Auxiliar 5

Precontrol

Profesor: Raimundo Undurraga
Auxiliares: Brandon Galarza, Camila Jáuregui, Leonardo Meneses, Francisca Monetta, Matías Reyes, Bastián Urzúa, Antonia Villegas.

Pregunta 1

a) Dado un estimador del parámetro poblacional θ, se define el error cuadrático medio como $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + (Sesgo(\hat{\theta}))^2$. Bajo este criterio, un estimador eficiente siempre es mejor que un estimador insesgado. Comente.

Respuesta:

Sea $\hat{\theta}$ un estimador del parámetro poblacional θ. Luego, $\hat{\theta}$ es eficiente si, dentro de todos los estimadores insesgados de θ, $\hat{\theta}$ es aquel con la menor varianza. En efecto, lo primero es reconocer que para que $\hat{\theta}$ sea eficiente es condición necesaria que sea insesgado. Luego, notar que si $\hat{\theta}$ es insesgado, entonces MSE($\hat{\theta}$) = Var($\hat{\theta}$). Pero como sabemos que $\hat{\theta}$ es eficiente, entonces sabemos que V ar($\hat{\theta}$) es mínima. En consecuencia, considerando MSE($\hat{\theta}$) como criterio para comparar la bondad entre estimadores, si $\hat{\theta}$ es eficiente, entonces tiene m´ınimo error cuadrático medio, y por tanto ser´a mejor que cualquier otro estimador insesgado (pues por definición ningún otro estimador insesgado puede tener menor varianza que Var($\hat{\theta}$)).

Nótese que la afirmación anterior excluye a estimadores sesgados. Estos pueden tener menor varianza que un estimador eficiente. En efecto, a pesar de que un estimador sesgado tenga por definición un sesgo distinto de cero, aun puede ser posible que tengan un menor MSE que un estimador eficiente, y por tanto, bajo ese criterio, ser mejores estimadores que un estimador eficiente.

b) Explique qué es la Ley de Esperanzas Iteradas (LIE) e ilustre gráficamente

Respuesta:

Sean X e Y dos variables aleatorias. La Ley de Esperanzas Iteradas se define como E(Y) = E(E(Y |X)), la cual indica que el promedio de Y puede ser descrito como el promedio 1 de las medias condicionales de Y en X. En otras palabras, para cada valor X = x, existe un promedio de Y que llamamos E(Y |X = x).

Luego, el promedio incondicional de Y es el promedio de esos promedios condicionales. Gráficamente, esto se puede ilustrar a través de un plano cartesiano que asocia coordenadas de X e Y. Supongamos que X es a~nos de educación (1 a 10) e Y el salario de la persona. El promedio de los salarios puede ser calculado como el promedio ponderado de los promedios salariales para cada nivel de educación, donde el ponderador es la proporción de observaciones en cada nivel educacional.

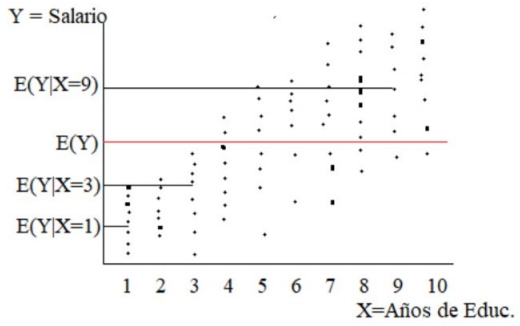


Figura 1: Gráfico Casen 2020

c) Sean W_i con $i = 1, ..., N$ un conjunto de variables aleatorias independientes e idénticamente distribuidas de media μ y varianza σ^2. Se define el siguiente estimador de la varianza: $S_*^2 = \frac{1}{N}\sum_{i=1}^{N}(W_i - \bar{W})^2$ donde $\bar{W} = \frac{1}{N}\sum_{i=1}^{N} w_i$ corresponde a la media muestral. Luego, $\frac{N}{N-1}S_*^2$ es un estimador insesgado de la varianza poblacional. Comente.

Respuesta:

Primero mostramos que S_*^2 es un estimador sesgado de la varianza poblaciona. A partir de la fórmula de del estimador de la varianza se desarrolla la demostración:

$$\begin{aligned}
S_*^2 &= \frac{1}{N}\sum_{i=1}^{N}(M_i - \bar{M})^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}(M_i^2 - 2M_i\bar{M} + \bar{M}^2) \\
&= \frac{1}{N}\left[\sum_{i=1}^{N} M_i^2 - 2\sum_{i=1}^{N} M_i\bar{M} + \sum_{i=1}^{N}\bar{M}^2\right] \\
&= \frac{1}{N}\left[\sum_{i=1}^{N} M_i^2 - 2\bar{M}\sum_{i=1}^{N} M_i + N\bar{M}^2\right]
\end{aligned}$$

Por definición del estadístico promedio: $\frac{1}{N}\sum_{i=1}^{N} M_i = M$

$$S_*^2 = \frac{1}{N}\sum_{i=1}^{N}(M_i^2 - \bar{M}^2)$$

Calculamos la esperanza:

$$E(S^2) = E\left(\frac{1}{N}\sum_{i=1}^{N} M_i^2 - \bar{M}^2\right) = \frac{1}{N}\sum_{i=1}^{N} E(M_i^2) - E(\bar{M}^2)$$

Recordando que la varianza para una variable aleatoria se puede calcular como $Var(X) = E(X^2) - [E(X)]^2$ para este caso se tiene:

$$E(M_i^2) = Var(M_i) + [E(M_i)]^2 = \sigma^2 + \mu^2 \Rightarrow \frac{1}{N}\sum_{i=1}^{N} E(M_i^2) = \frac{1}{N}\sum_{i=1}^{N}(\sigma^2 + \mu^2) = \sigma^2 + \mu^2$$

$$E(\bar{M}^2) = Var(\bar{M}) + [E(\bar{M})]^2 = \sigma^2\frac{1}{N} + \mu^2 = \frac{\sigma^2}{N} + \mu^2$$

Finalmente:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} E(M_i^2) - E(\bar{M}^2) &= \sigma^2 + \mu^2 - \frac{\sigma^2}{N} - \mu^2 \\ &= \sigma^2 - \frac{\sigma^2}{N} \\ &= \frac{N-1}{N}\sigma^2\end{aligned}$$

A partir del resultado anterior, tenemos que el estimador S^2 está sesgado en un factor $\frac{N-1}{N}$. En efecto, al multiplicarlo por $\frac{N}{N-1}$ se obtiene el valor de σ^2. Es decir, $\frac{N}{N-1}S^2$ es un estimador insesgado de la varianza poblacional.

Pregunta 2

a) Defina qué es un estimador insesgado. Defina qué es un estimador consistente. Explique cuidadosamente cuál es la diferencia entre ambos ¿Todos los estimadores insesgados son consistentes? ¿Todos los estimadores consistentes son insesgados? Ilustre con un ejemplo en cada caso.

Respuesta:

Sea $\hat{\theta}$ un estimador del parámetro poblacional θ. $\hat{\theta}$ es un estimador sesgado de θ si E($\hat{\theta}$) 6= θ. A su vez, $\hat{\theta}$ es un estimador inconsistente de θ si el límite de $\hat{\theta}$ conforme n tiende a infinito es distinto de θ. El insesgamiento es una propiedad

de los estimadores en muestras peque~nas, y por tanto no depende del tamaño muestral. En cambio, la consistencia es una propiedad asintótica que aplica a estimadores en muestras grandes, es decir, depende del tamaño muestral y tiene relación con la convergencia en probabilidad del estimador conforme el tamaño muestral crece.

Por otra parte, no todos los estimadores sesgados son inconsistentes, y no todos los estimadores inconsistentes son sesgados. Por ejemplo:

- El estimador $\hat{X}+\frac{1}{n}$ es un estimador sesgado del promedio poblacional ya que $E(\hat{X}+\frac{1}{n}) = \mu+\frac{1}{n} \neq \mu$. No obstante, el límite de $\hat{X}+\frac{1}{n}$ conforme n tiende a infinito es igual a μ, es decir, $\hat{X}+\frac{1}{n}$ es un estimador consistente del promedio poblacional.
- El estimador X_1 es un estimador insesgado del promedio poblacional ya que $E(X_1) = E(X_i) = \mu$. Sin embargo, X_1 no depende de n, es decir, no converge en probabilidad a algo distinto de X_1 conforme n tiende al infinito, y por tanto es un estimador inconsistente del promedio poblacional.

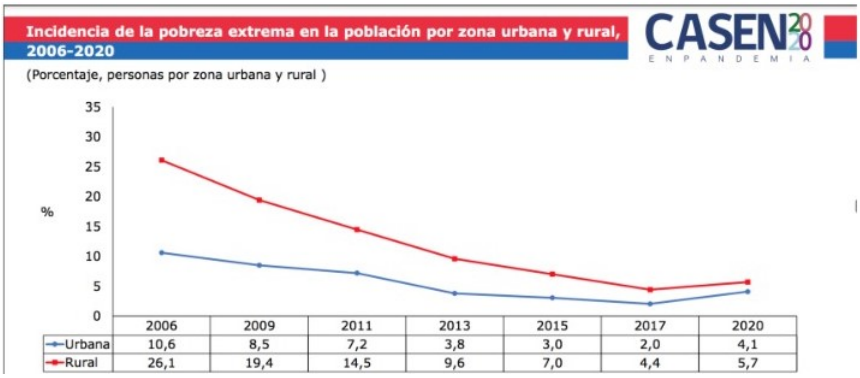
b) Suponga dos hipótesis respecto del valor poblacional μ, la hipótesis nula ($H_0 : \mu = \mu_1$) y la hipótesis alternativa ($H_1 : \mu = \mu_1$). Para testear cuál de estas dos hipótesis es la correcta, un aumento del valor crítico reduce el error tipo I y aumenta el error tipo II, mejorando así el poder estadístico del test. Verdadero o Falso? Defina y explique detalladamente.

Respuesta:

El error tipo I es la probabilidad de rechazar H_0 en circunstancias en las que H_0 es verdadera. Por otra parte, el error tipo II es la probabilidad de no rechazar H_0 en circunstancias en las que H_1 es verdadera. Finalmente, el poder estadístico es la probabilidad de rechazar H_0 en circunstancias en las que H_1 es correcta. O dicho de otra manera, el poder estadístico es igual a 1 menos la probabilidad de NO rechazar H_0 cuando H_1 es correcta, es decir, 1 menos el error tipo II. Un aumento del valor crítico k reduce el error tipo I (hace más improbable rechazar H_0 cuando H_0 es verdadera). A su vez, un aumento de k aumenta el error tipo II (es decir, hace más probable no rechazar H_0 cuando H_1 es verdadera), lo que significa que la expresión $1 -$ error tipo II se hace más pequeña, reduciendo así el poder estadístico del test. En efecto, el enunciado es Falso.

Pregunta 3

El Ministerio de Desarrollo Social está a cargo de implementar la Encuesta CASEN, la cual permite medir la incidencia de la pobreza extrema en la población nacional y su evolución en el tiempo. La pobreza extrema es una variable binaria que toma el valor 1 si el hogar está en pobreza extrema (debajo de un valor determinado de ingreso per capita) y 0 si no. En el gráfico a continuación se muestra la evolución de la pobreza extrema para población urbana y rural entre los años 2006 y 2020. Se denota una ca´ıda en la pobreza extrema a lo largo del período, tanto en zonas rurales como en zonas urbanas. Por ejemplo, tal como muestra el cuadro debajo del gráfico, la pobreza extrema en zonas rurales cae de un 26.1 % en 2006 a un 5.7 % en 2020. Por otra parte, la pobreza extrema en zonas urbanas cae de un 10.6 % en 2006 a un 4.1 % en 2020. Lo anterior sugiere que la brecha de pobreza extrema entre zonas rurales y zonas urbanas también ha disminuido. Mientras en 2006 la brecha era de 15.5 puntos porcentuales en favor de zonas rurales, dicha brecha en 2020 era sólo de 1.6 puntos porcentuales.



	2006	2009	2011	2013	2015	2017	2020
Urbana	10,6	8,5	7,2	3,8	3,0	2,0	4,1
Rural	26,1	19,4	14,5	9,6	7,0	4,4	5,7

Figura 2: Gráfico Casen 2020

a) Asumiendo una muestra de 1000 observaciones por zona por año, construya un intervalo de confianza al 95 % ($\alpha = 0.05$) de la tasa de pobreza extrema para zonas rurales en el año 2006 y otro para zonas urbanas en el año 2006. Puede concluir que las tasas de pobreza extrema entre zonas rurales y urbanas son estad´ısticamente distintas en ese año?

Respuesta:

De la tabla, tenemos que las tasas de pobreza extrema para el año 2006 son del 10.6 % y 26.1 % para las zonas urbana y rural, respectivamente.

Para construir un intervalo de confianza al 95 % de la tasa en cada zona, necesitamos construir el error estándar de la tasa y definir la distribución del test. Dado que es una muestra suficientemente grande para cada zona, podemos usar una distribución normal estándar, cuyo valor crítico para un $\alpha = 0.05$ es $z_{0.975} = 1.96$. Nótese que usar una distribución t de Student nos arrojaría resultados similares, pues en muestras grandes la distribución t de Student converge en distribución a una normal estándar.

Luego, el error estándar es la desviación estándar estimada dividida por la raíz cuadrada del número de observaciones:

$$SE_{urbano} = \sqrt{\frac{0.106 \cdot (1 - 0.106)}{1,000}}$$
$$SE_{urbano} = 0.009735$$
$$SE_{rural} = \sqrt{\frac{0.261 \cdot (1 - 0.261)}{1,000}}$$
$$SE_{rural} = 0.013888$$

En efecto, los intervalos de confianza en cada zona serían:

$$IC_{urbano} : [0.106 - 1.96 \cdot 0.009735; 0.106 + 1.96 \cdot 0.009735] = [0.08692; 0.1251] \approx [8.69\,\%; 12.51\,\%]$$

$$IC_{rural} : [0.261 - 1.96 \cdot 0.013888; 0.261 + 1.96 \cdot 0.013888] = [0.23378; 0.28822] \approx [23.38\,\%; 28.82\,\%]$$

Dado que el límite superior del IC en zonas urbanas no se intersecta con el límite inferior del IC en zonas rurales, se concluye que estad´ısticamente las tasas de pobreza extrema en 2006 son distintas entre zonas urbanas y rurales.

b) Asumiendo una muestra de 1000 observaciones por zona por año, testee a un 95 % de nivel de confianza ($\alpha = 0.05$) si la diferencia en las proporciones de pobreza extrema entre zonas rurales y urbana es mayor a cero. Explicitar hipótesis nula, hipótesis alternativa, y test a utilizar. Qué concluye?

Respuesta:

El test de hipótesis es de una cola. Situamos la hipótesis nula como la hipótesis a testear:

$$H0 : p_r - p_u > 0$$
$$Ha : p_r - p_u \leq 0$$

Este es un test de diferencia de proporciones. Dado que el tamaño muestral es suficientemente grande, podemos usar el estadístico Z:

$$Z = \frac{p_r - p_u}{\sqrt{\frac{p_r(1-p_r)}{n_r} + \frac{p_u(1-p_u)}{n_u}}}$$

Reemplazando para Z se tiene:

$$Z = \frac{0.155}{\sqrt{\frac{0.261 \cdot (1-0.261)}{1{,}000} + \frac{0.106 \cdot (1-0.106)}{1{,}000}}} = 9.139$$

Dado que la hipótesis nula indica que la diferencia es mayor que 0, entonces el margen de la distribución del test que nos interesa es el margen izquierdo, donde se ubica la zona de rechazo. En particular, el valor crítico Z que refleja el percentil 5 ($\alpha = 0.05$) de la distribución equivalente a -1.64. Dado que el valor calculado del estadístico es mayor al valor crítico (5, 6134 > -1.64), entonces no se rechaza la hipótesis nula. Concluimos entonces que, al menos para zonas urbanas, la tasa de pobreza extrema es estad´ısticamente mayor en 2006 que en 2020.

c) Explique en qué consiste el Teorema del Límite Central. Cuál es la relevancia del Teorema del Límite Central para llevar a cabo este test de hipótesis? Por qué es tan importante? Explique.

Respuesta:

El Teorema del Límite Central nos dice que para cualquier variable aleatoria X, independientemente de la distribución poblacional de la que provenga, al aplicarle la transformación estabilizadora $\sqrt{n}(\bar{X}-\mu)$ a su promedio $\bar{X}$, este convergerá en distribución a una distribución normal con media 0 y varianza σ^2. Luego, conforme n tiende a infinito, $\bar{X}$ converge asintóticamente a una distribución normal con media μ y varianza $\frac{\sigma^2}{n}$.

La importancia del Teorema del Límite Central radica en que nos permite hacer inferencia estadística sobre el comportamiento de una variable aleatoria sin necesidad de conocer la distribución poblacional de la que proviene. En el ejemplo mencionado, la tasa de pobreza extrema se interpreta como una proporción, y efectivamente podemos tratarla como una media, ya que se define como el número de casos que cumplen cierta condición entre el número total de casos.

Además, dado que suponemos que tenemos una muestra lo suficientemente grande, según el Teorema del Límite Central, sabemos que el estadístico del test de proporciones sigue una distribución normal estándar con media 0 y varianza 1. Esto nos permite hacer inferencia estadística sobre los parámetros de interés en el problema, incluyendo las tasas de pobreza en zonas rurales y urbanas.

En conclusión, sin el Teorema del Límite Central, no podríamos realizar inferencia estadística sobre los parámetros de interés en este problema.