

Resolución del práctico sobre estimación por intervalos y test de hipótesis

Juan Piccini

Ejercicio 1) Tenemos $\{X_1, \dots, X_{36}\}$ i.i.d. que según la letra siguen una distribución $N(\mu, \sigma)$ con $\sigma = 0.15$ conocida.

Nos piden un IC95 para μ , y se sabe que $\bar{X}_{36} = 2.25$.

Como ya vimos en clase, dicho intervalo es $[\bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}]$, donde $z_{\alpha/2}$ es el valor que deja a su derecha el 0.025 (o sea 2.5%) del área bajo la densidad de la $N(0, 1)$ (la llamada campana de Gauss), como muestra la figura (1).

Por la simetría de dicha campana, tenemos que entre $-z_{\alpha/2}$ y $z_{\alpha/2}$ se concentra el 95% de la masa, estando el 5% restante en las dos colas.

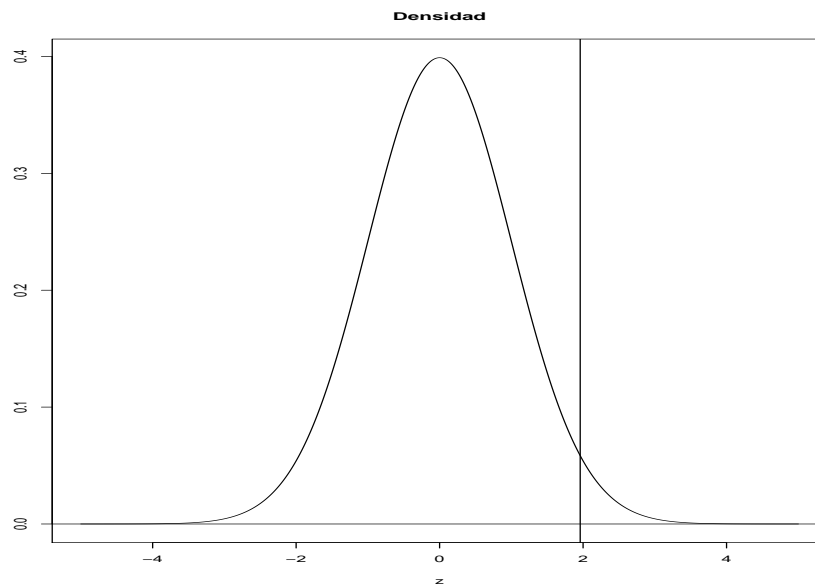


Fig. 1: Campana de Gauss con $z_{\alpha/2} = 1.96$. El área bajo la campana a la derecha de 1.96 es el 2.5% del total

Dicho valor es $z_{\alpha/2} = 1.959964$ que se aproxima por 1.96. Puede encon-

trarse en una tabla de percentiles para la distribución normal típica o standard.

También puede hallarse usando Octave con la sentencia “norminv(.975)”, obteniendo “ans = 1.9600”.

Tenemos entonces todo lo necesario para hallar dicho IC095: sabemos que está centrado en el promedio \bar{X}_n y que su radio es $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.

Operando tenemos que $z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{0.15}{\sqrt{36}} = 0.049$, de donde el IC095 es [2.2010, 2.2990].

En la segunda parte se pregunta sobre el tamaño que debe tener la muestra para tener un 95% de confianza en que \bar{X}_n no difiera de μ en más de 0.03.

En otras palabras buscamos que $|\bar{X}_n - \mu| \leq 0.03$. Esto es lo mismo que pedir que el radio del IC095 sea igual a 0.03.

Como dicho radio es $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, lo pedido se reduce a hallar el primer valor de n para el cual $z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq 0.03$.

Elevando al cuadrado y despejando n obtenemos $(\frac{1.96 \times 0.15}{0.03})^2 = 96.04 \leq n$, de donde $97 \leq n$, la muestra debe tener al menos 97 datos.

- 2) Es similar al anterior, nos dicen que tenemos el promedio de una muestra de 50 artículos cuya distribución es normal con σ conocido y nos piden un IC095 para la media μ .
- 3) Tenemos una muestra i.i.d. de 10 datos $\{X_1, \dots, X_{10}\} \sim N(\mu, \sigma)$ con $\sigma = 2.5$ conocido.

Nos piden implementar el TdH $\begin{cases} H0 & \mu = 104 \\ H1 & \mu > 104 \end{cases}$

trabajando con $\alpha = 0.05$.

El promedio de los 10 datos es $\bar{X}_{10} = 104.93$. Podemos hacer el test usando Octave: como es un test sobre la media para una población normal con varianza conocida (si nos dan el desvío σ entonces también nos dan la varianza σ^2).

La función `ztest` es la que hace dicho test. La sintaxis es “[h,pval, ci, z,zcrit]=ztest(x,104,2.5,”tail”,”right”)”, donde `x` es el vector con los 10 datos, 104 es el valor que propone H_0 para la media verdadera μ y 2.5 es el valor de σ .

Dicho test devuelve “h = 0,pval = 0.11947,ci =103.63 Inf,z = 1.1776, zcrit = 1.6449”.

h=0 nos dice que no rechaza H_0 , el p-valor es 0.11947, casi 0.12, muy por encima de 0.05. El IC 095 es que el el promedio sea mayor a 103.63, algo que nuestra muestra cumple.

Si hacemos el test “a mano”, la región de rechazo (región crítica) es $\bar{X}_{10} - 104 > z_{\alpha} \frac{\sigma}{\sqrt{n}}$.

Para $\alpha = 0.05$, tenemos $z_{0.05} = 1.6449$, de modo que rechazaremos H_0 si $\bar{X}_{10} - 104 > 1.6449 \frac{2.5}{\sqrt{10}=1.30}$.

Como $\bar{X}_{10} - 104 = 0.93$, no rechazamos H_0 .

Otra forma de hacer el test es mediante simulación: simulamos muchas muestras de 10 elementos c/u usando la distribución que propone H_0 ($N(\mu = 104, \sigma = 2.5)$), para c/u de dichas muestras calculamos su promedio y luego vemos si nuestro promedio de 104.93 es uno de los comunes o si es algo poco frecuente si H_0 fuese cierta.

Hacemos “muestra=normrnd(104,2.5,100000,10);”. Esta sentencia genera una matriz llamada muestra con 100000 filas y 10 columnas (cada fila es una muestra de 10 elementos), generada por una $N(\mu = 104, \sigma = 2.5)$.

Luego calculamos el promedio de cada fila mediante “medias=mean(muestra,2);”. Ahora tenemos los 100000 promedios guardados en la matriz columna llamada medias.

Ahora contamos la cantidad de promedios mayores al valor 104.93 de nuestros datos, “size(find(medias \geq 104.93))”, obteniendo “11971”. Si dividimos entre 100000, nos da la proporción de veces en que una muestra de 10 elementos generados tal y como H_0 lo propone da un promedio tanto o más extremo que el observado.

Esto no es otra cosa que el p-valor, que nos da 0.11971 (notemos lo cercano que está del p-valor exacto obtenido mediante el ztest). Como dicho p-valor es casi 0.12, superior a 0.05, no rechazamos H_0 .

- 4) Es similar al anterior, con 12 datos en lugar de 10 y con H_1 que involucra la cola izquierda en lugar de la cola derecha de la distribución.

La región crítica será $\bar{X}_n - \mu < z_\alpha \frac{\sigma}{\sqrt{n}}$. En esta ocasión se rechaza H_0 .

En esta ocasión si usamos ztest, debemos ingresar “[h,pval, ci, z,zcrit]=ztest(x,4,2,”tail”,,”left”)”. El p-valor ahora es 0.047, menor a 0.05 por lo que debemos rechazar H_0 .

Resultados similares se obtienen simulando muchas muestras de 12 elementos generados por una $N(\mu = 4, \sigma = 2)$, calculando el promedio de cada muestra y viendo la proporción de dichos promedios que es menor o igual a nuestro promedio $\bar{X}_{12} = 3.0358$.

Observaremos que dicha proporción o p-valor (para 100000 muestras) es 0.0472, casi igual al p-valor exacto.

- 5) Tenemos una muestra de 100 artículos elegidos al azar de entre una población mayor.

Si la máquina que produce dichos artículos arroja un porcentaje de defectuosos mayor al 10% entonces deberá ser reparada.

En esta muestra de 100 artículos, se encontraron 15 defectuosos (un 15% de la muestra). Se desea determinar si esto justifica que la máquina deba ser reparada. Esto es, se desea hacer el TdH $\begin{cases} H0 : p \leq 0.10 \\ H1 : p > 0.10 \end{cases}$

Nuestra muestra consta de 100 realizaciones $\{X_1, \dots, X_{100}\}$ de una variable aleatoria $X \sim Ber(p)$ (recordemos que la distribución de Bernoulli es la asociada a un experimento aleatorio con dos resultados posibles: éxito (con probabilidad p) y fracaso (con probabilidad $1-p$)).

En este caso el éxito es que el artículo sea defectuoso. Si codificamos éxito con 1 y fracaso con 0, entonces la cantidad de éxitos en las 100 realizaciones es $Y = \sum_{i=1}^{100} X_i$, que tiene distribución binomial de parámetros $(100, p)$.

Esto es, $y \sim Bin(100, p)$, cuenta la probabilidad de obtener k éxitos en 100 ensayos, donde la probabilidad de éxito en cada ensayo es p .

Debemos hacer un TdH sobre p , para ver si es creíble que una variable $Y \sim Bin(100, 0.10)$ (que es la cantidad de defectuosos en la muestra de 100 si en efecto fuese cierta $H0$) pueda arrojar un 15 defectuosos.

Para ello, simularemos muchas variables aleatorias generadas por la distribución Binomial de parámetros $n=100, p=0.10$ (que es lo que postula $H0$).

Cada una de estas variables nos dará la cantidad de defectuosos en una muestra de 100 artículos si la verdadera proporción de defectuosos en la población fuese 0.10.

Contamos luego cuantas veces dicha cantidad es mayor o igual a 15 (que fue lo que observamos en la realidad) y vemos que tan frecuente es observar 15 o más defectuosos en 100 casos.

Si dicha frecuencia es menor a 0.05 rechazaremos $H0$ ya que es poco creíble que una muestra de 100 artículos pueda tener tantos defectuosos.

Hacemos “`X=binornd(100,0.10,1000000,1);`”, lo que genera una matriz de 1 millón de filas y una columna, donde en cada fila tenemos un valor entero positivo, la cantidad de defectuosos en una muestra de 100 elementos donde cada elemento tiene un 10% de chance de ser defectuoso.

Contamos la cantidad de veces que obtuvimos valores ≥ 15 y lo dividimos entre el total: “`size(find(X \geq 15))/1000000`”, obteniendo 0.072, que es el p-valor obtenido mediante simulación.

Esto nos dice que si nuestros datos en efecto fuesen generados por la distribución Binomial con el p que dice $H0$, entonces la probabilidad de obtener 15 o más defectuosos entre los 100 artículos es del orden de 0.07, o 7%. No es alta, pero es mayor al 5% que nos fijamos para rechazar $H0$.

Por tanto, como trabajamos con $\alpha = 0.05$, no rechazamos $H0$.

- 6) 1) En una muestra de 400 individuos, tenemos 165 positivos. Nos piden estimar p (la proporción de enfermos en la población) y hallar un IC95 para dicha proporción.

Cada individuo puede o no estar enfermo, la probabilidad de que lo esté es p (desconocido). Por tanto podemos modelar cada individuo como un lanzamiento de moneda.

Puede salir cara (positivo) o cruz (sano). Tenemos entonces X_1, \dots, X_{400} i.i.d. $\sim \text{Ber}(p)$, con distribución Bernoulli de parámetro p .

Recordemos que si $X \sim \text{Ber}(p)$, entonces $E(X) = p$ y $\text{Var}(X) = p(1-p)$. Esto es, $\mu = p$ y $\sigma^2 = p(1-p)$.

Si denotamos cara con un uno y cruz con un cero, nuestros 400 lanzamientos darán una tira de 400 lugares, de los cuales en 165 obtuvimos unos.

$$\text{Entonces } \sum_{i=1}^{400} X_i = 165 \text{ y por tanto } \bar{X}_{400} = \frac{1}{400} \sum_{i=1}^{400} X_i = \frac{165}{400} = 0.412$$

Entonces el estimador puntual para p es 0.412.

Para obtener el IC95 podemos aplicar el Teorema Central del Límite o TCL.

El mismo dice que si X_1, \dots, X_n i.i.d. con $E(X_i) = \mu$ y $\text{Var}(X_i) = \sigma^2$, entonces la variable aleatoria $Z_n = \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$ tiene una distribución que tiende a la $N(0,1)$ cuando n tiende a infinito.

Entonces si buscamos un intervalo $[a, b]$ tal que $P(a \leq \mu \leq b) \geq 0.95$, si restamos \bar{X}_n y multiplicamos por $\frac{\sqrt{n}}{\sigma}$ obtenemos que $P(a \leq \mu \leq b) = P(\frac{\sqrt{n}}{\sigma}(\bar{X}_n - b) \leq \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \leq \frac{\sqrt{n}}{\sigma}(\bar{X}_n - a))$, o sea $P(\frac{\sqrt{n}}{\sigma}(\bar{X}_n - a) \leq Z_n \leq \frac{\sqrt{n}}{\sigma}(\bar{X}_n - a))$.

Entonces, hallar a, b tales que $P(a \leq \mu \leq b) \geq 0.95$ equivale a hallar a, b tales que $P(\frac{\sqrt{n}}{\sigma}(\bar{X}_n - a) \leq Z_n \leq \frac{\sqrt{n}}{\sigma}(\bar{X}_n - a)) \geq 0.95$.

Sabemos que para una variable aleatoria $Z \sim N(0, 1)$ se cumple $P(-1.96 \leq Z \leq 1.96) \geq 0.95$.

Como Z_n se parece mucho a Z , tenemos $P(-1.96 \leq \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \leq 1.96) \geq 0.95$, y despejando tenemos que $P(\bar{X}_n - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 1.96 \frac{\sigma}{\sqrt{n}}) \geq 0.95$.

El IC95 aproximado (porque Z_n se parece pero no es una $N(0,1)$) para μ sería pues $[\bar{X}_n - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96 \frac{\sigma}{\sqrt{n}}]$, pero no conocemos σ .

Lo que podemos hacer es sustituir σ por un estimador como sigue: Estamos asumiendo que $X_1, \dots, X_{400} \sim \text{Ber}(p)$, en cuyo caso debe ser $\sigma = \sqrt{p(1-p)}$,

No conocemos p pero tenemos un estimador que es $\bar{X}_{400} = 0.412$, entonces podemos estimar σ mediante $\sigma' = \sqrt{(\bar{X}_{400}(1 - \bar{X}_{400}))}$.

Hecho esto, las cuentas dan que $1.96 \frac{\sigma'}{\sqrt{400}} = 0.048$ y el IC queda $[0.412 - 0.048, 0.412 + 0.048]$, o sea $[0.364, 0.460]$.

2) En una muestra de 400 individuos, tenemos 165 positivos (un 0.4125 del total). Se pide un TdH $\begin{cases} H0 : p = 0.40 \\ H1 : p > 0.40 \end{cases}$

Al igual que en el ejercicio 5), bajo H0 la cantidad de positivos en 400 individuos puede modelarse como una variable aleatoria

$$Y \sim Bin(n = 400, p = 0.40).$$

Por tanto podemos simular muchas realizaciones de Y y luego ver qué tan frecuente es obtener 165 o más positivos en 400.

Simulando un millón de realizaciones de Y , obtenemos que se observan 322522 veces una cantidad de positivos mayor o igual a 165.

Esto nos da una proporción de 0.322, que es el p-valor empírico, la probabilidad de que una Y generada por la distribución que postula H0 produzca un resultado tanto o más extremo que el observado.

Como esta probabilidad es mayor al 0.05 habitual en estos tests, no rechazamos H0.

También podemos hacer el test usando el Teorema Central del Límite o TCL: Si tenemos X_1, \dots, X_n i.i.d. con $E(X_i) = \mu$ y desvío σ , entonces la variable aleatoria $Z_n = \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right)$ tiene una distribución que tiende a la distribución $N(0, 1)$ a medida que n crece.

En nuestro caso, bajo H0 tenemos X_1, \dots, X_{400} i.i.d. $\sim Ber(p = 0.40)$, o sea $\mu = 0.40$ y $\sigma = \sqrt{p(1-p)} = 0.4899$.

Tenemos entonces que $Z_{400} = \sqrt{400} \left(\frac{0.412 - 0.40}{0.4899} \right) = 0.51$.

Llamemos z_α al valor que deja a su derecha el 5% del área bajo la campana de Gauss, o sea $z_\alpha = 1.6445$

Rechazaremos H0 al nivel 0.05 si $Z_{400} > z_\alpha$, porque Z_{400} es casi $N(0, 1)$ y sabemos que $P(Z > 1.6445) \leq 0.05$.

Como 0.51 no es mayor que 1.6445, rechazamos H0 al nivel $\alpha = 0.05$.

Si calculamos el p-valor exacto (la probabilidad bajo H0 de obtener un valor tanto o más extremo que 0.51), hallamos que $P(N(0, 1) > 0.51) = 0.3050$, bastante cercano al p-valor hallado empíricamente mediante simulación.

7) Tenemos X_1, \dots, X_{200} i.i.d. $\sim U(0, b)$ con $\sum_{i=1}^{200} X_i = 928.68$, $\sum_{i=1}^{200} X_i^2 = 5726.77$.

Al igual que antes usaremos el IC095 que surge del TCL, y al igual que antes necesitamos σ , el desvío de nuestros datos.

Como no conocemos σ porque no conocemos b , vamos a sustituir σ por un estimador.

Recordemos que $\sigma^2 = E(X^2) - (E(X))^2$ y que la esperanza se puede aproximar por el promedio.

Entonces en lugar de $E(X^2)$ ponemos $\frac{1}{n} \sum_{i=1}^{200} X_i^2$ y en lugar de $E(X)$ ponemos $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{200} X_i$.

Entonces obtenemos $\sigma_n = \sqrt{\frac{1}{n} \sum_{i=1}^{200} X_i^2 - \bar{X}_n}$. Ahora podemos calcular el IC aproximado (aproximado porque hay un TCL en medio y además estamos aproximando el σ necesario).

El IC095 aproximado es $[\bar{X}_{200} - 1.96 \frac{\sigma_n}{\sqrt{200}}, \bar{X}_{200} + 1.96 \frac{\sigma_n}{\sqrt{200}}]$. Las cuentas restantes quedan para el lector.