Unidad 1. Revisión de conceptos estadísticos básicos (Parte 2)

Erika R. Badillo
erika.badilloen@unaula.edu.co

Facultad de Economía
Universidad Autónoma Latinoamericana

En este tema

- Poblaciones, parámetros y muestreo aleatorio
- Propiedades de muestra finita de los estimadores
- Propiedades asintóticas o de muestra grande de los estimadores

Lecturas

- Wooldridge, Jeffrey (2013). *Introducción a la econometría*. 5a edición, Cengage Learning. Apéndice C
- Gujarati, D. y Porter, D. (2010). *Econometría*. 5a edición, Mc Graw Hill. Apéndice A

Motivación

Algunos conceptos...

- Población: grupo bien definido (de personas, empresas, ciudades, etc.) objeto de estudio
- Muestra: parte de la población (representativa)
- Parámetro: característica de la población
- Estadístico: medida cuantitativa característica de la muestra
- Variable: características de cada uno de los elementos de la población (toma diferentes valores)
- Dato: valor de la variable asociada a un elemento de la población o muestra (diferentes tipos de datos)

Poblaciones, parámetros y muestreo aleatorio

- **Inferencia estadística**: involucra saber algo de una población, dada la disponibilidad de una muestra de esa población
  - **Población**: cualquier grupo bien definido de sujetos, que podrían ser individuos, empresas, ciudades o muchas otras posibilidades
    Ej.: estudiantes de econometría I, PEA, los homicidas de Medellín, empresas del sector industrial, equipos de fútbol que pertenecen a la UEFA
  - **Muestra**: Subconjunto que representa la población objetivo (representativa y aleatoria)
  - **Saber**: una estimación o prueba de hipótesis que se hace con la muestra para inferir sobre las características de la población

Ejemplos:

Mroz, T. A. (1987). "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions" *Econometrica*, 55(4):765–799

Población: mujeres casadas que participan en el mercado laboral en los Estados Unidos

Muestra: 753 mujeres casadas con edades entre 30 y 60 años que participan en el mercado laboral en los Estados Unidos en el año 1975

Saber: efectos de los salarios, la educación y la experiencia sobre la oferta laboral femenina, cálculo de los rendimientos a la educación y experiencia

Resultados: "We find substantial measurement error in the average hourly earnings, and this measurement error is negatively correlated with the woman's annual hours of work"

Angrist, J., Bettinger, E., y Kremer, M. (2006). "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia", *American Economic Review*, 96(3):847-862

Población: individuos que se graduaron de la primaria de un colegio público y aplicaron por la beca PACES (125,000 becas)

Muestra: 4044 individuos en Bogotá en 1994 en donde la muestra se asignaron de forma aleatoria. 59 % recibió la beca a través de una lotería

Saber: efecto del programa sobre graduación bachillerato, puntaje del ICFES, salarios, embarazo adolescente, formalidad, impuestos, etc

Resultados: "The empirical results point to an increase in high-school graduation rates of 5 to 7 percentage points, relative to a base rate of 25 to 30 percent"

Poblaciones, parámetros y muestreo aleatorio

- Para hacer inferencia se requiere un conocimiento claro de la población objetivo, y un modelo económico que describa la relación poblacional de interés
- Normalmente, dichos modelos económicos analizan variables aleatorias, las cuales se pueden caracterizar por el espacio muestral y la distribución de probabilidad
- Las distribuciones de probabilidad dependen de parámetros que son desconocidos. Econometría I nos enseña herramientas para estimar dichos parámetros y así saber sobre la población de interés, es decir, podamos hacer inferencia

Muestreo

- Para hacer inferencia estadística de una población objetivo es necesario recurrir al muestreo
- Supongamos que tenemos una va Y que representa una población con una función de densidad de probabilidad $f(y;\theta)$ que se supone conocida y que depende de un sólo parámetro desconocido θ (la distribución normal depende de 2 parámetros, la media y la varianza)
- Diferentes valores de θ implican diferentes distribuciones poblacionales $\Longrightarrow$ lo que interesa conocer es θ
- Obteniendo muestras de la población es posible conocer algo acerca de θ $\Longrightarrow$ el esquema de muestreo más sencillo de manejar es el muestreo aleatorio

Muestreo aleatorio

Si $Y_1, Y_2, ..., Y_n$ son va independientes con una función de densidad de probabilidad común $f(y;\theta)$, entonces se dice que $\{Y_1, ..., Y_n\}$ es una muestra aleatoria de la población representada por $f(y;\theta)$

- Cuando $\{Y_1, ..., Y_n\}$ es una muestra aleatoria de la densidad $f(y;\theta)$ se dice que Y_i son variables aleatorias independientes, identicamente distribuidas (o i.i.d) de $f(y;\theta)$

Propiedades de muestra finita de los estimadores

Muestra finita: proviene del hecho que las propiedades son válidas para muestras de cualquier tamaño, sin importar si es grande o pequeña $\Longrightarrow$ propiedades de muestras pequeñas

Estimador y estimaciones

Estimador: dada una muestra aleatoria $\{Y_1, Y_2, ..., Y_n\}$ extraída de una distribución de la población que depende de un parámetro desconocido θ, un estimador de θ es una regla que asigna a cada resultado posible de la muestra un valor de θ. Dicha regla se define antes de realizar el muestreo

Ejemplo: sea $\{Y_1, Y_2, ..., Y_n\}$ una muestra aleatoria de una población con media μ. Un estimador natural de μ es el promedio de la muestra aleatoria:

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$\bar{Y}$ recibe el nombre de promedio muestral, un estimador de μ. Dado cualquier resultado de las variables aleatorias $\{Y_1, Y_2, ..., Y_n\}$ se utiliza la misma regla para estimar μ: simplemente se promedian

Estimador y estimaciones

- En general un estimador W de un parámetro θ se puede expresar como una formula matemática abstracta que depende de la muestra aleatoria $\{Y_1, Y_2, ..., Y_n\}$:

$$W = h(Y_1, Y_2, ..., Y_n)$$

- W es una va pues depende de la muestra aleatoria: conforme se obtienen diferentes muestras aleatorias de la población, el valor de W puede cambiar

- Cuando un conjunto de números en particular, por ejemplo $\{y_1, y_2, ..., y_n\}$ se inserta en a la función h, se obtiene una estimación de θ, denotada por $w = h(y_1, y_2, ..., y_n)$

- W es una estimador puntual y w una estimación puntual

Estimador y estimaciones

- Para evaluar los procedimientos de estimación, se estudian las diferentes propiedades de la distribución probabilística de la va W
- La distribución de un estimador tiene por nombre distribución de muestreo
- Ya que existen ilimitadas reglas en la combinación de datos para estimar parámetros, se requieren algunos criterios para evaluar y elegir entre estimadores $\Longrightarrow$ en la elección de estimadores es más fácil centrarse en algunas características de la distribución de W que en toda la distribución:
  - Insesgadez
  - Eficiencia relativa

Insesgadez

- La primera propiedad importante de un estimador implica su valor esperado
- Estimador insesgado: un estimador, W de θ, es un estimador insesgado si

$$E(W) = \theta$$

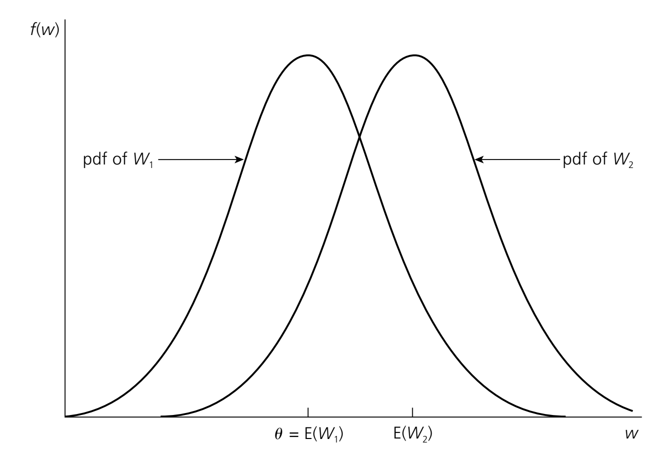
para todos los valores posibles de θ

- Sesgo de un estimador: si W es un estimador sesgado de θ, su sesgo se define como

$$Sesgo(W) = E(W) - \theta$$

Insesgadez

FIGURE C.1 An unbiased estimator, W_1, and an estimator with positive bias, W_2.



Eficiencia

- Eficiencia relativa: si W_1 y W_2 son dos estimadores insesgados de θ, W_1 es eficiente en relación con W_2 cuando $Var(W_1) \leq Var(W_2)$ para todo θ
- Una forma comparar los estimadores que no necesariamente son insesgados, es calcular el error cuadrático medio (ECM):

$$ECM(W) = E[(W - \theta)^2] = Var(W) + [Sesgo(W)]^2$$

- Mide, en promedio, qué tan alejado está el estimador de θ
- Permite comparar dos estimadores cuando uno o ambos son sesgados

Ahora vamos analizar dos propiedades asintóticas o de muestras grandes de los estimadores muy importantes:
- Consistencia
- Normalidad asintótica

Para entender las propiedades asintóticas, tenemos que entender tres conceptos muy importantes:
- La distribución normal
- Teorema del límite central
- La ley de los grandes números

El análisis asintótico implica aproximar las características de la distribución muestral de un estimador. Estas aproximaciones dependen del tamaño de la muestra

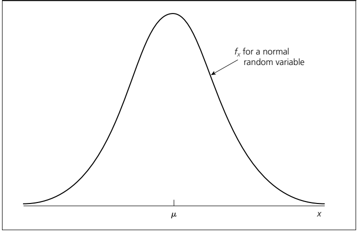
La distribución normal

- La distribución normal es la más utilizada en estadística y econometría
- El supuesto de que una va tiene una distribución normal simplifica el cálculo de probabilidades
- Se dice que X tiene una distribución normal cuyo valor esperado es μ y varianza σ^2
- Como la distribución normal es simétrica respecto a μ, μ es también la mediana de X
- La fdp de X se expresa como:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp[-(x-\mu)^2/2\sigma^2], \ -\infty < x < \infty$$

- Algunas va parecen seguir más o menos una distribución normal: las estaturas y los pesos de los seres humanos, las tasas de desempleo...

FIGURE B.7 The general shape of the normal probability density function.



Teorema del límite central

- Es uno de los resultados más poderosos en probabilidad y estadística
- Indica que el promedio de una muestra aleatoria para cualquier población (con varianza finita) tiene una distribución normal asintótica
- **TCL1**: Sean $X_1, X_2, ..., X_n$ n va independientes, las cuales tienen la misma fdp con media $= \mu$ y varianza $= \sigma^2$. Sea $\bar{X} = \sum X_i/n$ (es decir, la media muestral). Entonces, a medida que n aumenta indefinidamente (es decir, $n \to \infty$),

$$\overset{n \to \infty}{\bar{X}} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Teorema central del límite

Ejemplo: se generan n muestras aleatorias de tamaño 100 y en cada muestra se calcula la media muestral. Cuando n aumenta la distribución se acerca más a una distribución normal

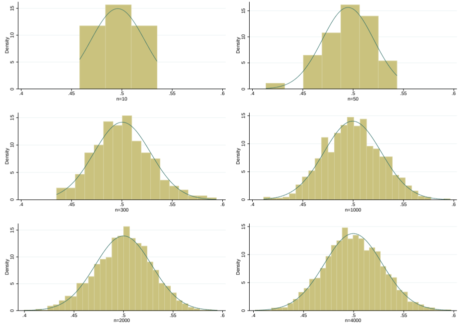
Simulando datos

```
cd "D:\Econometria I\Stata"

foreach rep in 10 50 300 1000 2000 4000 {
clear
save "simula`rep'.dta", replace emptyok
local j 1
while `j'<= `rep' {
clear
qui set obs 100
dis ""
dis as text in y "Repetición `j'"
gen rep = `j'
gen Y1 = runiform()
qui sum Y1
gen Ymedia = r(mean)
qui keep if _n==1
dis as text " Ymedia=" in y Ymedia
keep rep Ymedia
append using simula`rep'.dta
sort rep
qui save "simula`rep'.dta", replace
local j = `j' + 1
}
}
```

```
foreach rep in 10 50 300 1000 2000 4000 {
use simula`rep'.dta, clear
hist Ymedia, normal xlabel(.4(0.05).6) ylabel(0(5)15, gmax) ///
xtitle("n=`rep'") graphregion(color(white) fcolor(white)) ///
saving("n`rep'.gph", replace)
}

graph combine "n10.gph" "n50.gph" "n300.gph" ///
"n1000.gph" "n2000.gph" "n4000.gph", ///
col(3) iscale(*.6) graphregion(color(white) fcolor(white))
```



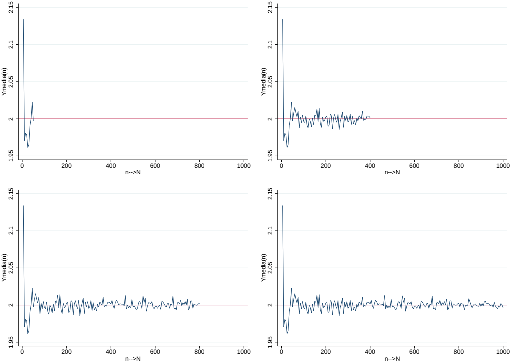
Teorema del límite central

- **TCL2**: Observe que este resultado se cumple sin importar la forma de la fdp. Como resultado, se deduce que
$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N\left(0,1\right)$$
Es decir, Z se acerca a la distribución normal estándar a medida que n aumenta

Ley de los Grandes Números

- La ley de los grandes números significa que, si interesa estimar el promedio poblacional μ, es posible aproximarse a μ si se elige una muestra suficientemente grande
- Sean $Y_1, Y_2, ..., Y_n$ va i.i.d con media μ. Entonces,

$$plim(\bar{Y}_n) = \mu$$



Consistencia

- Es razonable pensar que cuando el tamaño de la muestra aumenta cualquier procedimiento de estimación mejore
- Además se pueden descartar ciertos estimadores inútiles (no mejoran con el incremento de n) al estudiar las propiedades asintóticas o de muestras grandes de los estimadores
- El análisis asintótico implica aproximar las características de la distribución muestral de un estimador. Estas aproximaciones dependen del tamaño de la muestra
- Se sabe que las aproximaciones de muestras grandes funcionan bien para tamaños muestrales tan pequeñas como $n = 20$
- La primera propiedad asintótica de los estimadores concierne a qué tan lejos es probable que esté el estimador del parámetro que se supone estimará, a medida que el tamaño de la muestra aumenta indefinidamente

Consistencia

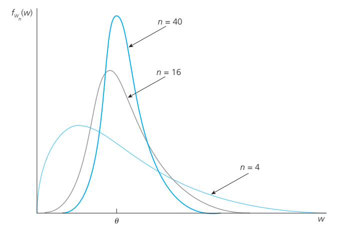
Consistencia

Sea W_n un estimador de θ basado en una muestra $Y_1, Y_2, ..., Y_n$ del tamaño n. Entonces, W_n es un estimador consistente de θ si para toda $\epsilon > 0$,

$$P(|W_n - \theta| > \epsilon) \to 0 \text{ cuando } n \to \infty$$

- Si W_n no es consistente para θ, entonces se dice que es inconsistente
- Cuando W_n es consistente, también se dice que θ es el límite en probabilidad de W_n, escrito como $plim(W_n) = \theta$
- La consistencia de un estimador implica que la distribución de W_n se concentra cada vez más en torno a θ, lo que significa que para tamaños de la muestra mayores, es cada vez menos probable que W_n se aleje mucho de θ

FIGURE C.3 The sampling distributions of a consistent estimator for three sample sizes.



Consistencia

Importante

En econometría o estadística solo nos interesan estimadores consistentes, es un requisito mínimo. Si el estimador W_n es inconsistente entonces no es útil para saber (inferir) sobre θ

En general, los "econometristas" proponen estimadores y argumentan que son consistentes bajo ciertos supuestos, y por lo tanto dichos estimadores serían inconsistentes si dichos supuestos fallan

Los estimadores insesgados no son necesariamente consistentes, y algunos estimadores sesgados pueden llegar a ser consistentes

Normalidad asintótica

- La consistencia es una propiedad que no expresa nada acerca de la forma de la distribución para un tamaño muestral dado
- La mayoría de los estimadores econométricos tienen distribuciones que se aproximan bien mediante una distribución normal para muestras grandes

Normalidad asintótica

Sea $Z_n = \frac{\bar{Y}_n - \mu}{\sigma^2/n}$, $n = 1, 2, ...$ una secuencia de va, tal que para todos los números z,

$$P(Z_n \leq z) \rightarrow \Phi(z) \text{ cuando } n \rightarrow \infty$$

donde $\Phi(z)$ es la función de distribución acumulada normal estándar

- Esta propiedad significa que la función Z_n se acerca cada vez más a la fda de la distribución normal estándar a medida que el tamaño de la muestra n aumenta