Econometría I

(MSc. Economía)

Damian Clarke[1]

Semestre 1 2018

[1]Profesor Asociado, Universidad de Santiago de Chile y Research Associate Centre for the Study of African Economies Oxford. Email damian.clarke@usach.cl.

Contents

Definiciones

Símbolo/Término	Definición
x	Un valor escalar
$\mathbf{x}$	Un vector
$\mathbf{X}$	Una matriz
iid	Independiente e idénticamente distribuida
$inid$	Independiente pero no idénticamente distribuida
$\mathbb{R}$	Número Real
$\mathbb{R}^N$	Una tupla de N números reales
$\langle \cdot, \cdot \rangle$	Espacio prehilbertiano para dos vectores
sii	si y solo si
$\lvert\mathbf{A}\rvert$	Determinante (de la matriz $\mathbf{A}$)
$\cdot \perp \cdot$	Ortogonalidad para dos vectores
$A \subset B$	Contención (el conjunto B contiene A)
$\Leftrightarrow$	Equivalencia
$A \Rightarrow B$	Implicancia (A implica B)
$\emptyset$	El conjunto vacío
iid	independiente e idénticamente distribuida
$inid$	independiente y no idénticamente distribuida
$\arg\min_x f(x)$	El valor de x que minimiza la función $f(x)$
$\min_x f(x)$	El valor de la función $f(x)$ en su punto mínimo evaluado sobre x

Sección 1

Introducción a la Econometría

1.1 Introducción al Curso de Teoria Econométrica

Estos apuntes fueron escritos para acompañar el curso "Teoría Econométrica" del Magíster en Ciencias Economícas en la Universidad de Santiago de Chile. Son apuntes nuevos (en 2018), y por lo tanto deben ser considerados como un trabajo en proceso. Cualquier comentario, sugerencia, o corrección está muy bienvendio.

La idea de los apuntes es complementar nuestra discusión en clases, su lectura de otros libros y papers, y los problemas aplicados que revisamos en clases y en ayudantías. En varias secciones de los apuntes hay en "Nota de Lectura". Éstas notas describen fuentes comprensivas para revisar el material de la sección. Las lecturas recomendadas aquí no son obligatorios, sin embargo pueden ayudar en fortalecer su aprendizaje cuando son leídos para complementar estos apuntes. Si hay temas de los apuntes que no quedan claros o que le gustaría revisar en más detalla, estas lecturas son la mejor fuente para resolver dudas (además de preguntas en clases). Si los libros indicados no están disponibles en la biblioteca, se los puede pedir del profesor.

1.2 Econometría

La econometría—más que ser una simple aplicación de los métodos de estadística a problemas economícas—tiene sus propios fundamentos y metas. La econometría une teoría estadística formal con datos empíricos del mundo real y con teoría y modelos económicos. A menudo cuando planetamos modelos econométricos nos interesa la relación *causal* entre variables. Una fortaleza de la econometría es que nos proporciona las herramientas necesarias para hablar en términos causales (o de qué pasaría *ceteris paribus* al variar una variable de interés) **si nuestros supuestos de identificación son correctos**.

Es común en la econometría trabajar con datos observacionales, en vez de datos experimentales (aunque también hay aplicaciones experimentales en la econométria). Esto implica llegar

con una pregunta de interés y datos que ya existen, cuando nos gustaría saber que pasaría al variar alguna variable sin cambiar ninguna otra variable. Como los datos observacionales generalmente no proporcionan variación de *una sola variable*, nuestros modelos econométricos tienen que intentar a encontrar una manera de aislar estos cambios únicos utilizando variación encontrada por sistemas naturales y económicos. Será una preocupación constante asegurar que nuestros estimadores capturan solo el cambio de una variable de interés, y no cambios simultáneos de otras variables. Cuando logramos hacer esto, nos permitirá hablar en términos de causalidad en vez de correlación.

La econometría es un campo bastante amplio, desde modelos estáticos con supuestos muy paramétricos, a modelos que siguen a sus observaciones sobre muchos periodos, o ponen muy poca estructura en sus supuestos (eg modelos no paramétricos). En este curso nos enfocaremos en las herramientas básicas que nos sirvirián en cualquier área de econometría (probabilidad y algébra lineal), y después introducimos una serie de modelos paramétricos. En cursos futuros del magíster, se examinará otros tipos de modelos y supuestos econométricos.

Aunque la econometría se basa en supuestos económicos y de probabilidad, los estudios aplicados de econometría se abarcan áreas *muy* amplias, incluyendo salud, educación, organización industrial, economía política, y cualquier otro área de estudio donde se interesa aislar el impacto causal de una(s) variable(s) independientes sobre otras variables de interés.... A continuación discutiremos un poco acerca de las aplicaciones y estudios empíricos en econometría.

1.3 Recursos

1.3.1 Estudios Teóricos y Aplicados

El canal de comunicación principal para compartir nuevos resultados en econometría (y en economía de forma más genérica) es a través de journals (o revistas) científicas. Cuando se concentra en el desarrollo de econometría teórica, una proporción importante de los avances en este campo se publica en journals como Econometrica, Journal of Econometrics, The Econometrics Journal, Journal of the American Statistical Association, y Econometric Theory. Estos journals publican ediciones varias veces por año con estudios que han sido juzgados por sus pares como teóricente correctos y como contribuciones importantes al campo de econometría teórica. Los journals una muy buena fuente para seguir el desarrollo de los temas áctivas en econometría moderna.

Además de articulos demostrando los avances en el campo de econometría teórica, hay una multiplicidad de journals que publican articulos que utilizan herramientas econométricas de forma aplicada. Existen muchos journals que publican papers de econometría aplicada incluyendo The Quarterly Journal of Economics, The American Economic Review, The Review of Economic Studies, The Journal of Political Economy, The Review of Economics and Statistics, The American Economic Journals, y The Economic Journal. Además, papers aplicados a ciertos sub-especialidades

de economía/econometría se publican en journals especializados en cada campo. Algunos ejemplos incluyen The Journal of Labor Economics (economía laboral), Journal of Health Economics (salud), Journal of Monetary Economics (economía monetaria), Journal of Development Economics (desarrollo económico), ... Es una buena idea seguir los papers que salen en estos journals, especialmente los journals en los campos que más le interésa, para ver el estado del arte de econométria aplicada. Aunque este curso se enfoca casi exclusivamente en la teoría econométrica en vez de aplicaciones empíricas, la lectura de estudios aplicados es una manera entretenida de entender como la teoría que revisamos en estos apuntes se traduce en aplicaciones reales.

Por último, los estudios más nuevos, antes de salir publicado en algún journal salen como un *working paper*. Los *working papers* son utilizados para comunicar resultados entre investigadores de economía/econometría y para recolectar comentarios, y además sirven como una manera para compartir resultados antes de que salgan definitivamente en el journal. Como el proceso de publicación en economía puede demorar bastante (no es atípica que un paper sale publicado dos o tres años después de salir por primera vez como un *working paper*), los *working papers* sirven como una manera más inmediata para hacer conocido resultados importantes. Si le interesa mantener al tanto de los desarrollos más nuevos en economía y econometría, es una buena idea inscribirse para recibir un resúmen temporal de *working papers*. Algunos buenas fuentes de estos papers son la serie del National Bureau of Economic Research, la serie de IZA Institute of Labor Economics, o inscribiéndose en los listados de NEP: New Economics Papers, que existen en casi 100 sub-especialidades de economía.

1.3.2 Herramientas

Al momento de estimar modelos econométricos con datos reales, es (casi siempre) necesario contar con un computador, e idealmente algún idioma computacional con implementaciones de modelos comúnes de econometría. Existen muchas opciones muy buenas de idiomas con fortalezas en econometría. Esto incluye idiomas como Python, Julia, R, Octave y Fortran (todos libres), y Stata, SAS, y MATLAB (pagados). Como la econometría se basa en mucha álgebra lineal, es particularmente conveniente contar con un idioma basada en matrices para simulaciones y aplicaciones para explorar resultados teóricos importantes. Idiomas que son especialmente fácil con matrices incluyen Julia, MATLAB, y Mata. Esta úlitma (Mata) es un tipo de sub-idioma que existe adentro de Stata con un syntaxis enfocado en algébra lineal y manipulación de matrices.

En este curso, generalmente utilizamos Stata y Mata. Este idiomas tienen muchas herraimentas muy desarrolladas enfocados en econometría, y una serie de documentación muy comprensiva. Pero un costo de Stata y Mata es, justamente, su costo! A diferencia de Python, Julia y otras, no es un idioma libre ni gratis. En la universidad tendrán acceso libre a Stata y Mata. Sin embargo, si le interesa trabajar con otros idiomas de los mencionados anteriormente (u otras), no hay problema! Existen muchos libros y materiales muy buenos con un enfoque de econometría computacional incluyendo Cameron and Trivedi (2009) (Stata), Adams, Clarke and Quinn (2015) (MATLAB), y el excelente curso de Stachurski y Sargent: https://lectures.quantecon.org/ (Python

y Julia) con un enfoque más amplio que solo econometría, a modelos economicos más generalmente.

1.4 Este Curso

Las detalles completas de este curso estarán disponibles en el siguiente sitio web:

https://sites.google.com/site/damiancclarke/econometria-i

Se sugiere revisar este sitio como la fuente oficial de la información principal del curso incluyendo el programa del curso, el calendario, ejercicios computacionales, trabajos, y pruebas pasadas. El curso de este año se difiere algo en cursos de años previos, incluyendo una sección nueva de repaso de herramientas algebraícas. Así, aunque las pruebas y exámenes anteriores pueden ser fuentes para repasar material, la estructura este año va a cambiar levamente. La actual versión de estos apuntes contiene la material para la primera mitad del curso. Una versión actualizada a mitad de semestre incluirá las detalles también de la segunda mitad del curso (temas 4-6 del programa).

Este curso está diseñado como el primer curso en un ciclo de hasta cuatro cursos de econometría en el Magíster en Ciencias Económicas. En el segundo semestre del magíster este curso será seguido por el curso de econometría II que introduce otros modelos y teécnicas, incluyendo modelos no-lineales. En el segundo año del magíster hay dos cursos electivos: uno "Topicos de Microeconometría" enfocado netamente de aplicaciones empíricas de modelos nuevos en microeconometría como modelos de regresión discontinua y modelos de diferencias-en-diferencias, y otro enfocado en modelos de series de tiempo, frecuentamente encontrado en aplicaciones *macro*-econométricas.

Sección 2

Un Repaso de Herramientas Algebraícas

Nota de Lectura: Para un repaso de algébra lineal existen muchas fuentes interesantes, tanto de matemática como de econometría. Un análisis de bastante alto nivel está disponible en Stachurski (2016), capítulos 2 y 3. Rao (1973) es una referencia clásica, con demostraciones muy elegantes. Generalmente libros de texto de econometría tienen un capítulo o apéndice de revisión, por ejemplo apéndice A de Hansen (2017).

2.1 Operaciones y Elementos Básicas

2.1.1 Elementos Básicos

Un valor escalar, escrito x es un único número. Un vector, genéricamente de $\mathbb{R}^N$, contiene N elementos, o escalares y se escribe como:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix},$$

con cada $x_n \in \mathbb{R}$. En esta notación $\mathbb{R}$ refiere a los números reales, y N a la cantidad de números naturales contenidos en el vector, o dimensiones. En el caso de que $N = 1$, $\mathbb{R} = \mathbb{R}^1$ es la línea de números reales, que es la unión de los números racionales e irracionales. Un único elemento de un vector $\boldsymbol{x}$ es un escalar.

Definimos una matriz de $N \times K$ dimensiones, llamado $\mathbf{X}$. La matriz $\mathbf{X}$ contiene números reales en N filas y K columnas. Las matrices son particularlmente útil para agrupar datos u operaciones algebraícas. En estos apuntes siempre utilicarémos letras minúsculas x para denotar valores escalares, letras minúsculas con negrita $\mathbf{x}$ para denotar vectores, y letras mayúsculas en

negrita $\mathbf{X}$ para denotar matrices. Entonces, una matríz de tamaño $N \times K$ refiere a una colección de número reales, y se escribe como:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix}.$$

Cuando escribimos x_{nk}, referimos a la entrada en la fila n-ésima y la columna k-ésima. En el caso de que $N = K$, la matriz $\mathbf{X}$ de $N \times N$ es conocido como una matriz cuadrada, y los elementos x_{nn} para $n = 1, 2, \ldots, N$ son conocidos como el diagonal principal de la matriz.

2.1.2 Operaciones Básicas

Las operaciones algebraícas básicas con matrices refieren a *adición*, y *multiplicación*. La adición, o sumatorias con matrices, es un proceso elemento por elemento. Para $\mathbf{W}, \mathbf{X} \in \mathbb{R}^{N \times K}$ su sumatoria es:

$$\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1K} \\ w_{21} & w_{22} & \cdots & w_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{NK} \end{pmatrix} + \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix} = \begin{pmatrix} w_{11} + x_{11} & w_{12} + x_{12} & \cdots & w_{1K} + x_{1K} \\ w_{21} + x_{21} & w_{22} + x_{22} & \cdots & w_{2K} + x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} + x_{N1} & w_{N2} + x_{N2} & \cdots & w_{NK} + x_{NK} \end{pmatrix}.$$

De esta sumatoria, es aparente que para sumar dos (o más) matrices es necesario que sean del mismo tamaño, y en este caso se dice que las matrices son conformables para adición. La sumatoria de vectores sigue exactamente la misma lógica, ya que una matriz de $N \times 1$ o de $1 \times K$ es un vector (llamado un vector de fila o vector de columna respectivamente), y la definición anterior cumple siempre cuando ambos vectores en la sumatoria son del mismo tamaño.

A continuación se resumen algunas de las propiedades de la adición de matrices. En este listado simplemente resumimos las propiedades, y dejamos la demostración de estas propiedades como un ejercicio.

Propiedades de Sumatorias de Matrices

1. Asociatividad: $(\mathbf{X} + \mathbf{Y}) + \mathbf{Z} = \mathbf{X} + (\mathbf{Y} + \mathbf{Z})$
2. Conmutatividad: $(\mathbf{X} + \mathbf{Y}) = (\mathbf{Y} + \mathbf{X})$
3. Existencia de un Elemento Nulo: $\mathbf{X} + \mathbf{0} = \mathbf{X}$

El elemento nulo en ítem 3 refiere a un vector conformable para la adición con $\mathbf{X}$ cuyos elementos x_{ij} son todos iguales a 0. Volveremos a una serie de matrices importantes, incluyendo el matriz nulo, en la sección 2.2.

Multiplicación La segunda operación básica de matrices es la multiplicación. La múltiplicación escalar—donde una matriz se multiplica con un valor escalar α—se define de la misma forma que múltiplicación en $\mathbb{R}$. Así, cuando una matriz se múltiplica con un único valor, se multiplica elemento por elemento para llegar a la solución:

$$\alpha \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix} = \begin{pmatrix} \alpha x_{11} & \alpha x_{12} & \cdots & \alpha x_{1K} \\ \alpha x_{21} & \alpha x_{22} & \cdots & \alpha x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha x_{N1} & \alpha x_{N2} & \cdots & \alpha x_{NK} \end{pmatrix}.$$

En la multiplicación de dos matrices para formar el producto matricial $\mathbf{Z} = \mathbf{XY}$, cada elemento z_{ij} se calcula de la siguiente forma:

$$z_{ij} = \sum_{k=1}^{K} w_{ik} x_{kj} = \langle \text{fila}_i \, \boldsymbol{W}, \text{columna}_j \, \boldsymbol{X} \rangle. \tag{2.1}$$

La segunda notación $\langle \cdot, \cdot \rangle$ refiere al espacio prehilbertiano, que es la suma del producto de los elementos de dos vectores. Este cálculo con dos matrices en forma extendida está presentado de forma esquemática en la ecuación 2.2.

$$\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1K} \\ w_{21} & w_{22} & \cdots & w_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{NK} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1K} \\ z_{21} & z_{22} & \cdots & z_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ z_{N1} & z_{N2} & \cdots & z_{NK} \end{pmatrix} \tag{2.2}$$

El elemento z_{11} se calcula a partir de la primera fila y la primera columna (los vectores azules):

$$z_{11} = w_{11}x_{11} + w_{12}x_{21} + \ldots + w_{1K}x_{K1},$$

que es el espacio prehilbertiano para fila 1 y columa 1. Todos los otros elementos se calculan de la misma forma, por ejemplo z_{2K} a partir de fila 2 y columna K. El espacio prehilbertiano es solo definido para vectores con la misma cantidad de elementos, que implica que las matrices solo son *conformables* (o pueden ser multiplicados) si la cantidad de elementos en las filas de $\mathbf{W}$ es igual a la cantidad de elementos en las columnas de $\mathbf{X}$. Si $\mathbf{W}$ es $N \times K$ y $\mathbf{X}$ is $J \times M$, esto implica que una matriz es conformable solo si $K = J$. En este caso, el producto $\mathbf{Z}$ será de $N \times M$

Por lo general, multiplicación matricial no es conmutativa: $\mathbf{XY} \neq \mathbf{YX}$. En algunos casos, una matriz que puede ser pre-multiplicado con otra matriz ni siquera es conformable cuando se post-multiplica con la misma matriz. Por ejemplo, multiplicando una matriz de 5×2 con otro de 2×4 resulta en una matriz de 5×4, pero al revés no es conformable, dado que hay 4 columnas en la primera matriz, y 5 filas en la segunda. Y en otros casos, aunque dos matrices son conformables,

el resultado no es lo mismo multiplicando de ambas formas. Por ejemplo, consideremos:

$$\mathbf{A} = \begin{pmatrix} 2 & 3 \\ 1 & 0 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 4 & 1 \\ 2 & 1 \end{pmatrix}$$

Es fácil confirmar que:

$$\mathbf{AB} = \begin{pmatrix} 14 & 5 \\ 4 & 1 \end{pmatrix} \qquad \mathbf{BA} = \begin{pmatrix} 9 & 12 \\ 5 & 6 \end{pmatrix},$$

y en este caso $\mathbf{AB} \neq \mathbf{BA}$.

Las otras propiedades de multiplicación de matrices son parecidas a las propiedades de multiplicación de números reales. Específicamente, para matrices conformables $\mathbf{X}$, $\mathbf{Y}$, y $\mathbf{Z}$, y un escalar α:

Propiedades de Multiplicación de Matrices

1. Asociatividad: $(\mathbf{XY})\mathbf{Z} = \mathbf{X}(\mathbf{YZ})$
2. Distribuitividad por la izquierda: $\mathbf{X}(\mathbf{Y} + \mathbf{Z}) = \mathbf{XY} + \mathbf{XZ}$
3. Distribuitividad por la derecha: $(\mathbf{X} + \mathbf{Y})\mathbf{Z} = \mathbf{XZ} + \mathbf{YZ}$
4. Multiplicación Escalar: $\mathbf{X}\alpha\mathbf{Y} = \alpha\mathbf{XY}$
5. Existencia de un Elemento Neutro: si $\mathbf{X}$ es una matriz cuadrada $\mathbf{XI} = \mathbf{X}$ y $\mathbf{IX} = \mathbf{X}$

El elemento neutro en el ítem 5 refiere al matriz de identidad: una matriz cuadrada con valores 1 en el diagonal principal, y valores de 0 en cada otra posición. Describimos esta matriz con más detalle en sección 2.2.

Trasposición La traspuesta de una matriz $\mathbf{X}$ de $K \times N$, escrito $\mathbf{X}^\top$ o $\mathbf{X}'$ (en estos apuntos preferimos $\mathbf{X}'$), es una matriz de $N \times K$ donde $x_{nk} = x_{kn}$ para cada k y n. Las columnas de $\mathbf{X}$ se convierten en las filas de $\mathbf{X}'$, y visualmente se ve de la siguiente forma:

$$\mathbf{X} = \begin{pmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \end{pmatrix} \qquad \mathbf{X}' = \begin{pmatrix} 0 & 3 \\ 1 & 4 \\ 2 & 5 \end{pmatrix}.$$

Si la matriz $\mathbf{X}$ es cuadrada, su traspuesta se produce rotatndo la matriz alrededor del diagonal principal.

Existen varias propiedades de traspuestas, resumidas en el siguiente listado, donde supongamos que $\mathbf{X}$ y $\mathbf{Y}$ son conformables, y α es un escalar.

Propiedades de Trasposición

1. $(\mathbf{X}')' = \mathbf{X}$
2. Traspuesta de un producto: $(\mathbf{XY})' = \mathbf{Y}'\mathbf{X}'$
3. Traspuesta de un producto extendido: $(\mathbf{XYZ})' = \mathbf{Z}'\mathbf{Y}'\mathbf{X}'$

4. $(\mathbf{X} + \mathbf{Y})' = \mathbf{X}' + \mathbf{Y}'$
5. $(\alpha\mathbf{X})' = \alpha\mathbf{X}'$

De nuevo, estas propiedades están presentadas sin demostración, y la demostración se deja como un ejercicio.

2.2 Matrices Importantes

Cuando revisamos las propiededes de multiplicación y sumatoria de matrices, conocimos dos matrices importantes. Estos son la matriz de identidad, $\mathbf{I_k}$, y la matriz nula, $\mathbf{0}$:

$$\mathbf{I_k} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \qquad \mathbf{0} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

La matriz identidad es una matriz cuadrada de $k \times k$ con unos en el diagonal principal, y ceros en todas las demás posiciónes. La matriz identidad es un tipo de matriz diagonal (una matriz cuadrada cuyas únicas elementos no-nulos aparecen en el diagonal principal), y un tipo de matriz escalar (una matriz diagonal con un único constante en cada posición del diagonal principal). La matriz nula es una matriz con ceros en cada posición. A diferencia de la matriz identidad, no es necesariamente una matriz cuadrada.

Una matriz cuadrada puede ser *triangular* si solo tiene elemento nulos (i) arriba o (ii) abajo del diagonal principal. En el primer caso la matriz se conoce como una matriz triangular inferior, y en el segundo caso la matriz se conoce como una matriz triangular superior. A continuación, se presenta la versión inferior ($\mathbf{L}$), y la versión superior ($\mathbf{U}$).

$$\mathbf{L} = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{k1} & l_{k2} & \cdots & l_{kk} \end{pmatrix} \qquad \mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1k} \\ 0 & u_{22} & \cdots & u_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{kk} \end{pmatrix} \tag{2.3}$$

Como veremos cuando hablamos de Invertibilidad y descomposiciones de Cholesky y QR en la sección 2.3, una propiedad muy conveniente de las matrices triangulares es que permiten soluciones muy simples a sistemas de ecuaciones. Por ejemplo, consideramos un sistema de ecuaciones del estilo $\mathbf{Lx} = \mathbf{b}$, donde $\mathbf{L}$ es una matriz triangular inferior (conocido), $\mathbf{b}$ es un vector de constantes conocidos, y $\mathbf{x}$ es un vector de incógnitas. Entonces:

$$\begin{pmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} L_{11}x_1 \\ L_{21}x_1 + L_{22}x_2 \\ L_{31}x_1 + L_{32}x_2 + L_{33}x_3 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}, \tag{2.4}$$

y existe una manera simple de resolver para las incógnitas. La primera ecuación sólo incluye un incógnita x_1, y se puede encontrar su valor directamente. Después se sustituya el valor de x_1 en la segunda línea, para resolver la segunda incógnita x_2. Se sigue este proceso recursiva hasta resolver para todos los elementos del vector $\mathbf{x}$.

2.2.1 Formas Cuadráticas, y Formas Definidas

Cuando se habla de la forma cuadrática de una matriz $\mathbf{A}$ o su función cuadrática, se refiere a la forma:

$$Q = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j a_{ij}. \tag{2.5}$$

Aquí $\mathbf{A}$ es una matriz cuadrada (de $N \times N$), y $\mathbf{x} \in \mathbb{R}^N$ es un vector (de $1 \times N$). Formas cuadráticas de este estilo se encuentran frecuentemente en problemas de optimización, y en algunas clases, existen resultados muy elegantes acerca de la forma cuadrática. Específicamente, nos interesa saber si las formas definidas de estas matrices son definidas en alguna forma. Existen cuatro tipos de matrices definidas, que se definen a continuación:

1. Si $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ para cualquier candidato $\mathbf{x} \neq \mathbf{0}$, $\mathbf{A}$ se conoce como una matriz definida positiva
2. Si $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ para cualquier candidato $\mathbf{x}$, $\mathbf{A}$ se conoce como una matriz semi-definida positiva
3. Si $\mathbf{x}'\mathbf{A}\mathbf{x} < 0$ para cualquier candidato $\mathbf{x} \neq \mathbf{0}$, $\mathbf{A}$ se conoce como una matriz definida negativa
4. Si $\mathbf{x}'\mathbf{A}\mathbf{x} \leq 0$ para cualquier candidato $\mathbf{x}$, $\mathbf{A}$ se conoce como una matriz semi-definida negativa

Si $\mathbf{A}$ no cumple con ninguno de los puntos 1-4, la matriz se conoce como una matriz indefinida. La matriz definida positiva (y negativa) parecen, de alguna forma, a los números reales positivos (negativos). Si se suma dos matrices definidas positivas (negativas), la matriz resultante tiene que ser positiva (negativa). Existen muchos resultados muy útiles si se sabe que una matriz es definida de cierta forma. Por ejemplo, una matriz definida positiva siempre es convexo, que tiene implicancias importantes para optimización. Conoceremos otro resultado importante relacionado con matrices definidas positivas en la sección 2.3.4 cuando revisamos la inversión de matrices y la descomposición Cholesky. En la econometría, existen muchos matrices importantes que son definida de cierta forma, por ejemplo la matriz de covarianza, que es una matriz definida positiva.

Notemos que aquí, aunque la matriz $\mathbf{A}$ es conocida en ecucación 2.5, los vectores $\mathbf{x}$ refieren a *cualquier* vector posible. A primera vista, podría parecer bastante difícil determinar el signo de Q para cualquier vector $\mathbf{x}$, pero dada la forma cuadrática en que $\mathbf{x}$ entra la ecuación, no es tan restrictiva que parece. Como ejemplo, consideramos la matriz identidad $\mathbf{I}_3$, y cualquier vector no cero $\mathbf{x} = [x_1\ x_2\ x_3]'$. Se puede mostrar fácilmente que la matriz de identidad $\mathbf{I}_3$ (y cualquier

matriz de identidad) es positiva definida, dado que:

$$Q = \mathbf{x}'\mathbf{I}_3\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1^2 + x_2^2 + x_3^2,$$

y cada elemento $x_i^2 > 0 \; \forall \; i \in 1, 2, 3 \Rightarrow Q > 0$.

2.3 El Inverso de Una Matriz

2.3.1 Definición y Uso de Inversión

Se dice que una matriz cuadrada de $N \times N$, $\mathbf{A}$ es invertible (o no singular, o no degenerada) si existe otra matriz $\mathbf{B}$ de tal forma que:

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I_N}. \tag{2.6}$$

Si la ecucación 2.6 cumple, entonces $\mathbf{B}$ es única, y se conoce como el inverso de $\mathbf{A}$. El inverso se escribe como $\mathbf{A}^{-1}$. De la definición en 2.6, tenemos que $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$, donde el tamaño de $\mathbf{I}$ es igual al tamaño de la matriz original $\mathbf{A}$ (y su inverso $\mathbf{A}^{-1}$).

Consideramos un sistema de ecuaciones lineales de la forma:

$$\mathbf{Ax} = \mathbf{b}. \tag{2.7}$$

Aquí, $\mathbf{A}$ es un matriz de $N \times N$ y $\mathbf{b}$ es un vector de $N \times 1$, ambas conocidas. El vector $\mathbf{x}$ (de $N \times 1$) contiene los valores desconocidos, que estamos buscando resolver en la ecuación. Si $\mathbf{A}$, $\mathbf{x}$, y $\mathbf{b}$ eran valores escalares, sería fácil encontrar la solución para $\mathbf{x}$, simplemente dividiendo ambos lados de 2.7 por $\mathbf{A}$. Sin embargo, para encontrar la solución a un sistema de ecuaciones matricial, necesitamos utilizar la idea del inverso de una matriz. Supongamos que existe una matríz $\mathbf{B}$ de la forma descrita en la ecuación 2.6. Entonces, podemos encontrar la solución $\mathbf{x}$ de la siguiente manera:

$$\mathbf{BAx} = \mathbf{Bb} \tag{2.8}$$
$$\mathbf{Ix} = \mathbf{Bb} \tag{2.9}$$
$$\mathbf{x} = \mathbf{Bb} \tag{2.10}$$
$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \tag{2.11}$$

En la ecuación 2.8 pre-multiplicamos ambos lados de 2.6 por $\mathbf{B}$. Por la naturaleza del inverso, $\mathbf{AB} = \mathbf{AB} = \mathbf{I}$, y $\mathbf{Ix} = \mathbf{x}$ ya que la matriz identidad es un elemento neutro. Por último, en 2.11 reemplazamos $\mathbf{B}$ por su notación común, llegando a la solución única para ecuación 2.7. Demostraciones de la unicidad del inverso (si una matriz es invertible) están disponibles en varios libros de econometría. Por ejemplo, una demostración de Greene (2002, §A.4.2) refiere a la matriz

A y su inverso **B**. Ahora, supongamos que existe otro inverso **C**. En este caso:

$$(\mathbf{CA})\mathbf{B} = \mathbf{IB} = \mathbf{B} \tag{2.12}$$

$$\mathbf{C}(\mathbf{AB}) = \mathbf{CI} = \mathbf{C}, \tag{2.13}$$

y 2.12-2.13 son inconsistentes si **C** no es igual a **B**.

Algunas propiedades de los inversos de matrices simétricas e invertibles **A** y **B**, ambos del mismo tamaño de $N \times N$ y un escalar no igual a cero α son:
Propiedades de Inversión de Matrices

1. $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
2. $(\alpha\mathbf{A})^{-1} = \alpha^{-1}\mathbf{A}^{-1}$ para un escalar $\alpha \neq 0$
3. $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$
4. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
5. $(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1}(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}\mathbf{B}^{-1}$

2.3.2 El Determinante de una Matriz

Para calcular algebraícamente el inverso de una matriz, se requiere calcular el determinante. El determinante de una matriz cuadrada **A** (el determinante solo existe para matrices cuadradas) se escribe $|\mathbf{A}|$, o det **A**. El determinante es un valor única para cada matriz, y captura el volumen de la matriz.Como veremos más adelante, el valor del determinante interviene en mucho resultados algebraícos. Por ejemplo, una matriz es invertible si y solo si (*ssi*) su determinante no es igual a cero.

Computacionalmente, para encontrar el determinante se sigue un proceso recursiva para considerar sub-bloques en cada matriz. Sin embargo, existe una fórmula analítica para el determinante de una matriz. Siguiendo la definición de Hansen (2017, p. 460), el determinante de una matriz de $k \times k$ $|\mathbf{A}| = a_{ij}$ se puede calcular utilizando las permutaciones de $\ldots \pi = (j_1, \ldots, j_k)$, que son todas las formas posibles para reorganizar los valores $1 \ldots, k$ (existen $k!$ posibles permutaciones). Y definimos como $\varepsilon_\pi = 1$ si la cantidad de inversiones en el orden de $1 \ldots, k$ para llegar a $\pi = (j_1, \ldots, j_k)$ es un número par, y como $\varepsilon_\pi = -1$ si la cantidad de inversiones es un número impar. Entonces, definimos a la determinante de una matriz **A** como:

$$|\mathbf{A}| = \sum_{\pi} \varepsilon_\pi a_{1j_1} a_{2j_2} \cdots a_{kj_k}. \tag{2.14}$$

Para la matriz **A**, decimos que el menor para cada elemento a_{ij}, denotado M_{ij}, es el determinante de la matriz, una vez que hemos eliminido la fila i y la columna j de **A**. Y definimos como al cofactor del mismo elemento $C_{ij} = (-1)^{i+j} M_{ij}$. Estos cofactores tienen un vínculo importante con el determinante de la matriz entero resumido en el Teorema de Laplace. Este teorema relaciona

$|\mathbf{A}|$ con sus cofactores mediante la fórmula:

$$|\mathbf{A}| = \sum_{j=1}^{k} a_{ij} C_{ij},$$

y esta fórmula es conveniente para computar el determinante en pasos sucesivos para llegar a una matriz de 3×3 or 2×2. En el caso de una matriz de 2×2, se puede demostrar que (utilizando ecuación 2.14), que el determinante es:

$$\left| \begin{pmatrix} a & b \\ c & d \end{pmatrix} \right| = ab - bc.$$

2.3.3 Encontrando el Inverso de Una Matriz

Por lo general, al momento de encontrar el inverso de una matriz, se utiliza un algoritmo conocido (por ejemplo la eliminación de Gauss-Jordan) y un computador, aunque también es un proceso que se puede calcular 'a mano'. No revisaremos estos algoritmos aquí (si le interesa, una descripción está disponible en Simon and Blume (1994, §7.1).

Pero en casos de matrices pequeñas, es sencillo expresar la fórmula para el inverso. Por ejemplo, en el caso de una matriz $\mathbf{A}$ de 2×2, tenemos que:

$$\mathbf{A}^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{|\mathbf{A}|} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

y para un $\mathbf{A}$ de 3×3 el inverso se calcula como:

$$\mathbf{A}^{-1} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}^{-1} = \frac{1}{|\mathbf{A}|} \begin{pmatrix} ei - fh & -(bi - ch) & bf - ce \\ -(di - fg) & ai - cg & -(af - cd) \\ dh - eg & -(ah - bg) & ae - bd \end{pmatrix}.$$

Adicionalmente, el determinante en el denominador se puede calcular utilizando la regla de Sarrus, que da el determinante de una matriz de 3×3 como $|\mathbf{A}| = aei + dhc + gbf - ceg - fha - ibd$. Por el momento no discutiremos los requisitos para saber si una matriz es invertible o no. Volveremos a examinar los requisitos de invertibilidad en la sección 2.4.2, después de introducir la idea de independencia en la sección 2.4.

2.3.4 ⋆La Descomposición de Cholesky

La descomposición de Cholesky es una descomposición de una matriz simétrica definida positiva ($\mathbf{A}$). La descomposición consiste en encontrar una matriz $\mathbf{L}$ (y su traspuesta $\mathbf{L}'$) de forma que:

$$\mathbf{A} = \mathbf{L}\mathbf{L}'.$$

Aquí $\mathbf{L}$ es una matriz triangular como en la ecuación 2.3, y cada elemento del diagonal principal de $\mathbf{L}$ es estríctamente positivo. Esta descomposición es particularmente útil[1] por su uso como una manera más (computacionalmente) eficiente de resolver sistemas de ecaciones sin la necesidad de invertir una matriz. Para ver esto, partimos con una ecuación lineal de la misma forma que en 2.7. Calculamos la descomposición de Cholesky de la matriz definida positiva $\mathbf{A}$, dando $\mathbf{A} = \mathbf{LL}'$. Con esto, se puede re-escribir ecuación 2.7 como:

$$\mathbf{LL}'\mathbf{x} = \mathbf{b}. \tag{2.15}$$

Ahora, si definimos $\mathbf{z} = \mathbf{L}'\mathbf{x}$, finalmente se puede escribir 2.15 como:

$$\mathbf{Lz} = \mathbf{b}. \tag{2.16}$$

Es fácil resolver 2.16 para el desconocido $\mathbf{z}$ ya que $\mathbf{L}$ es una matrix triangular inferior, y una vez que se sabe $\mathbf{z}$, también podemos volver a $\mathbf{z} = \mathbf{L}'\mathbf{x}$ para encontrar el desconocido $\mathbf{x}$ (de interés) fácilmente, dado que $\mathbf{L}'$ es una matriz triangular superior. En ambos casos, el hecho de que las matrices en las ecuaciones son triangulares implica que se puede resolver la ecuación utilizando sustitución recursiva, un proceso simple y rápido de resolver un sistema de ecuaciones como revisamos en la ecucación 2.4.[2]

Para una matriz $\mathbf{A}$ definida positiva (y de $k \times k$), la descomposición de Cholesky es única. Las demostraciones de unicidad son inductivas. Para el caso de $k = 1$, simplemente se toma la raíz cuadrada (que es única). Para un k arbitrario, la demostración formal está disponible en Golub and Van Loan (1983, p. 88). En el texto de Hansen (2017, §A.14), se presenta una derivación muy intuitiva de la descomposición única cuando $k = 3$.

2.3.5 ⋆La Descomposición QR

La descomposición QR es otra descomposición en el estilo de Cholesky, donde se descompone la matriz $\boldsymbol{A}$ (simétrica, semi-definida positiva) como:

$$\mathbf{A} = \mathbf{QR}.$$

Esta proceso descompone cualquier $\boldsymbol{A}$ invertible en dos matrices. La primera, $\boldsymbol{Q}$, es una matriz ortogonal, que implica que sus columnas y filas son vectores ortogonales unitarios, y por ende $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}$. La segunda matriz, $\boldsymbol{R}$, es una matriz triangular superior. Existen una serie de algoritmos comúnes para realizar esta descomposición.

[1] En realidad, la descomposición de Cholesky es muy útil para varias razones, pero en este curso referimos principalmente a la descomposición cuando pensamos en una manera para resolver sistemas de ecuaciones en mínimos cuadrados ordinarios. Adicionalmente, la descomposición de Cholesky es de utilidad en simulaciones de números pseudo-aleatorios cuando se quiere simular múltiples variables correlacionadas. Con una base de una matriz de variables no correlacionadas $\mathbf{U}$, se puede simular una matriz de variables con una correlación deseada, $\mathbf{Z}$, utilizando la descomposición Cholesky de la matriz de covarianza $\mathbf{\Sigma} = \mathbf{LL}'$, mediante $\mathbf{Z} = \mathbf{LU}$.

[2] Refiere al programa `cholesky.do` para un ejemplo de la descomposición de Cholesky con datos simulados en Mata.

De nuevo, esta descomposición proporciona una manera eficiente para resolver sistemas de ecuaciones lineales como ecuación 2.7. Para ver porque, observamos:

$$\begin{aligned} \mathbf{Ax} &= \mathbf{b} \\ \mathbf{QRx} &= \mathbf{b} \\ \mathbf{Q'QRx} &= \mathbf{Q'b} \\ \mathbf{Rx} &= \mathbf{Q'b}, \end{aligned} \tag{2.17}$$

y notamos que la ecuación 2.17 se puede resolver simplemente para $\mathbf{x}$ utilizando sustitución recursive dado que $\mathbf{R}$ es una matriz triangular superior.

2.4 Independencia y Ortogonalidad (de vectores)

2.4.1 Independencia

Consideremos una serie de vectores $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K\}$. Siempre se puede formar la matriz nula como una combinación lineal de los vectores de la forma:

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \ldots + \alpha_K\mathbf{x}_K = \mathbf{0}$$

si definimos a cada valor escalar α como 0. Decimos que X es linealmente independente se ésta es la *única* combinación de valores de α posible para formar el vector nulo. Es decir, la independencia lineal implica:

$$\sum_{k=1}^{K} \alpha_k\mathbf{x}_k = \mathbf{0} \qquad \Rightarrow \qquad \alpha_1 = \alpha_2 = \cdots = \alpha_K = 0, \tag{2.18}$$

donde el símbolo $\Rightarrow$ significa que si la primera ecuación cumple, entonces la segunda ecuación tiene que cumplir. De otra forma—si hay otras soluciones potenciales para el vector de valores α—se dice que X es linealmente dependiente.

La dependencia lineal implica que por en un set de $k \geq 2$ vectores, por lo menos uno de los vectores se puede escribir como una combinación lineal de los otros vectores. Para un caso muy simple, consideremos los vectores $\mathbf{x}_1 = (2\ 1\ 3)'$ y $\mathbf{x}_2 = (-6\ -3\ -9)'$. Es fácil comprobar que los vectores $\mathbf{x}_1$ y $\mathbf{x}_2$ son linealmente dependientes, ya que $\mathbf{x}_2 = -3 \times \mathbf{x}_1$. En este caso, se puede llegar al vector $\mathbf{0} = (0\ 0\ 0)'$ a partir de la ecuación $\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2$ de varias formas:

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 = 0\begin{pmatrix}2\\1\\3\end{pmatrix} + 0\begin{pmatrix}-6\\-3\\-9\end{pmatrix} = 3\begin{pmatrix}2\\1\\3\end{pmatrix} + 1\begin{pmatrix}-6\\-3\\-9\end{pmatrix} = \begin{pmatrix}0\\0\\0\end{pmatrix},$$

y por ende la matriz $\mathbf{X} = [\mathbf{x_1} \quad \mathbf{x_2}]$ es linealmente dependiente.

Independencia de Los Vectores de la Base Canónica Consideremos los vectores $\{\mathbf{e_1}, \ldots, \mathbf{e_K}\}$ donde cada $\mathbf{e_k}$ tiene cada elemento igual a 0, con la excepción de un valor de 1 en elemento k. Estos vectores se conocen como la "base canónica" de $\mathbb{R}^K$, ya que con una combinación linea de estos vectores de la forma:

$$\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \ldots + \alpha_K \mathbf{x}_K,$$

y valores judicialmente elegidos para cada α_k, se puede producir cualquier vector posible $\in \mathbb{R}^K$.

Se puede demostrar que los K vectores de la base canónica son linealmente independientes en $\mathbb{R}^K$. Replicando Stachurski (2016, p. 18), definimos a los vectores canónicas $\{\mathbf{e_1}, \ldots, \mathbf{e_K}\}$, y una serie de coeficientes $\alpha_1, \ldots, \alpha_K$, de tal forma que $\sum_{k=i}^{K} \alpha_k \mathbf{e_k} = \mathbf{0}$. Ahora, esto implica:

$$\alpha_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \alpha_K \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_K \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{2.19}$$

y $\alpha_k = 0$ para cada k, que es la única solución, cumpliendo con la definición de independencia lineal en la ecuación 2.18.

Independencia y Unicidad Anteriormente, vimos que la independencia implica por definición una sola solución para una ecuación lineal que forma el vector nulo (ecuación 2.18). En realidad, esta condición es mucho más generalizable y la independencia y unicidad son estrechemente vínculados. La independencia implica que la solución para *cualquier* variable $\mathbf{y}$ que es la suma de una ecuación de la forma $\sum_{k=1}^{K} \alpha_k \mathbf{x}_k$ es única (si la solución existe).

Específicamente, consideramos una colección de vectores en $\mathbb{R}^N$, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$. Se puede demostrar que es equivalente decir: (a) $\mathbf{X}$ es linealmente independiente, y (b) que para cada $\mathbf{y} \in \mathbb{R}^N$, existe como máximo un grupo de escalares $\alpha_1, \ldots, \alpha_K$ de tal forma que:

$$\mathbf{y} = \alpha_1 \mathbf{x}_1 + \cdots + \mathbf{x}_K. \tag{2.20}$$

La demostración formal de esta equivalencia se deja como un ejercicio.

El Rango de una Matriz El rango de una matriz $\mathbf{A}$ de $N \times K$ (con $K \leq N$) es la cantidad de columnas linealmente independientes, y lo escribimos aquí como $rango(\mathbf{A})$. Se dice que $\mathbf{A}$ tiene rango completo si $rango(\mathbf{A}) = K$. Una matriz cuadrada de $K \times K$ es conocida como una matriz no-singular si tiene rango completo. De otra forma, se conoce como una matriz singular, que implicar que por lo menos dos de las columnas son linealmente dependientes.

2.4.2 La Relación entre Independencia e Invertibilidad

Cuando definimos el inverso de las matrices en la sección 2.3, nunca definimos los requisitos precisos para saber si una matriz es invertible. Ahora con la definición de independencia y el rango de una matriz, tenemos todas las detalles necesarias para introducir los requisitos de invertibilidad para cualquier matriz cuadrada $\mathbf{A}$.

Una matriz $\mathbf{A}$ de $K \times K$ es invertible si la matriz es linealmente independiente. Como hemos viston en las secciones anteriores, independencia lineal implica varias cosas:

1. La ecuación $\mathbf{Ax} = \mathbf{0}$ tiene una sola solución, de $\mathbf{x} = \mathbf{0}$
2. La ecuación $\mathbf{Ax} = \mathbf{b}$ tiene una sola solución para cada $\mathbf{b} \in \mathbb{R}^K$
3. La matriz tiene rango completo, $rango(\mathbf{A}) = k$
4. $|\mathbf{A}| \neq 0$
5. $\mathbf{A}$ es invertible (no-singular)
6. Existe una matriz $\mathbf{B}$ de $K \times K$ de tal forma que $\mathbf{AB} = \mathbf{I_K} = \mathbf{BA}$.
7. La traspuesta $\mathbf{A}'$ es invertible

Estos hechos son parte del teorema de inversión matricial. Y cada hecho es equivalente—si uno cumple, todos cumplen, y si uno no cumple, ninguno cumple. Por lo tanto, mostrar que uno de estos hechos cumple basta para mostrar que la matriz es invertible.

2.4.3 Ortogonalidad de Vectores

La ortogonalidad es un concepto clave al momento de considerar relaciones entre variables, vectores, y matrices. Si $\mathbf{x}$ y $\mathbf{u}$ son dos vectores en $\mathbb{R}^N$, decimos que son vectores ortogonales si:

$$\langle \mathbf{x}, \mathbf{u} \rangle = 0 \tag{2.21}$$

donde la notación aquí sigue la definición en la ecuación 2.1. Esto también se puede escribir de la forma $\mathbf{x} \perp \mathbf{u}$. En dos dimensiones, la ortogonalidad implica que dos vectores son perpendiculares, o que cruzan para formar una intersección con angúlos internos de 90 grados. Volveremos a la ortogonalidad de vectores en mucha detalle cuando nos encontramos con la regresión lineal más tarde en el curso.

Sección 3

Un Repaso de Herramientas Probabilísticas

Nota de Lectura: Se sugiere leer capítulo 1 de Goldberger (1991) para una introducción interesante. Una buena cobertura de todas las materiales presentado en esta sección está disponible en Casella and Berger (2002) capítulos 1-3. Sin embargo existen otros libros de probabilidad que contienen presentaciones apropiadas, como DeGroot and Schervish (2012). Un alternativa avanzada es Stachurski (2016), capítulos 8–10.

3.1 Elementos Básicos de Probabilidad

3.1.1 Una Introducción a la Probabilidad

La probabilidad es el estudio del certidumbre que se puede asociar con eventos futuros inciertos. Hay varias concepciones de probabilidad, incluyendo una interpretación frecuentista (que considera la relativa frecuencia de distinto eventos para definir sus probabilidades), la interpretación clásica que parte de la base de iqualdad de probabilidad de eventos, para así definir probabilidades iguales, y una interpretación subjetiva, que considera la probabilidad que una persona (en particular) asigna a los eventos. En estos apuntes no examinaremos la historia o filosofía detras de la probabilidad, pero existen muchos libros de interés si le gustaría profundizar más en este tema. Una opción para partir es capítulo 1 de DeGroot and Schervish (2012) y sus referencias.

La teoría de probabilidad es la base de toda estadística, y por ende fundamental para nuestro estudio de econometría. Y la base de la probabilidad es el teoría de conjuntos. Partimos en esta sección introduciendo la noción del teoría de conjuntos y otros aspectos funamentales de probabilidad, antes de desarrollar algunas herramientas que serán fundamental en nuestros modelos econométricos, como los estimadores, intervalos de confianza, y contrastes de hipótesis.

Algunas Definiciones Preliminares

Para poder introducir la notación básica del teoría de conjuntos, primero definimos la idea de un **experimento** y un **evento**. Un experimento se define como cualquier proceso, verdadero o hipotético, en que se sabe con anterioridad todos los resultados potenciales (definición 1.3.1 de DeGroot and Schervish (2012)). Un experimento es aleatorio (o *estocástico*) si hay varios posibles resultados, y *deterministico* si solo hay un resultado potencial. Un evento refiere a un conjunto bien definido de los resultados potenciales del experimento.

Esta definición de un 'experimento' y un 'evento' es muy amplio, y nos permite examinar muchos procesos de interés (con incertidumbre) en la econometría. Aunque un experimento puede ser tan simple como tirar un dado, la definición también incluye procesos como observar una persona para ver la educación total acumulada durante su vida, o su eventual salario laboral.

El Teoría de Conjuntos

El teoría de conjuntos sirve para una base de la teoría de probabilidad. El teoría de conjuntos es una herramienta de lógica que clasifica a elementos como perteneciente o no perteneciente a espacios—o conjuntos—particulares.

Volviendo a la idea de un experimento definido anteriormente, referimos al conjunto de todos los resultados potenciales de un experimento como el espacio muestral, o S. Por ejemplo, si el experimento consiste simplemente en lanzar un dado, definimos al espacio muestral $S \in \{1, 2, \ldots, 6\}$, y si el experimento consiste en observar el nivel de educación alcanzada en la vida de una persona elegido al azar de una población específica, el espacio muestral consistiría de $S \in \{\text{sin educación, básica, secundaria, terciaria}\}$.

Un evento, como lo definimos anteriormente, es cualquier resultado potencial del experimento de interés, y por ende es un sub-conjunto de S. Si definimos un evento A, decimos que A ocurre cuando el resultado aleatorio del experimento es contenido en el sub-conjunto A. Definimos contención de la siguiente forma:

$$A \subset B \Leftrightarrow x \in A \Rightarrow x \in B, \tag{3.1}$$

y definimos igualdad como:

$$A = B \Leftrightarrow A \subset B \text{ y } B \subset A. \tag{3.2}$$

Por último, definimos el conjunto vacío $\emptyset$ como el conjunto que consiste de ningún elemento. Trivialmente, $\emptyset \in A$ donde A es cualquier evento. Dado que el conjunto vacío no contiene ningún elemento, es cierto que todos los elementos que pertenecen a $\emptyset$ también pertenecen a A.

Los operaciones básicas de conjuntos se resumen a continuación (para dos conjuntos arbitrarias A y B), donde la nomenclatura sigue Casella and Berger (2002, §1.1).

Operaciones Básicas de Conjuntos

1. Unión: Los elementos que pertenecen a A o a B. Se define $A \cup B = \{x : x \in A \text{ o } x \in B\}$
2. Intersección: Los elementos que pertenecen a A y a B. Se define: $A \cap B = \{x : x \in A \text{ y } x \in B\}$
3. Complemento: El complemento de A son todos los elementos que no pertenecen a A. Se define $A^c = \{x : x \notin A\}$.

Se dice que dos eventos son eventos disjuntos (o mutuamente excluyentes) si $A \cap B = \emptyset$.

La Teoría de Probabilidad

La teoría de probabilidad intenta asignar a cada evento en un experimento un valor para capturar la frecuencia del evento si el experimento fue repetido muchas veces. La teoría de probabilidad define una serie de axiomas que caracterizan al número que captura esta frecuencia. En lo que sigue consideramos un evento A en el espacio S, y definimos a $P(A)$ como la probabilidad que el evento A ocurre. Hay varias consideraciones técnicas acerca de exáctamente cuáles son los subconjuntos de S sobre cual se define una probabilidad, pero no las examinamos en este curso. Detalles más comprensivos están disponibles en Stachurski (2016) o Casella and Berger (2002, §1.2).

Los axiomas de probabilidad define las propiedades para una función de probabilidad. Estos axiomas son:

Los Axiomas de Probabilidad

1. Para cada evento A, $P(A) \geq 0$.
2. $P(S) = 1$
3. Para cada secuencia infinita de eventos dijuntos $A_1, A_2, \ldots$, $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

Los tres axiomas refieren al evento arbitrario A, y el espacio muestral S. Entonces, el segundo axioma declara que si un evento es cierto (el resultado de un experimento por definición siempre cae en el espacio muestral), su probabilidad es igual a 1. El último axioma sigue la definición de unión (y eventos disjuntos) descritas anteriormente, pero aquí la unión es sobre una cantidad infinita de eventos.

A partir de las axiomas de probabilidad, se puede derivar una serie de otras propiedades de funciones de probabilidad. Resumimos algunas de estas propiedades aquí para una función de probabilidad P, y dos conjuntos A y B. Dejamos como un ejercicio la demostración de estas propiedades.

Otras Propiedades de Probabilidad

1. $P(\emptyset) = 0$
2. $P(A) \leq 1$
3. $P(A^c) = 1 - P(A)$
4. Si $A \subset B$, entonces $P(A) \leq P(B)$.

Notemos que con estos axiomas y propiedades adicionales, no hay una restricción acerca de exáctamente qué función de probabilidad P se debe elegir, sino se define una serie de condiciones básicas para tener un P válido. Para cualquier espacio muestral existen muchas posibles funciones de probabilidad que cumplen con los axiomas 1-3. La definición de una función de probabilidad específica depende de la naturaleza del experimento bajo estudio, y una serie de otras propiedades que definimos a continuación.

Sin embargo, podemos definir una serie de condiciones que—si cumplen—aseguran que la función de probabilidad siempre satisface los axiomas 1-3. Consideramos un experimento con una cantidad finita de resultados potenciales. Esto implica que el espacio muestral S contiene una cantidad finita de puntos $s_1, \ldots, s_n$. En este tipo de experimento, para definir una función de probabilidad necesitamos asignar una probabilidad p_i a cada punto $s_i \in S$. Para asegurar que los axiomas de probabildad se satisfagan, los valores para $p_1, \ldots, p_n$ deben cumplir con las sugientes dos condiciones:

$$p_i \geq 0 \quad \text{para } i = 1, \ldots, n, \text{ y} \tag{3.3}$$

$$\sum_{i=1}^{n} p_i = 1. \tag{3.4}$$

Una demostración formal de este resultado está disponible en Casella and Berger (2002, Teorema 1.2.6).

Probabilidad Condicional

En nuestro análisis econométrico, a menudo cuando enfrentamos incertidumbre, contaremos con algo de información preliminar. Por ejemplo, nos podría intersar la probabilidad de que una persona tengo empleo condicional en el hecho de que la persona tiene una educación secundaria. En este caso, más que pensar en probabilidades incondicionales, vamos a querer hacer análisis condicional.

Al contar con más información acerca de un experimento, tenemos que considerar un espacio muestral alterado. Para fijar ideas, imaginamos que nos intersa un evento A en un experimento con espacio muestral S. Ahora, imaginamos que aprendimos que ocurrió otro evento B, que reduce el espacio muestral en alguna forma. Ahora, dado que sabemos que B ocurrió, en vez de estar tratando de encontrar $P(A)$—la probabilidad incondicional de A—queremos encontrar $P(A|B)$, que es la probabildad condicioanl del evento A, dado el evento B. Computamos esta probabilidad condicional de la siguiente forma:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{3.5}$$

Notemos dos cosas de esta ecuación. Primero, si la probabilidad que $P(B) = 0$, la probabilidad condicional no existe. Intuitivamente, esto no es muy problemático, dado que estamos condicionando en el hecho de que B *ocurrió*, que implica que *ex-ante* hubo una probabilidad no nulo que

esto podría pasar. Y segundo, notamos que estamos buscando la unión de A y B. Esta probabilidad condicional pregunta cuál es la probabilidad de que A y B ocurrieron juntos, y lo divide por la probabilidad total de B, que de alguna forma es el espacio muestral reducido una vez que se sabe que B ha ocurrido.

Podemos invertir la ecuación 3.5 para calcular probabilidades de intersecciones entre eventos:

$$P(A \cap B) = P(B)P(A|B). \tag{3.6}$$

Este cálculo es particularmente útil en casos cuando la probabilidad condicional es fácil de calcular o asignar. Y, por simetria, si $P(A) > 0$, tenemos:

$$P(A \cap B) = P(A)P(B|A). \tag{3.7}$$

Esto se conoce como la ley de múltiplicación para probabilidades condicionales.

Una otra ley importante que se basa en la probabilidad condicional es la ley de probabilidad total. Para introducir la ley de probabilidad total, necesitamos definir la idea de una partición. Una partición es una separación de un estado muestral en una serie de áreas mutuamente excluyentes, que cubren todo el espacio. Formalmente (DeGroot and Schervish, 2012, Definición 2.1.2) consideramos k eventos de un experimento (con espacio muestral S) llamados $B_1, \ldots, B_k$, de tal forma que $B_1, \ldots, B_k$ son disjuntos, y $\bigcup_{i=1}^{k} B_i = S$. Entonces, se dice que los eventos $B_1, \ldots, B_k$ forman una partición de S.

La *ley de probabilidad total* parte con una partición de eventos $B_1, \ldots, B_k$ del espacio S. Asumimos que $P(B_j) > 0$ para cada j. Entonces, para cada evento A en S:

$$P(A) = \sum_{j=1}^{k} P(A|B_j)P(B_j). \tag{3.8}$$

Examinamos algunos ejemplos como un ejercicios en clase.

Si combinamos los resultados anteriores (ecuaciones 3.6-3.8), llegamos al famoso Teorema de Bayes. Notemos que el lado izquierdo de ecucación 3.6 y 3.7 son idénticos, de tal forma que:

$$\begin{aligned} P(A)P(B|A) &= P(B)P(A|B) \\ P(B|A) &= \frac{P(B)P(A|B)}{P(A)} \end{aligned} \tag{3.9}$$

La ecuación 3.9 es la versión más simple del teorema de Bayes, que vincula la probabilidad condicional de un evento a su probabilidad incondicional, y información preliminar acerca de la probabilidad de la condición (revisamos algunos ejercicios aplicados). En análisis Bayesiano, la probabilidad de $B|A$ se conoce como la probabilidad posterior, ya que es la probabilidad de ocurrencia de B una vez que sabemos que A ocurrió. La probabilidad incondicional, o $P(B)$, se conoce como la probabilidad *a priori* en la ausencia de saber algo de A.

Sin embargo, también hay una versión del Teorema de Bayes para todos los eventos en una partición, y se llega a esta versión del teorema reemplazando el denominador de 3.9 por la ley de probabilidad total. Consideramos la partición $B_1, \ldots, B_k$ con $P(B_j) > 0$ para cada $j \in \{1, \ldots, k\}$, y el evento A con $P(A) > 0$. Entonces, para cada $i = 1, \ldots, k$, el Teorema de Bayes dice:

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{k} P(A|B_j)P(B_j)}. \tag{3.10}$$

Independencia

A veces, al haber aprendido que un evento A ha ocurrido no nos hace cambiar nuestra creencia acerca de la probabilidad de otro evento B. En estos casos decimos que A y B son eventos independientes. Trivialmente, consideramos un caso cuando lanzamos dos dados tradicionales, una tras otro. Al observar un cierto resultado tras lanzar el primer dado, esto no nos entrega más información relevante acerca del resultados probable del segundo lanzamiento. En este caso, utilizando la notación anterior, tenemos que $P(B|A) = P(A)$.

Formalmente, dos eventos A y B son independientes si:

$$P(A \cap B) = P(A)P(B). \tag{3.11}$$

Para ver porqué, simplemente tenemos que volver a la ecuación 3.6 (o 3.7) y reemplazar la probabilidad condicional $P(B|A)$ con su probabilidad incondicional $P(B)$, dado que condicionar en A no cambia la probabilidad de ocurrencia. De la definición aquí, y las ecuaciones 3.6-3.7 se ve que dos eventos A y B son independientes si y solo si (*sii*) $P(A|B) = A$, y $P(B|A) = B$. Esta fórmula de independencia también se extiende para independencia múltiple. En particular, para tres eventos A, B y C, decimos que son mutuamente independientes si:

$$P(A \cap B \cap C) = P(A)P(B)P(C). \tag{3.12}$$

3.1.2 Variables Aleatorias

Una variable aleatoria toma valores que son determinandos por un proceso aleatorio, que muchas veces es un proceso natural estocástico.[1] Una variable aleatoria es una representación númerica de los resultados potenciales de un experimento, y formalmente es una función que mapea los resultados potenciales en el espacio muestral S a un número real $X \in \mathbb{R}$. La variable aleatoria captura la información contenido en todo el espacio muestral del experimento en una cantidad numérica—una manera conveniente para seguir con análisis posterior.

[1] Un proceso estocástico simplemente refiere a un proceso cuyo resultado no es conocido con certeza *ex-ante*. Es el opuesto a un proceso deterministico, que es un proceso que siempre produce el mismo resultado dado una condición inicial específica. Por ejemplo, lanzar una moneda es un proceso estocástico, y sumar dos números específicos es un proceso deterministo.

El valor de un experimento aleatorio es, por definición, no conocido antes de realizar un experimento, pero dado que S es conocido, todos los valores potenciales de la variable aleatoria son conocidas. Posterior al experimento se observa el valor que la variable tomó en esta realización particular. A menudo, se refiere a una variable aleatoria por una letra mayúscula: X, y realizaciones específicas de la variable por una letra minúscula; x. Entonces, cuando se escribe $X = x$, o $X = 500$ (o la probabilidad $\mathrm{P}(X = x)$, o $\mathrm{P}(X = 500)$) es una combinación de la variable aleatoria con una realización específica de la variable aleatoria (x, o 500). Aquí, hemos saltado de hablar de la probabilidad de observer una realización particular de una *variable aleatoria*, pero hasta ahora solo sabemos que las funciones de probabilidad cumplen con los axiomas de probabilidad cuando consideramos los eventos en el espacio muestral original. Explícitamente, cuando hablamos de la probabilidad de observar un resultado particular, x_i, de una variable aleatoria X (con rango $\mathcal{X} = \{x_1, \ldots, x_k\}$), referimos a la función P_X:

$$P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\}).$$

Aquí, observamos el resultado x_i, *sii* el resultado del experimento inicial era $s_j \in S$. Dada que P_X satisface los tres axiomas de probabilidad[2], P_X es una función de probabilidad, y podemos simplemente escribir $P()$ en vez de $P_X()$.

La definición específica de una variable aleatoria depende del experimento de interés, y el resultado particular de interés. Por ejemplo, si un experimento consiste en lanzar una moneda 25 veces, una variable aleatoria podría ser la cantidad total de 'caras' que salen, o podría ser la cantidad de lanzamientos hasta que salga una cara, o una variable binaria que toma el valor 1 si la cantidad de caras es mayor a 12, etc. Aquí se puede ser como una variable aleatoria puede resumir mucha información del espacio muestral subyacente. Consideramos el experimento de lanzar una moneda 25 veces y observar la cantidad de caras. El espacio muestral S consiste de 2^{25} elementos: cada uno un vector ordenado de 'caras' y 'sellos' de tamaño 25. Sin embargo, al definir una variable aleatoria X = Cantidad total de caras, hemos reducido el espacio muestral a una serie de números enteros con rango $\mathcal{X} = \{0, 2, \ldots, 25\}$.

Las variables aleatorias pueden ser discretas, cuando toman una cantidad finita de posibles valores, o continuas, cuando pueden tomar infinitos posibles valores. En el caso de variables discretas, un caso especial es cuando toman solo dos valores (variables binarias) como sexo, o si una persona está empleado o no, o pueden tomar más valores, por ejemplo edad en años. A pesar de tomar infinitos valores posibles, las variables continuas pueden ser limitado en algún sentido, como por ejemplo el peso de un objeto, que no puede tomar valores negativos.

[2]Para ver esto, consideramos los tres axiomas. (1) Para cualquier evento A, $P_X(A) = P(\bigcup_{x_i \in A}\{s_j \in S : X(s_j) = x_i\}) \geq 0$ dado que $P()$ es una función de probabilidad; (2) $P_X(\mathcal{X}) = P(\bigcup_{i=1}^{k}\{s_j \in S : X(s_j) = x_i\}) = P(S) = 1$, y (3) si $A_1, A_2, \ldots$ son eventos disjuntos, $P_X(\bigcup_{m=1}^{\infty} A_m) = P(\bigcup_{m=1}^{\infty}\{\bigcup_{x_i \in A_m}\{s_j \in S : X(s_j) = x_i\}\}) = \sum_{m=1}^{\infty} P\{\bigcup_{x_i \in A_m}\{s_j \in S : X(s_j) = x_i\} = \sum_{m=1}^{\infty} P_X(A_m)$. Dado que los tres axiomas cumplen con P_X, se concluye que P_X es una función de probabilidad válida.

3.1.3 Esperanza, Momentos y Medidas Simples de Asociación

Existen una serie de valores particularmente interesante para caracterizar una variable aleatoria. Una de ellos es la esperanza, o expectativa, de la variable aleatoria X. La expectativa es una medida del valor promedio de la variable aleatoria, o al valor esperado de un valor típico de la variable. Esta promedio considera todas los valores posibles de X, ponderando por la frecuencia de ocurrencia del valor. La esperanza de una variable aleatoria X se denota $E(X)$, donde:

$$E(X) = \sum_{j=1}^{J} x_j P(X = x_j).$$

Aquí X tiene una cantidad finita de valores potenciales, y por lo tanto se puede sumar sobre todos los valores x_j. En el caso de una variable continua, definimos a la esperanza como:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{3.13}$$

donde aquí $f(x)$ refiere a la masa de probabilidad concentrada en un rango de x. Como veremos con más detalle en la sección 3.1.4, esta función $f()$ es conocida como la función de densidad de probabilidad.

Algunas Propiedades de la Esperanza

1. Si $Y = aX + b$ donde a y b son constantes, entonces $E[Y] = aE(X) + b$
2. Si $X = c$ con probabilidad 1, entonces $E(X) = c$
3. Si $X_1, \ldots, X_n$ son n variables aleatorias, entonces, $E(X_1 + \ldots + X_k) = E(X_1) + \ldots + E(X_k)$
4. Si $X_1, \ldots, X_n$ son n variables aleatorias, entonces, $E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i)$
5. Para funciones de variables aleatorias: $E[g(X)] = \sum_{j=1}^{k} g(x_j)P(X = x_j)$. Por lo general, $E[g(X)] \neq g(E(X))$. Una excepción son la clase de funciones lineales.

Las demostraciones de éstas propiedades se encuentran en DeGroot and Schervish (2012, §4.2).

Varianza

La esperanza proporciona *un* valor para caracterizar una variable aleatoria. Sin embargo, esconde mucha información acerca de la variable aleatoria. Si queremos saber algo acerca del nivel de variablidad de una variable aleatoria alrededor de su expectativa, un otro valor característico importante es la varianza. La varianza de una variable aleatoria X se define como:

$$Var(X) = \sigma^2 = E[X - E(X)]^2,$$

o la distancia (en promedio) de todas las realizaciones de la variable aleatoria X de su valor esperado, al cuadardo. La desviación estandar también es una medida de dispersión, pero medido

en la misma unidad que la variable original:

$$sd(X) = \sigma = \sqrt{(Var(X))}$$

Una otra manera de representar la fórmula de varianza que utilizaremos a veces en estos apuntes es $Var(X) = E[X^2] - [E(X)]^2$. Para ver por qué notamos:

$$\begin{aligned} Var(X) &= E[X - E(X)]^2 \\ &= E(X^2 - 2E(X)X + E[X]^2) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2. \end{aligned} \tag{3.14}$$

Resumimos algunas propiedades de la varianza a continuación, para una variable aleatoria X, y dos constantes a y b.

Propiedades de la Varianza

1. Para cada X, $\text{Var}(X) \geq 0$
2. $\text{Var}(X) = 0$ sii existe una constante c tal que $Pr(X = c) = 1$
3. Si $Y = aX + b$ donde a y b son constantes, entonces $Var[Y] = a^2 Var(X)$
4. Si $X_1, \ldots, X_n$ son n variables aleatorias *independientes*, entonces, $Var(X_1 + \ldots + X_k) = Var(X_1) + \ldots + Var(X_k)$

Nuevamente, las demostraciones formales de estas propiedades se encuentran en DeGroot and Schervish (2012), sección 4.3.

Los Momentos de una Variable Aleatoria

Para cada variable aleatoria X y cada número entero k, la esperanza $E(X^k)$ se conoce como el momento k-ésimo de X. La esperanza, como lo definimos anteriormente, es el primer momento de la distribución de X. El momento existe si $E(|X^k|) < \infty$. Ahora, supongamos que X es una variable aleatoria, y definimos a $E(X) = \mu$. Para cada número entero positivo k la esperanza $E[(X - \mu)^k]$ es conocido como el k-ésimo momento central de X. Siguendo esta notación, se observa que la varianza es el segundo momento central de X.

Los momentos sirven como medidas de distintas características de una variable aleatoria. Varios momentos o momentos centrales son cantidades muy conocidas, como por ejemplo la expectativa y la varianza. El tercer y cuarto momento central son conocidos como asimetría estadística, y kurtosis respectivamente. Como veremos más adelante, a veces vamos a definir nuestros estimadores en base de los momentos de la distribución observada de datos. Lo llamaremos "metodo de los momentos" o "metodo de los momentos generalizados" (Sección 3.3.3 de los apuntes.).

Asociación: Covarianza, Correlación

Cuando trabajamos con más de una variable aleatoria, con frecuencia vamos a estar interesado en saber cómo se mueven en conjunto. La covarianza y la correlación son dos medidas simples de la asociación entre variables. Entre otras cosas, estas estadísticas son útiles para saber qué hace una variable aleatoria Y, si otra variable aleatoria X aumenta. Definimos la covarianza y la correlación ente dos variables X y Y a continuación:

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E[XY] - E[X]E[Y]$$

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{sd(X) \cdot sd(Y)}.$$

A veces la covarianza entre X y Y se denota σ_{xy}, y la correlación se denota ρ_{xy}.

La covarianza mide cómo las dos variables se mueven en conjunto, y está medida en las unidades de las dos variables de interés. Por ejemplo, si una variable X es medido en tiempo, y otro Y en peso, la covarianza estará expresado en tiempo×peso. A veces esto será una medida útil, pero muchas veces sería más útil tener una medida estandarizada. La versión estandarizada es la correlación, que tiene un rango de $[-1, 1]$, con -1 implicando una correlación negativa perfecta y 1 una correlación positiva perfecta. Si X e Y son independientes, $\text{Corr}(X, Y) = 0$ (pero el revés no es necesariamente cierto)[3].

Esperanzas Condicionales y Esperanzas Iteradas

Cuando consideramos múltiples variables aleatorias, también podemos considerar algunas propiedades de una variable *condicional* en ciertas carecterísticas de la otra variable. Un ejemplo de este tipo proceso sería considerar el salario promedio de personas con educación secundaria. En términos generales, si dos variables aleatorias X e Y no son independientes, conocer algo acerca de X puede aportar información acerca de Y.

Sean X e Y dos v.a. La esperanza condicional de Y dado que $X = x$ se denota $E(Y|X = x)$, o simplemente como $E(Y|x)$. Para una variable Y contínua, se define la esperanza condicional de la siguiente forma:

$$E[Y|X = x] = \sum_{j=1}^{m} y_j P_{Y|X}(Y = y_j|X = x).$$

Para una variable continua, definimos,

$$E[Y|x] = \int_{-\infty}^{\infty} y g_2(y|x) dy$$

donde utilizamos la misma idea de la ecuación 3.13 de una función que define la masa de probabil-

[3]Para ver un caso simple en que dos variables no son independientes, pero sí tienen una correlación igual a cero, consideremos las variables Y y X, donde $Y = X^2$. En esta función cuadrática, Y claramente depende de X, pero la correlación entre las dos variables es igual a cero.

idad localizada en cada punto de y (para un valor dado de $X = x$), que se conoce como la función de densidad condicional. Definamos la función de densidad condicional de forma completa en la sección 3.1.4.

La Ley de las Esperanzas Iteradas (Ley de Probabilidad Total) Existe una clara relación entre la esperanza global y la esperanza condicional mediante la ley de las esperanzas iteradas, o la ley de probabildad total. La ley de esperanzas iteradas dice que la esperanza global se puede calcular tomando el promedio de todos los promedios condicionales! En notación, tenemos que:

$$E[Y] = E_X[E[Y|X]],$$

donde destacamos que la primera expectativa en el lado derecho es sobre X. Las otras dos expectatvias son sobre Y, y la segunda expectativa en el lado derecho es una expectativa condicional. Con un X discreto tenemos, entonces:

$$E[Y] = \sum_{x_i} E[Y|X = x_i] \cdot Pr(X = x_i).$$

Esto implica que si ponderamos a las expectivas condicionales de Y sobre X para *todas* las valores posibles de X, volvemos a la expectativa global de Y. Para un ejemplo sencillo, si estamos interesados en el salario (Y) y la educación de las personas (X), una manera de calcular el salario promedio sería calcular el promedio de salario condicional en no tener educación; el salario promedio condicional en tener educación básica; en tener educación secundaria; y en tener educación terciaria, y después ponderamos todos estos promedios por la probabilidad de tener cada nivel de educación para llegar al salario promedio simple. Por supuesto en realidad, sería más directo simplemente calcular el salario promedio, pero la ley de esperanzas iteradas es un resultado fundamental que utilizaremos al momento de estar trabajando con análsis multivariada en econometría.[4] Revisaremos en ejemplo (sencillo) trabajado en clases.

3.1.4 Distribuciones

Una distribución para una variable X asigna probabilidades al evento que X cae en subespacios en $\mathbb{R}$. Cuando definimos probabilidades de realizaciones de variables aleatorias anteriormente, definimos la probabilidad de observar un resultado particular Las distribuciones consideran probabilidades para un conjunto de valores de X.

La Funcion de Densidad Acumulada

Una clase importante de distribuciones de probabilidad son las funciones de densidad acumulada. La función de densidad acumulada (fda), denotada F, de una variable aleatoria X es la

[4]Una demostración simple de la ley de esperanzas iteradas está disponible en DeGroot and Schervish (2012, Theorem 2.1.4). Otra demostración está disponible en Casella and Berger (2002, Theorem 4.4.2).

función:

$$F(x) = Pr(X \leq x) \quad \text{para toda } x$$

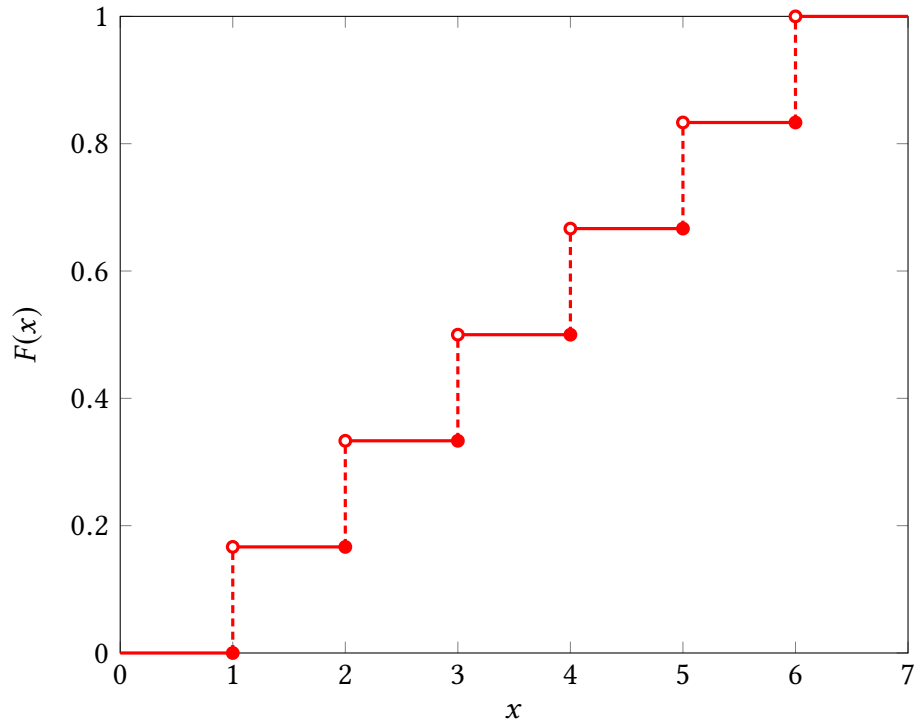
y esta función está definida así tanto para variables continuas como variables discretas. Una fda satisface las siguientes propiedades:

Propiedades de la Función de Densidad Acumulada

1. No decreciente: si $x_1 < x_2$, entonces $F(x_1) \leq F(x_2)$
2. Limites a $\pm\infty$: $\lim_{x\to-\infty} F(x) = 0$ y $\lim_{x\to\infty} F(x) = 1$
3. Continuidad desde la derecha: $\lim_{x\downarrow x_0} F(x) = F(x_0)$ en cada x_0

Consideramos la variable aleatoria que registra el valor que sale del lanzamiento de un dado. El ejemplo de la fda asociada con esta variable está presentada en la Figura 3.1. La probabildad de observar un valor de $x < 1$ es igual a 0. En la Figura 3.1 el eje horizontal está acotada entre 0 y 7, pero en realidad extiende entre $-\infty$ y ∞ (con $F(x) = 0$ cuando $x < 0$, y $F(x) = 1$ cuando $x > 6$). La probabildad acumulada de observar un número inferior a x salta a cada número entero, ya que la existe una probabildad positiva de que salga este número al lanzar el dado.

Figure 3.1: Función de Densidad Acumulada de Una Variable Discreta



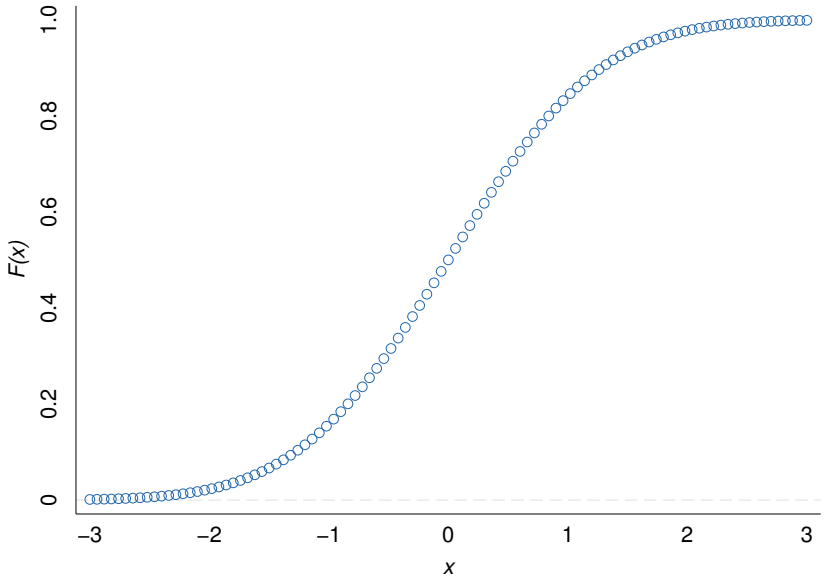
Para considerar el caso de una fda de una variable continua, consideramos el caso de la distribución normal, o distribución Gaussiana. La distribución normal es potencialmente la distribución más frecuentamente encontrada en la econometría, reflejando su frecuencia en variables observado en el mundo natural. La cdf de una variable normal o Gaussiana tiene la siguiente

distribución:

$$F(x) = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)}{2\sigma^2}\right) \mathrm{d}t \tag{3.15}$$

donde μ indica el promedio de la distribución, y σ su desviación estándar. Más adelante, vamos a llamar esta distribución como $\mathcal{N}(\mu, \sigma^2)$, y en el caso especial cuando $\mathcal{N}(0, 1)$, la distribución es conocida como la distribución normal estandarizada, y se denota Φ. Esta fda se grafica en la Figura 3.2. Nuevamente, aunque el eje horizontal está acotado (entre -3 y 3), su dominio es entre $-\infty$ y ∞, como estipulado en la propiedad 2 de la fda. Volvemos a examinar más distribuciones como un ejercicio computacional, y resumimos algunas propiedades básicas de distribuciones importantes en la Tabla 3.1.

Figure 3.2: La Distribución Normal Estandarizada – fda



Funciones de Densidad

Las funciones de densidad presentan las probabilidades puntuales de observar un resultado x en el rango de X. La manera que se describen estas funciones depende de si la variable subyacente X es discreta o continua. En el caso de una variable discreta, son conocidas como una función de probabilidadx, y en el caso de una variable continua, son conocidas como funciones de densidad de probabilidad. Examinamos ambos tipos de distribuciones (y variables) a continuación

Distribuciones Discretas Una variable aleatoria X tiene una *distribución discreta* si X sólo puede tomar un número finito de valores distintos $x_1, \ldots, x_k$, o una secuencia infinitia de valores distintos $x_1, x_2, \ldots$. Si una variable aleatoria X tiene una distribución discreta, se define la función de probabilidad, f, como la función que para cada número real x:

$$f(x) = P(X = x)$$

con el soporte $\{x : f(x) > 0\}$. Esta función $f(x)$ presenta la masa de probabildad asociado con cada punto x. Las propiedades de la función de probabilidad (y los axiomas de probabilidad) implican (i) que si x no es uno de los valores posibles de X entonces $f(x) = 0$. Y (ii) si la sequencia $x_1, x_2, \ldots$ contiene todos los valores posibles de X entonces $\sum_{i=1}^{\infty} f(x_i) = 1$.

Un ejemplo simple de una variable discreta y su función de probabilidad asociada es una variable aleatorio Z que solo toma dos valores: 0 y 1, con $P(Z = 1) = p$. Esta variable tiene una distribución Bernoulli con parametro p, que se caracteriza de la siguiente forma:

$$f(z; p) = \begin{cases} p & \text{if } z = 1 \\ 1 - p & \text{if } z = 0. \end{cases}$$

En esta función se escribe $f(z; p)$ para denotar que la función depende de la realización de la variable aleatoria z, pero también del parametro p. Con frecuencia omitimos los parametros de la definición de la función cuando es claro que refieren a parametros que dependen del contexto del experimento. Otro ejemplo de una función de probabilidad corresponde a la Distribución Uniforme en Números Enteros. Sean $a \leq b$ dos números enteros. Supongamos que es igualmente probable que el valor de la variable aleatoria X toma el número entero $a, \ldots, b$. Entonces decimos que X tiene una distribución uniforme en los número enteros $a, \ldots, b$, y se escribe la densidad:

$$f(x) = \begin{cases} \frac{1}{b-a+1} & \text{si } x = a, \ldots, b \\ 0 & \text{si no.} \end{cases}$$

En ambos casos, es fácil confirmar que las dos propiedades descritas anteriormente en (i) y (ii) cumplen con la función de probabilidad definida.

Distribuciones Continuas En el caso de una variable continua, la probabilidad de observar cualquier realización *particular* de X es igual a cero (ver Casella and Berger (2002, p. 35)), y por lo tanto tenemos que definir la distribución utilizando integrales en vez de probabilidades puntuales. En este caso, definimos a la función de densidad de probabilidad de una variable aleatoria X de tal forma que:

$$Pr(a \leq X \leq b) = \int_a^b f(x)dx$$

para cada intervalo cerrado $[a, b]$. Si una variable aleatoria X tiene una distribución continua, la función f se llama la función de densidad de probabilidad (fdp) de X, y el conjunto tiene el soporte $\{x : f(x) > 0\}$. En el caso de una variable continua, la fdp tiene que cumplir con las siguientes dos condiciones:

Propiedades de una Función de Densidad de Probabilidad

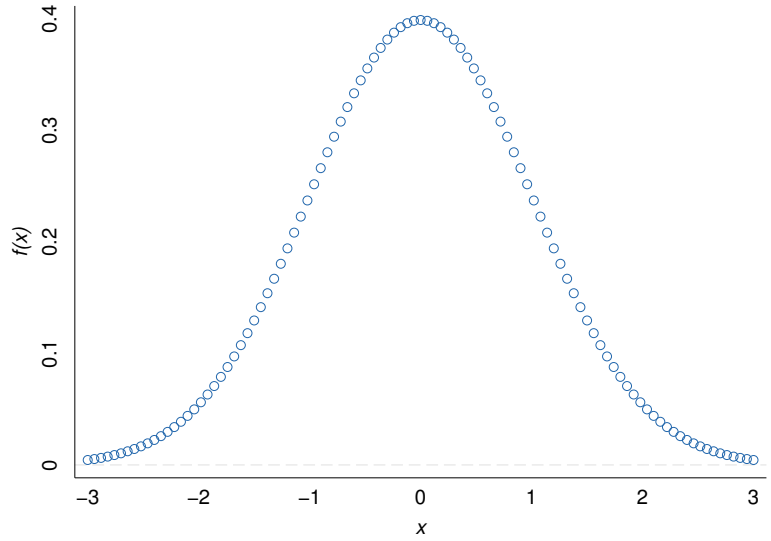
1. $f(x) \geq 0$ para toda x,
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

La función de densidad de probabilidad de una variable normal $\mathcal{N}(\mu, \sigma^2)$ está presentada en

la Figura 3.3. En este caso, $\mu = 0$ y $\sigma = 1$, que es la densidad de la normal estandarizada, denotada ϕ. Para cualquier valor μ y σ, esta fdp se escribe:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)}{2\sigma^2}\right). \tag{3.16}$$

Figure 3.3: La Función de Densidad de Probabilidad de la Normal Estandarizada



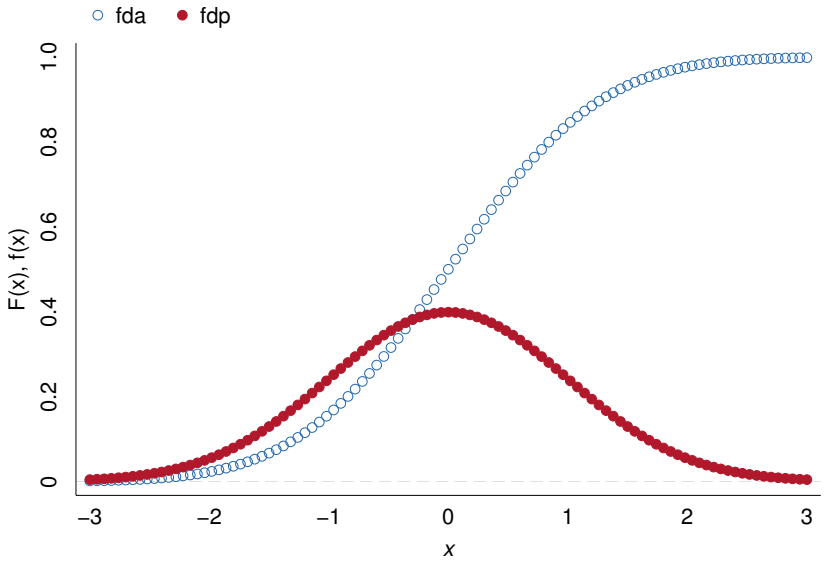
La Relación Entre la fda y la fdp Hay un vinculo muy estrecho y aparente entre las funciones de densidad acumulada, y las funciones de densidad de probabilidad (o funciones de densidad). Dado que las fdp demuestran la masa de probabilidad en puntos de x, y las fda la probabilidad acumulada *hasta* cada punto x, se llega a la fda integrando sobre la fdp, y se llega la fdp diferenciando la fda:

$$\frac{dF(x)}{dx} = f(x) \tag{3.17}$$

Para ver un caso particular, consideramos la fda y la fdp de la distribución normal estandardizada. Se presentan las dos funciones en la Figura 3.4. La utilidad de esta relación es aparente cuando se quiere responder a preguntas como ¿cuál es la probabildad de observar un valor de $X < 1$?; ¿cuál es la probabildiad de observar un valor de $X \geq 2$?; o ¿cuál es la probabilidad de observar un valor $-1 < X \leq 1$? Estas preguntas refieren a la masa de probabilidad baja la función de densidad de probabilidad (refiere a la Figura 3.5 para la visualización gráfica de la tercera pregunta). Sin embargo, la respuesta se obtiene sencillamente a partir de función de densidad *acumulada*. En el primer caso, o más generalmente casos cuando se quiere calcular $P(X < a)$ para algún escalar a, simplemente buscamos $F(a)$, que es justamente toda la masa de probabilidad inferior a a. En el segundo caso, o de nuevo, más generalmente casos cuando se quiere calcular $P(X > b)$ para un valor escalar b, se requiere calcular $(1 - F(b))$, donde utilizamos el hecho que $\lim_{x\to\infty} F(x) = 1$. Por último, consideramos un caso de querer calcular $P(a < X \leq b)$ para cualquier dos valores escalares a y b con $a < b$. Aquí, calculamos la probabilidad como:

$F(b) - F(a)$. Notamos de la figura 3.5 (con $b = 1$ y $a = -1$), que aquí estamos simplemente calculando toda la masa acumulada hasta el valor $b = 1$, y después restando la masa de probabilidad entre $-\infty$ y b.

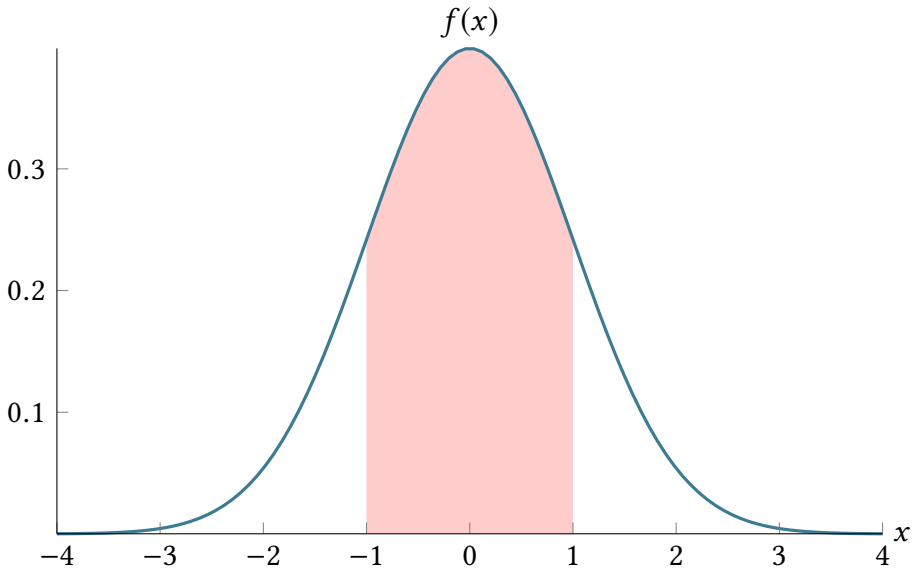
Figure 3.4: La fda y la fdp de la Distribución Normal Estandarizada



Para los cálculos anteriores, hemos considerado valores como $F(a)$, que son valores específicas de la función de densidad acumuladad. Aunque sabemos algunas condiciones básicas de las fda (como por ejemplo que son limitado por 0 y 1), no es necesariamente trivial calcular un valor como $F(a)$ para alguna función de densidad acumulada específica. Por ejemplo, en el caso de la fda de la distribución normal, no hay una solución de forma cerrada para calcular el valor de la función. Una manera de calcular el valor de la fda sería algún tipo de integración númerica sobre la fdp, por ejemplo utilizando la Cuadratura de Gauss. Por tipícamente, lenguajes de computación estadística proveen fórmulas fáciles para calcular valores específicas de funciones de densidad acumulada. Y en algunos casos muy comunes, por ejemplo la normal estandarizada, a menudo se ven tablas estadísticas que resumen los valores de la fda en muchos puntos de la distribución. Por ejemplo, en la Tabla 3.2 de estos apuntes, resumimos los valores de la fda de la normal estandarizada en muchos puntos (positivos). Dado que la fdp es una distribución simétrica, se puede inferir los valores de $F(x)$ para un $x < 0$. Esta tabla demuestra (por ejemplo) que 50% de la masa de probabilidad cae abajo del valor de 0.00, y 97.5% de la masa cae abajo de 1.96. Dejamos como ejercicios la revisión del cálculo de varias valores específicas, y una extensión a normales con media y desviación estándar distinto a 0 y 1.

La Función Cuantil Se utiliza la función de densidad acumulada de esta forma para responder a preguntas de la probabilidad de observar un valor de X superior, o inferior a cierto punto de interés. Pero también hay casos en que nos interesa invertir la pregunta, y saber el valor exacto de x donde la probabilidad de ocurrencia es igual a algún valor p. Por esto, utilizamos la

Figure 3.5: Area Bajo la Curva en una fdp



función cuantil. Sea X una variable aleatoria con un fda F. Para cada p estrictamente entre 0 y 1, definamos $F^{-1}(p)$ como el menor valor de x tal que $F(x) \geq p$. Entonces, $F^{-1}(p)$ es el cuantil p de X, y la función está definido sobre el intervalo (0,1). Esta función F^{-1} es la función cuantil, y como la fda, generalmente no tiene una solución de forma cerrada. De igual modo que la fda, la manera más simple para calcular un valor específica (para una distribución particular) de la función cuantil es utilizando un rutina estadística de un programa computaciónal. Y también se puede encontar el valor de una función cuantil para la distirbución normal estandarizada a partir de tablas como la Tabla 3.2. Por ejemplo, imaginamos que estamos interesados en saber el valor mínimo de la distribución normal abajo de donde cae 97.5% de la masa de probabilidad. En la Tabla 3.2 observamos que $F^{-}1(0.975) = 1.96$ para una variable normal estandarizada.

Distribuciones Bivariadas

En situaciones cuando trabajamos con dos variables, además de considerar medidas de asociación común entre variables (como la correlación y la covarianza), podemos considerar toda la distribución conjunta de ambas variables. La distribución bivariada considera el soporte de dos variables en conjunto. Para definir la distribución bivariada, consideramos dos variables aleatorias, X e Y. La distribución bivariada es la colleción de probabilidades de la forma $\mathrm{P}[(X, Y) \in C]$ para todos los pares de números reales tal que $\{(X, Y) \in C\}$ es un evento. También como las funciones de distribución univariable, las distribuciónes bivariadas pueden ser discretas o continuas (o mesclazdas cuando una variable es discreta y otra es continua).

La función de probabilidad bivariada de X e Y, está definido como la función f tal como para cada punto (x, y) en el plano xy,:

$$f(x, y) = Pr(X = x \text{ y } Y = y),$$

que cumple las siguientes condiciones:

1. Si (x, y) no es uno de los posibles valores de los pares (X, Y), entonces $f(x, y) = 0$
2. $\sum_{Cada(x,y)} f(x, y) = 1$

Y las dos variables aleatorias X e Y tienen una distribución bivariada continua si existe una función no negativa f definido en todo el plano xy tal como para cada subconjunto C en el plano:

$$Pr[(X, Y) \in C] = \int_C \int f(x, y)dx\ dy$$

Aquí la función f se llama la función de densidad de probabilidad bivariada cuyas condiciones son:

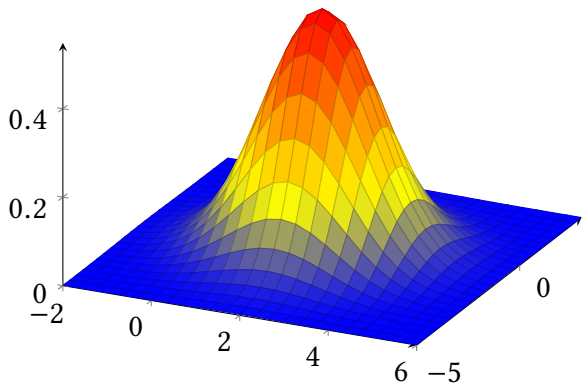
1. $f(x, y) \geq 0$ para $-\infty < x < \infty$ y $-\infty < y < \infty$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dx\ dy = 1$

Un ejemplo sencillo de una fdp bivariada contínua es la distribución normal bivariada, representada en la Figura 3.6. En forma general, la fdp de la Normal bivariada tiene la siguente fórmula, caracterizada por el promedio de cada variable X e Y, sus desviaciones estándares respectivos, y la correlación entre X e Y:

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right).$$

El ejemplo a continuación (Figura 3.6) representa dos variables normales, uno con promedio 2 y otro con promedio 1, ambos con desviación estándar igual a uno, y con una correlación $\rho = 0$.

Figure 3.6: Función de Densidad de Probabilidad Normal Bivariada



Distribuciones Condicionales

Las distribuciones condicionales describen las probabilidades asociadas con una variable aleatoria, condicionando en eventos determinadas por otras variables aleatorias. Después de observar los resultados de una (o varias) variables aleatorias, nos gustaría poder actualizar las probabildades asociadas con variables que aún no hemos observado. Por ejemplo, sabiendo que alguien estudio una carrera universitaria nos entrega información relevante para la distribución de su salario laboral. O sabiendo que una empresa tiene 10 trabajadores probablemente impacta la distribución posible de su producción total.

Supongamos que X e Y son dos variables aleatorias con una distribución bivariada cuya función de probabilidad es f. Ahora, f_1 y f_2 son las funciones de probabilidad marginales (individuales) de x e y. Si observamos que $Y = y$, la probabilidad que una variable aleatoria X tomará un valor específico x está dado por la probabildad condicional:

$$\begin{aligned} Pr(X = x|Y = y) &= \frac{Pr(X = x) \text{ and } Y = y}{Pr(Y = y)} \\ &= \frac{f(x, y)}{f_2(y)} \end{aligned}$$

Ahora, en vez de una probabilidad para un $X = x$, podemos escribir la función de probabilidad condicional entera:

$$g_1(x|y) = \frac{f(x, y)}{f_2(y)}$$

y la llamamos "la distribución condicional de X dado que $Y = y$."

Sean X e Y dos variables aleatorias con una distribución bivariada cuyo función de probabilidad es f y con funciones de probabilidad marginal f_1. Sea y un valor para cual $f_2(y) > 0$. Entonces, definamos la fdp condicional de X dado que $Y = y$ está definido como:

$$g_1(x|y) = \frac{f(x, y)}{f_2(y)} \text{ para } -\infty < x < \infty$$

De la misma forma cuando trabajamos con probabilidades condicionales y independia de variables en la ecuación 3.11, podemos hablar de independencia entre distribuciones enteras. Específicamente, tenemos que dos variables aleatorias X y Y con una fdp bivariada $f(x, y)$ son independientes *sii* para cada valor de y con $f_2(y) > 0$ y cada valor de x:

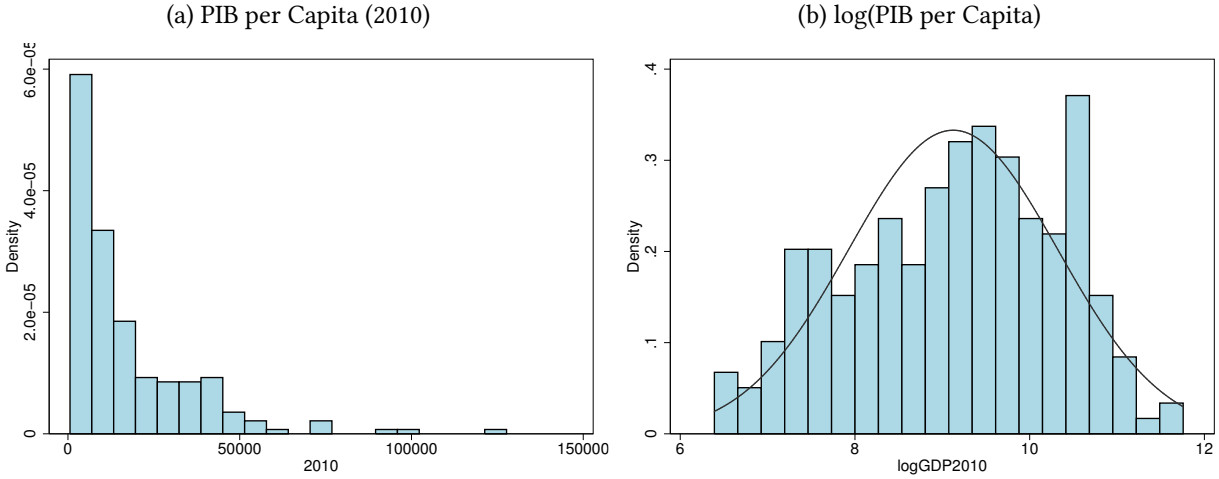
$$g_1(x|y) = f_1(x).$$

Algunas Distribuciones de Interés Para cerrar nuestra discusión de distribuciones de probabilidad, introducimos una serie de distribuciones conocidas que nos serán de utilidad durante este curso. Estos incluyen la distribución log-normal, la distribucíon chi-cuadrada (o χ^2), la dis-

tribución t de Student, y la distribución F.

La Distribución log-Normal: El uso de logaritmos para modelar variables es común (crecimiento, escala Richter, ...). Nos permite hablar en términos de cambios constantes (eg un log(PIB per capita) de 4.77 (Sweden) es 10 veces mayor que el de Vietnam (3.77) que es 10 veces mayor que el de la República Centroafricana). Los logaritmos tienen un soporte sobre $\mathbb{R}^+$. Decimos que si $\log(X)$ tiene una distribución normal, entonces la variable no transformada X es log-normal. En la Figura 3.7, illustramos una distribución empírica que parece ser log-Normal.

Figure 3.7: PIB per Capita (ajustado por poder de paridad de compra)



Si $\log(X)$ tiene una distribución normal con media μ ($-\infty < \mu < \infty$) y varianza σ^2 ($\sigma > 0$) decimos que X tiene una distribución log-normal.

$$f(\log(x); \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{\log(x) - \mu}{\sigma}\right)^2\right] \tag{3.18}$$

Esta distribución tiene una FDA de la siguiente forma:

$$F(x) = \Phi([log(x) - \mu]/\sigma)$$

y una función cuantil:

$$F^{-1}(p) = exp(\mu + \sigma\Phi^{-1}(p))$$

La Distribución Chi-cuadrado (χ^2): Sumando N variables normales (iid) al cuadrado resulta en una distribución χ^2 con N grados de libertad. Decimos que si $\{Y_i\}_{i=1}^N$ son $\mathcal{N}(0,1)$, iid. Entonces, $X = \sum_{i=1}^N Y_i^2 \sim \chi_N^2$. Esta distribución tiene un soporte sobre $\mathbb{R}^+$.

La Distribución t de Student La distribución t de "Student" (ver Figura 3.8) es la distribución que describe una variable aleatoria formada por la ratio de una variable aleatoria normal, y el raíz cuadrado de una variable Chi-cuadrado. Si $Y \sim \mathcal{N}(0,1)$, $X \sim \chi_N^2$, y Y, X son independientes,

entonces

$$T = \frac{Y}{\sqrt{(X/N)}} \sim t_N$$

Esta distribución nos sirve para inferencia estadística en muestras pequeñas. Cuando $N \to \infty$, $t \to \mathcal{N}(0, 1)$.

Figure 3.8: "Student" Biometrika, 1908.

PROBABLE ERROR OF A CORRELATION COEFFICIENT.

By STUDENT.

At the discussion of Mr R. H. Hooker's recent paper "The correlation of the weather and crops" (*Journ. Royal Stat. Soc.* 1907) Dr Shaw made an enquiry as to the significance of correlation coefficients derived from small numbers of cases.

His question was answered by Messrs Yule and Hooker and Professor Edgeworth, all of whom considered that Mr Hooker was probably safe in taking ·50 as his limit of significance for a sample of 21. They did not, however, answer Dr Shaw's question in any more general way. Now Mr Hooker is not the only statistician

La Distribución F (Fisher): La distribución de Fisher con N_1 grados de libertad en el numerador y N_2 grados de libertad en el denominador es el ratio de dos variables distribuidas como un χ^2, divididas por sus respectivas grados de libertad. si $X_1 \sim \chi^2_{N_1}$ y $X_2 \sim \chi^2_{N_2}$, y X_1 y X_2 son independientes. entonces,

$$F = \frac{X_1/N_1}{X_2/N_2} \sim F(N_1, N_2)$$

Nos encontraremos con esta distribución cuando implementamos contrastes de hipótesis múltimples en la segunda mitad del curso.

Table 3.1: Distribuciones Comúnes con su Esperanza y Varianza

Distribución	FDP	$E(X)$	$V(X)$
Distribuciones Discretas			
Bernoulli	$f(x;p) = \begin{cases} p^x(1-p)^{1-x} & \text{si } x = 0,1 \\ 0 & \text{si no} \end{cases}$	p	$p(1-p)$
Uniforme Discreta	$f(x;n) = 1/n \quad \text{si } x = 0,1,2\ldots,n$	$(N+1)/2$	$(N^2-1)/12$
Binomial	$f(x;n,p) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x} & \text{si } x = 0,1,2\ldots,n \\ 0 & \text{si no} \end{cases}$	np	$np(1-p)$
Poisson	$f(x;\lambda) = \begin{cases} \frac{\exp(-\lambda)\lambda^x}{x!} & \text{si } x = 1,2,\ldots \\ 0 & \text{si no} \end{cases}$	λ	λ
Distribuciones Continuas			
Rectangular en intervalo $[a,b]$	$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \le x \le b \\ 0 & \text{si no} \end{cases}$	$(a+b)/2$	$(b-a)^2/12$
Exponencial	$f(x) = \lambda \exp(-\lambda x)$	$1/\lambda$	$1/\lambda^2$
Normal Estandarizada	$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x}{2}\right).$	0	1
Logística Estandarizada	$f(x) = \frac{\exp(x)}{(1+\exp(x))^2}$	0	$\pi^2/3$

Refiere a Tabla 3.1 de Goldberger (1991, p. 28)

3.2 Comportamiento Asintótico

Nota de Lectura: Cameron and Trivedi (2005) tienen un apéndice que ofrece un tratamiento de los elementos más centrales de teoría asintótica aplicado a estimadores econométricos. Los capítulos de Stachurski (2016) son muy buenos, y ofrecen más detalle de los resultados que revisamos en estos apuntes. Demostraciones formales para los resultados de esta sección están disponibles en Rao (1973) (las demostraciones formales no son examinables, pero en cada caso se indica la referencia exacta a su ubicación en Rao (1973)). Para un tratamiento muy extensivo, existe el libro de texto (avanzado) de Davidson (1994). En esta sección, vamos a seguir la notación de Cameron and Trivedi (2005).

Consideramos el comportamiento de una secuencia de variables aleatorias b_N en la medida que $N \to \infty$. Utilizamos el subíndice N para indicar que esta estadística depende de la cantidad de observaciones N a partir de que se calcula. Por ejemplo, si nos interesa hablar de la cantidad promedio de educación en la población de Chile, el valor que estimamos con $N =\sim 15.000.000$ observaciones en el Censo es, probablemente, distinto al valor que estimamos con $N = 30.000$ observaciones aleatorias en la encuesta CASEN. Vamos a tener dos preocupaciones principales cuando hablamos del comportamiento asintótica. La primera es la **convergencia en probabilidad** de la cantidad b_N a algun límite b, que podría ser un valor escalar, o una variable aleatoria. El segundo, si el límite b es una variable aleatoria, es la **distribución límite**.

Convergencia en Probabilidad Decimos que la secuencia de variables aleatorias b_N converge en probabilidad a b si para toda $\delta > 0$:

$$Pr\left(|b_N - b| > \delta\right) \to 0 \qquad \text{a la medida que } N \to \infty. \tag{3.19}$$

Cameron and Trivedi (2005, p. 945) dan una definición más formal como su Definición A.1. Ésta nos permite elegir cualquier valor arbitrariamente pequeño para δ, y sabemos que si el N de la muestra es suficientemente grande, la probabilidad de que b_N difiere de b incluso en solo δ unidades, es nula. Eso nos permite escribir $\text{plim}_{N\to\infty} b_N = b$, donde plim refiere a la **probabilidad límite** (o a veces por simplicidad $\text{plim}\, b_N = b$) o $b_N \xrightarrow{p} b$. Esta definición sirve para variables aleatorias escalares. También existe una definición básicamente idéntica para variables aleatorias vectoriales, donde se reemplaza $|b_N - b|$ de la ecuación 3.19 con $||\boldsymbol{b}_N - \boldsymbol{b}|| = \sqrt{(b_{1N} - b_1)^2 + \ldots + (b_{KN} - b_K)^2}$. Volveremos a la Convergencia en Probabilidad pronto cuando hablamos de la consistencia de estimadores.

En la econometría, generalmente nos va a interesar trabajar con promedios o sumatorias sobre variables aleatorias en una muestra de interés. Por lo tanto, nos va a interesar considerar resultados límites considerando el comportamiento de promedios. Por suerte, existen dos clases de teoremas que son de importancia fundamental en probabilidad que hablan del comportamiento de promedios en el límite. Éstos son la **Ley de los Grandes Números** (LGN), y el **Teorema del**

Límite Central (TLC). El LGN describe el comportamiento del promedio de N variables cuando N crece sin límites, el y TLC describe el comportamiento de la distribución del promedio cuando N crece sin límites. Vamos a considerar estos dos resultados con más detalle en las sub-secciones que vienen.

3.2.1 La Ley de los Grandes Números

La ley de los Grandes Números describe un caso especial de convergencia en probabilidad, cuando la variable b_N refiere a una media muestral, o:

$$b_N = \bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i. \tag{3.20}$$

Aquí X_i refiere a una única realización de una variable aleatoria. Si cada X_i comparte el mismo promedio μ definimos $E(X) = \mu$ como el promedio poblacional, y la Ley Débil de los Grandes Números dice:

$$\bar{X}_N \to \mu \qquad \text{a la medida que } N \to \infty.$$

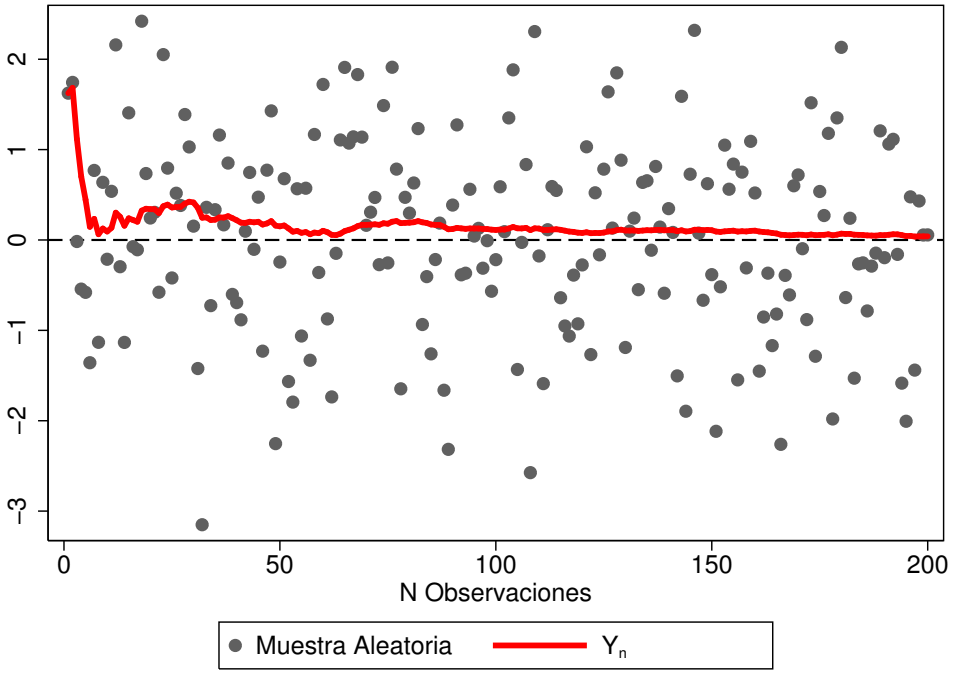
También se puede escribir de la misma forma como plim $\bar{X}_N = \mu$. Existe varias Leyes de Grandes Números, dependiendo de la distribución de variable que entra en el promedio, y el tipo de convergencia deseada. Detalles adicionales y definiciones formales están disponibles en Cameron and Trivedi (2005, pp. 947-948), pero brevamente, consideramos dos LGN. La LGN de Kolmogorov es la Ley relevante cuando cada X_i es iid. Si asumimos que X_i es iid, entonces la LGN de Kolmogorov demuestra convergencia al promedio μ con el único otro supuesto necesario siendo que $E(X) = \mu$. La LGN de Markov es una LGN que no requiere un μ común. Esta LGN sólo asume que cada X_i es independiente, pero no idénticamente distribuida ($inid$). A diferencia de la LGN de Kolmogorov, la LGN de Markov da como resultado que $\bar{X}_N$ converge en probabilidad a $E[\bar{X}_N]$. En el caso de la LGN de Markov, aunque se elimina el supuesto de una distribución común, es necesario agregar un supuesto adicional, de la existencia de un momento más alto que el primer momento.

La figura 3.9 demuestra una versión simulada de la LGN con variables iid. Consideramos una serie de observaciones X_i para toda $i \in \{1, \ldots, 200\}$, (que vienen de una distribución $\mathcal{N}(0, 1)$). La línea roja presenta el promedio $\bar{X}_N$ de todas las observaciones hasta N en el eje horizontal. En la medida que N aumenta, observamos que $\bar{X}_N$ se acerca cada vez más a la línea punteada igual a $\mu = 0$. De la LGN de Kolmogorov, sabemos que cuando $N \to \infty$ la probabilidad de que $\bar{X}_N$ difiere de μ para cualquier valor $\delta > 0$ se acerca a 0.

3.2.2 El Teorema del Límite Central

Antes de describir el Teorema del Límite Central (TLC) y sus implicancias, vamos a introducir una otra definición fundamental cuando consideramos el comportamiento asintótico de estadísticas de interés. Esto es la **Convergencia en Distribución**. Decimos que una secuencia

Figure 3.9: La Ley de los Grandes Números



de variables aleatorias b_N converge en distribución a la *variable aleatoria* b si:

$$\lim_{N\to\infty} F_N = F \tag{3.21}$$

en cada punto de continuidad de F. Aquí F_N es la distribución de b_N y F la distribución de b. Notemos que esta definición es la contraparte distribucional de la convergencia en probabilidad que definimos antes. Ahora, escribimos $b_N \xrightarrow{d} b$, y la distribución F se conoce como la distribución límite de b_N.

La ley de los grandes número solo nos entrega información acerca de cómo cambia la esperanza con un aumento del N. El teorema central del límite entrega información acerca de la *distribución entera* de la esperanza estimada, y como esta distribución se comporta en el límite. La TLC demuestra la existencia de una relación muy regular (y sorprendente) *sin importar* la distribución original de donde se extraye la muestra (y por ende la esperanza).

El Teorema del Límite Central (Lindeberg y Lévy) Si las variables aleatorias $X_1, \ldots, X_N$ son una muestra aleatoria de tamaño N de una distribución dada con media μ y varianza finita σ^2, entonces el Teorema de Límite Central implica:

$$Z_N = \frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}} \xrightarrow{d} \mathcal{N}(0,1). \tag{3.22}$$

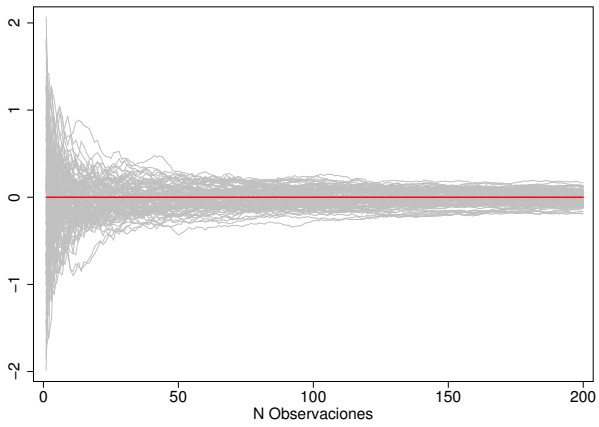
Este resultado dice que si obtenemos una muestra aleatoria (*iid*) grande de *cualquiera* distribución con media μ y varianza σ^2 (discreta o continua), la distribución de la variable aleatoria $N^{1/2}(\bar{X}_n - \mu)/\sigma$ será aproximadamente una normal estandarizada. O también, expresada en otra forma, implica que la distribución de $\bar{X}_N$ será una normal con media μ y varianza σ^2/N. Notamos que aquí la variable $(\bar{X}_N - \mu)/(\sigma/\sqrt{N})$ tendrá por definición en el límite un promedio igual a cero (ya que a $\bar{X}_N$ restamos μ), y una desviación estándar de 1, ya que se divide esta cantidad por su desviación estándar, pero la parte más importante es que sin importar la distribución base de la variable X, la distribución límite será normal. La demostración de este teorema se encuentra en Rao (1973, §2c.5)

En el mundo real, muchas variables siguen una distribución normal (altura, peso, ...). El TLC proporciona una posible explicación de este fenómeno. Si estas variables vienen de la influencia de muchos otros factores, entonces la distribución de la variable debería ser normal.

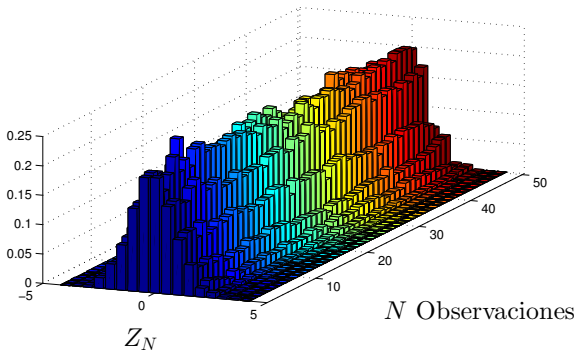
En la Figura 3.10 vemos un ejemplo de este idea. En la figura (a), observamos procesos repetetidos como el proceso observado en la Figura 3.9. Aquí sabemos (por la ley de los grandes números) que en el límite, este promedio debería acercar al valor verdadero de μ. Pero cuando repetimos este proceso muchas veces (cada línea gris es un promedio distinto), también observamos que aparece una distribución alrededor de $\bar{X}_N$. En la medida que el tamaño de la muestra (N) aumenta, esta distribución vuelve cada vez más acotada alrededor de μ. En la figura (b) observamos primero, que la distribución vuelve cada vez más precisa, y segundo, que sigue una distribución normal. Ésta distribución en la figura (b) está justamente descrito por el TLC. Cuando el tamaño de la muestra crece, la varianza cae en el orden de magnitud N^{-1}.[5] Este resultado se

Figure 3.10: El Vínculo Entre la LGN y el TLC

(a) Repeticiones de la LGN



(b) fdp de Repeticiones de la LGN

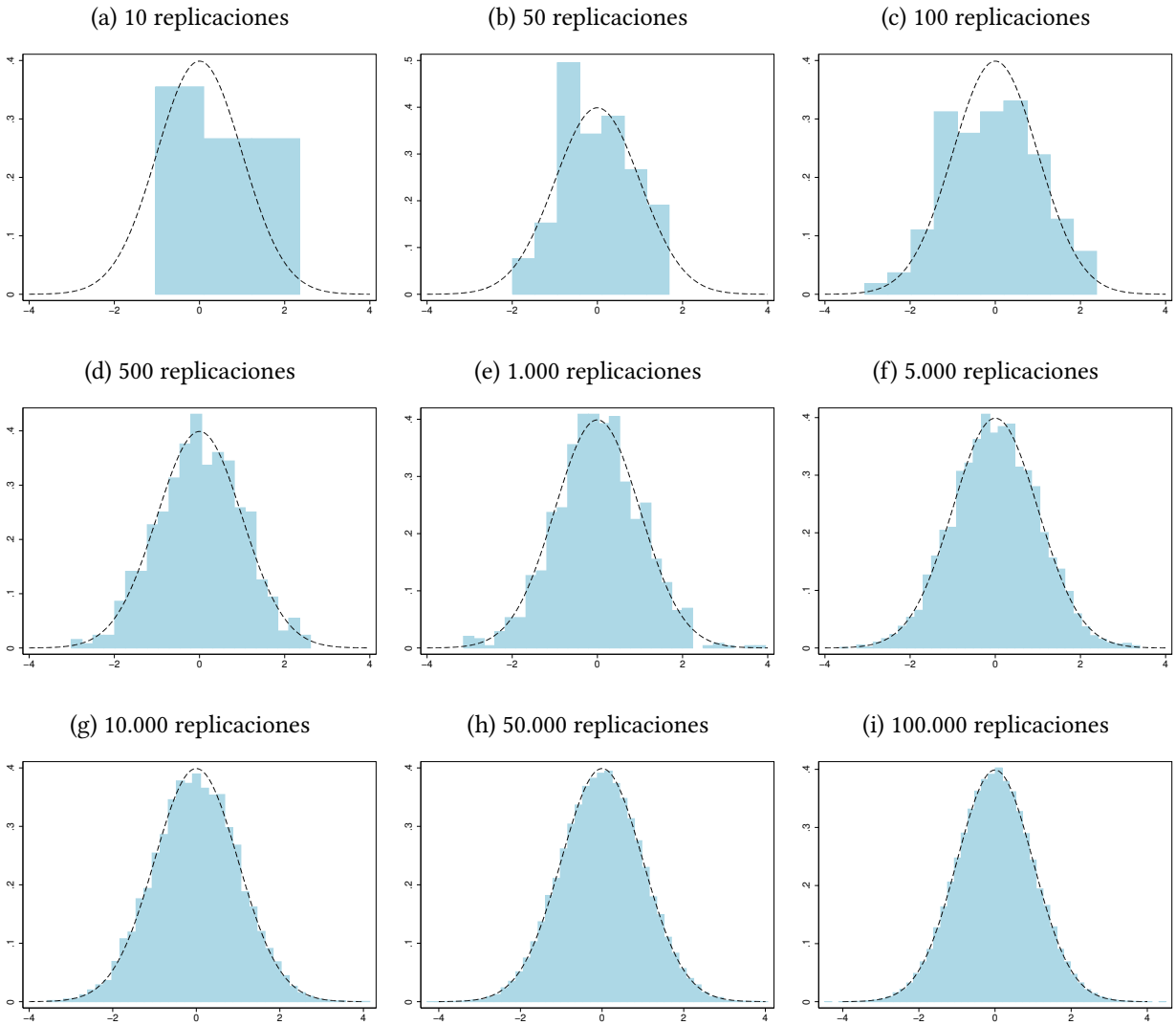


traduce en el teorema del límite central. Si consideramos la transformación descrita en ecuación 3.22, observamos como la distribución de esta variable se va regularizando (hacía la distribución $\Phi(x)$) en la medida que se observan más replicaciones del proceso subyacente.

[5]Para ver esto, notemos que la ecuación 3.22 implica que $\bar{X}_N \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/N)$, donde el denominador de la varianza es N.

La Figura 3.11 presenta una otra ilustración del TLC. En esta figura observamos justo la aparencia de la distribución normal prometida en la medida que nos acercamos a un distribución límite. Los paneles iniciales desmuestra la distribución de Z_N de 3.22 cuando consideramos una cantidad pequeña de replicaciones del proceso generador de Z_N (con un N fijo). En la primera fila observamos que esta distribución se asemeja bastante poco a una distribución normal estandarizada con 10, 50 y 100 replicaciones. A partir de 500 replicaciones, la distribución de Z_N empieza a se bastante parecida a una distribución normal estandarizada analítica, y una vez que se observa 100.000 replicaciones del proceso, la distribución empírica y la distribución analítica casi no se diferencian. En cada caso, la distribución subyacente que genera $\bar{X}_N$ en este proceso está basada en una serie de simulaciones de números pseudo-aleatorios que son normales. En la clase computacional a final de esta sección, dejamos como un ejercicio demostrar que este resultado se obtiene a partir de *otras* distribuciones también.

Figure 3.11: El Teorema del Límite Central



Nota: Consideramos la sumatoria de $N = 100$ muestras de una variable normal. Las distintas figuras demuestran la variación en $Z_N = \frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}}$ cuando Z_N se calcula una distinta cantidad de veces.

TLC de Liapounov (independencia) El TLC de Lindeberg Lévy asume que cada observación es independiente e idénticamente distribuida. Típicamente en econometría no tenemos observaciones con esta regularidad. Pero también existe un teorema del límite central para una secuencia de variables aleatorias que son independientes, pero no necesariamente identicamente distribuidas. Asumimos que $E(X_i) = \mu_i$ y $Var(X_i) = \sigma_i^2$. Definimos:

$$Z_N = \frac{\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \mu_i}{\left(\sum_{i=1}^{N} \sigma_i^2\right)^{1/2}},$$

y entonces $E(Y_n) = 0$, y $Var(Y_n) = 1$. Suponemos que las variables aleatorias $X_1, X_2, \ldots$ son independientes y $E(|X_i - \mu_i|^3) < \infty$ for $i = 1, 2, \ldots$. Entonces, $Z_N \xrightarrow{d} \mathcal{N}(0, 1)$. Notemos que aquí el resultado final es igual que el caso *iid*, pero en este caso *inid* (independiente y no idénticamente distribuida), requerimos el supuesto adicional $E(|X_i - \mu_i|^3) < \infty$. Esta versión del TLC está basado en Rao (1973). Cameron and Trivedi (2005, p. 950) presenta una versión generalizada, en base a White (2001).

Computer Class: Aplicación Simulada Refiere al código TLC.do. Este código proporciona un programa para examinar el Teorema del Límit Central variando el tamaño de la muestra (N observaciones) la cantidad de repeticiones, y la distribución de la variable subyacente. Su sintaxis es:
`tlc rnormal, observaciones(#) reps(#) mu(#) sigma(#).`

1. Examine el funcionamiento del programa cambiando el tipo de variable aleatoria, la cantidad de observaciones, y la cantidad de replicaciones.

2. (DIFÍCIL) Este código funciona para el TLC de Lindeberg–Lévy con variables iid. Escribe una versión para examinar el TLC de Liapunov para variables inid.

3.3 Estimadores

Nota de Lectura: En esta sección, examinamos estimación paramétrica, es decir, asumiendo que nuestros datos vienen de una distribución con una cantidad finita de parametros. Detalles de estimadores en este contexto están presentados en Casella and Berger (2002, Capítulo 7) y DeGroot and Schervish (2012, Capítulo 7). Ambos textos son una buena referencia. La presentación en Stachurski (2016, Capítulo 8) es más general, presentando una teoría de estimación para situaciones paramétricas y no-paramétricas. Este libro presenta una sintésis muy comprensiva, y más detallada que la presentación aquí.

3.3.1 Una Introducción y Descripción Generalizado

Consideramos un experimento, y asumimos que podemos repetir este experimento N veces. Como resultado, observemos N variables aleatorias: $(Y_1, \ldots, Y_N) = \mathbf{Y}$. Esta muestra de N realizaciones viene de una población descrita por alguna función de densidad o función de densidad de probabilidad: $f(\mathbf{y}|\boldsymbol{\theta})$. Aquí no suponemos nada acerca de los parametros que describen esta distribución $\boldsymbol{\theta}$, pero asumimos que la forma general de la distribución es conocida.

Cuando hablamos de estimación paramétrica, nuestra meta es "encontrar" los parametros $(\theta_1, \ldots, \theta_K) = \boldsymbol{\theta}$ que describen esta distribución.[6] Partimos sabiendo que estos parametros vienen del espacio Ω, que con frecuencia es un espacio poco restringido por ejemplo, podría ser $\mathbb{R}^K$. A veces vamos a querer poner restricciones en esta espacio. Un ejemplo es si vamos a estimar una varianza, que probablemente queremos limitar a un valor positivo. El desafío de la inferencia estadística es en cómo utilizar los valores de la muestra $\boldsymbol{Y}$ para estimar valores plausibles de $\theta_1 \ldots, \theta_K$.

Para estimar los parametros de interés, entonces necesitamos contar con (a) una muestra de datos extraídos de la población, (b) un modelo estadístico, y (c) una manera de llevar a cabo nuestra estimación puntual. Cuando hablamos de un modelo estadístico (o economíco), referimos a la manera en que llegamos a la clase general de distribuciones que supongamos que describe nuestra poblacon, $f(\mathbf{y}|\boldsymbol{\theta})$. Y cuando hablamos de una manera de estimar, referimos al proceso de encontrar los parametros $\boldsymbol{\theta}$ *una vez* que tengamos restringido las funciones de densidad de población a una clase paramétrica. En esta sección vamos a enfocar principalmente en el paso (c), y más adelante veremos algunos ejemplos de cómo formamos un modelo estadístico para un problema específica.

Formalmente, definimos a un **estimador puntual** como cualquier función

$$\hat{\theta} = \hat{\theta}(Y_1, \ldots, Y_N),$$

[6] O a veces nos interesa una función de estos parametros, $\tau(\boldsymbol{\theta})$. Veremos que el desarrollo aquí también para encontrar funciones de los parametros estimados.

es decir un estadístico de la muestra de datos. Notemos que esta definición es muy general, y de hecho, ni siquera implica alguna correspondencia entre el estimador y el parametro que se intenta estimar. Veremos más adelante (en la sección 3.3.5), que esta correspondencia viene más bien por condiciones que vamos a exigir de nuestra estimador para considerarlo un buen estimador. De esta definición podemos desprender varios hechos. Primero, notamos que el estimador es una función de los datos $Y_1, \ldots, Y_N$, que son variables aleatorias, y por ende el estimador es una variable aleatoria. Segundo, el estimador es una regla deterministica de cómo convertir la muestra de datos disponibles para cualquier valores de esta muestra al vector o escalar $\hat{\theta}$. Una vez que substituyemos los valores de las variables aleatorias específicas en la ecuación, tenemos una **estimación**. Y por último, notamos el uso de un 'gorro' para indicar que estamos trabajando con un estimador (o estimación). Utilizaremos esta notición de forma consistente en estos apuntes.

En algunos casos puede parecer bastante obvio al momento de determinar un estimador por un estadístico de interés. Por ejemplo, si nos interesa estimar el medio o algún cuantil de una población de interés, la contraparte *muestral* es probablemente un buen candidato. Pero en otras situaciones puede ser no tan obvio cómo definir cuál es un buen candidato. En la sección 3.3.5 de estos apuntes volvemos a definir una series de características que nos permiten juzgar a los estimadores, y comparar varios estimadores potenciales para un mismo parametro.

En esta sección de los apuntes vamos a examinar a dos clases grandes de estimadores. Estas clases nacen de distintos supuestos, y permiten estimar parametros en un rango amplio de situaciones. Además, vamos a utilizar estos estimadores a lo largo del curso. Específicamente, vamos a introducir el estimador de Máxima Verosimilitud, y el estimador de Método de Momentos. Esto no es para sugerir que estos estimadores sean las únicas posibilidades, o incluso los más utilizados en aplicaciones econométricas, pero los introducimos aquí dada su flexibilidad y utilidad en un importante grupo de problemas.

Por último, antes de seguir, destacamos la importancia de nuestros supuestos acerca del estado del mundo cuando definimos nuestros estimadores. A la base de cada estimador es un modelo de probabilidad que permite definir las clases de distribuciones que consideramos al momento de estimar nuestros parametros de interés. Cuando partimos con supuestos muy fuertes, esto nos puede proporcionar una solucón muy fácil al estimador, o producir un resultado matemático muy elegante. Sin embargo, al estar interesado en fenómenos del mundo natural, y especificamente de comportamiento humano, estos supuestos también deben ser capaces de capturar la complejidad de los fenómenos de interés. Es importante que siempre recordamos esta dualidad al momento plantear nuestros estimadores, y más generalmente cuando pensamos en inferencia estadística. A continución, reproducimos un párrafo que captura esta idea de forma muy simple de Manski (2003):

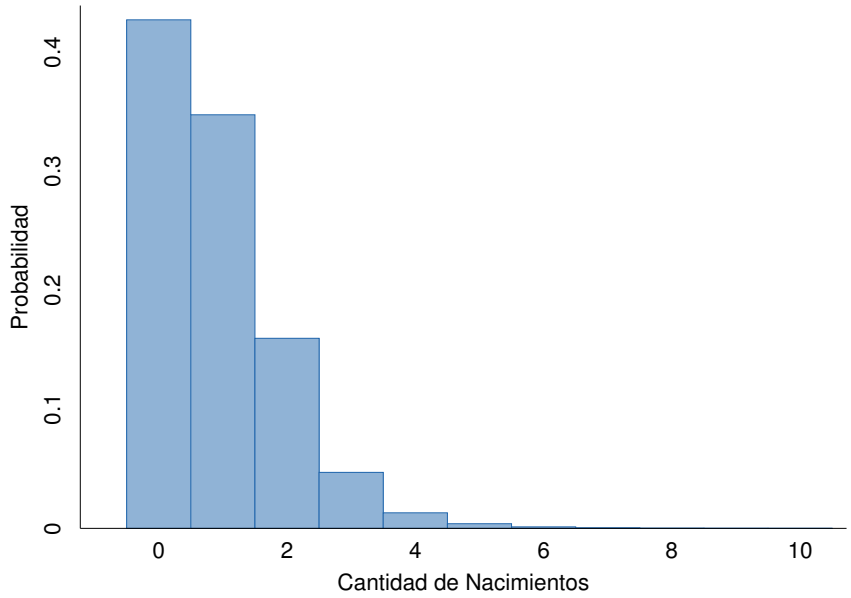
> "*The Law of Decreasing Credibility*: The credibility of inference decreases with the strength of the assumptions maintained.
>
> This principle implies that empirical researchers face a dilemma as the decide what

> assumptions to maintain: Stronger assumptions yield inferences that may be more powerful but less credible. Statistical theory cannot resolve the dilemma but can clarify its nature."
>
> Manski (2003, p. 1)

3.3.2 Una Aclaración: La Población

Cuando referemios a "la población" referimos al universo de objetos (personas, unidades, etc.) en que estamos interesado/as hacer inferencia estadística. Es una construcción teórica, y es infinita. Por ejemplo, si queremos saber algo acerca de las Pequeñas y Medianas empresas (PYMES) en Chile, la población es todas las PYMES que podrían hipotéticamente existir. Podemos observar todas las PYMES ahora en Chile, pero esto es solo parte de la población infinita. Por lo tanto, incluso cuando tenemos un censo, hablamos de una muestra de la población. Aunque podría parecer curioso no tratar el censo como el universo de la población estadística cuando efectivamente *es* toda la población del país, esto se debe a la diferencia en la terminología de poblacón estadística, y población como típicamente se utiliza la palabra en la vida cotidiana. Nuestro universo o población estadística refiere a toda la población que podría existir a la raíz del proceso generador de datos.

Figure 3.12: Función de Probabilidad: Número de Nacimientos Anteriores de Madres, Chile 2015

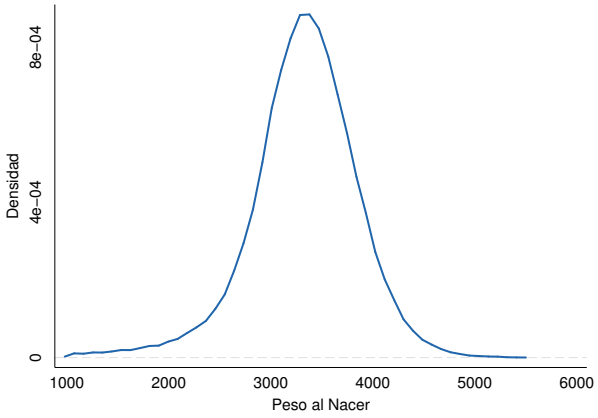


Cuando trabajamos con una muestra de la población, referimos a funciones de distribución *empíricas*. En las figuras 3.12-3.13 consideramos una serie de distribuciones empíricas. La figura 3.12 presenta una representación gráfica de una función de probabilidad en base de la representación empírica de la cantidad de hermano/as de todos los bebés que nacieron en Chile en 2015. Cada resultado potencial (tener una cantidad x de hermana/os mayor o igual a cero) está asociado con una probabilidad correspondiente.

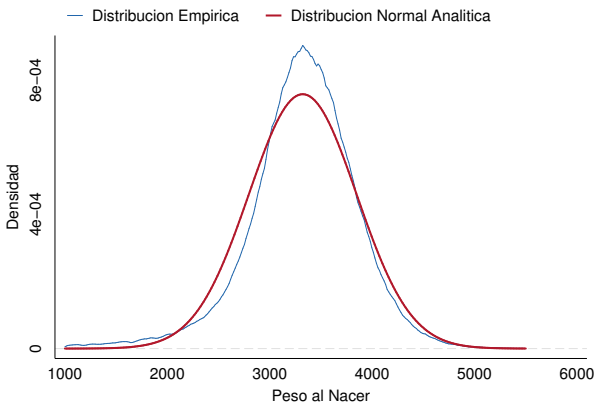
Y en el segundo caso, observamos una función de densidad de probabilidad empírica: la distribución de peso al nacer de todos los nacimientos ocurridos en Chile en el año 2015. Aquí, aunque son todos los nacimientos, sigue siendo una muestra, dada la definición de población a que referimos antes. Y aquí se ve el vínculo con los modelos estadísticos que discutimos en la sección anterior. Aunque no sabemos los parametros de la distribución poblaciónal que produjo la densidad en La Figura 3.13, al observar la distribución empírica podría ser razonable suponer que esta distribución empírica está extraída de una distribución poblacional normal. En la figura (b) se compara la distribución empírica con una distribución analítica con la misma media y desviación estándar. En las secciones que vienen, veremos con más detalle los supuestos y pasos necesarios para estimar parametros de una distribución poblacional, a partir de una muestra empírica.

Figure 3.13: Función de Densidad: Peso al Nacer, Chile 2015

(a) Función de Densidad Empírica



(b) Distribución Empírica y Analítica



Nota: Distribuciones se basan en los 244.670 nacimientos que ocurrieron en Chile en 2015. En el segundo panel, también se presenta una distribución normal analítica con la misma media y desviación estándar que la distribución empírica.

3.3.3 Método de Estimación: Método de Momentos

En las dos sub-secciones que siguen, vamos a introducir una serie de métodos de estimación: primero método de momentos, y segundo máxima verosimilitud. Para introducir estas técnicas, partiremos con un caso simple: imaginemos que existe alguna variable Y que es una variable aleatoria con:

$$Y \sim \mathcal{N}(\mu, \sigma^2) \tag{3.23}$$

y vamos a suponer que tenemos una muestra de realizaciones Y_i extraída de la población. Nuestro interés es un estimar $\widehat{\mu}$ y $\widehat{\sigma}$. Esta descripción es un tipo de "modelo estadístico" (aunque uno bastante sencillo). Estamos asumiendo que tenemos un proceso generador de datos (poblacional) donde Y sigue la distribución normal. Así, estamos limitando la clase de distribuciones posibles para Y, pero no estamos restringiendo los parametros que describen Y dentro de esta clase de

distribuciones.[7]

Ahora, nosotros contamos con una muestra Y de tamaño N (Y es un vector de $N \times 1$). Nuestro interés es inferir algo acerca de los parametros $\boldsymbol{\theta} = (\mu, \sigma^2)$ que describen el modelo poblacional 3.23. Sin embargo, solo tenemos nuestra muestra finita (de tamaño N) para hacer deducciones. Entonces, necesitamos una técnica de estimación que vincula los datos de muestra, con nuestro modelo poblacional para estimar $\boldsymbol{\theta}$.

La técnica de Método de Momentos parte con el principio de analogía. Este principio sugiere que debemos estimar nuestros parametros poblacionales utilizando estadísticas muestrales que tienen las mismas características en la muestra, que los parametros poblacionales tienen en la población. Entonces, por ejemplo, si estamos interesados en estimar la media poblacional (algo que no observamos), el principio de analogia sugiere utilizar la media muestral, algo que *sí* observamos. Este es una técnica simple, y muy poderosa y generalizable. Específicamente, un estimador de método de momentos busca definir los *momentos poblacionales* que caracterizan el proceso generador de datos, y después estimarlos utilizando **los momentos análogos de la muestra**.

En el caso de interés aquí (el proceso generador de datos 3.23), tenemos dos momentos de interés (refierese a la sección 3.1.3 para una discusión acerca de los momentos de una distribución). Estos primeros dos momentos son la esperanza y la varianza, que en la población se escriben:

$$E(Y) = \mu \tag{3.24}$$

$$Var(Y) = E[Y^2] - (E[Y]^2) = \sigma^2. \tag{3.25}$$

Referimos a las ecuaciones 3.24 y 3.25 como los momentos poblacionales.

La idea del método de momentos es simplemente utilizar los contrapartes muestrales para estimar la respectiva cantidad poblacional. En este caso, podemos escribir los momentos muestrales como:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{3.26}$$

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2 \tag{3.27}$$

donde ahora sí los momentos son observables. En términos mecánicos, siempre escribimos a nuestros momentos como alguna cantidad igual a cero, dando que las condiciones de momentos

[7] En realidad, tenemos *una* restricción sobre uno de los parametros. Sabemos que la varianza σ^2 es igual o superior a 0.

poblacionales son:

$$E[Y] - \mu = 0 \tag{3.28}$$
$$E[Y^2] - (E[Y]^2) - \sigma^2 = 0. \tag{3.29}$$

Hasta ahora hemos referido a estos momentos como "momentos centrales". Y entonces, las vamos a estimar usando las condiciones de momentos centrales análogas de la muestra que observamos:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i - \widehat{\mu} = 0 \tag{3.30}$$
$$\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2 - \widehat{\sigma}^2 = 0 \tag{3.31}$$

donde este proceso de estimación consiste en resolver dos ecuaciones (ecuación 3.30-3.31) con dos incógnitas, que es un proceso simple de eliminación. Más adelante en el curso, vamos a considerar estimadores de métodos de momentos bastante más complejos con más momentos que parametros a estimar, y en este caso el problema de estimación es un problema de minimización, la técnica de estimación es conocida como **método de momentos generalizados**.

El Estimador de Forma General Una distribución normal y $\boldsymbol{\theta} = (\mu, \sigma)$ es solo uno de *muchos* posibles ejemplos de método de momentos. En forma general, se puede representar el método de momentos con: (a) Los momentos poblacionales (b) Los momentos muestrales correspondientes, y (c) La solución de los momentos muestrales. Los momentos poblacionales se escriben de forma genérica como:

$$E[\boldsymbol{h}(\boldsymbol{w}_i, \boldsymbol{\theta}_0)] = \mathbf{0}$$

donde:

- $\boldsymbol{\theta}$ es un vector de $K \times 1$ de los parametros a estimar
- $\boldsymbol{\theta}_0$ es el valor de $\boldsymbol{\theta}$ en la población
- $\boldsymbol{h}(\cdot)$ en una función vectoral de $r \times 1$ de los momentos ($r \geq K$)
- $\boldsymbol{w}$ es un vector que incluye todos los datos observables

Y de forma parecida, escribimos los momentos muestrales como:

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{h}(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}) = \mathbf{0} \tag{3.32}$$

donde aquí la única diferencia es que reemplazamos los momentos poblacionales para su contraparte muestral, y denotamos al vector de parametros que resuelve 3.32 como $\widehat{\theta}$. Este vector es la solución a esta ecuación, y el estimador de método de momentos. A veces denotamos al estimador como $\hat{\boldsymbol{\theta}}_{MM}$.

Cuando la cantidad de momentos es igual a la cantidad de parametros que se busca estimar ($r = K$), se puede simplemente resolver el sistema de ecuaciones en 3.32. Sin embargo, muchas veces es más facil (computacionalmente) minimizar una función de la siguiente función.

$$Q_N(\boldsymbol{\theta}) = \left[\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{h}(\boldsymbol{w}_i, \boldsymbol{\theta})\right]' \left[\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{h}(\boldsymbol{w}_i, \boldsymbol{\theta})\right]$$

El estimador de método de momentos $\hat{\boldsymbol{\theta}}_{MM}$ es (también) la cantidad que minimiza Q_N:

$$\hat{\boldsymbol{\theta}}_{MM} = \arg\min_{\boldsymbol{\theta}} Q_N(\boldsymbol{\theta}).$$

3.3.4 Método de Estimación: Máxima Verosimilitud

El estimador de método de momentos tiene una lógica bastante razonable: nuestro valor estimado de θ es el valor de θ que resuelve la contraparte de los momentos poblacionales en la muestra. Sin embargo, no es el único estimador posible, y no utiliza todo la información disponible en nuestro modelo estadístico. En particular, método de momentos extraye toda la informacíon para estimar de algunos pocos puntos en la distribución de probabilidad (los momentos). En algunos casos esto puede ser conveniente, dado que es más fácil plantear momentos que toda una distribución de probabilidad (y los supuestos son más flexibles), pero en otras casos (como el caso de interés aquí), ya contamos con más información que solo los momentos. Esto nos trae a atra manera de estimar, que es la técnica de máxima verosimilitud. Con máxima verosimilitud utilizamos *más información* para estimar que solo los momentos.

La idea de Máxima Verosimilitud (MV) viene de Fisher (1922) y el principio de verosimilitud:

> Eligimos como estimador para el vector de parametros $\boldsymbol{\theta}_0$ el valor de $\boldsymbol{\theta}$ que maximiza la probabilidad de observar la muestra actual.

Notamos que no estamos diciendo nada acerca de hacer cumplir los momentos de la distribución (y de hecho, pueden haber buenos estimadores de máxima verosimilitud que no hacen cumplir los momentos), pero ahora tenemos que tener una manera de calcular la probabilidad de haber observado una serie de parametros específicos, dado un modelo de probabilidad.

De nuevo en esta sección vamos a asumir que tenemos una muestra aleatoria (*iid*) extraíd de una distribución normal:

$$Y_i \sim \mathcal{N}(\mu, \sigma^2)$$

La idea de MV es que podemos utilizar toda la información de la fdp (y no solo los momentos centrales). Utilizando la misma notación anterior, escribimos la función de densidad de probabilidad como $f(\boldsymbol{y}|\boldsymbol{\theta})$, o también como $f(\boldsymbol{y}|\mu, \sigma)$ si queremos destacar que los parametros que describen la fdp en este caso son μ, su promedio, y σ, su desviación estándar.

Como tenemos una muestra de tamaño N extraída de una distribución normal, por la naturaleza de un proceso estocástico, habrán algunas observaciones que son bastante cercanos a la media μ, y otras observaciones que son más alejadas. Pero nosotros observamos todas las observaciones en la muestra, y queremos inferir a partir de estas observaciones cuál eran los parametros poblacionales que los generó. Para poder considerar la muestra entera, tenemos que partir con la fdp conjunta de todas las variables aleatorias. Esta fdp conjunta nos dice la probabilidad de haber observado toda la muestra en conjunto. Dado que contamos con una muestra de variables independientes, tenemos que:

$$\begin{aligned} f(y_1, ..., y_N|\mu, \sigma) &= f(y_1|\mu, \sigma) \times \cdots \times f(y_N|\mu, \sigma) \\ &= \prod_{i=1}^{N} f(y_i|\mu, \sigma), \end{aligned} \tag{3.33}$$

donde el lado derecho de la ecuación 3.33 viene de la independencia de observaciones, y la ecuación 3.12 discutido anteriormente.

Aquí en palabras tenemos la probabilidad de observar una combinación $y_1, \ldots, y_N$, dado los parametros verdaderas μ y σ. Sin embargo, lo que queremos es algo un poco distinto. Queremos saber cuál es la probabilidad de tener parametros verdaderos μ y σ dado el conjunto de datos que hemos observado. Es decir, en vez de considerar a la función con parametros dados y con observaciones de y_i que cambian, queremos imaginar que los y_i son dados, y que vamos variando los parametros μ y σ.

Por esto, escribimos la **función de verosimilitud**:

$$\mathcal{L}(\mu, \sigma|y_1, \ldots, y_n) = \prod_{i=1}^{N} f(y_i|\mu, \sigma) \tag{3.34}$$

Ahora, aunque la función en el lado izquierdo igual, la interpretación es distinto. Con la función de verosimilitud, calculamos la probabilidad conjunto de haber observado el vector Y variando los parametros μ y σ. Por ende, la función de verosimilitud cuantifica cuán probable es que hubiesemos observado los datos que observamos, para un μ y σ determinado. A partir de esta función, se estima $\hat{\mu}$ y $\hat{\sigma}$ ($\hat{\theta}$). Son los valores que maximizan la función de verosimilitud, o los valores que hacen lo más probable haber observado la muestra que observamos:

$$\hat{\theta}_{MV} = \arg\max_{\theta \in \Omega} \mathcal{L}(\theta|y_1, \ldots, y_N).$$

En la práctica, maximizando la función de verosimilitud puede ser computacionalmente (o algebraícamente) difícil. En términos computacionales, dado que la función de verosimilitud se forma multiplicando N funciones de densidad (con un rango de entre 0 y 1), cuando N es grande, este producto vuelve muy pequeño. Por eso definimos la función de log verosimilitud, que tiene

varias propiedades muy convenientes.

$$\begin{aligned} \ell(\mu, \sigma | y_1, \ldots, y_n) &\equiv \ln \mathcal{L}(\mu, \sigma | y_1, \ldots, y_n) \\ &= \sum_{i=1}^{N} \ln f(y_i | \mu, \sigma) \end{aligned} \tag{3.35}$$

Lo más conveniente es que dado que ℓ es una función monótona creciente de $\mathcal{L}$, la función ℓ y la función $\mathcal{L}$ alcanzan sus máximos en el mismo valor de $\boldsymbol{\theta}$. Así, estimamos MV de la misma manera:

$$\hat{\theta}_{MV} = \arg\max_{\theta \in \Omega} \ell(\theta | y_1, \ldots, y_N).$$

Generalmente (pero con algunas excepciones), las funciones de verosimilitud no tienen una solución de forma cerrada. En este caso, se puede estimar usando métodos computacionales. En su forma más simple, estos métodos prueban todas las combinaciones posibles de $\hat{\theta}$ y $\hat{\sigma}$ hasta encontrar el máximo. Y en programas como Stata, existen librerias que pueden maximizar una función de manera mucho más rápido, por ejemplo utilizando algoritmos como Newton-Raphson. Para nuestros requisitos, basta saber que una vez que escribimos una función de máximia verosimilitud, tenemos que utilizar el computador para maximizarlo! Si le interesaría leer más de métodos computacionales para maximizar una función, refiere al capítulo 10 de Cameron and Trivedi (2005).

Las derivaciones anteriores han presentado de forma general el proceso de MV para cualquier modelo de probabilidad, y cualquier fdp. Pero para el caso de una distribución normal, podemos ir más lejos. En este caso, sabemos cuál es la fdp para una sola realización de la variable aleatoria (de la ecuación 3.16). Así, escribimos:

$$\begin{aligned} f(y_1, ..., y_N | \mu, \sigma) &= f(y_1 | \mu, \sigma) \times \cdots \times f(y_N | \mu, \sigma) \\ &= \prod_{i=1}^{N} f(y_i | \mu, \sigma) \\ &= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{(y_i - \mu)^2}{2\sigma^2} \end{aligned} \tag{3.36}$$

donde el último paso viene de la distribución normal 3.16. Dado esto, nuestra función de verosimilitud *para una muestra de variables extraídas de una distribución normal* es:

$$\mathcal{L}(\mu, \sigma | y_1, \ldots, y_N) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{(y_i - \mu)^2}{2\sigma^2}$$

y el logaritmo de la función de verosimilitud es:

$$\begin{aligned}\ell(\mu, \sigma | y_1, \ldots, y_N) &= N \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{N} \frac{(y_i - \mu)^2}{2\sigma^2}. \\ &= -\frac{N}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(y_i - \mu)^2}{2\sigma^2}. \end{aligned} \tag{3.37}$$

Esta última ecuación 3.37 es una función que podríamos entregar a un programa computacional para maximar, y encontrar los parametros estimados $\hat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\sigma})$:

$$\hat{\boldsymbol{\theta}}_{MV} = \arg\max_{\boldsymbol{\theta}} \ell(\mu, \sigma | y_1, \ldots, y_N).$$

Examinaremos un ejemplo de este proceso como un código computacional, y el comando `mlexp` de Stata.

3.3.5 Propiedades de Estimadores

Como hemos definido, la estimación consiste en utilizar una muestra de datos y otra información que tenemos *a priori* para producir un valor que es en algún sentido nuestra "mejor estimación" de un parametro desconocido. Pero un estimador es simplemente una regla determinística para definir una estadística a partir de datos observados. En realidad, podemos imaginar muchos estimadores. Una posibilidad (bastante absurda) sería simplemente elegir nuestro número favorito para todos los parametros estimados. Aunque hemos conocido un par de técnicas basadas en principios de estimación que parecen razonables (MV y MM), no hemos definido ninguna manera para definir cuáles son los mejores estimadores, ni comparar técnicas de estimación. Aunque lo que es "mejor estimación" depende de nuestra definición y metas al momento de estimar, podríamos esperar que nuestros estimadores cumplen con ciertas condiciones o criteros para ser consideradas buenas.

En las muestras finitas (o muestras pequeñas), hay dos propiedades que son particularmente relevantes. Éstas son **Sesgo** y **Eficiencia**. Veremos ambas propiedades en un poco más de detalle.

Sesgo Una propiedad intuitiva, y una manera clásica de evaluar un estimador es sesgo. Un estimador insesgado cumple con:

$$E[\hat{\theta}] = \theta.$$

Si $E[\hat{\theta}] \neq \theta$, decimos que el estimador $\hat{\theta}$ está sesgada. Definimos el sesgo como la diferencia entre la expectativa del estimador, y el valor verdadero del parametro poblacional que estamos intentando estimar: $E[\hat{\theta}] - \theta = \delta$. Aquí cuando hablamos de la expectativa del estimador, estamos refiriendo al valor promedio si podríamos repetir el experimento subyacente infinitas veces, cada vez sacando una muestra de tamaño N, y estimando nuestro estimador $\hat{\theta}$. Si el estimador es insesgado, encontraremos que en promedio, este estimador es igual a θ. Pero esto es un constructo

teórico. En la realidad, no tenemos infinitos experimentos, sino uno. Por esto, también pensamos en la eficiencia de los estimadores.

Eficiencia Dado que generalmente observamos una sola serie de datos, el insesgadez de un estimador no es nuestra única preocupación. Si tenemos un estimador insesgado pero muy impreciso, una realización en particular de $\hat{\theta}$ podría estar bastante lejos de θ. La eficiencia considera la precisión de un estimador, y la criteria de eficiencia consiste en considerar solo los estimadores insesgados, y elegir él que tiene menor varianza. Decimos que un estimador $\hat{\theta}$ es eficiente si:

$$\text{Var}(\hat{\theta}) \leq \text{Var}(\tilde{\theta}) \tag{3.38}$$

donde $\tilde{\theta}$ es cualquier otro estimador insesgado de θ. Una manera de comprobar que un estimador sea el estimador más eficiente de todos los estimadores posibles, es utilizando la cota inferior de Cramér-Rao. La cota inferior de Cramér-Rao, demuestra que:

$$\text{Var}(\hat{\theta}) \leq \frac{1}{-E\left[\frac{d^2\ell(\theta)}{d\theta^2}\right]}$$

donde ℓ es el logaritmo de la función de verosimilitud. Esto implica que si podemos mostrar que la varianza de un estimador es igual al término de la derecha en la ecuación 3.38, sabemos que es el mejor estimador insesgado posible.

Así aunque la eficiencia también considera la varianza del estimador, se limita solo a estimadores insesgados. En la práctica, si estamos frente a un estimador insesgado e impreciso, y otro sesgado, pero preciso, ¿cuál es mejor? Una solución potencial viene del error cuadrático medio, que es una función que define la 'perdida' asociada con cualquier estimador $\hat{\theta}$. El ECM se define de la siguiente forma:

$$ECM(\hat{\theta}) = E[\hat{\theta} - \theta]^2 \tag{3.39}$$

que incorpora un castigo por sesgo, y además por precisión. Para ver esto, notamos que podemos re-escribir la función de pérdida 3.39 como:

$$\begin{aligned}
ECM(\hat{\theta}) &= E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\
&= E\{[\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta]\}^2 \\
&= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 + 2E[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta] \\
&= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 \\
&= \text{var}(\hat{\theta}) + [\text{sesgo}(\hat{\theta})]^2.
\end{aligned}$$

Propiedades Asintóticas En varias estimadores que encontraremos durante el curso, no va a ser posible derivar propiedades en mustras finitas. En este caso, es necesario considerar cómo el

estimador comporta en muestras asintóticas (o muestras grandes). Hablaremos de las siguientes propiedades, que juegan el mismo papel de sesgo y eficiencia en muestras chicas:

1. Consistencia
2. Eficiencia asintótica

La consistencia—como el sesgo en muestras finitas—asegura que un estimador $\hat{\theta}$ estará "cerca" al valor θ cuando la muestra está suficientemente grande. Formalmente, decimos que θ_N es consistente si:

$$\lim_{N\to\infty} P(|\hat{\theta}_N - \theta| < \varepsilon) = 1$$

donde ε es un número positivo y pequeño. Generalmente, escribimos lo anterior como:

$$\text{p lim}\, \hat{\theta}_T = \theta.$$

Exigir que un estimador sea insesgado es más exigente que exigir que un estimador sea consistente. Al tener un estimador insesgado, sabemos que $E(\hat{\theta}) = \theta$, incluso con un N pequeño. En el caso de un estimador consistente, solo sabemos que $\lim_{N\to\infty} E[\hat{\theta}] = \theta$ (es decir, cuando el N va hacia infinito). Sin embargo, muchas veces vamos a tener estimadores que no son insesgados, pero son consistentes. Un ejemplo de ello es el $\hat{\sigma}$ que estimamos esta clase con MV (más detalles en clase).

Convergencia en Distribución Como vimos en la sección 3.2, cuando el tamaño de una muestra aumenta, aunque la distribución subyacente no es necesariamente regular, la distribución límite sí lo es! Este teorema (el teorema del límite central), nos sirve bastante con estimadores en muestras grandes.

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Este hecho es muy conveniente cuando tenemos un estimador con una distribución desconocido en muestras pequeñas. Sin embargo, con una muestra grande sabemos:

$$\begin{aligned} \sqrt{N}(\hat{\theta} - \theta) &\dot{\sim} \mathcal{N}(0, \sigma^2) \\ \Rightarrow \hat{\theta} &\dot{\sim} \mathcal{N}(\theta, \sigma^2/N) \end{aligned}$$

donde $\dot{\sim}$ es "aproximadamante distribuida"

Eficiencia Asintótica Como último, tenemos la eficiencia asintótica, que cumple el mismo rol que la eficiencia en muestras pequeñas. Decimos que un estimador $\hat{\theta}$ es relativamente más eficiente que otro, $\tilde{\theta}$, si se cumplen las siguientes tres propiedades:

1. $\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma_1^2)$

2. $\sqrt{N}(\tilde{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma_2^2)$
3. y: $\sigma_2^2 \geq \sigma_1^2$.

3.4 Inferencia

Cuando hablamos de sesgo y eficiencia en la sección anterior, era aparente que un estimador viene con su propia distribución. Cuando tenemos una muestra dada, una determinada estimación siempre será lo mismo. Pero si tuviesemos otra mustra (representativa) de la misma población, la estimación probablemente será un poco distinto, simplemente por el hecho de contar con una distinta muestra extraída de la población. En la práctica, generalmente contaremos con una sola muestra para estimar parametros de interés. Pero para reconocer que nuestra estimación viene con su propia distribución, también nos interesaría estimar su varianza. Una vez que hemos estimado un estimador puntual, y además su varianza, podemos hacer enunciados acerca de la probabilidad que el parametro poblacional (verdadero) cae en un cierto rango alrededor de nuestra estimador puntual. Se refiere a este proceso de formar enunciados de probabilidad como Inferencia estadística. La inferencia puede consistir en la formación de intervalos de confianza alrededor de un estimador puntual, o a testear formalmente hipótesis estadísticos. Estudiamos ambos casos en más detalle ahora.

3.4.1 Estimación de Intervalos

Estimación con Varianza Conocida

Partimos considerando un caso simple. Supongamos que tenemos una muestra aleatoria $Y_1, \ldots, Y_N$ de una población con $\mathcal{N}(\mu, \sigma^2)$ con σ^2 conocida. Y supongamos que nos interesa estimar el parametro μ. Por ahora no nos preocupamos del por qué σ^2 es conocida. Simplemente notaremos que probablemente no es tan realística pensar que podríamos saber σ^2, así que puede ser un supuesto que queremos eliminar en el futuro! Podesmos mostrar que el estimador de maxima verosimilitud para μ es $\hat{\mu} = \sum_{i=1}^{N} Y_i/N$. Dejamos como un ejercicio esta demostración.

Ahora, para considerar la distribución de esta estimador, partimos considerando las propiedades de estimador y varianza. Tenemos que la expectativa del estimador es:

$$E[\hat{\mu}] = \sum_{i=1}^{N} E[Y_i/N] = \sum_{i=1}^{N} \mu/N = N\mu/N = \mu,$$

(que implica que $\hat{\mu}$ es un estimador insesgado), y que la varianza es:

$$Var[\hat{\mu}] = \sum_{i=1}^{N} Var[Y_i/N] = \frac{1}{N^2} \sum_{i=1}^{N} Var[Y_i] = N\sigma^2/N^2 = \sigma^2/N.$$

Dado que cada Y_i es *iid* de una distribución normal, podemos definir la distribución *exacta* del

estimador $\hat{\mu}$ como:

$$\mathcal{N}(\mu, \sigma^2/N) \tag{3.40}$$

La ecuación 3.40 es un resultado muy conveniente, ya que víncula por primera vez la distribución del estimador con el parametro poblacional en sí. Y esto nos permite hacer enunciados de probabilidad acerca del parametro desconocido μ utilizando el estimador $\hat{\mu}$ (conocido) y la varianza σ^2 (conocido por nuestro supuesto anterior).

Para ver cómo podemos seguir con estos enunciados de probabilidad, partimos con una fórmula conocida de la normal estandarizada:

$$z = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0, 1) \tag{3.41}$$

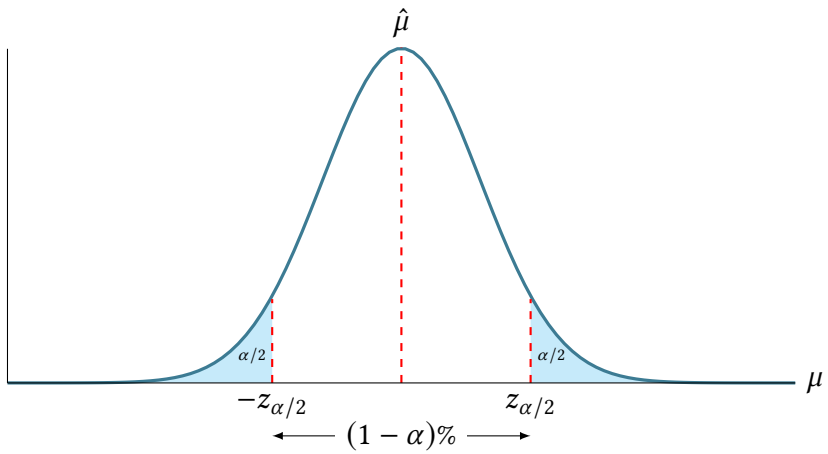
Esta variable z es una transformación de cualquier variable aleatoria normal, restando el promedio y dividiendo por su desviación estándar para expresar la variable como una normal con medio 0 y desviación estándar 1. Hacemos esta transformación dada la facilidad de trabajar con la normal estandarizada. La probabilidad de masa de probabilidad baja la curva normal en cada punto se resume en tablas estadísticas como la Tabla 3.2.

Por la naturaleza de la normal estandarízada, podemos calcular la probabildad que z cae entre cualquier dos valores simétricas $-z_{\alpha/2}$ y $z_{\alpha/2}$:

$$Pr[-z_{\alpha/2} \leq z \leq z_{\alpha/2}] = 1 - \alpha \tag{3.42}$$

Por ejemplo, utilizando la Tabla 3.2, la probabilidad que z cae entre -1.96 y 1.96 es 95%. Este cálculo se realiza sumando la masa de probabilidad que cae *afuera* del rango de interés, como resumido en la Figura 3.14.

Figure 3.14: Intervalos de la Normal Estándarizada



Ahora, con un poco de álgebra con las ecuaciones 3.41 y 3.42, tenemos:

$$\begin{aligned} Pr[-z_{\alpha/2} \leq z \leq z_{\alpha/2}] &= 1-\alpha \\ Pr\left[-z_{\alpha/2} \leq \frac{\hat{\mu}-\mu}{\sigma/\sqrt{N}} \leq z_{\alpha/2}\right] &= 1-\alpha \\ Pr\left[\hat{\mu} - z_{\alpha/2}(\sigma/\sqrt{N}) \leq \mu \leq \hat{\mu} + z_{\alpha/2}(\sigma/\sqrt{N})\right] &= 1-\alpha \end{aligned} \tag{3.43}$$

Los puntos finales de la ecuación 3.43 son variables aleatorias ya que $\hat{\mu}$ es una variable aleatorio. Entonces el intervalo aleatorio:

$$[\hat{\mu} - z_{\alpha/2}(\sigma/\sqrt{N}), \hat{\mu} + z_{\alpha/2}(\sigma/\sqrt{N})]$$

es un estimador de intervalo, y contiene el parametro poblacional μ con probabilidad $1-\alpha$.

Una Aclaración Importante Cuando hablamos de un estimador de intervalo, **No** podemos decir que el intervalo calculado en un caso particular contiene el parametro verdadero con una probabilidad de $1-\alpha$. Dependiendo del parametro poblacional μ, el intervalo de confianza o lo contiene, o no lo contiene. Cuando calculamos intervalos de confianza, estamos haciendo enunciados de probabilidad acerca del estimador del intervalo, y por lo tanto es correcto decir que tenemos un *intervalo de confianza* de $1-\alpha\%$ para el parametro poblacional μ.

Estimación con Varianza Desconocida

En la derivación del intervalo de confianza anterior, asumimos que σ era una cantidad conocida. En la práctica, generalmente tenemos que estimar la cantidad σ también. Por ejemplo, si nos interesa estimar la latura promedio de una población, es curioso imaginar que no sabemos el promedio, pero sí sabemos su desviación estándar. En lo que queda de esta sección, consideramos el caso mucho más realística de determinar un intervalo de confizana cuando σ es una cantidad desconocida.

En este caso, además de estimar $\hat{\mu}$, necesitamos estimar $\hat{\sigma}$. Cuando estimamos un parametro adicional, ocuapmos un "grado de libertad" adicional. También, ahora si formamos una variable z (parecida a la ecuación 3.42), va a depender de dos variables aleatorias ($\hat{\mu}$ y $\hat{\sigma}$). Una variable $z(\hat{\mu}, \hat{\sigma})$ no va a tener las propiedades convenientes de la $z(\hat{\mu})$ que estamos acustombrudo/as a utilizar.

Para partir, notamos una serie de resultados importantes:

1. El estimador de la varianza: $\tilde{\sigma}^2 = \sum_{i=1}^{N}(Y_i - \hat{\mu})^2/N$ será sesgado (más detalles en clases)
2. Sin embargo, $\hat{\sigma}^2 = \sum_{i=1}^{N}(Y_i - \hat{\mu})^2/(N-1)$ es un estimador insesgado para σ
3. Una variable que es la suma de k variables aleatorias al cuadrado sigue una distribución chi cuadrado con k grados de libertad (χ_k^2)

4. Y una variable aleatoria normal estándarizada divida por la raíz cuadrada de una variable χ^2_k divida por sus grados de libertad (es decir $\sqrt{(\chi^2_k/k)}$) sigue una distribución t con k grados de libertad

Ahora, sea $Y_1, \ldots, Y_N$ una muestra aleatoria de una población con $\mathcal{N}(\mu, \sigma^2)$. Esta vez necesitmos estimar tanto μ como σ. Tenemos las siguientes estimadores insesgados para μ y σ^2, y la distribución entera para $\hat{\mu}$:

$$\begin{aligned} \hat{\mu} &= \sum_{i=1}^{N} Y_i/N \sim \mathcal{N}(\mu, \sigma^2/N) \\ \hat{\sigma}^2 &= \sum_{i=1}^{N} (Y_i - \hat{\mu})^2/(N-1). \end{aligned}$$

Ahora, si definimos la siguiente variable aleatoria, por el hecho 3 del listado anterior tendrá una distribución chi-cuadrado:

$$\frac{(N-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(N-1)}.$$

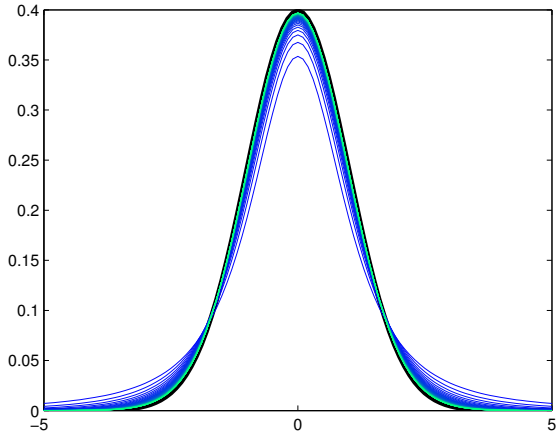
Con todas esas detalles, podemos derivar una variable aleatoria dependiendo solo de cantidades observables (o estimables) y el parametro poblacional de interés que sigue una distribución t de Student:

$$\begin{aligned} t &= \frac{\frac{\hat{\mu}-\mu}{\sigma/\sqrt{N}}}{\left\{\left[\frac{(N-1)\hat{\sigma}^2}{\sigma^2}\right]/(N-1)\right\}^{1/2}} \\ &= \frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{N}} \end{aligned} \tag{3.44}$$

Por el hecho 4, sabemos que t tiene una distribución t con $N-1$ grados de libertad. La distribución t, como la distribución normal estandarizada, tiene una función de densidad de probabilidad estándar que frecuentemente se tabula (ver por ejemplo la Tabla 3.3), y que permite cuantificar de forma muy fácil la masa de probabilidad contenido dentro de cualquier dos valores. Y de hecho, la distribución t se asemaja bastante a la distribución normal estandarizada. En la Figura 3.15, comparamos la función de densidad de la normal estandarizada (la linea negra), y la distribución t de Student variando la cantidad de grados de libertad. Como se observa, la distribución t concentra más masa de probabilidad en las colas de la función de densidad, reconociendo el aumento de la varianza por la estimación de σ^2. La distribución límite de la distribución t (cuando $N \longrightarrow \infty$), es la distribución normal estandarizada.

A partir de la ecuación 3.44 ahora por fin podemos derivar un estimador de intervalo con μ

Figure 3.15: Distribución Normal versus Distribución t con k grados de libertad



Nota: Curvas corresponden a funciones de densidad para la normal estándar (linea negra) y distribuciones t con distintas cantidad de libertad (lineas azules/verdes). Lineas máz azules tienen una cantidad menor de grados de libertad.

y σ desconocida:

$$\begin{aligned} Pr[-t_{(N-1,\alpha/2)} \leq t \leq t_{(N-1,\alpha/2)}] &= 1-\alpha \\ Pr\left[-t_{(N-1,\alpha/2)} \leq \frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{N}} \leq t_{(N-1,\alpha/2)}\right] &= 1-\alpha \\ Pr\left[\hat{\mu} - t_{(N-1,\alpha/2)}(\hat{\sigma}/\sqrt{N}) \leq \mu \leq \hat{\mu} + t_{(N-1,\alpha/2)}(\hat{\sigma}/\sqrt{N})\right] &= 1-\alpha. \end{aligned}$$

Y ésta nos da nuestro estimador de intervalo:

$$\hat{\mu} \pm t_{(N-1,\alpha/2)}(\hat{\sigma}/\sqrt{N})$$

que contiene el parametro desconocido μ con probabilidad $(1-\alpha)$.

3.4.2 Contrastes de Hipótesis

La lógica del intervalo de confianza es encontrar un rango donde, si se replica el experimento subyacente infinitas veces, en $1-\alpha\%$ de las replicaciones el parametro verdadero caerá en el intervalo. Sin embargo, existen otras maneras de considerar un parametro estimado con incertidumbre. Uno que encontraremos en varios momentos de nuestro curso son los contrastes de hipótesis. Un contraste estadístico es un problema de decisión acerca de un parametro θ que está contenido en el conjunto Ω. Este espacio Ω puede estar particionado en dos distintos (y mutuamente excluyentes) sub-espacios Ω_0 y Ω_1. Utilizando nuestra muestra de datos, tenemos que decidir si θ pertenece a Ω_0 o Ω_1. Dado que son espacios mutuamente exluyentes, el parametro tiene que pertenecer a uno, y solo uno de los espacios.

Los contrastes de hipótesis parten de la base de dos hipótesis:

1. **La Hipotesis Nula**: Denotamos a H_0 como la hipótesis nula que $\theta \in \Omega_0$
2. **La Hipotesis Alternativa**: Denotamos a H_1 como la hipótesis alternativa que $\theta \in \Omega_1$.

Y el test estadístico (o contraste de hipótesis) entonces consiste en aceptar una hipótesis (y rechazar la otra) tomando en cuenta (a) Los costos de una decisión incorrecta, y (b) Todos los datos disponibles.

Suponemos que tenemos una muestra aleatoria de datos $\boldsymbol{Y}$ que tiene fdp conjunta $f(\boldsymbol{y}|\boldsymbol{\theta})$. El conjunto de todos los valores posibles de $\boldsymbol{Y}$ es el espacio muestral del experimento. Definamos un procedimiento de prueba que divide el espacio en dos partes: uno conteniendo todos los valores de $\boldsymbol{Y}$ que hacen que se acepta H_0, y el otro donde H_1 será aceptada (y H_0 rechazada). El segundo conjunto se llama la **región de rechazo** del test, ya que nos hará rechazar la hipótesis nula. Este espacio de $\boldsymbol{Y}$ es de dimensión N, y por lo general será muy complejo abarcar un contraste de hipótesis considerando todas las observaciones por separado. Por lo tanto, cuando realizamos un contraste de hipótesis, buscamos formar una **estadística de prueba**, que reduce este espacio de N dimensiones en un valor escalar en $\mathbb{R}$. Por definición una estadística de prueba tiene una distribución conocida bajo la hipótesis nula, y la región de rechazo es el conjunto de valores de la estadística de prueba para las cuales se rechzará la hipótesis nula.

Los Elementos Básicos de un Test de Hipótesis consisten de:

1. Una hipótesis nula, que será tratado como cierto hasta que haya evidencia al contrario
2. Una hipótesis alternativa que se adoptará si la nula está rechazada
3. Una estadística de prueba
4. Una región de rechazo (o "valor crítico")

La hipótesis nula es, de cierta forma, nuestra posición *ex ante*. Es la posición que mantenemos si no encontramos evidencia que sugiere el contrario. La idea de tener una hipótesis que se supone que cumple hasta encontrar evidencia en contra es análgo al principio jurídico de la presunción de inocencia. En econometría, generalmente la nula considera si un parametro (o parametro) toma un valor (o valores) específico(s), que con frecuencia es 0. En este caso, la alternativa es simplemente que el (los) parametro(s) no es (son) igual al valor. Ejemplos comunes incluyen contrastes de un solo parametro: $H_0 : \theta = 0, H_1 : \theta \neq 0$; contrastes de una combinación lineal de parametros: $H_0 : \theta_1 + \theta_2 = 1, H_1 : \theta_1 + \theta_2 \neq 1$; contrastes acerca de varios parametros $H_0 : \theta_1 = 0$ y $\theta_2 = 0, H_1$: por lo menos uno de los parametros no es igual a cero, o contrastes en base a desigualdades: $H_0 : \theta \geq 0, H_1 : \theta < 0$.

Para ilustrar la idea de un test de hipótesis, examinamos un caso particular. Más adelante en el curso, veremos otros ejemplos cuando llegamos a analisar modelos de regresión. Imaginamos que queremos evaluar una hipótesis acerca de un promedio desconocido μ de una población con distribución normal y varianza conocida de $\sigma^2 = 10$. Como vimos anteriormente, aunque el supuesto de varianza conocida es bastante poco creíble, se puede eliminar el supuesto sin tantos

problemas. En este caso, formamos nuestra hipótesis nula y alternativa:

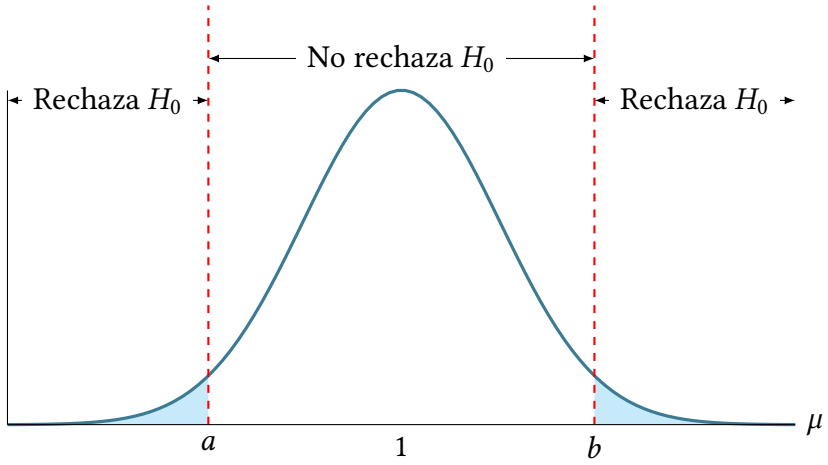
$$H_0 : \mu = 1$$
$$H_1 : \mu \neq 1$$

Aquí elegimos el valor 1 de forma arbitraria, pero por lo general, cuando realizamos un test de hipótesis la nula tiene un vínculo a alguna base teórica. Y imaginamos que tenemos un muestra de tamaño $N = 10$: $y_1, \ldots, y_{10}$. Dada está muestra, sabemos que:

$$\hat{\mu} = \sum_{i=1}^{N} Y_i/N \sim \mathcal{N}(\mu, \sigma^2/N) = \mathcal{N}(\mu, 1)$$

Bajo nuestra hipótesis nula, es decir *imponiendo* que la nula es cierta, tenemos una distribución para $\hat{\mu}$ de la forma presentada en la Figura 3.16. Con la distribución (bajo la nula) en mano,

Figure 3.16: La Distribución de $\widehat{\mu}$ bajo H_0



tenemos que elegir valores para a o b que sugieren que debemos rechazar la nula. Estos valores son elegidos de la base de que si la nula fuese cierta, parece bastante poco probable haber visto un valor tan extremo. Por ejemplo, si observamos un valor de $\hat{\mu} = 100$, parece muy poco probable que este valor habría salido de la distribución en la Figura 3.16, aunque teóricamente *es* probable. Además, los valores a y b dependen de nuestra deseada nivel de significancia α. Por ejemplo, si pensamos que debemos exigir mucha evidencia para poder rechazar la nula, vamos a fijar a y b de tal modo que hay poco masa de probabilidad en las colas de la distribución. Aquí, elegimos a y b de tal modo que:

$$Pr[\hat{\mu} \leq a] = Pr[\hat{\mu} \geq b] = \alpha/2$$

Y, definimos nuestra regla de rechazo:

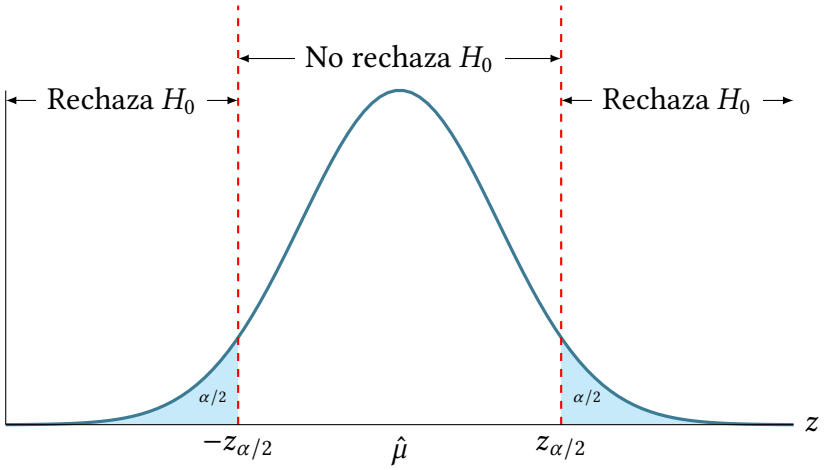
- Rechaza H_0 si $\hat{\mu} \leq a$ o $\hat{\mu} \geq b$
- Acepta (o por lo menos no rechaza) H_0 si $a < \hat{\mu} < b$

Por último, definimos nuestra estadístico de prueba. En el caso de este ejemplo, trabajamos

con la distribución normal estándarizada que conocemos bastante bien:

$$z = \frac{\hat{\mu} - 1}{\sigma/\sqrt{N}}$$

y: $z \sim \mathcal{N}(0, 1)$ si H_0 es cierto. Ahora, utilizando nuestro valor para α, calculamos los valores que corresponden a z...



Por ejemplo, si fijamos $\alpha = 0.05$, podemos buscar los valores críticos en la Tabla estadística 3.2, encontrando valores criíticos de ± 1.96.

¿Por qué rechazamos si $|z| > z_{\alpha/2}$? Dada la naturaleza de una distribución normal, nunca vamos a poder asegurar con certeza que un parametro no es igual a algún valor. Pero aquí, si rechazamos la nula, implica uno de dos casos. Si la nula es cierta, la probabilidad de obtener un valor de z en la región de rechazo es sólo α. Y dado que este evento es poco probable (elegimos un α pequeña), y por lo tanto concluimos que la estadística de prueba probablemente no tiene una distribución $\mathcal{N}(0, 1)$.

Tipos de Errores en un Contraste de Hipótesis Cuando hacemos un test de hipótesis, idealmente rechazamos la hipótesis nula cuando la nula sea falsa, y no rechazamos la nula cuando la alternativa sea falsa. Definimos una funcion $\Pi(\theta)$ donde:

$$\Pi(\theta) = Pr[\text{rechazar } H_0 | \theta]$$

La función $\Pi(\theta)$ es conocido como la función de potencia. Idealmente, tendremos que $\Pi(\theta) = 0 \; \forall \; \theta \in \Omega_0$ y $\Pi(\theta) = 1 \; \forall \; \theta \in \Omega_1$. Sin embargo, generalmente no existen test de hipótesis ideales. En particular, hay dos tipos de errores que podríamos cometer:

Error Tipo I: La probabilidad de rechazar H_0 cuando H_0 es verdadera.

$$P[\text{Error Tipo I}] = P[\text{rechazar } H_0 | H_0 \text{ es cierto}] \leq \alpha$$

Aquí α se conoce como el nivel de significancia del test, y α es el valor más grande de $\Pi(\theta)$ para cualquier valor de θ.

Error Tipo II: La probabilidad de aceptar una hipotesis falsa.

$$\begin{aligned} \beta &= P[\text{Error Tipo II}] = P[\text{aceptar } H_0 | H_1 \text{ es cierto}] \\ &= 1 - \Pi(\theta) \quad \text{para} \quad \theta \in \Omega_1 \end{aligned}$$

No existen test de hipótesis para eliminar (o hacer arbitrariamente pequeño) ambos tipos de errores. Generalmente α está definida como un valor fijo (y pequeño) ya que un error tipo I es más grave que un error tipo II.

3.4.3 Test de la Razón de Verosimilitudes

Por último, antes de terminar esta sección, consideramos el test de Razón de Verosimilitudes. Ésta es na clase generalizada de test de hipótesis, que viene de la estimación por maxima verosimilitud. El proceso general de este test es:

1. Estimar un parametro (o parametros) utilizando máxima verosimilitud
2. Estimar el model utilizando máxima verosimilitud, pero con la restricción definida por la test de hipótesis
3. Comparar los valores de las dos cantidades de la máximia verosimilitud
4. Si el valor de 1 es muy distinto al valor de 2, es probable que el parametro restringido no es el valor correcto para el parametro poblacional

Imaginamos que estamos intersado/as en estimar un modelo con un solo parametro μ. Queremos testear la hipótesis:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned}$$

donde μ_0 es simplemente un valor particular. Entonces, la idea del test de razón de verosimilitudes es que debemos maximizar dos funciones de verosimilitud: una función no restringida, que nos da el estimador $\hat{\mu}_{ML}$, y otra función restringida, que restringimos para dar el valor μ_0. Estos valores estimados traen consigo un valor de la función de verosimilitud, $\ell(\hat{\mu}_{ML})$ y $\ell(\mu_0)$ respectivamente. ¿Qué podemos decir acerca de los dos valores? $\ell(\hat{\mu}_{ML}) \overset{?}{\lesseqgtr} \ell(\mu_0)$

Y la lógica del test es que si las funciones $\ell(\hat{\mu}_{ML})$ y $\ell(\mu_0)$ son muy distintos, haber impuesto que $\mu = \mu_0$ fue una restricción fuerte, y probablemente no tan realista dadas los datos observados. Sin embargo, si $\ell(\hat{\mu}_{ML})$ y $\ell(\mu_0)$ son muy parecidos, es razonable pensar que el valor del parametro podría ser μ_0. La única parte que queda es saber cómo formar la estadística de prueba. El test de

la razón de verosimilitudes sugiere utilizar:

$$2\left[\ell(\hat{\mu}_{ML}) - \ell(\mu_0)\right] \sim \chi^2_{(1)}$$

Y ahora, con la distribución para el estadístico de prueba (una distribución $\chi^2_{(k)}$ donde k es la cantidad de restricciones impuestas por la hipótesis nula), podemos calcular la región de rechazo para cualquier test dado los valores calculados para cada ℓ, una tabla estadística de la distribución χ^2 y el nivel de significancia deseada (α).

Bibliography

Adams, Abi, Damian Clarke, and Simon Quinn. 2015. *Microeconometrics and MATLAB: An Introduction.* Oxford University Press.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications.* Cambridge University Press.

Cameron, A. Colin, and Pravin K. Trivedi. 2009. *Microeconometrics Using Stata.* Stata Press.

Casella, George, and Roger L Berger. 2002. *Statistical Inference.* . 2 ed., Duxberry Thomson.

Davidson, James. 1994. *Stochastic Limit Theory.* Oxford University Press.

DeGroot, Morris H., and Mark J. Schervish. 2012. *Probability and Statistics.* . 4 ed., Addison-Wesley.

Fisher, R. A. 1922. "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 222(594-604): 309–368.

Goldberger, Arthur S. 1991. *A Course in Econometrics.* Harvard University Press.

Golub, G.H., and C.F. Van Loan. 1983. *Matrix Computations.* Johns Hopkins University Press.

Greene, William H. 2002. *Econometric Analysis.* . 5 ed., Pearson.

Hansen, Bruce. 2017. *Econometrics.* Online Manuscript, http://www.ssc.wisc.edu/~bhansen/econometrics/, descargado 26/12/2017.

Manski, Charles F. 2003. *Partial Identification of Probability Distributions.* Springer.

Rao, C. Radhakrishna. 1973. *Linear Statistical Inference and its Applications.* John Wiley and Sons.

Simon, Carl P., and Lawrence Blume. 1994. *Mathematics for Economists.* New York, N.Y.:W. W. Norton & Company, Inc.

Stachurski, John. 2016. *A Primer in Econometric Theory.* The MIT Press.

White, Halpert. 2001. *Asymptotic Theory for Econometricians.* San Diego, Academic Press.

Table 3.2: fda Normal Estandarizada

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 3.3: Valores Críticos de la Distribución t de Student

k	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307
45	0.255	0.434	0.680	0.850	1.165	1.301	1.679	2.014	2.412	2.690	3.281
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090