Nota 1
Repaso de estadística

1 Introducción

La **econometría** consiste, principalmente, en el uso de métodos estadísticos para estimar relaciones económicas, evaluar teorías y analizar el impacto de políticas públicas.

Un **análisis empírico** consiste en utilizar datos y aplicar dichos métodos estadísticos. En la mayoría de los casos, debe tenerse un modelo teórico para sustentar la relación o hipótesis que se pretende analizar empíricamente.

El uso de datos frecuentemente conlleva ciertas restricciones en el alcance que el análisis empírico puede tener. Por ejemplo, las bases de datos en algunas ocasiones no incluyen información relevante para llevar a cabo algunos análisis, los individuos encuestados pueden no ser representativos de la población que se pretende analizar, los individuos encuestados pueden reportar incorrectamente la información, etc.

Las preguntas de mayor interés en el análisis econométrico suelen involucrar el análisis del **efecto causal** que tiene una variable sobre otra. Por ejemplo, podemos estar interesados en el efecto de la implementación de una política educativa sobre la escolaridad, el efecto de shocks en el tipo de cambio sobre decisiones de consumo de los hogares, o el efecto de aplicar fertilizante sobre la eficiencia en producción agrícola de los hogares. Para esto, es importante poder aislar el efecto del cambio en una variable sobre otra sin que cambios en variables adicionales puedan estar afectando dicha relación. En otras palabras, estamos interesados en analizar la relación entre dos variables *caeteris paribus*.

2 Probabilidad

Esta sección consiste es un repaso muy breve y general de conceptos de probabilidad que se utilizarán a lo largo del curso.[1] En general, cuando hablamos de probabilidad es usual distinguir entre variables discretas y continuas. En términos prácticos, las **variables discretas** son aquellas que toman un pequeño número de valores posibles. Por su parte, las **variables continuas** son aquellas que pueden tener un número infinito de valores posibles. Usualmente, se les identifica como variables que toman un valor dentro de cierto rango de valores posibles.

Las variables suelen ser descritas por una **función de densidad (pdf)**, la cual describe la probabilidad de que una variable tome cierto valor: $f(x_j) = P(X = x_j) = p_j$. Dicha función de densidad suele estar ligada a la **función de densidad acumulada (cdf)**, que describe la probabilidad de que una variable tenga un valor menor o igual a cierto número: $F(x_j) = P(X \leq x_j)$.

Cuando involucramos en nuestro análisis más de una variable, es común hacer referencia a la **función de densidad conjunta**: $f_{X,Y}(x_j, y_k) = P(X = x_j, Y = y_k)$ y a la **función de densidad marginal**: $f_X(x_j) = P(X = x_j)$. En el caso en que las variables X y Y sean independientes: $f_{X,Y}(x_j, y_k) = f_X(x_j)f_Y(y_k)$. Asimismo, con más de una variable, será común referirnos a la **densidad condicional**: $f_{Y|X}(y_k|x_j) = \frac{f_{X,Y}(x_j,y_k}{f_X(x_j)}$. [En adelante, para simplificar únicamente utilizaremos x en vez de x_j y y en vez de y_k. El único propósito de untilizar x_j y y_k era para indicar que son valores específicos, i.e. algún número. En adelante, cuando utilicemos mínúsculas estaremos haciendo referencia a valores específicos.]

Valor esperado. El valor esperado es una medida de tendencia central. Si X es una variable discreta que toma k distintos posibles valores, tendremos que:

[1] Aquellos alumnos que sientan que necesitan un repaso más a detalle, se recomienda que revisen el capítulo 2 del *Stock y Watson* o el apéndice B del *Wooldridge*.

$$E(X) = \sum_{j=1}^{k} x_j f(x_j) \tag{1}$$

Asimismo, en el caso continuo:

$$E(X) = \int_{-\infty}^{\infty} x f(x) \tag{2}$$

Propiedades del valor esperado:

a) $E(a) = a$, donde a es una constante

b) $E(aX + bY + c) = aE(X) + bE(Y) + c$, donde b, c son constantes

Varianza. La varianza describe, en promedio, que tan lejos suele estar una variable del valor esperado ($\mu = E(X)$).

$$\sigma^2 = Var(X) = E[(X - \mu)^2] = E(X^2) - \mu^2 \tag{3}$$

La desviación estándar es simplemente: $\sigma_X = \sqrt{Var(X)}$.

Propiedades de la varianza:

a) $Var(a) = 0$

b) $Var(aX + b) = a^2 Var(X)$

c) $Var(aX \pm bY) = a^2 Var(X) + b^2 Var(Y) \pm abCov(X, Y)$

Covarianza. La covarianza mide la relación entre dos variables. Una covarianza positiva indica que ambas variables suelen moverse en la misma dirección. Una covarianza negativa indica lo contrario. (Sea $E(X) = \mu_X$ y $E(Y) = \mu_Y$)

$$\sigma_{X,Y} = Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{4}$$

Propiedades de la covarianza:

a) $Cov(X, Y) = 0$ si X y Y son independientes

b) $Cov(aX + b, cY + d) = acCov(X, Y)$, donde a, b, c, d son constantes

El coeficiente de correlación es simplemente: $\rho_{X,Y} = corr(X, Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

3 Notación

El análisis econométrico empírico generalmente involucra llevar a cabo una inferencia a partir de los datos a los cuales se tiene acceso (la **muestra**) acerca del comportamiento de cierta población.

La **población** es un grupo definido de agentes que pueden ser individuos, empresas, ciudades u otra unidad de observación.

La inferencia generalmente conlleva hacer una estimación o una prueba de hipótesis utilizando datos de la muestra con el objetivo de obtener conclusiones acerca de la población.

Es importante conocer los detalles de la estrategia utilizada para recabar la muestra. Generalmente, nos referiremos al uso de muestras aleatorias donde cada uno de sus componentes se selecciona de forma independiente y proviene de una distribución común $\{Y_1, \ldots, Y_n\}$. En este caso se dice que Y_i es una **variable aleatoria independiente e idénticamente distribuida (i.i.d.)**. Las variables aleatorias $\{Y_1, \ldots, Y_n\}$ son variables desconocidas. Una vez que la muestra es recabada tendremos un conjunto de números $\{y_1, \ldots, y_n\}$, los cuales utilizaremos para llevar a cabo inferencia.

Ejemplo: Supón que nuestra variable de interés es la media de la edad de los alumnos cursando educación superior en México.

- ¿Cuál es la población?
- ¿Cómo recabarías una muestra para poder obtener hacer una estimación válida?

- ¿Qué sucedería si yo recabo una muestra basada en los alumnos del ITAM o de la clase?

Algunas otras definiciones que serán de utilidad en el curso:

Un **parámetro** es una constante poblacional que describe la relación entre una o más variables. Frecuentemente, será la variable de interés que buscaremos estimar utilizando los datos de la muestra.

Un **estimador** es una fórmula que asigna un valor a cualquier posible combinación que resulte de la muestra recabada. El estimador, al igual que la variable aleatoria, no tiene un valor específico.

Un **valor estimado** resulta de tomar los datos observados y aplicar la fórmula del estimador. El valor estimado es la contraparte muestral del estimador.

Ejemplo: Continuando con el ejemplo de la media de edad de alumnos en educación superior en México: supongamos a partir de aquí que se recaba una muestra es aleatoria $\{Y_1, \ldots, Y_n\}$

- Parámetro: μ. Es un número específico que típicamente no es conocido. En este caso, es la media de la edad de todos los alumnos cursando educación media superior en México.
- Estimador: $W = h(Y_1, \ldots, Y_n)$. Es una fórmula que genera un estimador del parametro.
- Valor estimado: $w = h(y_1, \ldots, y_n)$. Es un número específico que resulta de aplicar la fórmula del estimador a la muestra recabada.

El estimador (W) de un parámetro (μ) es **insesgado** si su valor esperado es igual al parámetro: $E(W) = \mu$

Si un estimador es insesgado, esto no quiere decir que el valor estimado será igual al parámetro (o incluso cercano a éste), ya que esto dependerá de la muestra que sea recabada.

Un estimador (W_1) es más **eficiente** relativamente a otro (W_2) si $Var(W_1) \leq Var(W_2)$ para todos los valores del parámetro, con desigualdad estricta para al menos un valor del parámetro.

Ejemplo: Consideremos dos estimadores para la media:

a) El valor promedio: $W_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$

b) El promedio de edad de dos personas de mi muestra elegidas al azar: $W_2 = \frac{Y_a + Y_b}{2}$

¿Son estos estimadores insesgados?, ¿cuál es más eficiente?, ¿qué valor estimado estará más cerca de la media poblacional?

4 Propiedades asintóticas de los estimadores

En el ejemplo anterior podemos deducir que el estimador W_1 se vuelve más eficiente conforme el tamaño de muestra aumenta. Las propiedades asintóticas de los estimadores son aquellas que aplican cuando se tienen muestras *grandes*. Sin embargo, no es claro de qué tamaño necesita ser el número de observaciones (n) para que la muestra sea considerada como *grande* y sea correcto aplicar las propiedades asintóticas a los estimadores. Generalmente, esto depende de la distribución poblacional de la variable de interés, pero en la mayoría de los casos en los que utilizamos encuestas, aplicar propiedades asintóticas será razonable.

Ejemplo: Dar un ejemplo mostrando dos distribuciones normales, una más dispersa que la otra.

Consistencia. Sea W_n un estimador del parámetro μ basado en una muestra $\{Y_1, \ldots, Y_n\}$ de tamaño n. W_n será un estimador consistente de μ si para todo $\epsilon > 0$:

$$P(|W_n - \mu| > \epsilon) \to 0 \quad \text{mientras } n \to \infty \tag{5}$$

Si W_n es un estimador consistente de μ también se dice que la probabilidad límite de W_n es μ:

$$plim(W_n) = \mu \tag{6}$$

La consistencia se refiere al comportamiento de la distribución muestral del estimador conforme el tamaño de la muestra se incrementa. Esto, traducido a términos más intuitivos, quiere decir que conforme aumentamos el tamaño de la muestra, la distribución de W_n se volverá más concentrada alrededor de μ. Por lo tanto, mientras mayor es la muestra, menos probable es que nuestro estimador se ubique lejos de μ.

Ejemplo: Es el estimador W_2 consistente?, es el estimador W_1 consistente?

Ley de Grandes Números (LGN). Sean $\{Y_1, \ldots, Y_n\}$ variables aleatorias i.i.d. con media μ. Entonces,

$$plim(\overline{Y}_n) = \mu \tag{7}$$

donde $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Intuitivamente, la ley de grandes números nos dice que si queremos aproximarnos a la media poblacional, podemos hacerlo en gran medida si elegimos muestras lo suficientemente grandes. Sin embargo, utilizando la LGN únicamente obtenemos estimadores puntuales y no tenemos información acerca de su distribución.

Algunas propiedades útiles ligadas a LGN:

1. Sea θ un parámetro. Sea $\kappa = g(\theta)$ un nuevo parámetro, donde $g(.)$ es una función continua. Supongamos que W_n es un estimador tal que $plim(W_n) = \theta$. Si definimos un estimador $G_n = g(W_n)$, dicho estimador será un estimador consistente de κ:

$$plim(G_n) = \kappa$$

2. Si suponemos que $plim(T_n) = \theta$ y $plim(U_n) = \kappa$, entonces:

   (a) $plim(T_n + U_n) = \theta + \kappa$

   (b) $plim(T_n * U_n) = \theta * \kappa$

   (c) $plim(T_n/U_n) = \theta/\kappa$, si $\kappa \neq 0$

Teorema Central del Límite (TCL). Sea $\{Y_1, \ldots, Y_n\}$ una muestra aleatoria con media μ y varianza σ^2. Entonces,

$$Z_n = \frac{\sqrt{n}(\overline{Y}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1) \text{ conforme } n \to \infty \tag{8}$$

Intuitivamente, este resultado indica que, sin importar la distribución poblacional de Y, la distribución de la variable Z_n (que es la versión estandarizada de $\overline{Y}_n$) se aproxima en gran medida a una distribución normal estándar (N(0,1)) conforme el tamaño de la muestra (n) aumenta.

Algunas propiedades útiles ligadas al TCL:
Teorema de Slutsky. Si X_n converge en distribución a X y $plim(Y_n) = \theta$, entonces:

1. $Y_n X_n$ converge en distribución a θX
2. $X_n + Y_n$ converge en distribución a $X + \theta$

5 Pruebas de hipótesis

En la gran mayoría de las aplicaciones empíricas de econometría tendremos que llevar a cabo pruebas de hipótesis. Generalmente, durante el curso asumiremos que las muestras son grandes y, por tanto, podemos aplicar propiedades asintóticas. Esto quiere decir que en la mayoría de los casos podremos utilizar la distribución normal y ji-cuadrada para llevar a cabo las pruebas de hipótesis.

Para repasar cómo llevar a cabo pruebas de hipótesis supongamos que estamos interesados en evaluar si la media de edad de los alumnos en educación media superior en México es igual a 20. Cabe señalar que nuestra hipótesis es acerca del valor de un parámetro poblacional y utilizaremos una muestra para evaluar dicha hipótesis. En nuestro ejemplo, nuestra prueba de hipótesis establece la siguiente hipótesis nula:

$$H_0 : \mu = 20$$

La hipótesis alternativa se establece para especificar la zona de rechazo de la hipótesis nula. Generalmente, en una prueba se busca rechazar la hipótesis nula en favor de la

alternativa. En nuestro caso, la hipótesis alternativa será:

$$H_1 : \mu \neq 20$$

La hipótesis alternativa puede también ser establecida como $\mu > 20$ (o $\mu < 20$). Este sería el caso si lo que nos interesa es evaluar si la media poblacional es mayor (menor) a 20. Es importante recordar que como resultado de la prueba de hipótesis, la hipótesis nula puede ser rechazada o no rechazada. Sin embargo, es incorrecto decir que es aceptada. Formalmente, en el ejemplo anterior podríamos concluir ya sea que: (i) hay evidencia suficiente para rechazar que la media poblacional es igual a 20 con $x\%$ de significancia, o que (ii) no hay evidencia suficiente para rechazar que la media poblacional es igual a 20 con $x\%$ de significancia.

Para establecer el nivel de significancia $x\%$ (o alternativamente el nivel de confianza $[1-x]\%$), hay que tomar en cuenta los dos tipos de errores que podemos cometer al evaluar pruebas de hipótesis:

a Error tipo I: Podemos rechazar la hipótesis nula siendo que esta es verdadera

b Error tipo II: Podemos no rechazar la hipótesis nula siendo esta falsa

Típicamente, el nivel de significancia se establece basado en el error tipo I, que generalmente busca reducirse en las pruebas de hipótesis. Dada nuestra notación, el nivel de significancia se define como:

$$x\% = Pr(\text{Rechazar } H_0 | H_0) = Pr(\text{Error tipo I})$$

El valor de $x\%$ será una valor que tendremos que asumir para llevar a cabo la prueba de hipótesis. El valor más común es de 0.05 (o 5%) de significancia, seguido de 0.01 y 0.1 (lo cual es equivalente a 95%, 99% y 90% de nivel de confianza, respectivamente).

El error tipo II suele estar ligado al poder estadístico. Más adelante en el curso discutiremos cómo utilizar el poder estadístico para determinar el número de observaciones que se requieren para llevar a cabo un análisis estadístico experimental.

Supongamos por el momento que elegimos un nivel de significancia de 5% y que nuestra muestra de estudiantes mexicanos es aleatoria y consta de 1000 individuos. Supongamos que el promedio muestral de edad es de 21.5 y la varianza de las edades de 500.

Tenemos 3 alternativas para llevar a cabo la prueba de hipótesis:

5.1 Estadístico t

Para utilizar este método utilizamos la media y la desviación estándar estimada. La idea es asumir que la hipótesis nula es verdadera. Dado que asumimos esto, querremos determinar qué tan probable es que observemos la variable aleatoria $\overline{Y}$ suponiendo que dicha variable tiene una distribución normal (por propiedades asintóticas) con media 20 y desviación estándar $\sqrt{500/1000}$.

Tomando dichos supuestos estandarizamos la observación de $\overline{Y}$ que tenemos y esto es nuestro estadístico t:

$$\begin{aligned} t =& \frac{(\overline{Y} - \mu_0)}{\sqrt{S^2/n}} \\ & \frac{(21.5 - 20)}{\sqrt{500/1000}} = 2.1213 \end{aligned} \tag{9}$$

Una vez que tenemos dicho estadístico utilizamos la distribución de la normal y comparamos este valor con un valor que esta en el límite de ser razonable (el valor crítico). Para ello empleamos el nivel de significancia. Utilizando un 5% de significancia determinamos que valor crítico (en términos absolutos) representa el 95% del cdf de la distribución normal estándar. Dicho valor (1.96) se compara con el estadístico t para determinar qué tan probable es observar dicho estadístico t dada la distribución que hemos asumido. Dado que el estadístico-t es mayor que el valor crítico rechazamos la hipótesis nula con un 5% de significancia.

5.2 Intervalo de confianza

Si nuevamente utilizamos un nivel de significancia de 5%, necesitaremos determinar un intervalo de confianza del 95%. Dicho intervalo de confianza se genera para el parámetro poblacional μ:

$$\begin{aligned} Pr\left(-1.96 < \frac{\sqrt{n}(\overline{Y}-\mu)}{S} < 1.96\right) = 0.95 \\ Pr\left(\overline{Y} - 1.96 * \frac{S}{\sqrt{n}} < \mu < \overline{Y} + 1.96 * \frac{S}{\sqrt{n}}\right) = 0.95 \end{aligned} \tag{10}$$

Todos los valores anteriores son conocidos, excepto μ. El intervalo lo podemos interpretar como: de cada 100 muestras aleatorias que obtengamos, 95% de ellas tendrán al valor real del parámetro poblacional μ. No podemos decir que una vez que calculemos el intervalo, con 95% de probabilidad éste contendrá el valor real del parámetro. Recordemos que el parámetro es un valor específico (no aleatorio), por lo tanto, se encuentra o no en el intervalo.

En nuestro caso, el intervalo de 95% será: $[20.114, 22.88]$. Dado que 20 no se encuentra dentro del intervalo, rechazamos la hipótesis nula con un 5% de significancia.

5.3 Valor-p

El valor-p nos dice hasta qué nivel de significancia la hipótesis nula sería rechazada. Siempre que el nivel de significancia sea mayor al valor-p, la hipótesis nula sería rechazada. Por lo tanto, este valor usualmente se compara con 0.01, 0.05 y 0.1. En nuestro ejemplo:

$$\text{valor-p} = 2 * (1 - F(|t|)) = 0.034 \tag{11}$$

[Es importante, tener en consideración que en el caso en que la hipótesis alternativa sea unilateral (one-sided), valor-p=$(1 - F(|t|))$.] En el ejemplo, la hipótesis nula no sería rechazada con un 1% de significancia, pero como vimos en los incisos anteriores, sí es rechazada con un 5% de significancia.

6 Bootstrap

En el caso de las pruebas de hipótesis anteriores estamos basando los resultados en el teorema central del límite. Para aplicar el TCL calculamos analíticamente la varianza de la media como:

$$Var(\overline{Y}) = \frac{\sigma^2}{n} \tag{12}$$

donde σ^2 es la varianza de Y_i.

En el caso de TCL asumimos que una versión estandarizada de $\overline{Y}$ (Z_n) converge en distribución a una normal estándar. Una alternativa a este procedimiento consiste en generar una distribución empírica de $\overline{Y}$ y utilizar dicha distribución para calcular la varianza. Un problema con esta idea radica en que para generar una distribución empírica de $\overline{Y}$ necesitamos varias observaciones de $\overline{Y}$.

El método de bootstrap genera diversas observaciones partiendo de una muestra aleatoria $\{Y_1, \ldots, Y_n\}$ siguiendo los siguientes pasos:

1. Utilizando las observaciones de la muestra, elige una submuestra aleatoria de tamaño n (mismo tamaño que la muestra) con reemplazo. Esto quiere decir que habrá observaciones que se repitan más de una vez
2. Usando la submuestra calcula el estimador ($\overline{Y}$ en nuestro ejemplo)
3. Repite los pasos anteriores M veces. Con esto tendrás M observaciones para $\overline{Y}$: $\{\overline{Y}_1, \ldots, \overline{Y}_M\}$ y habrás generado una distribución empírica
4. Genera los estimadores:

$$\begin{aligned} E(\overline{Y}) =& \frac{1}{M}\sum_{k=1}^{M}\overline{Y}_k \\ Var(\overline{Y}) =& \frac{1}{M}\sum_{k=1}^{M}(\overline{Y}_k - E(\overline{Y}))^2 \end{aligned} \tag{13}$$

5. Utiliza dichos estimadores para llevar a cabo pruebas de hipótesis

Este método puede ser aplicado con la mayoría de los estimadores que veremos durante el curso. Es un método de gran utilidad siempre que sea dificil calcular una varianza para llevar a cabo pruebas de hipótesis. En particular podría utilizarse para calcular errores estándar de los coeficientes en una regresión. En dicho caso los pasos a seguir son los mismos que los descritos anteriormente. Lo que sucedería en el caso de una regresión de mínimos cuadrados ordinarios es que se llevaría a cabo una regresión con cada una de las submuestras elegidas en el primer paso. Con ello se obtendrían M posibles coeficientes para cada variable. El error estándar podría calcularse como la desviación estándar para cada uno de los coeficientes.

Existe también la posibilidad de utilizar una submuestra de tamaño menor al tamaño de la muestra original (n). En dicho caso, tendría que llevarse a cabo un ajuste al cálculo de la varianza. Supongamos que se eligen submuestras de tamaño L. Todos los pasos serían los mismos que los descritos anteriormente, con la diferencia que el estimador de la varianza se calcularía como:

$$Var(\overline{Y}) = \frac{L}{n}\frac{1}{M}\sum_{k=1}^{M}(\overline{Y}_k - E(\overline{Y}))^2 \tag{14}$$