



CONTROL # 3

Raúl Gouet, Jorge Lemus.

1. Considere una partición del intervalo $[0, 1]$ dada por los puntos $a_0 = 0 < a_1 < \dots < a_{n-1} < a_n = 1$, en donde el subintervalo $[a_{i-1}, a_i]$ tiene largo $p_i = a_i - a_{i-1}$. Se define la entropía de esta partición por

$$h = - \sum_{i=1}^n p_i \ln p_i$$

Sean X_1, X_2, \dots v.a. independientes con distribución uniforme $[0, 1]$. Sea $Z_m(i)$ la cantidad de variables aleatorias en el conjunto $\{X_1, \dots, X_m\}$ que están en el intervalo $[a_{i-1}, a_i]$. Sea

$$R_m = \prod_{i=1}^n p_i^{Z_m(i)}.$$

Probar que

$$\frac{1}{m} \ln R_m \rightarrow -h \quad \text{casi seguramente}$$

Para esto, siga los siguientes pasos:

- (a) (1 pto.) Se define $I_{ij} = \mathbb{1}_{[a_{i-1}, a_i]}(X_j)$ (también denotado como $\mathbb{1}_{\{X_j \in [a_{i-1}, a_i]\}}$) la variable aleatoria que vale 1 si $X_j \in [a_{i-1}, a_i]$ y 0 si no. Expresa $Z_m(i)$ como suma de variables I_{ij} .

Solución:

$$Z_m(i) = \sum_{j=1}^m \mathbb{1}_{\{X_j \in [a_{i-1}, a_i]\}} = \sum_{j=1}^m I_{ij}.$$

- (b) (2 ptos.) Defina $Y_j = \sum_{i=1}^n I_{ij} \ln p_i$ y pruebe que $\frac{1}{m} \ln R_m = \frac{1}{m} \sum_{j=1}^m Y_j$.

Solución:

$$\begin{aligned} \frac{1}{m} \ln R_m &= \frac{1}{m} \ln \left(\prod_{i=1}^n p_i^{Z_m(i)} \right) \\ &= \frac{1}{m} \sum_{i=1}^n Z_m(i) \ln p_i \\ &= \frac{1}{m} \sum_{i=1}^n \sum_{j=1}^m I_{ij} \ln p_i \\ &= \frac{1}{m} \sum_{j=1}^m \sum_{i=1}^n I_{ij} \ln p_i \\ &= \frac{1}{m} \sum_{j=1}^m Y_j \end{aligned}$$

(c) (3 ptos.) Concluya la convergencia pedida.

Solución:

Notemos que Y_j es independiente de Y_k para $j \neq k$, pues I_{ji} es independiente de I_{ki} , $\forall i$, ya que X_j es independiente de X_k para $j \neq k$.

Además

$$\mathbb{E}(Y_j) = \sum_{i=1}^n \mathbb{E}(I_{ij}) \ln p_i = \sum_{i=1}^n \mathbb{P}(X_j \in [a_{i-1}, a_i]) \ln p_i = \sum_{i=1}^n p_i \ln p_i.$$

Por otra parte,

$$\text{Var}(I_{ij}) \leq \max_{p \in [0,1]} p(1-p) = \frac{1}{4}$$

con lo que podemos acotar

$$\text{Var}(Y_j) \leq \frac{1}{4} \sum_{i=1}^n \ln p_i < \infty.$$

Por lo tanto, aplicando la ley fuerte de los grandes números se tiene que

$$\frac{1}{m} \ln R_m \rightarrow -h \quad \text{casi seguramente}$$

2. (a) (3 ptos.) Se desea determinar lo más precisamente posible el área de un rectángulo de lados a, b con largos desconocidos l_a y l_b respectivamente. Para ello se realizan n mediciones independientes de ambos lados (a y b) obteniendo los pares de va $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n)$. Suponga que las Z_i son independientes e idénticamente distribuidas (iid) con densidad $f(x, y; \theta)$, donde $\theta = (l_a, l_b) \in \Theta = \mathbb{R}_+^2$ es el parámetro desconocido. Suponga además que $E_\theta(X_i) = l_a$, $E_\theta(Y_i) = l_b$, $\text{Var}_\theta(X_i) = \text{Var}_\theta(Y_i) = 1$, y que X_i es independiente de Y_i , para $i = 1 \dots n$. Interesa estimar $g(\theta) = l_a l_b$ (el área del rectángulo), para lo cual se proponen los estimadores $\hat{g} = \bar{X} \cdot \bar{Y}$ y $\tilde{g} = \frac{1}{n} \sum_{i=1}^n X_i Y_i$.

- i. Muestre que ambos estimadores son insesgados para $g(\theta)$.

Solución:

Usaremos que \bar{X} es independiente a \bar{Y} .

$$\mathbb{E}(\hat{g}) = \mathbb{E}(\bar{X} \cdot \bar{Y}) = \mathbb{E}(\bar{X})\mathbb{E}(\bar{Y}) = l_a l_b.$$

$$\mathbb{E}(\tilde{g}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i \cdot Y_i) = \frac{1}{n} \sum_{i=1}^n l_a l_b = l_a l_b.$$

- ii. Calcule las varianzas de los estimadores e indique cuál de los dos estimadores es preferible.

Solución:

Notemos que $\text{Var}(\bar{X}) = \text{Var}(\bar{Y}) = \frac{1}{n}$. Luego,

$$\text{Var}(\hat{g}) = \mathbb{E}(\bar{X}^2 \bar{Y}^2) - l_a^2 l_b^2 = (\text{Var}(\bar{X}) + l_a^2)(\text{Var}(\bar{Y}) + l_b^2) - l_a^2 l_b^2 = \frac{1}{n^2} + \frac{l_a^2 + l_b^2}{n}$$

$$\text{Var}(\tilde{g}) = \frac{1}{n^2} \sum_{i=1}^n (\mathbb{E}(X_i^2 Y_i^2) - l_a^2 l_b^2) = \frac{1}{n^2} \sum_{i=1}^n ((\text{Var}(X_i) + l_a^2)(\text{Var}(Y_i) + l_b^2) - l_a^2 l_b^2) = \frac{1 + l_a^2 + l_b^2}{n}$$

Como $\frac{1}{n^2} < \frac{1}{n}$ para $1 < n$, se preferiría el estimador \hat{g} , pues tiene menor varianza.

(b) Considere el experimento que consiste en hacer lanzamientos independientes de una moneda, con probabilidad de cara θ desconocida, hasta que hayan salido n caras. De este experimento se han registrado los largos, eventualmente nulos, de las rachas de sellos entre las n caras X_1, X_2, \dots, X_n , es decir, X_1 es el número de sellos desde el comienzo hasta la primera cara; X_2 es el número de sellos entre la primera y la segunda cara, etc.

- i. (1 pto.) Muestre que las v.a. X_1, X_2, \dots, X_n son una M.A.S. de la v.a. discreta X con valores $x = 0, 1, 2, \dots$ y función de probabilidad $p(x; \theta) = (1 - \theta)^x \theta; \theta \in [0, 1]$.

Solución:

Como los lanzamientos de la moneda son independientes y la probabilidad de que salga cara en cada lanzamiento es θ , las v.a. X_i serán todas independientes con distribución geométrica. Esto es, $X_1 = x$ cuando han habido x sellos seguidos y en el lanzamiento siguiente sale una cara. Esto ocurre con probabilidad $p(x; \theta) = (1 - \theta)^x \theta$.

- ii. (2 pts.) Calcule el Estimador de Máxima Verosimilitud de θ .

Solución:

Supongamos que se tiene una muestra $X_1 = x_1, \dots, X_n = x_n$. Buscamos el parámetro θ que maximiza la probabilidad conjunta de que ocurra esta muestra. Usando la independencia de las variables X_i , la probabilidad conjunta es:

$$f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n p(x_i; \theta) = \prod_{i=1}^n (1 - \theta)^{x_i} \theta = \theta^n (1 - \theta)^{\sum_{i=1}^n x_i}$$

Además, $f(x_1, \dots, x_n; \theta)$ tendrá el mismo punto de máximo que $\ln f(x_1, \dots, x_n; \theta)$.

Así encontramos θ como el parámetro que resuelve el problema

$$\max_{\theta \in [0, 1]} \ln \left[\theta^n (1 - \theta)^{\sum_{i=1}^n x_i} \right] = \max_{\theta \in [0, 1]} n \ln \theta + \ln(1 - \theta) \sum_{i=1}^n x_i$$

Por lo tanto, la condición de primer orden queda:

$$\frac{n}{\theta} - \frac{\sum_{i=1}^n x_i}{1 - \theta} = 0$$

de donde se despeja $\hat{\theta} = \frac{1}{1 + \bar{x}}$.

3. (**Sólo sección de Jorge Lemus**) La duración de unas determinadas baterías es una variable aleatoria normal, cuya media se desea estimar, para lo cual se toma una muestra de 16 baterías. El promedio de la duración es $\bar{x} = 7$ y la varianza muestral $S^2 = 0,9$.

- (a) (1.5 puntos) Encontrar un intervalo de confianza al 95% para estimar la media.

Solución: El intervalo de confianza está dado por

$$\left[\bar{x} - t_{n-1, \frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{x} + t_{n-1, \frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right] = \left[7 - 2,13 \frac{0,949}{4}, 7 + 2,13 \frac{0,949}{4} \right] = [6,49, 7,51]$$

- (b) (1.5 puntos) Encontrar un intervalo de confianza al 95% estimar la varianza.

Solución:

$$\left[(n-1) \frac{S^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}, (n-1) \frac{S^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right] = \left[\frac{13,5}{27,5}, \frac{13,5}{6,26} \right] = [0,49, 2,16]$$

- (c) (1.5 puntos) Suponga que se sabe que la varianza poblacional es $\sigma^2 = 1$ ¿cuál es el intervalo de confianza para la media en este caso?

Solución:

$$\left[\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right] = \left[7 - 1,96 \frac{1}{4}, 7 + 1,96 \frac{1}{4} \right] = [6,51, 7,49]$$

- (d) (1.5 puntos) Si se desea reducir un 20% el largo intervalo anterior, manteniendo el nivel de confianza, ¿ cuántas baterías adicionales se deberían probar?

Solución: El largo del intervalo está dado por $\ell(n) = 2z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$. Si se quiere reducir el largo en 20%, usando A baterías adicionales, se debe tener que $\ell(n+A) \leq 0.8\ell(n)$. De esta ecuación se despeja $A = \left\lceil \frac{n}{4} \right\rceil$, como el mínimo número de baterías que se deberían usar para que el largo del nuevo intervalo de confianza sea igual a un 80% del largo original.

Indicaciones: Si $Z \sim \mathcal{N}(0, 1)$, $\mathbb{P}(Z < 1.96) = 0.975$. Si $X \sim \chi_{15}^2$, $\mathbb{P}(X < 6, 26) = 0.025$, $\mathbb{P}(X < 27, 5) = 0.975$. Si $T \sim t_{15}$, $\mathbb{P}(T < 2, 13) = 0.975$

4. (**Sólo sección de Raúl Gouet**) Sea X_1, \dots, X_n una MAS del modelo uniforme sobre el intervalo $(0, \beta)$, donde $\beta > 0$ es un parámetro desconocido. Escribir un intervalo de confianza para β con nivel de confianza $1 - \alpha$. En particular, calcular un intervalo de confianza para $\alpha = 0.1$ si se obtiene la siguiente muestra de 4 elementos: 1.13, 0.67, 1.32, 0.27. Indicación: considere como pivote a la función $T(X, \beta) = \max\{X_1, \dots, X_n\}/\beta$

Tiempo: 3 horas.

Solución problema 4 Control 3

MA-3403 Prof. R. Gouet, 8/06/09

P4. Sea X_1, \dots, X_n una MAS del modelo uniforme sobre el intervalo $(0, \beta)$. Escribir un intervalo de confianza para β con nivel de confianza $1 - \alpha$. En particular, calcular un intervalo de confianza para $\alpha = 0.1$ si se obtiene la siguiente muestra de 4 elementos: 1.13, 0.67, 1.32, 0.27. Indicación: considere como pivote a la función $T(X, \beta) = \max\{X_1, \dots, X_n\}/\beta$.

Solución: T es una función pivote apropiada porque depende monótonamente del parámetro desconocido β y su distribución no depende de β . En efecto

$$P_\beta(\max\{X_1, \dots, X_n\}/\beta \leq t) = \prod_{i=1}^n P_\beta(X_i \leq \beta t) = F_\beta(\beta t)^n,$$

donde F_β designa la función de distribución de una va uniforme en $(0, \beta)$.

Por otra parte, resulta que

$$F_\beta(x) = \frac{x}{\beta} \mathbb{1}_{(0, \beta)}(x) + \mathbb{1}_{[\beta, \infty)}(x),$$

de manera que

$$F_\beta(\beta t) = \frac{\beta t}{\beta} \mathbb{1}_{(0, \beta)}(\beta t) + \mathbb{1}_{[\beta, \infty)}(\beta t) = t \mathbb{1}_{(0, 1)}(t) + \mathbb{1}_{[1, \infty)}(t) = F_1(t).$$

Obtenemos entonces

$$P_\beta(\max\{X_1, \dots, X_n\}/\beta \leq t) = F_1(t)^n.$$

Habiendo comprobado que es pivote, buscamos $t_{\alpha 1}, t_{\alpha 2}$ tales que

$$P_\beta(t_{\alpha 1} \leq \max\{X_1, \dots, X_n\}/\beta \leq t_{\alpha 2}) = F_1(t_{\alpha 2})^n - F_1(t_{\alpha 1})^n = 1 - \alpha.$$

Esta ecuación tiene múltiples soluciones, por ejemplo $t_{\alpha 2} = 1, t_{\alpha 1} = \alpha^{1/n}$. O bien, $t_{\alpha 1} = (\alpha/2)^{1/n}, t_{\alpha 2} = (1 - \alpha/2)^{1/n}$. En cualquier caso, el intervalo de confianza para β se obtiene despejando la desigualdad dentro de la probabilidad, para llegar a

$$\max\{X_1, \dots, X_n\}/t_{\alpha 2} \leq \beta \leq \max\{X_1, \dots, X_n\}/t_{\alpha 1}.$$

Para la aplicación numérica supongamos que se escoge la segunda solución, de manera que $t_{\alpha 1} = (\alpha/2)^{1/n} = (0.05)^{1/4} = 0.4728$ y $t_{\alpha 2} = (1 - \alpha/2)^{1/n} = (0.95)^{1/4} = 0.9873$. Además $\max\{1.13, 0.67, 1.32, 0.27\} = 1.32$, luego los extremos del intervalo son $1.32/0.9873 =$

1.337 y $1.32/0.4728 = 2.792$. Finalmente, el intervalo con nivel $100(1 - \alpha)\% = 90\%$ de confianza para β es

$$[1.337, 2.792].$$

Si escogemos la segunda solución encontramos $t_{\alpha 1} = (0.1)^{1/4} = 0.5623, t_{\alpha 2} = 1$ y el intervalo para β sería

$$[1.32, 2.35].$$

Comparando los intervalos en términos de sus largos, vemos que el segundo es mejor. Nota a los correctores: cualquier solución correcta vale como respuesta y no se espera que el estudiante compare soluciones ni comente.