PUBLICACIONES DE 4er CURSO

Curso: 4º

Grado: Economía

Asignatura: ECONOMETRÍA III

TRANSPARENCIAS PARTE 1
TEMA 1: Fundamentos
TEMA 2: Revisión de conceptos estadísticos

Profesores: Antonio Aznar

Departamento de ANÁLISIS ECONÓMICO

Curso Académico 2015/16

Facultad de Economía y Empresa
Universidad Zaragoza

TEMA 1: FUNDAMENTOS

1. Preguntas relevantes
2. Efectos causales y experimentos
3. Tipos de datos
4. Usos de los modelos econométricos

1. Preguntas relevantes

Pregunta #1¿Mejora la reducción del tamaño de las clases la educación en la escuela primaria?

Datos sobre 420 distritos escolares en California en 1999.

Pregunta #2 ¿Existe discriminación racial en el mercado de prestamos para la vivienda?

Datos del banco de la Reserva Federal de Boston. Se observó que se rechazó el 28% de las solicitudes de negros y el 9% de blancos.

Pregunta #3 ¿Cuánto reducen los impuestos sobre cigarrillos el tabaquismo?

Datos: ventas de cigarrillos, precios, impuestos y renta personal

Pregunta #4 ¿Cuál será la tasa de inflación del año que viene? Datos de cualquier país.

Añadir más preguntas

2. **Efectos causales y Experimentos**

- Se dice que una acción causa un resultado si el efecto es el resultado directo, o consecuencia, de esta acción. La causalidad significa que una acción específica (aplicar fertilizante) conlleva una consecuencia específica (más tomates)
- Un modo de medir este efecto causal es mediante un experimento aleatorizado controlado. Es controlado porque existe un grupo, que se llama de control, que no recibe tratamiento, y un grupo de tratamiento, que si que lo recibe. **En este experimento, la única razón sistemática para la diferencia en los resultados entre los grupos de tratamiento y de control es el tratamiento en sí mismo.**

- En la práctica es difícil llevar a cabo experimentos ideales.
- Si el tratamiento en un experimento aleatorizado es binario, entonces el efecto causal puede ser estimado por la diferencia en las medias muestrales entre los grupos de tratamiento y de control. La hipótesis de que el tratamiento es ineficaz es equivalente a la hipótesis de que ambas medias son iguales. Por ejemplo, en medicina se hace así. Pero en economía es difícil llevar a cabo este tipo de experimentos. Por esta razón, los económetras a veces realizan "experimentos naturales" o cuasi experimentos que tratan de aproximarse a los experimentos ideales. **Ver recuadro de jubilaciones en página 64, SW.**

3. Tipos de datos

- **Datos Experimentales versus Datos Observacionales.** Experimento del estado de Tenessee.

- **Datos de Sección Cruzada.** Calificaciones de examen en California. Tabla 1.1

- **Datos de Series Temporales.** Inflación y desempleo en USA. Tabla 1.2.

- **Datos de Panel o Logitudinales.** Venta de cigarrillos en diferentes estados de USA en los años 1985-1995.Tabla 1.3.

3. Uso de los modelos

Suplemento docente 1: Uso de los modelos econométricos.

TEMA 2. Revisión de conceptos estadísticos.
Indice

1.Variable Aleatoria y Distribución de Probabilidad.

2. Momentos: Esperanza y Varianza.

3. Distribución condicional.

4. Estimación.

5. Contrastes e intervalos de confianza.

1. Variable Aleatoria y Distribución de Probabilidad.

- **Probabilidad:** de un resultado es la proporción de veces que el resultado ocurre en el largo plazo.
- **Espacio muestral y sucesos:** El conjunto de todos los posibles resultados se denomina espacio muestral. Un suceso es un subconjunto del espacio muestral.
- **Variable Aleatoria:** Es un resumen numérico de un resultado aleatorio. Discretas o continuas.
- **Distribución de Probabilidad:** de una variable aleatoria discreta es una relación de todos los valores posibles de la variable junto con la probabilidad de que ocurra cada valor. Si continua, **Función de Densidad.**

- **Distribución de Probabilidad Acumulada:** es la probabilidad de que la variable aleatoria sea menor o igual a un valor concreto.

- **Distribución de Bernouilli:** Es una variable aleatoria discreta binaria, es decir, los valores son 1 y 0. Toma el valor 1 con probabilidad p y el valor 0 con probabilidad 1-p.

2. Momentos: Esperanza y Varianza

- **Esperanza:** de una variable aleatoria Y, denominada E(Y) es el valor medio de largo plazo de la variable aleatoria a lo largo de muchos intentos.

Discreta: $$E(Y) = \sum_{1}^{n} p_i Y_i$$

Continua: $$E(Y) = \int_{0}^{1} Yf(Y)dY$$

Esperanza de una variable Bernouilli:

$$E(G) = 1 \times p + 0(1 - p) = p$$

La Desviación Típica y la Varianza: La varianza y la desviación típica miden la dispersión de una distribución de probabilidad.

$$Var(Y) = \sigma_Y^2 = E\left[\left(Y - \mu_Y\right)^2\right]$$

$$DT(Y) = \sqrt{\sigma_Y^2}$$

Varianza de una variable Bernouilli

$$Var(G) = \left(0 - p\right)^2 \times (1 - p) + \left(1 - p\right)^2 \times p = p(1 - p)$$

Media y Varianza de una función lineal de variables aleatorias.

Otras Medidas de Forma

$$\textbf{\textit{Skewness (Asimetría)}} = \frac{E\left[\left(Y-\mu_Y\right)^3\right]}{\sigma_Y^3}$$

= medida de asimetría de una distribución

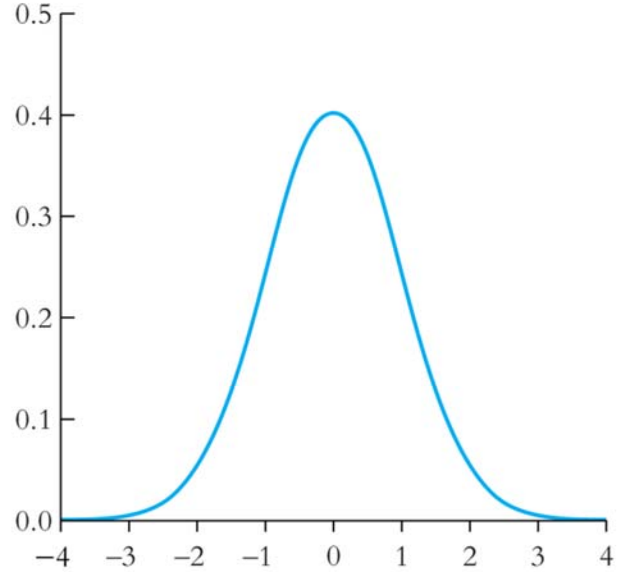
- *skewness* = 0: la distribución es simétrica
- *skewness* > (<) 0: la distribución tiene larga cola a la derecha (izquierda)

$$\textbf{\textit{Kurtosis (Apuntamiento)}} = \frac{E\left[\left(Y-\mu_Y\right)^4\right]}{\sigma_Y^4}$$

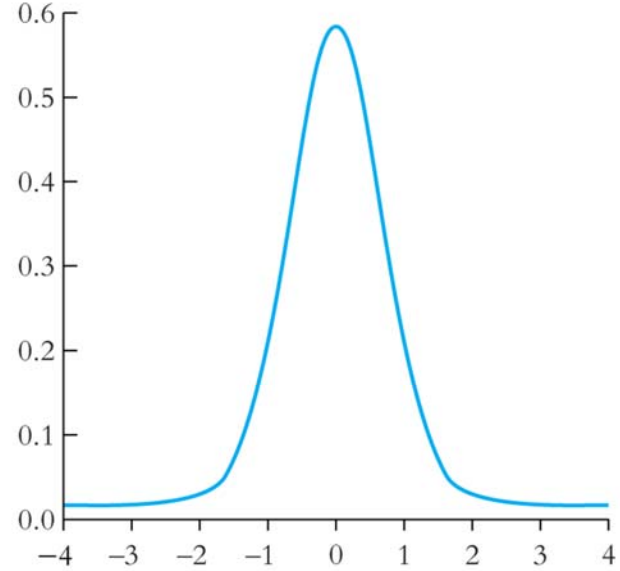
= medida de probabilidad en las colas

= medida de probabilidad de valores grandes

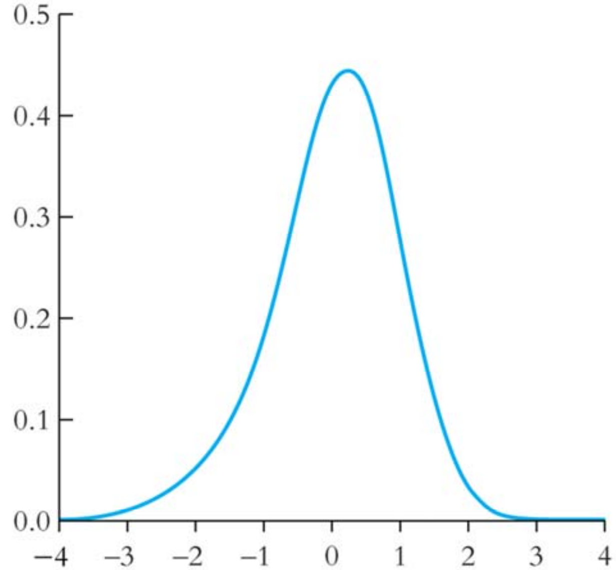
- *kurtosis* = 3: distribución normal
- *skewness* > 3: colas pesadas (“***leptokurtotic***”**)**



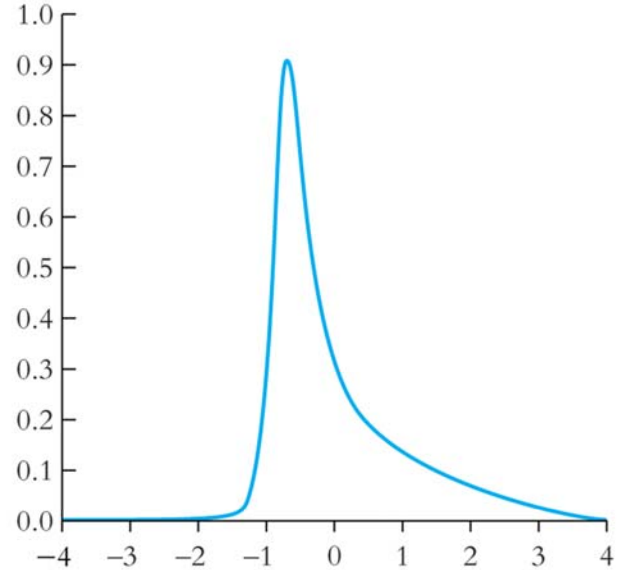
(a) Skewness = 0, kurtosis = 3



(b) Skewness = 0, kurtosis = 20



(c) Skewness = −0.1, kurtosis = 5



(d) Skewness = 0.6, kurtosis = 5

Dos variables aleatorias

Distribución Conjunta y Distribución Condicional. La distribución de probabilidad conjunta de dos variables aleatorias discretas, X e Y, es la probabilidad de que las dos variables aleatorias tomen valores concretos de forma simultanea.

	Lluvia (X = 0)	**Sin lluvia (X = 1)**	**Total**
Desplazamiento largo (Y = 0)	0,15	0,07	0,22
Desplazamiento Corto (Y = 1)	0,15	0,63	0,78
Total	0,30	0,70	1,0

Distribución conjunta: Pr[X=x,Y=y]; Pr[X=0,Y=0]=0,15

Distribución Marginal: Pr[Y=y]=$\sum_{1}^{k} \Pr[X = x_i, Y = y]$. Las

Distribuciones marginales aparecen en la fila y columna Total.

3. Distribución condicional

Distribución Condicional: Es la distribución de una variable aleatoria Y condicionada a que otra X tome un valor específico.

$$\Pr\left[Y = y / X = x\right] = \frac{\Pr\left(X = x, Y = y\right)}{\Pr\left(X = x\right)}$$

En el ejemplo, la distribución condicional de Y dado que X=1

$$\Pr(Y=0/X=1) = 0,07/0,70 = 0,1$$

$$\Pr(Y=1/X=1) = 0,63/0,70 = 0,9$$

Esperanza Condicional: o media condicional de Y dado X es la media de la distribución condicional de Y dado X. En el ejemplo,

$$E(Y/X=1) = 0 \times 0,1+1 \times 0,9 = 0,9$$

Varianza Condicional: Es la varianza de la distribución condicional de Y dado X. En el ejemplo,

$$Var\left(Y / X = 1\right) = \left(0 - 0,9\right)^2 \times 0,1 + \left(1 - 0,9\right)^2 \times 0,9 = 0,081 + 0,009 = 0,09$$

La ley de esperanzas iteradas

La media de Y es la media ponderada de la esperanza condicional de Y dado X, ponderada por la distribución de probabilidad de X.

$$E(Y)=\sum_{1}^{n} E(Y / X = x_i)P(X = x_i)$$

Una expresión equivalente es:

$$E(Y) = E(E(Y/X))$$

La ley de esperanzas iteradas implica que si la media condicional de Y dado X es cero, entonces la media de Y es cero. Esta ley se extiende al caso en que la esperanza condicional es sobre varias variables aleatorias.

Covarianza y Correlación

- Suponemos que las variables aleatorias X y Z tienen una distribución conjunta de probabilidad.
- La covarianza entre X y Z es:

$$Cov\left(X,Z\right)=\sigma_{XY}=E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]$$

- La covarianza es una medida de asociación lineal entre X y Z; sus unidades son unidades de $X \times$ unidades de Z.
- cov(X,Z) > 0 significa que existe una relación positiva entre X y Z.
- Si X y Z se distribuyen independientemente, entonces cov(X,Z) = 0 (pero no vice versa!!).
- La covarianza de una v.a. consigo mismo es su varianza:

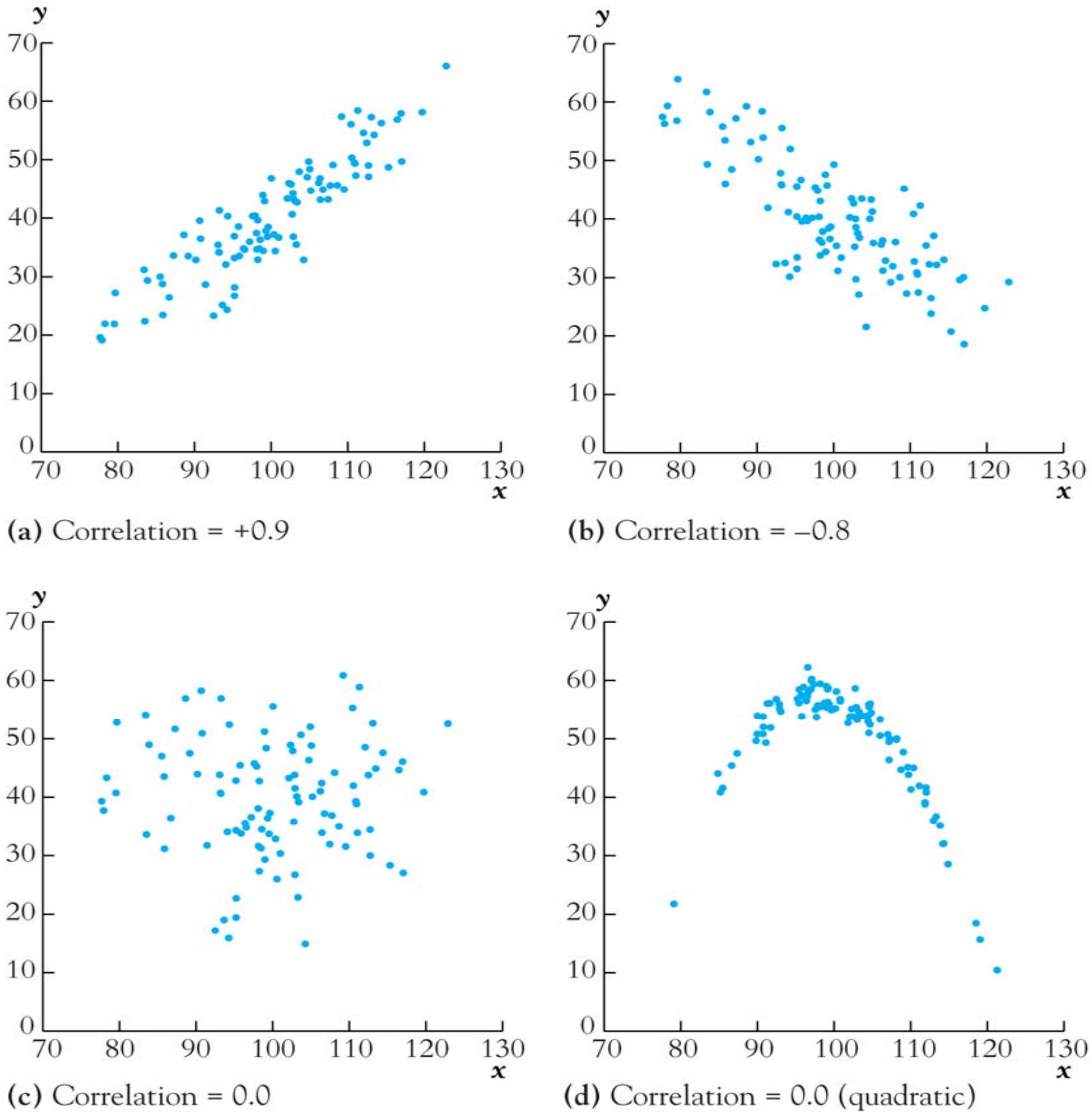
$$Cov(X, X) = \sigma_X^2 = E\left[(X - \mu_X)(X - \mu_X)\right]$$

El coeficiente de correlación se define en términos de la covarianza:

$$\text{corr}(X,Z) = \frac{\text{cov}(X,Z)}{\sqrt{\text{var}(X)\,\text{var}(Z)}} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z}$$

- $-1 \leq \text{corr}(X,Z) \leq 1$
- corr(X,Z) = 1 significa perfecta asociación lineal positiva
- corr(X,Z) = –1 significa perfecta asociación lineal negativa
- corr(X,Z) = 0 significa ningún tipo de asociación lineal

El coeficiente de correlación mide la asociación lineal



(a) Correlation = +0.9

(b) Correlation = –0.8

(c) Correlation = 0.0

(d) Correlation = 0.0 (quadratic)

4. Estimación

- Obtención de la muestra
- Definición del estimador

• Obtención de la muestra

Muestreo Aleatorio Simple (MAS)

Suponer una muestra de tamaño n**,** $Y_1, Y_2 \ldots Y_n$ obtenida a partir de una población. Se supone que el proceso seguido para extraer los elementos muestrales es el llamado MAS que cumple dos condiciones:

1. Todos los elementos de la muestra tienen la misma distribución de probabilidad.

2. Todos los elementos son variables independientes entre sí.

Se dice que los elementos de la muestra están idénticamente e independientemente distribuidos, i.i.d.

•Definición del estimador

- Criterios para definir estimadores
- Propiedades de los estimadores
  - Todo tamaño muestral
  - Muestras grandes

Criterios para definir estimadores

- **MCO**

- $\bar{Y}$ es el estimador de los "mínimos cuadrados" de μ_Y; $\bar{Y}$ minimiza,

$$\min_m \sum_{i=1}^{n} (Y_i - m)^2$$

$$\frac{d}{dm}\sum_{i=1}^{n}(Y_i - m)^2 = \sum_{i=1}^{n}\frac{d}{dm}(Y_i - m)^2 = -2\sum_{i=1}^{n}(Y_i - m)$$

Igualando a cero la derivada se obtiene el estimador $\hat{m}$:

$$\sum_{i=1}^{n} Y = \sum_{i=1}^{n}\hat{m} = n\hat{m} \text{ or } \hat{m} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

- **Máximo-verosímil (**Suplemento docente nº 2: Estimación Máximo-verosímil)

-MCO

Distribución Muestral de la Media Poblacional

La media muestral se define como:

$$\bar{Y} = \frac{1}{n}\left(Y_1 + Y_2 + \ldots + Y_n\right) = \frac{1}{n}\sum_{1}^{n} Y_i$$

$\bar{Y}$ es una variable aleatoria, y sus propiedades están determinadas por la distribución muestral de $\bar{Y}$

Media y Varianza de $\bar{Y}$

Caso general – esto es, para Y_i i.i.d. de cualquier distribución, no necesariamente de Bernoulli:

Media: $E(\bar{Y}) = E(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_Y = \mu_Y$

Varianza: $\text{var}(\bar{Y}) = E[\bar{Y} - E(\bar{Y})]^2$

$$= E[\bar{Y} - \mu_Y]^2$$

$$= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) - \mu_Y\right]^2$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)\right]^2$$

Por lo que

$$\operatorname{var}(\overline{Y}) = E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)\right]^2$$

$$= E\left\{\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)\right]\times\left[\frac{1}{n}\sum_{j=1}^{n}(Y_j - \mu_Y)\right]\right\}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left[(Y_i - \mu_Y)(Y_j - \mu_Y)\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}(Y_i, Y_j)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2$$

$$= \frac{\sigma_Y^2}{n}$$

Media y Varianza de $\bar{Y}$

$$\boxed{E(\bar{Y}) = \mu_Y}$$

$$\boxed{\operatorname{var}(\bar{Y}) = \frac{\sigma_Y^2}{n}}$$

Implicaciones:

1. $\bar{Y}$ es un estimador insesgado de μ_Y (esto es, $E(\bar{Y}) = \mu_Y$)
2. $\operatorname{var}(\bar{Y})$ es inversamente proporcional a n
   - la dispersión de la distribución muestral es proporcional a$1/\sqrt{n}$
   - Por lo tanto, la incertidumbre muestral asociada con $\bar{Y}$ es proporcional a $1/\sqrt{n}$ (cuanto mayor es la muestra menor es la incertidumbre)

Distribución muestral de $\bar{Y}$ cuando n es grande

Para muestras pequeñas, la distribución de $\bar{Y}$ es complicada, pero si *n* es grande, la distribución muestral es simple!

1. Conforme n crece, la distribución de $\bar{Y}$ se concentra cada vez mas en torno a μ_Y (la Ley de los Grandes Numeros)
2. Además, la distribución de $\sqrt{n}(\bar{Y} - \mu_Y)$ se hace normal (el Teorema Central del Límite)

La Ley de los Grandes Números:

Un estimador es consistente si la probabilidad de estar dentro de un intervalo de longitud dada en torno al verdadero valor poblacional tiende a 1 conforme el tamaño muestral crece.

Si $(Y_1,\ldots,Y_n)$ son i.i.d. y $\sigma_Y^2 < \infty$, entonces $\bar{Y}$ es un estimador consistente de μ_Y, esto es,

$$\Pr[|\bar{Y} - \mu_Y| < \varepsilon] \to 1 \text{ conforme } n \to \infty$$

Que se escribe como, $\bar{Y} \xrightarrow{p} \mu_Y$

("$\bar{Y} \xrightarrow{p} \mu_Y$" significa "$\bar{Y}$ converge in probabilidad a μ_Y").

(matemáticas: conforme $n \longrightarrow \infty$, var($\bar{Y}$) = $\frac{\sigma_Y^2}{n}$ $\longrightarrow$ 0, que implica que Pr[$|\bar{Y} - \mu_Y| < \varepsilon$] $\longrightarrow$ 1.)

Dos condiciones que son suficientes para que $\bar{Y}$ sea un estimador consistente de μ_Y:

1. Sesgo $\left(\bar{Y}\right) \to 0$. En este caso se cumple porque el estimador es insesgado.
2. Var$\left(\bar{Y}\right) \to 0$. Esta condición también se cumple porque, como ya hemos visto, en el denominador de la varianza aparece el tamaño muestral que cada vez tiende a hacerse mayor.

Teorema Central del Límite (TCL):

Si $(Y_1,\ldots,Y_n)$ son i.i.d. y $0 < \sigma_Y^2 < \infty$, entonces cuando n es grande la distribución de una versión normalizada de $\bar{Y}$, $\sqrt{n}\left(\bar{Y}-\mu_Y\right)/\sigma_Y$, se aproxima a una distribución normal.

- $\bar{Y}$ se ditribuye aproximadamente como $N(\mu_Y, \dfrac{\sigma_Y^2}{n})$ ("distribución normal con media μ_Y y varianza σ_Y^2/n")
- $\sqrt{n}(\bar{Y} - \mu_Y)/\sigma_Y$ se distribuye como $N(0,1)$ (normal estandar)
- **Cuanto mayor es *n*, mejor es la aproximación.**

Resumen: La Distribución muestral de $\bar{Y}$

Para $Y_1,\ldots,Y_n$ i.i.d. con $0 < \sigma_Y^2 < \infty$,

- La distribución muestral exacta (muestra finita) de $\bar{Y}$ tiene de media μ_Y ("$\bar{Y}$ es un estimador insesgado de μ_Y") y varianza σ_Y^2/n.
- Aparte de la media y varianza, la distribución exacta de $\bar{Y}$ es complicada y depende de la distribución de Y (la distribución de la población)
- Cuando n es grande, la distribución muestral se simplifica:
  - $\boxed{\bar{Y} \xrightarrow{p} \mu_Y}$ (Ley de los Grandes Números)
  - $\boxed{\dfrac{\bar{Y} - E(\bar{Y})}{\sqrt{\text{var}(\bar{Y})}} \text{ es aproximadamente } N(0,1)}$ (TCL)

¿Por qué usar $\bar{Y}$ para estimar μ_Y?

- $\bar{Y}$ es insesgados: $E(\bar{Y}) = \mu_Y$
- $\bar{Y}$ es consistente: $\bar{Y} \xrightarrow{p} \mu_Y$
- $\bar{Y}$ es el estimador de los "mínimos cuadrados" de μ_Y; $\bar{Y}$ minimiza,

$$\min_m \sum_{i=1}^{n}(Y_i - m)^2$$

$$\frac{d}{dm}\sum_{i=1}^{n}(Y_i - m)^2 = \sum_{i=1}^{n}\frac{d}{dm}(Y_i - m)^2 = -2\sum_{i=1}^{n}(Y_i - m)$$

Igualando a cero la derivada se obtiene el estimador $\hat{m}$:

$$\sum_{i=1}^{n}Y = \sum_{i=1}^{n}\hat{m} = n\hat{m} \text{ or } \hat{m} = \frac{1}{n}\sum_{i=1}^{n}Y_i = \bar{Y}$$

- $\bar{Y}$ tiene la menor varianza entre los estimadores que son lineales e insesgados: considerar el estimador,

  $\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} a_i Y_i$, donde $\{a_i\}$ son tales que $\hat{\mu}_Y$ es insesgado;

  entonces var($\bar{Y}$) $\leq$ var($\hat{\mu}_Y$) (prueba: SW, Ch. 17)

- $\bar{Y}$ no es el único estimador de μ_Y

5. Contraste e intervalo de confianza

Contraste de Hipótesis

- Definiciones
  - Tipos de hipótesis
  - Tipos de errores
  - Función de potencia y potencia de un contraste
  - Contraste UMP
  - Contraste consistente
- Proceso para definir un contraste

- Contraste de hipótesis: Definiciones

El problema del contraste de Hipótesis (para la media): se toma una decisión provisional con la información a mano de que la hipótesis nula es verdadera frente a la hipótesis alternativa. Esto es, se contrasta

H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) > \mu_{Y,0}$ (1-sided, >)

H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) < \mu_{Y,0}$ (1-sided, <)

H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) \neq \mu_{Y,0}$ (2-sided)

- **Contraste de hipótesis: Proceso para definir un contraste**.

  Etapas:

  1. Definición del estadístico del contraste

     Siempre la diferencia normalizada entre una estimación sin restricción y una estimación restringida de la combinación de parámetros a los que se refiere la H_0.

  2. Derivación de la distribución del estadístico bajo H_0.

  3. Determinación de la regla de decisión.

     - p-valor
     - Región crítica

p-valor = es la probabilidad de que el estadístico de contraste (en nuestro caso $\overline{Y}$) tome un valor de que sea al menos tan adverso para la hipótesis nula como el valor calculado de dicho estadístico. O, lo que es lo mismo, la probabilidad de que el estadístico tome un valor de la(s) cola(s) de la distribución determinada por el valor calculado.

El **nivel de significación** de un contraste es la probabilidad especificada a priori de rechazar incorrectamente la hipótesis nula. Es decir, la probabilidad de rechazarla cuando es cierta.

Calculando el p-valor para $\bar{Y}$:

$$\boxed{p\text{-valor} = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]}$$

donde $\bar{Y}^{act}$ es el valor calculado de $\bar{Y}$

Calculando el *p-valor:*

- Para calcular el *p*-valor, necesitas conocer la distribución muestral de $\bar{Y}$, que es complicada si n es pequeño.
- Si n es grande, puedes hacer uso de la aproximación normal (TCL):

$$p\text{-valor} = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|],$$

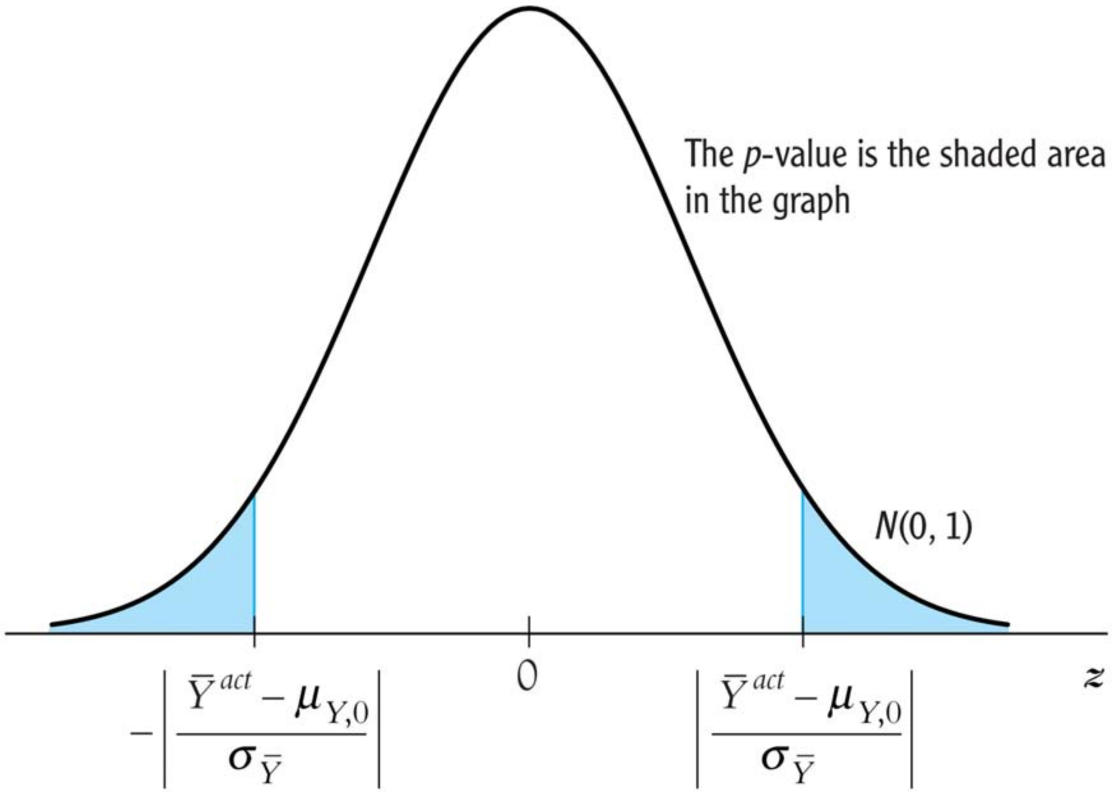
$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right|\right]$$

$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right]$$

$$\cong \text{probabilidad en las colas de la } N(0,1)$$

donde $\sigma_{\bar{Y}}$ = des. est. de la distribución de $\bar{Y}$ = $\sigma_Y/\sqrt{n}$.

Calculando el ***p-valor con σ_Y conocida:***



- Para n grande, *p*-valor = probabilidad de que una variable aleatoria $N(0,1)$ tome valores fuera de $|(\bar{Y}^{act} - \mu_{Y,0})/\sigma_{\bar{Y}}|$
- En la práctica, $\sigma_{\bar{Y}}$ no es conocido – debe ser estimado.

***Estimador de la varianza de Y*:**

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$ = “varianza muestral de *Y*”

Resultado:

If ($Y_1,\ldots,Y_n$) son i.i.d. y $E(Y^4) < \infty$, entonces $s_Y^2 \xrightarrow{p} \sigma_Y^2$

¿Por qué se aplica la ley de los grandes números?

- Porque s_Y^2 es un promedio muestral; ver Apéndice 3.3(S.W.)
- Nota técnica: asumimos que $E(Y^4) < \infty$ porque aquí el promedio no es de las Y_i, sino de sus cuadrados; ver Ap. 3.3 (S.W.)

Calculando el ***p-valor con*** σ_Y^2 ***estimado***:

$$p\text{-valor} = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|],$$

$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}\right|\right]$$

$$\cong \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y / \sqrt{n}}\right|\right] \text{ (large } n\text{)}$$

so

$$\boxed{p\text{-valor} = \Pr_{H_0}[|t| > |t^{act}|]} \quad (\sigma_Y^2 \textbf{ estimado})$$

$\cong$ probabilidad de las colas de la normal fuera de $|t^{act}|$

donde $t = \dfrac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}}$ (el estadístico-t usual)

¿Cuál es la relación entre el *p*-valor y el nivel de significación?

El nivel de significación se especifica a priori. Por ejemplo, si el nivel es el 5%,

- Rechazas la hipótesis nula si $|t| \geq 1.96$.
- Equivalentemente, tu rechazas si $p \leq 0.05$.
- El *p*-valor también se llama el nivel de significación marginal.
    - A menudo, es mejor informar el *p*-valor que dar cuenta de si se rachaza o no – el *p*-valor contiene más información que la afirmación "si/no" acerca de si el contraste se rechaza.

- En este punto os podeis preguntar…..

¿Qué le ocurrió a la tabla-t y a los grados de libertad?

<u>La distribución t de Student</u>

Si Y_i, i = 1,…, n es i.i.d. $N(\mu_Y,\sigma_Y^2)$, entonces el estadístico-t tiene la distribución t de Student con n – 1 grados de libertad.

Los puntos críticos de la distribución t de Student pueden encontrarse en la parte final de los libros de estadística. El proceso es:

1. Calcula el estadístico-t
2. Calcula los grados de libertad, que son n-1.
3. Determina el punto crítico al 5%.
4. Si el estadístico-t supera (en valor absoluto) este punto crítico la hipótesis nula se rechaza.

Si el tamaño muestral es moderado (algunas docenas) o grande (cientos o más), la diferencia entre valores críticos de la distribución t y N(0,1) es despreciable. En la siguiente tabla aparecen los puntos críticos para contrastes de dos colas:

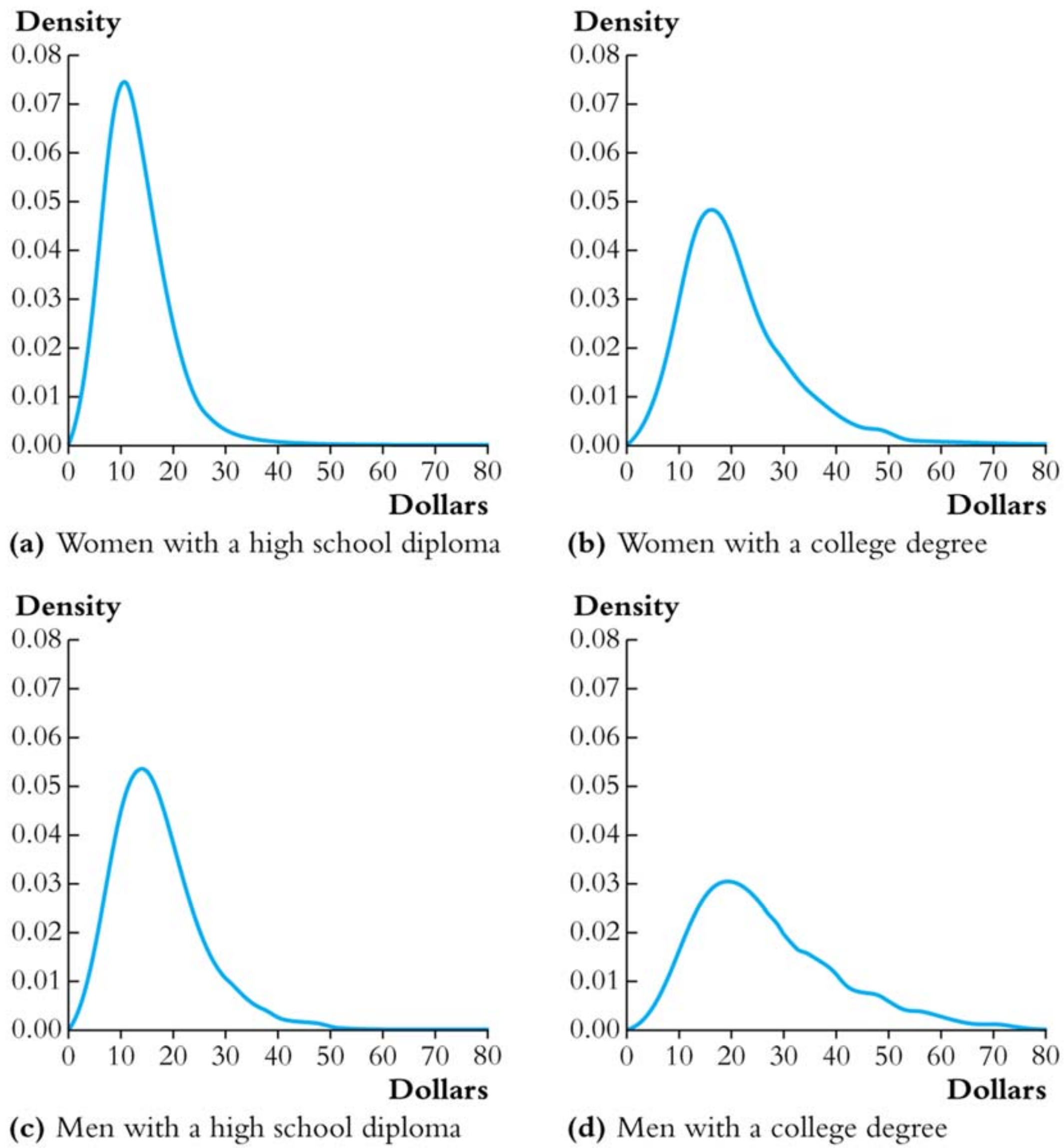
Grados de libertad (n-1)	Valor crítico del 5% de la distribución t
10	2,23
20	2,09
30	2,04
60	2,00
∞	1,96

Comentarios sobre la distribución t de Student

- Así, la distribución t de Student solamente es relevante si el tamaño de la muestra es muy pequeño; Pero en ese caso, para que sea aceptable, tienes que estar seguro que la distribución poblacional de Y es normal. Esto, para los datos económicos, en muchas ocasiones no es creíble.

FIGURE 2.4 **Conditional Distribution of Average Hourly Earnings of U.S. Full-Time Workers in 2004, Given Education Level and Gender**

The four distributions of earnings are for women and men, for those with only a high school diploma (a and c) and those whose highest degree is from a four-year college (b and d).



Resumen sobre la Distribución t de Student

- La hipótesis de que Y sigue una distribución $N(\mu_Y, \sigma_Y^2)$ no es muy plausible en la práctica (¿Renta? ¿Número de niños?)
- Para $n > 30$, la distribución-t y la $N(0,1)$ están muy próximas (conforme n crece, la distribución t_{n-1} converge a $N(0,1)$)
- La distribución t estaba justificada en aquellos días en que las muestras eran pequeñas y los ordenadores no existían.
- Por esta razón, en este curso nos centraremos en la aproximación de muestras grandes dada por el CLT.

Intervalos de confianza

Un intervalo de confianza del 95% se puede construir siempre como el conjunto de valores μ_Y que no es rechazado con un contraste de hipótesis con un 5% de nivel de significación.

$$\{\mu_Y\text{:}\ \left|\frac{\bar{Y}-\mu_Y}{s_Y/\sqrt{n}}\right| \leq 1.96\} = \{\mu_Y\text{: } -1.96 \leq \frac{\bar{Y}-\mu_Y}{s_Y/\sqrt{n}} \leq 1.96\}$$

$$= \{\mu_Y\text{: } -1.96\frac{s_Y}{\sqrt{n}} \leq \bar{Y} - \mu_Y \leq 1.96\frac{s_Y}{\sqrt{n}}\}$$

$$= \{\mu_Y \in (\bar{Y} - 1.96\frac{s_Y}{\sqrt{n}}\ ,\ \bar{Y} + 1.96\frac{s_Y}{\sqrt{n}})\}$$

Este intervalo de confianza descansa en el hecho de que con muestras grandes $\bar{Y}$ *sigue una distribución normal y* $s_Y^2 \xrightarrow{p} \sigma_Y^2$.