones es 0.00807, ciertamente muy similar al obtenido con la aproximación asintótica.[22]

[22] Este error estándar será diferente cada vez que repliquemos el ejercicio, dado que las pseudo-muestras que le dan origen son elegidas aleatoriamente. Con un número grande de réplicas la diferencia debería ser mínima.

Un resultado de la implementación del *bootstrap* es la distribución empírica de la tasa de pobreza, es decir, 500 pseudo-estimaciones de la tasa de pobreza en base a 500 pseudo-muestras de la muestra original. Una forma simple de construir un intervalo de confianza al 95% es tomar los percentiles 0.025 y 0.975 de estas estimaciones *bootstrap*, es decir los valores que dejan al 95% central de las observaciones. En nuestro caso el intervalo obtenido es (0.3295, 0.3619), bastante similar al obtenido con la fórmula asintótica.[23] Este intervalo puede ser utilizado para evaluar algunas hipótesis. Por ejemplo, la hipótesis nula de que la tasa de pobreza es 0.35 no es rechazada, ya que este valor cae dentro del intervalo de confianza antes construido.

### 2.6.3. Igualdad de distribuciones

En algunas situaciones puede resultar relevante plantear la hipótesis nula de que dos distribuciones son iguales versus la alternativa de que no lo son. Este es un problema clásico en estadística, y la naturaleza de la solución depende de cuánto se conozca de antemano el problema en cuestión. En un extremo, si ambas distribuciones fuesen normales y con idéntica varianza, un test de diferencia de medias es suficiente para el problema. Pero, como señalamos anteriormente, las cuestiones distributivas operan en un contexto de tal incertidumbre que puede ser costoso hacer supuestos funcionales, por lo que es deseable disponer de algún método que permita evaluar la hipótesis de interés sin recurrir a supuestos funcionales restrictivos.

Supongamos que $F(x)$ y $G(x)$ son las funciones de distribución acumuladas para dos variables aleatorias y estamos interesados en la hipótesis nula $H_0$: $F(x) = G(x)$ para todo $x$, es decir, ambas funciones coinciden en todo el soporte. Un estadístico útil para esta hipótesis es el de *Kolmogorov-Smirnov*:

$$J = \frac{mn}{d} \max_x \left[ \left| F_m(x) - G_n(x) \right| \right] \tag{2.37}$$

donde $m$ y $n$ son los tamaños de muestra para las poblaciones cuyas distribuciones son, respectivamente, $F$ y $G$, $d$ es el mayor divisor común entre $m$ y $n$, y $F_m(.)$ y $G_n(.)$ son las funciones de distribución empíricas discutidas anteriormente. Intuitivamente, el estadístico se basa en la máxima discrepancia posible entre ambas distribuciones y, naturalmente, la regla consiste en rechazar la hipótesis nula si $J$ es demasiado grande. Existen tablas apropiadas para este estadístico y también aproximaciones asintóticas a los valores críticos de su distribución. La mayoría de los paquetes estadísticos - incluyendo Stata - proveen este test y sus correspondientes "valores $p$".[24]

[23] Se aplica acá la misma aclaración que en la nota de pié de página anterior.

[24] Ver Hollander y Wolfe (1999) para mayores detalles.

## 2.7. Formas funcionales

En la sección 2.3 mencionamos que todas las distribuciones del ingreso del mundo real tienen algunas características comunes; en particular, son asimétricas, con una cola superior desproporcionadamente larga. El famoso economista italiano Vilfredo Pareto (1848-1923) fue uno de los primeros en notar y estudiar estas similitudes. De hecho, Pareto (1895, 1897) sostuvo que todas las distribuciones del ingreso reales podían ser adecuadamente aproximadas mediante la función

$$F(x) = 1 - Kx^{-\alpha} \tag{2.38}$$

donde $K$ y $\alpha$ son dos parámetros positivos.[25] La ecuación (2.38) es una forma funcional *paramétrica*: la forma de la función está enteramente determinada por un número de parámetros, en este caso sólo dos. El trabajo pionero de Pareto despertó la curiosidad de los investigadores: ¿responden las distribuciones del mundo real a formas funcionales paramétricas? ¿Es la función propuesta por Pareto la mejor representación de las distribuciones reales?

En principio, es claro que ninguna distribución real responde exactamente a una forma funcional dada. El proceso por el cual se generan los ingresos de una población es tan complejo y con tantas aleatoriedades que es imposible representarlo perfectamente mediante alguna forma funcional paramétrica manejable. Por esta razón, el objetivo empírico no reside en encontrar una forma funcional que reproduzca exactamente los datos, sino una que los aproxime razonablemente bien; es decir, que constituya un "modelo razonable" de la realidad. [26]

### 2.7.1. Funciones paramétricas

El modelo más habitualmente utilizado para representar a la distribución del ingreso es el *log-normal* (Gibrat, 1931). Una variable aleatoria $x$ se distribuye en forma log-normal si $\ln(x)$ tiene distribución normal. La función de densidad para una variable aleatoria log-normal, definida en el soporte $[0, \infty)$, está dada por

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} \tag{2.39}$$

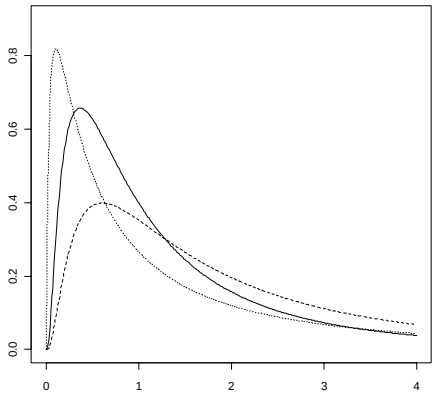
Nótese que esta función depende solamente de dos parámetros $\mu$ y $\sigma$. Estos parámetros se relacionan con el centro y la dispersión, respectivamente, del logaritmo de los ingresos. De hecho, si $x$ tiene distribución log-normal, entonces $\mu = E(\ln x)$ y $\sigma^2 = V(\ln x)$, donde E(.) y V(.) denotan esperanza y varianza, respectivamente. La figura 2.27

[25] Pareto fue más allá y sostuvo que había evidencia sobre la estabilidad del parámetro $\alpha$ - que aproxima el grado de desigualdad en la distribución - en el tiempo y en el espacio; lo que llevaba a pensar que la magnitud de las desigualdades en una sociedad era consecuencia de la naturaleza humana más que de la forma como se organizaba esa sociedad. Esta idea, naturalmente, generó un arduo debate con quienes subrayaban la relevancia de los sistemas económicos en moldear la distribución del ingreso y la riqueza.

[26] Quizás en los términos del notable estadístico George Box "todos los modelos están mal, pero algunos son útiles".

muestra la función de densidad de tres variables log-normales, para combinaciones alternativas de $\mu$ y $\sigma$.

**Figura 2.27**
**Función de densidad de variables log-normales**

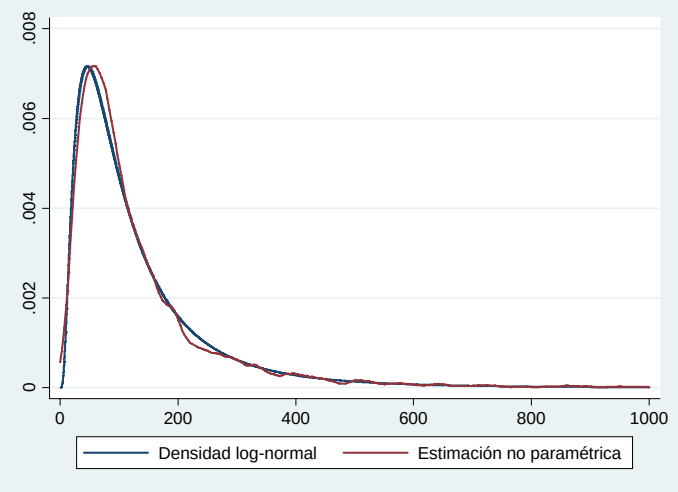



Nota: La línea sólida representa la densidad de una variable log-normal con $\mu = 0$ y $\sigma = 1$.
La función que comienza por debajo de la sólida corresponde a $\mu = 0.5$ y $\sigma = 1$
y la que comienza por encima corresponde a $\mu = 1$ y $\sigma = 0.5$.

La línea sólida representa la densidad de una variable log-normal con $\mu = 0$ y $\sigma = 1$. Recordar que en este caso 0 es la esperanza del logaritmo de la variable de interés y 1 será el desvío estándar del logaritmo. La misma figura presenta dos líneas punteadas. La función que se encuentra en principio debajo de la función de densidad (antes de 1.5, aproximadamente) corresponde a $\mu = 0.5$ y $\sigma = 1$, y la que comienza por encima de la original (más pegada al eje vertical) se corresponde a $\mu = 1$ y $\sigma = 0.5$. Es muy importante notar que la distribución log-normal es asimétrica y que estos parámetros no controlan la media y la varianza de la distribución sino su transformación logarítmica.

Esta característica asimétrica de la distribución log-normal parece proveer una representación adecuada para la distribución de los ingresos de una sociedad. La figura 2.28 muestra la función de densidad de una distribución log-normal con parámetros $\mu$ y $\sigma$ iguales a la media y desvío del logaritmo de los ingresos efectivamente observados en la encuesta de hogares de El Salvador, 2008.

**Figura 2.28**
**Función de densidad de una distribución log-normal**
**Ingreso per cápita familiar**
**El Salvador, 2008**




Fuente: elaboración propia sobre la base de datos de la EHPM

A efectos comparativos la figura también incluye una estimación no paramétrica (por el método de *kernels* discutido arriba) de la distribución del ingreso. Aunque menos rica que la estimación no paramétrica, visualmente la función log-normal parece ser una aproximación razonable de la distribución. Naturalmente, esta apreciación visual debe ser corroborada con rigurosidad analítica. Desafortunadamente, no existe una forma conclusiva de evaluar log-normalidad; una estrategia simple está basada en un test de normalidad para el logaritmo del ingreso. Por ejemplo, el test de Jarque y Bera (1982), de uso frecuente en econometría, puede ser utilizado para evaluar la hipótesis nula de normalidad del logaritmo del ingreso.[27]

El modelo log-normal es el más popular, dada su simplicidad analítica, pero no es la única opción disponible. Como mencionamos, una alternativa utilizada es la distribución de Pareto descripta en (2.38). Una forma funcional alternativa es la de Singh-Maddala

$$F(x) = 1 - [x + \delta x^{\beta}]^{-\alpha} \tag{2.40}$$

donde $\alpha$, $\beta$ y $\delta$ son parámetros que garantizan que la función de distribución parta de 0 y termine en 1 y que su función de densidad sea positiva. Una ventaja de esta función es que incluye a varios casos conocidos. Por ejemplo, las distribuciones de Pareto, Weibull o la exponencial se obtienen para configuraciones específicas de los parámetros.[28]

Una función también comúnmente utilizada es la propuesta por el investigador argentino Camilo Dagum, formalmente expresada como

[27] López y Servén (2006), en un estudio sobre cerca de 800 encuestas de hogares en el mundo, concluyen que la hipótesis nula de que el ingreso per cápita sigue una distribución log-normal no puede ser rechazada.

[28] Ver Cowell (1995) para más detalles.

$$F(x) = \left[1 - \left(\frac{x}{\beta}\right)^{-\alpha}\right]^{-\gamma} \tag{2.41}$$

donde $\alpha$, $\beta$ y $\gamma$ son parámetros positivos.

La literatura sobre funciones paramétricas es fecunda, técnicamente elegante y académicamente prestigiosa.[29] Son numerosos los trabajos donde se proponen funciones más complejas que las mencionadas, o formas generales abarcativas de muchas funciones,[30] o que pueden aproximar distribuciones con peculiaridades, como truncamientos o multimodalidades.[31] Sin embargo, a nuestro juicio la utilidad de estas aproximaciones es acotada a algunos usos particulares, por lo que preferimos no extendernos en su desarrollo.

### 2.7.2. El uso de las formas funcionales

Las parametrizaciones de las distribuciones del ingreso tienen una utilidad limitada. Gran parte del análisis distributivo empírico está basado en los microdatos reales, sin necesidad de conocer la forma funcional que mejor los aproxima. Como veremos a lo largo del libro, la pobreza o la desigualdad se calculan sin ninguna necesidad de saber si los datos subyacentes responden a alguna forma funcional determinada. Inclusive en la etapa exploratoria que hemos abordado en este capítulo, los métodos paramétricos tienen limitaciones frente a un examen no paramétrico más flexible. Al estar basados en formas funcionales preestablecidas, por construcción no pueden ser completamente informativos acerca de la forma de la distribución.

Existen al menos dos áreas en las que el uso de las formas funcionales adquiere relevancia: la modelización teórica y la estimación con disponibilidad de pocos datos agregados.

*Modelos*

Los modelos teóricos son estilizaciones de la realidad destinadas a ilustrar algún fenómeno. Existen modelos económicos que predicen reglas de generación de los ingresos de las personas, y por ende distribuciones. Aunque lo más usual es que estas predicciones no involucren formas funcionales específicas, en ocasiones lo hacen. Supóngase un modelo donde el logaritmo del ingreso se comporta como un *random walk*, es decir el valor hoy es igual al de ayer más un término aleatorio *iid*. En este caso,

[29] Pareto (1897), Gibrat (1931), Kalecki (1945), Rutherford (1955) y Singh y Maddala (1976) son algunos de los antecedentes ilustres.

[30] Por ejemplo, la función beta generalizada de cinco parámetros abarca como casos particulares a las funciones de Pareto, log-normal, gamma, Weibull, Fisk y Singh-Maddala.

[31] Ver, por ejemplo, Jenkins(2007) y Pinkovskiy (2008); y Botargues y Petrecolla (1999) para un país de América Latina.

es sabido que con el paso del tiempo la distribución del *random walk* (apropiadamente normalizado) se vuelve normal. En nuestro caso, eso implica que el logaritmo del ingreso se distribuye normalmente, y por ende el ingreso tiene una distribución log-normal. El reciente artículo de Battistin, Blundell y Lewbel (2009) argumenta que un modelo log-normal puede proveer una representación adecuada de la distribución del ingreso y el consumo. El argumento parte de la llamada "Ley de Gibrat", que postula que el ingreso es una acumulación de shocks multiplicativos, de modo que apelando al Teorema Central del Límite, el mismo es asintóticamente normal.[32] Las distribuciones Pareto también pueden surgir de modelos simples alternativos de generación de ingresos.

*Información limitada*

Uno de los principales usos empíricos de las formas funcionales para las distribuciones del ingreso consiste en estimar parámetros en situaciones en que contamos sólo con algunos pocos datos agregados. En esta situación asumir una determinada forma funcional puede ayudar a llenar el vacío de datos. Por ejemplo, las formas funcionales paramétricas son usadas frecuentemente para estimaciones de la distribución del ingreso mundial o de alguna de sus características, como la tasa de pobreza global. Si bien lo ideal para este caso es agregar los microdatos de las encuestas de todos los países, este procedimiento resulta engorroso o impracticable por falta de información. En su lugar, varios investigadores han asumido que las distribuciones del ingreso nacionales siguen una forma funcional paramétrica simple y aproximan los parámetros requeridos con datos agregados de fuentes secundarias. El procedimiento típico es asumir distribuciones nacionales lognormales, donde la media es aproximada con el ingreso o PIB per cápita a PPA, y el desvío es estimado a partir de datos del coeficiente de Gini o estimado por mínimos cuadrados de información de participaciones de centiles (*e.g.* Pinkovskiy y Sala-i-Martin, 2009).

[32] Battistin *et al*. (2009) argumentan que en un contexto dinámico de optimización intertemporal del bienestar, las ecuaciones de Euler que caracterizan a las condiciones de primer orden de dicho proceso, implican que el ingreso permanente y el consumo deberían obedecer una Ley de Gibrat. Este hecho explica también porque el modelo log-normal ajusta mejor al consumo que al ingreso corriente: este último está "contaminado" por discrepancias transitorias. Adicionalmente, estos autores sugieren que las discrepancias con respecto al ideal log-normal pueden deberse a las propias inexactitudes de un modelo simple de optimización intertemporal, tales como la presencia de restricciones de liquidez, horizontes finitos o errores de medición.

# Apéndice: en la práctica

Los apéndices con aplicaciones prácticas asumen cierta familiaridad con el software estadístico-econométrico Stata. El apéndice I del libro introduce un conjunto básico de comandos de Stata que el lector debería conocer para seguir con relativa facilidad los apéndices "en la práctica" del libro.[33] Además, en el sitio web que acompaña a este libro, ponemos a disposición del lector un conjunto de encuestas de hogares procesadas que contienen todas las variables que se requieren para implementar los códigos de Stata. Ciertamente, el procesamiento de una encuesta implica un sinnúmero de decisiones para las que no siempre existe consenso. En consecuencia, los resultados que surgen de los ejemplos que se implementan utilizando las bases de datos disponibles en el sitio web del libro pueden no coincidir con las estadísticas oficiales, o las que derive el lector empleando criterios alternativos de procesamiento.

### Ejemplo Brasil

En este apartado se muestra cómo replicar los resultados que fueron presentados en el cuadro 2.1 del texto. El primer paso que debe seguir el lector es obtener la versión procesada de la PNAD (Pesquisa Nacional por Amostra de Domicílios) de Brasil para el año 2007. Es decir, conteniendo las variables que se emplean a continuación. Para ello, puede dirigirse a la sección Encuestas de Hogares del sitio web del libro.

El código siguiente asume que el archivo con extensión `.dta` fue descargado en el directorio `C:\libro-distribucion\cap2`. Como se explica en el apéndice I del libro, las sentencias de Stata a continuación pueden introducirse de a una por vez en la línea de comando de Stata o, alternativamente, todas juntas en un archivo **do** que luego Stata ejecuta completo, línea por línea desde arriba hacia abajo.[34] En términos generales, esta segunda alternativa es más recomendable porque nos permite reutilizar el código con mucha facilidad. En Stata, las líneas que inician con asterisco (`*`) son comentarios; es decir, se trata de líneas que - en general - documentan el código pero que Stata ignora. Por último, antes de comenzar con nuestro primer ejemplo, cabe aclarar que los números de línea que se muestran no forman parte del código que debe introducirse en Stata; aquí se los emplea para facilitar la explicación.

```
1 * cap2-ejemplo.do
2
3 clear all
```

[33] Las aplicaciones han sido desarrolladas empleando la versión 11.2 del Stata, pero en su gran mayoría también funcionarán con versiones anteriores del software.

[34] Los archivos **do** son archivos de texto plano con extensión `.do`. En general, una forma útil de trabajar con Stata es utilizando la línea de comando para chequear lo que queremos hacer, copiando luego las sentencias que sirvieron en el archivo **do**. Alternativamente, puede emplearse el editor de archivos do de Stata para ejecutar partes de un archivo **do**. En el sitio web del libro se sugieren editores de texto alternativos más poderosos para emplear con Stata (ver también el apéndice I).

```
 4 set mem 250m
 5 cd "C:\libro-distribucion\cap2"
 6
 7 * cargar encuesta Brasil 2007
 8 use "bra07.dta"
 9
10
11 * total
12 summ ipcf [w=pondera], detail
13
14 * región norte
15 summ ipcf [w=pondera] if region==1, detail
16
17 * región nordeste
18 summ ipcf [w=pondera] if region==2, detail
```

El comando `clear all` (línea 3) elimina, si existe, la base de datos actualmente cargada.[35] En la cuarta línea se asignan 250 MB de memoria RAM para almacenar la base de datos - el lector puede comprobar que el comando `memory` muestra como está asignada la memoria.[36] El comando `cd` se utiliza para determinar cuál es el directorio en el que se guardan los archivos que se están utilizando (ver línea 5); así, cualquier comando de Stata que trabaje con archivos lo hará en esa carpeta, a menos que se especifique lo contrario. La línea 8 carga en la memoria el contenido del archivo `bra07.dta` utilizando el comando `use`.

Las encuestas de hogares, al igual que cualquier otra base de datos, se organizan en Stata como una tabla donde las filas representan observaciones o registros y las columnas variables o campos. A su vez, por tratarse de una encuesta, cada observación representa a varios individuos, tantos como indica el factor de expansión o variable de ponderación. En nuestro caso, todas las encuestas que utilizaremos contienen una variable de nombre `pondera` que almacena el factor de expansión. Para más detalles sobre el uso de ponderadores, consultar la sección 3.6 del capítulo 3.

Las encuestas de hogares procesadas que se utilizan a lo largo del libro sólo contienen observaciones que denominamos coherentes (ver capítulo 3).[37] Por último, el comando `summarize` con la opción `detail` de la línea 12 muestra estadísticos básicos (ponderados) para el ingreso per cápita familiar (ver variable `ipcf`); en particular, nos muestra la media, el desvío estándar, algunos percentiles y el número de observaciones.[38]

En el caso de la PNAD 2007 de Brasil, la variable región puede tomar los valores 1, 2, 3, 4 o 5 dependiendo de si la observación corresponde a la región Norte, Nordeste,

---

[35] Además, elimina todos los elementos de Stata definidos por el usuario (por ejemplo, matrices).

[36] El comando `set mem` es innecesario a partir de la versión 12 de Stata. Para que Stata funcione a una velocidad razonable, es necesario que la base de datos que estamos utilizando pueda almacenarse completamente en la memoria RAM. De lo contrario, el rendimiento disminuye de manera considerable porque se utiliza el disco rígido como memoria RAM.

[37] En pocas palabras, se trata de observaciones válidas que utilizamos en el cálculo de los ingresos familiares.

[38] En general, el nombre de los comandos de Stata puede abreviarse. En el caso de `summarize`, puede emplearse `su`, `sum`, `summ`, etc. En la ayuda de Stata se muestra cuál es la abreviación mínima que puede emplearse para cada uno de los comandos utilizados en este libro.

Sudeste, Sur o Centro-Oeste, respectivamente.[39] Así, las líneas 15 y 18 pueden utilizarse para computar las columnas "Norte" y "Nordeste" del cuadro 2.1.

Para computar el coeficiente de variación de la distribución del ingreso per cápita familiar es necesario conocer la media y el desvío estándar de la variable `ipcf`. En el ejemplo, una forma de hacerlo para el total nacional es escribir en la línea de comando de Stata

```
. display 970.2443/574.3455
1.6893043
```

Sin embargo, esto resulta poco práctico si queremos utilizar el mismo código para procesar otra base de datos, o la misma base de datos pero con algunas observaciones eliminadas.

En general, luego de ejecutar un comando, Stata guarda varios de los resultados que presenta en pantalla. Para ver todos los resultados que Stata almacena luego de un comando como el `summarize`,[40] puede utilizarse el comando `return list` - cabe recalcar que el comando `return list` es sólo informativo; es decir, no es necesario introducirlo para que Stata almacene los resultados luego del comando `summarize`. En nuestro ejemplo, luego de ejecutar el comando `summarize`, Stata guarda los siguientes valores en `r(`*resultado*`)`, donde *resultado* es cada uno de los elementos que se muestran a continuación.

```
. summarize ipcf [w=pondera]
(analytic weights assumed)

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------------------
        ipcf |  394551   186985040    574.3455   970.2443          0      66000

. return list

scalars:
                  r(N) =  394551
              r(sum_w) =  186985040
               r(mean) =  574.3455226749792
                r(Var) =  941373.9479314596
                 r(sd) =  970.2442723002592
                r(min) =  0
                r(max) =  66000
                r(sum) =  107394020531.2019
```

donde `r(N)` es número de observaciones sin incluir observaciones con *missing*[41] en `ipcf` o `pondera`, `r(sum_w)` es la suma de la variable `pondera`, `r(mean)` es la

[39] Naturalmente, el contenido de la variable región difiere entre encuestas.

[40] En la terminología de Stata, el comando `summarize` es de tipo "`r`", por lo que sus resultados se almacenan en `r(`*resultado*`)`. Como veremos más adelante, los comandos de estimación econométrica son de tipo "e", por lo que resultados se almacenan en `e(`*estimación*`)`.

[41] El significado de las observaciones *missing* se explica en el apéndice I.

media del `ipcf`, `r(Var)` es la varianza del `ipcf`, `r(sd)` es el desvío estándar del `ipcf`, `r(min)` es el valor mínimo del `ipcf`, `r(max)` es el valor máximo `ipcf`, y `r(sum)` es la suma ponderada de la variable `ipcf`.

Los valores almacenados en `r(`*resultado*`)` son reemplazados cada vez que se ejecuta una nueva sentencia de Stata que también utiliza `r(`*resultado*`)`. Como consecuencia, en los próximos ejemplos veremos cómo se pueden conservar los valores `r(`*resultado*`)` de forma tal que puedan ser reutilizados. Por el momento, para calcular el coeficiente de variación alcanza con introducir, luego de cada comando `summarize`,

```
. display r(sd)/r(mean)
1.6893041
```

Por su parte, la población de referencia o número de observaciones expandidas como se dice en el cuadro 2.1 puede mostrarse con

```
. display r(sum_w)
1.870e+08
```

donde vemos que la población de referencia de la PNAD 2007 es 187 millones de persona aproximadamente. En resumen, el código para reproducir la primera columna del cuadro 2.1 del texto quedaría como se muestra a continuación.

```
 1 * cap2-ejemplo.do
 2
 3 clear
 4 set mem 250m
 5 cd "C:\libro-distribucion\cap2"
 6
 7 capture log close "cap-ejemplo.log"
 8 log using "cap2-ejemplo.log", replace
 9
10 * cargar encuesta de Brasil 2007
11 use "bra07.dta"
12
13 * total país
14 summ ipcf [w=pondera], detail
15 display "poblacion de referencia = " r(sum_w)
16 display "cv = " r(sd)/r(mean)
17
18 log close
```

Comentario: clear all?

A diferencia del código anterior, en este caso se genera un archivo **log** que contiene un "eco" de todo lo que Stata va mostrando en la ventana de resultados. La creación de dicho tipo de archivo se realiza con el comando `log using` (ver línea 8).[42] En nuestro caso, el nombre del archivo **log** que se está creando es `"cap2-ejemplo.log"`. La extensión `.log` hace que el archivo que se crea sea de texto plano, por lo que puede

[42] En todo el libro sólo empleamos un único archivo **log** a la vez. Sin embargo, Stata permite utilizar hasta cinco archivos **log** de forma simultánea.

examinarse con cualquier editor de texto. El comando `log close` cierra el último archivo log abierto (ver línea 18). En la línea 7 también se cierra, en caso de que este abierto, el archivo **log**. En caso de que no exista un archivo **log** abierto, el comando `capture` evita que se genere un error. La sentencia `capture` puede anteponerse a cualquier comando de Stata; lo que hace es capturar un eventual código de error evitando, en nuestro caso, que se interrumpa la ejecución del archivo **do** `cap2-ejemplo.do`. En general, nos interesa conocer si una sentencia genera un error, por lo que el comando `capture` debe emplearse con sumo cuidado.

El cómputo de la participación de cada decil de `ipcf` en el ingreso total se pospone para una sección posterior de este apéndice, donde se muestra cómo graficar curvas de Lorenz.

El ejemplo del texto finaliza con el cómputo de la pobreza en Brasil para el año 2007 utilizando una línea de pobreza de 130 reales mensuales. Una forma sencilla de computar la proporción de individuos con ingresos mensuales menores a 130 reales es mediante el bloque de código siguiente, que puede agregarse a continuación del anterior.

```
 1 * cap2-ejemplo-pobreza.do
 2
 3 * identificar individuos pobres
 4 gen pobre = 1 if ipcf < 129.883
 5 replace pobre = 0 if pobre!=1
 6
 7 * total país
 8 summ pobre [w=pondera]
 9 display "shr pobres = " r(sum)/r(sum_w)
10
11 * región Norte
12 summ pobre [w=pondera] if region==1
13 display "shr pobres = " r(sum)/r(sum_w)
```

Las líneas 4 y 5 generan la variable `pobre` que vale 1 para los individuos con `ipcf` menor a 130 y 0 para el resto. El comando `generate` (abreviado `gen`) puede emplearse para agregar variables a la base de datos (ver apéndice I). Por su parte, el comando `replace` permite reemplazar el contenido de una variable; se usa, en general, con alguna condición *if*.[43] La línea 9 muestra el cociente entre la suma ponderada de la variable `pobre` (es decir, el número de pobres expandido por el factor de ponderación) y la población de referencia (es decir, la suma total de la variable `pondera`). Así, las líneas 7 a 9 computan la proporción de individuos pobres en Brasil.

### Histograma

En primer lugar mostramos cómo puede graficarse un histograma de la distribución del `ipcf` en México para el año 2006 (ver figuras 2.2 a 2.8 del texto). Al igual que en el

[43] Típicamente, una condición *if* permite limitar el rango de observaciones que se utilizan en un determinado comando.

ejemplo de Brasil, el lector puede obtener la ENIGH (Encuesta Nacional de Ingresos y Gastos de los Hogares) mexicana de 2006 del sitio web del libro. La misma también cuenta con las variables `ipcf` y `pondera` que utilizaremos en lo que resta de este apéndice.[44] El código siguiente asume que la base de datos ya se encuentra cargada en Stata.[45]

El comando que se emplea para graficar un histograma es, justamente, `histogram` (ver línea 5). Igual que antes, `[w=pondera]` indica que cada observación de la encuesta debe expandirse según la cantidad de individuos que representa. La opción `bin(100)` del comando `histogram` especifica que el histograma debe identificar 100 grupos - 100 barras. Por último, la opción `frac` indica que el eje vertical debe mostrar la fracción del ingreso total que recibe cada uno de los 100 grupos. El comando `more` (línea 6) suspende la ejecución del archivo **do** `cap2-hist.do` hasta que el usuario oprima una tecla. Las líneas 8-11 grafican el mismo histograma pero sólo considerando a los individuos con `ipcf` menor a 15000.

```
 1 * cap2-hist.do
 2
 3 * figura 2.2
 4 * histograma ipcf
 5 hist ipcf [w=pondera], bin(100) frac
 6 more
 7
 8 * figura 2.3
 9 * histograma ipcf
10 hist ipcf [w=pondera] if ipcf < 15000, bin(100) frac
11 more
```

La figura 2.4 del texto puede replicarse utilizando el bloque de código siguiente. En la línea 14 se genera la variable `lipcf` como igual al logaritmo de la variable `ipcf`. Las interpretaciones de las demás líneas de código no deberían representar mayor dificultad para el lector.

**Comentario:** ¿Las figuras 2.4 y 2.5?

```
12 * figura 2.4
13 * histograma logaritmo ipcf
14 gen lipcf = log(ipcf)
15 hist lipcf, bin(100) frac
16 more
17
18 * figura 2.5
19 * histograma ipcf
20 hist lipcf [w=pondera], bin(10) frac
21 more
22 hist lipcf [w=pondera], bin(100) frac
23 more
24 hist lipcf [w=pondera], bin(500) frac
25 more
```

**Comentario:** ¿no pondera?

**Comentario:** Una pavada: En realidad la figura está hecha para 10, 50, 100 y 1000 intervalos.

**Comentario:** logaritmo ipcf

[44] Asimismo, también fueron eliminadas las observaciones que consideramos incoherentes.

[45] En el resto de los apéndices prácticos también se omiten las líneas de código que cargan la base de datos en la memoria de Stata.

El bloque de código siguiente agrega al histograma anterior una estimación no paramétrica por el método de kernels de la función de densidad del logaritmo del ingreso per cápita familiar (ver opción `kdensity` en línea 28). Las líneas 31-36 grafican, superpuestos, los "histogramas suavizados" de las funciones de densidad del `ipcf` para las regiones Noroeste y Sur de México (ver figura 2.7).

**Comentario:** logaritmo?

La línea 34 almacena en la macro local `lp` el logaritmo de la línea de pobreza mexicana de 2.5 dólares por día por persona. Como se explica en el apéndice I, una macro local puede emplearse para almacenar un número o una frase,[46] a diferencia de una variable que almacena una lista de valores. Cuando el nombre de una macro local se encierra entre comillas simples (la de apertura inclinada a la izquierda, `` ` ``, la de cierre vertical, `'`), Stata reemplaza el nombre de la macro local por su contenido. Así, la opción `xline(`lp')` del comando `twoway` agrega una línea vertical en el valor `log(608.245)` del eje horizontal.

Por último, la línea 40 utiliza la opción `normal` del comando `hist` para superponer al histograma de la variable `lipcf` una distribución normal con media y varianza iguales a las observadas.

```
26 * figura 2.6
27 * histograma ipcf
28 hist lipcf [w=pondera], bin(100) frac kdensity
29 more
30
31 * figura 2.7
32 * región 1 = Noroeste
33 * región 6 = Sur
34 local lp = log(608.245)
35 twoway (kdensity lipcf [w=pondera] if region==1) ///
36        (kdensity lipcf [w=pondera] if region==6), xline(`lp')
37 more
38
39 * figura 2.8
40 hist lipcf [w=pondera], bin(100) frac normal
```

**Comentario:** Logaritmo?

### Función de distribución

En este apartado se muestra cómo pueden graficarse las funciones de distribución presentadas en la sección 2.3.2 del cuerpo principal del capítulo. El primer paso para construir una función de distribución es ordenar - de menor a mayor - las observaciones de la encuesta según la variable de ingreso elegida, `ipcf` en nuestro caso (ver línea 4). En la línea 7 se crea la variable `shrpop` para almacenar la suma acumulada de la variable `pondera`. Así, la última observación de dicha variable (ver `shrpop[_N]`) contiene la población de referencia, computada como la suma de los factores de expansión individuales.[47] La línea 8 computa la proporción de la población que se

[46] En el segundo caso, se dice que el contenido de la macro local es una cadena de caracteres o *string*. En general, las cadenas de caracteres no pueden utilizarse en operaciones matemáticas. Además, su contenido suele definirse encerrado entre comillas dobles.

[47] La expresión `_N` hace referencia al número de observaciones en la base de datos. De forma más general, la expresión `shrpop[#]` hace referencia a la observación número # de la variable `shrpop`.

acumula hasta cada observación de la encuesta. En otras palabras, la variable `shrpop` se genera en dos pasos a partir de la variable `pondera`; empleando notación matemática,

(1) paso 1 $$shrpop_i = \sum_{j \le i} pondera_j$$

(2) paso 2 $$shrpop_i = \frac{shrpop_i}{\sum_j shrpop_j}$$

donde $i = j$ se refieren a cada uno de los individuos de la encuesta de hogares. En términos de nuestro código de Stata, el denominador de la expresión (2) se encuentra en `shrpop[_N]` luego de ejecutar la línea 7.

La función de distribución presenta las variables `shrpop` e `ipcf` en los ejes vertical y horizontal, respectivamente (ver línea (12)). Se deja como ejercicio para el lector elaborar las otras funciones de distribución presentadas en la sección 2.3.2. Por su parte, la curva de Pen (ver figuras 2.12 y 2.13) se construye igual que la función de distribución pero se grafica invirtiendo los ejes.

```
 1 * cap2-func-dist.do
 2
 3 * ordenar según ipcf
 4 sort ipcf
 5
 6 * población acumulada ordenamiento ipcf
 7 gen shrpop = sum(pondera)
 8 replace shrpop = shrpop/shrpop[_N]
 9
10 * figura 2.9
11 * función de distribución acumulada
12 line shrpop ipcf
```

La elaboración de un gráfico con dos o más funciones de distribución superpuestas se pospone hasta el capítulo 4 del libro.

### Diagrama de Pareto

En esta sección se muestra cómo replicar la figura 2.14 del texto, que muestra los diagramas de Pareto para las regiones Noroeste y Sur de México. En primer lugar, se ordenan las observaciones primero por región y luego en orden creciente según `ipcf` (ver línea 4). En la línea 7 se computa, para cada región, la suma acumulada de la variable `pondera`. El prefijo `by region:` ejecuta el comando a la derecha de los dos puntos para cada grupo en que puede dividirse la base de datos según la variable `region`, 8 regiones en nuestro caso.[48] De hecho, el prefijo `by region:` funciona

[48] La utilización del prefijo `by` *varlist*`:` requiere que la base de datos esté ordenada según la lista de variables *varlist*. Alternativamente, puede emplearse el prefijo `bysort` *varlist*`:`.

dividiendo la base de datos en tantas partes como valores distintos tome la variable `region`. Por lo tanto, la última observación perteneciente a cada región (es decir, la número `_N` de cada región) contiene la población de referencia regional. Luego, se utiliza la sentencia `replace` para reemplazar el contenido de la variable `shrpop` por el resultado de dividir cada una de sus observaciones por la población de referencia regional (ver línea 8).

La línea 10 genera la variable `lpareto` a partir de la variable `shrpop`, siguiendo la explicación de la sección 2.3.4 del texto. En las líneas 12-15 se grafican, superpuestas, las curvas de Pareto correspondientes a las regiones Noroeste y Sur de México. Las líneas 18-21 repiten el ejercicio pero dejando de lado al 1% más rico de la población.[49]

**Comentario:** ¿? Saca al 1% de la región 1 y al 1% de la región 6.

```
 1 * cap2-pareto.do
 2
 3 * ordenar según región + ipcf
 4 sort region ipcf
 5
 6 * población acumulada por región
 7 by region: gen shrpop = sum(pondera)
 8 by region: replace shrpop = shrpop/shrpop[_N]
 9
10 gen lpareto = log(1-shrpop)
11
12 * función de distribución acumulada
13 twoway (line lpareto lipcf if region==1)  ///
14        (line lpareto lipcf if region==6), ///
15        legend(label(1 "Noroeste") label(2 "Sur"))
16 more
17
18 local cutoff = 0.99
19 twoway (line lpareto lipcf [w=pondera] if region==1 & shrpop<=`cutoff')  ///
20        (line lpareto lipcf [w=pondera] if region==6 & shrpop<=`cutoff'), ///
21        legend(label(1 "Noroeste") label(2 "Sur"))
```

**Comentario:** Diagrama de Pareto

**Comentario:** ¿? Ya está ponderada la variable lpareto

**Comentario:** ¿? Ya está ponderada la variable lpareto

### Box-Plot

Aquí se muestra cómo elaborar diagramas de caja o box-plot como los presentados en la sección 2.3.5 del texto. Como muestran los siguientes ejemplos, este tipo de gráfico es muy sencillo de construir utilizando Stata; ver comando `graph box`. La opción `nooutsides` deja de lado las observaciones no adyacentes a los percentiles 25 por abajo y 75 por arriba.[50] Se deja como ejercicio para el lector el análisis del código siguiente.

```
1 * cap2-box-plot.do
2
3 * generar log ipcf
4 gen lipcf = log(ipcf)
5
6 * figura 2.15
7 * box-plot
8 graph box ipcf [w=pondera], nooutsides
9 graph box lipcf [w=pondera], nooutsides
```

[49] En el apéndice I se explica la utilización de macros locales (ver línea 18).

[50] La forma de computar los valores no adyacentes puede consultarse en el manual sobre gráficos de Stata.

```
10 more
11
12 * figura 2.16
13 * box-plot
14 graph box ipcf [w=pondera]
15 graph box lipcf [w=pondera]
16 more
17
18 * figura 2.17
19 * box-plot
20
21 graph box ipcf [w=pondera] if region==1 | region==6, ///
22       over(region, relabel(1 "Noroeste" 2 "Sur"))
```

**Comentario:** logaritmo?

### Curva de Lorenz

En este apartado se muestra cómo pueden construirse las curvas de Lorenz introducidas en la sección 2.3.6 del capítulo. El primer paso consiste en ordenar a los individuos de menor a mayor según su ingreso, en nuestro caso contenido en la variable `ipcf` (ver línea 4). Las líneas 6-8 generan la variable `shrpop` de la misma forma en la que fue generada más arriba para construir el eje vertical de la función de distribución; contiene la proporción de la población que se acumula hasta cada observación de la encuesta - así, la última observación de la encuesta tendrá el valor 1 (100%).

Las líneas 10-12 crean la variable `shrinc` que contiene la proporción del ingreso que se acumula hasta cada observación de la encuesta. El ingreso acumulado se computa en la línea 11 teniendo en cuenta a los ponderadores. En la línea 12 el ingreso acumulado se expresa como proporción del ingreso total que registra la encuesta de hogares, contenido en la última observación de la variable `shrinc` (es decir, en `shrinc [_N]`), luego de ejecutar la línea 11 pero antes de ejecutar la línea 12. La línea 16 emplea el comando `line` para graficar la variables `shrinc` y `shrpop` en los ejes vertical y horizontal, respectivamente. Cabe hacer notar que el comando `line` se emplea sin ponderadores porque los mismos fueron empleados en la construcción de las variables `shrinc` y `shrpop`. Las líneas 18 a 35 se emplean para graficar, superpuestas, las curvas de Lorenz de las regiones Noroeste y Sur de México; la explicación detallada de este fragmento de código se pospone hasta el capítulo 4.

**Comentario:** Ya se explicó todo.

```
 1 * cap2-lorenz.do
 2
 3 * ordenar según ipcf
 4 sort ipcf
 5
 6 * población acumulada ordenamiento ipcf
 7 gen shrpop = sum(pondera)
 8 replace shrpop = shrpop/shrpop[_N]
 9
10 * ingreso acumulado
11 gen shrinc = sum(ipcf*pondera)
12 replace shrinc = shrinc/shrinc[_N]
13
14 * figura 2.18
15 * curva de lorenz
16 twoway line shrinc shrpop
17
18 * figura 2.19
19 * curva de lorenz dos regiones
20 drop shrpop shrinc
21
22 * ordenar según región + ipcf
```

```
23 sort region ipcf
24
25 * población acumulada por región
26 by region: gen shrpop = sum(pondera)
27 by region: replace shrpop = shrpop/shrpop[_N]
28
29 * ingreso acumulado por región
30 by region: gen shrinc = sum(ipcf*pondera)
31 by region: replace shrinc = shrinc/shrinc[_N]
32
33 twoway (line shrinc shrpop if region==1)  ///
34        (line shrinc shrpop if region==6), ///
35        legend(label(1 "Noroeste") label(2 "Sur"))
```

Se deja como ejercicio para el lector agregar a los gráficos que genera el código anterior las líneas de perfecta igualdad.

### Curva generalizada de Lorenz

La curva generalizada de Lorenz se construye a partir de la curva de Lorenz pero multiplicando su eje vertical por el ingreso promedio (ver sección 2.3.6 en el cuerpo del capítulo).

```
 1 * cap2-lorenz-generalizada.do
 2
 3 * figura 2.20
 4 * curva lorenz generalizada dos regiones
 5
 6 * ordenar según región + ipcf
 7 sort region ipcf
 8
 9 * población acumulada por región
10 by region: gen pop = sum(pondera)
11 by region: gen shrpop = pop/pop[_N]
12
13 * lorenz generalizada por región
14 by region: gen glorenz = sum(ipcf*pondera)
15 by region: replace glorenz = glorenz/pop[_N]
16
17 twoway (line glorenz shrpop if region==1)  ///
18        (line glorenz shrpop if region==6), ///
19        legend(label(1 "Noroeste") label(2 "Sur"))
```

Las líneas 1 a 11 no se modifican respecto de las utilizadas para estimar una curva de Lorenz; notar, sin embargo, que la variable `shrpop` la generamos ahora en base a la variable `pop`. Las líneas 13-15 se emplean para construir el eje vertical de la curva generalizada de Lorenz. La línea 14 aplicada a cada región puede escribirse, utilizando notación matemática estándar, como

$$glorenz_i = \sum_{j \le i} ipcf_j \, pondera_j$$

donde la sumatoria a la derecha del igual se realiza para todos los individuos *j* con ingreso inferior o igual al del individuo *i* - al igual que para construir una curva de Lorenz, el primer paso consiste en ordenar a los encuestados según su `ipcf`. Luego, la línea 15 puede expresarse como

$$glorenz_i = \frac{\sum_{j \le i} ipcf_j \, pondera_j}{ipcf^T} \overline{ipcf}$$

donde $\overline{ipcf}$ es el ingreso per cápita familiar promedio e $ipcf^T$ es el ingreso per cápita familiar total; operando sobre la expresión anterior se obtiene la fórmula utilizada en el código para computar la variable `glorenz`,

$$glorenz_i = \frac{\sum_{j \le i} ipcf_j \, pondera_j}{ipcf^T} \frac{ipcf^T}{pop^T}$$

donde $pop^T$ es la población total o de referencia. El resto del código que se emplea para graficar las curvas generalizadas de Lorenz es relativamente sencillo.

### Curvas de incidencia del crecimiento

En este apartado se muestra cómo pueden estimarse las curvas de incidencia del crecimiento que aparecen en la figura 2.21 del texto. A modo de ejemplo, se computa la curva de incidencia del crecimiento para Argentina 1992-2006, utilizando percentiles del ingreso per cápita familiar.

El código siguiente asume que los archivos `arg92.dta` y `arg06.dta` se encuentran en la carpeta indicada con el comando `cd`. El primero (segundo) contiene la EPH (Encuesta Permanente de Hogares) de Argentina para el año 1992 (2006). En la línea 3 se crea un bucle a través de los valores 92 y 06, correspondientes a los años de las encuestas que se emplean en el ejemplo.[51] La línea 5 carga una base de datos cuyo nombre comienza con "`arg`" y se completa con el valor de la macro local `i` (i.e., 92 en la primera iteración, y 06 en la segunda). La opción `clear` del comando `use` elimina las variables y etiquetas de la base de datos antes de abrir el archivo **dta** que se indica. En las líneas 7-9 se realiza un ajuste por inflación si la base de datos que se abre es la correspondiente a 1992; en particular, se expresa el `ipcf` de dicho año a precios de 2006. El ajuste se realiza multiplicando el valor de `ipcf` por 2.1, que representa un incremento de 110% del índice de precios al consumidor entre septiembre de 1992 y septiembre de 2006, meses a los que se refiere la información en las respectivas encuetas.[52] La línea 12 ordena la base de datos de menor a mayor según la variable `ipcf`. Las líneas 14-16 computan el porcentaje de población - notar el empleo de ponderadores - que se acumula hasta cada observación de la encuesta; la misma porción de código se utilizó más arriba para construir las curvas de Lorenz. La líneas 18-22 identifican el percentil de ingreso al que pertenece cada observación. La línea 20 itera, utilizando la macro local `j` como contador, desde uno

[51] En el apéndice I del libro se explica con detalle cómo pueden implementarse bucles en Stata.

[52] En realidad, en el caso de la encuesta de 2006, la información fue recolectada durante todo el segundo semestre de dicho año.

hasta cien a intervalos de 1; en cada iteración se ejecuta el código contenido entre las llaves - notar que estas iteraciones se realizan para cada uno de los valores que puede tomar la macro local `i` (ver `foreach` en línea 3). La línea 24 utiliza el comando `table` para computar el `ipcf` promedio para cada percentil de ingreso; la opción `replace` reemplaza la base de datos en memoria por el resultado del tabulado. Así, se genera una nueva base de datos con 100 observaciones que tiene dos variables: (1) `percentil`, y (2) `table1` que contiene el `ipcf` promedio de cada percentil.[53] La línea 25 renombra la variable `table1`. En la línea 26 se ordena la nueva base de datos según la variable `percentil`; este paso es necesario para realizar - en un paso posterior - la unión entre las bases de datos con `ipcf` promedio por percentil de 1992 y 2006. La línea 27 almacena dicha base de datos con un nombre que se completa con el contenido de la macro local `i`; la opción `replace` del comando `save` reemplaza la base de datos del mismo nombre si ya existe. La línea 30 agrega a la base de datos con los `ipcf` promedio de 2006 la base de datos con los `ipcf` promedio de 1992 (ver comando `merge`). En la línea 31 se genera la variable `chg` con el cambio en el `ipcf` promedio para cada percentil del ingreso per cápita familiar. Por último, la línea 32 grafica la curva de incidencia del crecimiento para Argentina 1992-2006.

```
 1 * cap2-incidencia-crecimiento.do
 2
 3 foreach i in 92 06 {
 4
 5   use "arg`i'.dta", clear
 6
 7   if "`i'" == "92" {
 8     replace ipcf = ipcf * 2.0994
 9   }
10
11   * ordenar por ipcf
12   sort ipcf
13
14   * computar porcentaje acumulado población
15   gen shrpop = sum(pondera)
16   replace shrpop = shrpop/shrpop[_N]
17
18   * identificar percentil de ipcf
19   gen percentil = .
20   forvalues j = 1(1)100 {
21     replace percentil = `j' if shrpop > (`j'-1)*0.01 & shrpop <= `j'*0.01
22   }
23
24   table percentil [w=pondera], c(mean ipcf) replace
25   rename table1 ipcf`i'
26   sort percentil
27   save "percentil_arg`i'", replace
28 }
29
30 merge 1:1 percentil using "percentil_arg92"
31 gen chg = 100 * (ipcf06/ipcf92 - 1)
32 twoway line chg percentil, xlabel(#10)
```

[53] El nombre `table1` lo elige Stata por defecto. La opción `name` del comando `table` puede emplearse para elegir un nombre distinto.

Se deja como ejercicio para el lector agregar al gráfico que genera el bloque de código anterior intervalos de confianza del 95% para la curva de incidencia del crecimiento.