Una ley de arco seno discreto

José Heber Nieto   Gonzalo Pérez Iribarren

Resumen

Una ley de arco seno discreto por J. H. Nieto y G. Pérez Iribarren. Universidad del Zulia, Facultad Experimental de Ciencias.

Sea X_1, X_2,..., X_{2p+1} una muestra aleatoria simple de una distribución uniforme en algún intervalo $[a, b]$. Consideremos la muestra ordenada $X^{(1)} < X^{(2)} < \cdots < X^{(2p+1)}$ y las diferencias $e_i = X^{(p+i+1)} - X^{(p+1)}$, $h_i = X^{(p+1)} - X^{(p-i+1)}$, $(i = 1, 2, \ldots, p)$. Sea además M el número de índices i tales que $e_i < h_i$. En este trabajo se prueba que la distribución de M es la del *arco seno discreto*, es decir que $\mathrm{P}(M = k) = \binom{2k}{k}\binom{2p-2k}{p-k}2^{-2p}$, $k = 0, 1, \ldots, p$. También se consideran algunas variantes de este resultado y posibles aplicaciones estadísticas.

Abstract

A discrete arcsin law by J. H. Nieto y G. Pérez Iribarren. Universidad del Zulia, Facultad Experimental de Ciencias.

Let X_1, X_2,..., X_{2p+1} be independent random variables uniformly distributed in some interval $[a, b]$. Consider the order statistic $X^{(1)} < X^{(2)} < \cdots < X^{(2p+1)}$ and the differences $e_i = X^{(p+i+1)} - X^{(p+1)}$, $h_i = X^{(p+1)} - X^{(p-i+1)}$, $(i = 1, 2, \ldots, p)$. Let M be the number of indexes i such that $e_i < h_i$. In this paper it is proved that the distribution of M follows the *discrete arcsin law*, i.e., $\mathrm{P}(M = k) = \binom{2k}{k}\binom{2p-2k}{p-k}2^{-2p}$, $k = 0, 1, \ldots, p$. Some variants of this result and possible statistic applications are considered.

1. Introducción

Sea X_1, X_2,..., X_{2p+1} una muestra aleatoria simple de una cierta distribución F. Sea $X^{(1)} < X^{(2)} \cdots < X^{(2p+1)}$ la muestra ordenada (o estadístico de orden) y definamos también:

$$\begin{aligned} e_i &= X^{(p+i+1)} - X^{(p+1)}, \\ h_i &= X^{(p+1)} - X^{(p-i+1)}, \qquad (1 < i < p) \end{aligned}$$

Por último sea M el número de índices i para los cuales se cumple $e_i < h_i$.

La motivación para estudiar los estadísticos e_i, h_i y M surgió del problema de hallar pruebas estadísticas para la simetría de una densidad respecto a un punto desconocido (ver [3]). Como M es invariante por traslaciones de la distribución F, parece razonable que refleje de algún modo cualquier posible simetría. Sin embargo el problema se dificulta por la no-independencia de los e_i y h_i. Como primera aproximación realizamos algunas simulaciones con computadora. Escribimos una rutina en assembler para generar números pseudo-aleatorios en el intervalo [0,1] y ordenarlos empleando el algoritmo de inserción en listas múltiples ([2], vol. 3, p. 99). De esta manera logramos altas velocidades de ejecución, que nos permitieron trabajar con altos valores de p. y miles de ensayos, para diversas distribuciones F. En el caso de la distribución uniforme obtuvimos un claro predominio de valores extremos de M (es decir, cercanos a 0 o a p). Es más, la representación gráfica de los resultados obtenidos nos hizo pensar en la densidad

$$f(x) = \frac{1}{\pi\sqrt{x(1-x)}}, \quad x \in (0,1),$$

la cual aparece en la teoría de paseos al azar (ver [1] vol. I, Cap. 3) como límite de la distribución llamada *arco seno discreto*, es decir

$$a_{2k,2n} = \binom{2k}{k}\binom{2n-2k}{n-k}2^{-2n}$$

(en efecto, se puede probar que $\sum_{k\leq nx} a_{2k,2n} \approx \int_0^x f(u)\,du = \frac{1}{\pi}\mathrm{Arcsen}(\sqrt{x})$). Estas consideraciones son las que nos condujeron al teorema A.

2. Resultados

Teorema A. *Sean X_1, X_2,..., X_{2p+1} variables aleatorias independientes con distribución uniforme en en un intervalo $[a,b]$. Entonces*

$$\mathbf{P}(M=k) = \binom{2k}{k}\binom{2p-2k}{p-k}2^{-2p}, \quad k = 0, 1, \ldots, p.$$

Demostración. Si X_1, X_2,..., X_{2p+1} es una muestra aleatoria simple proveniente de una densidad f es sabido que la densidad conjunta correspondiente al estadístico de orden respectivo es

$$(2p+1)! \prod_{i=1}^{2p+1} f(X^{(i)}), \qquad (X^{(1)} < X^{(2)} \cdots < X^{(2p+1)})$$

(ver por ejemplo [4]).

Si introducimos las nuevas variables $e_i = X^{(p+i+1)} - X^{(p+1)}$, $h_i = X^{(p+1)} - X^{(p-i+1)}$ para $i = 1, \ldots, p$ y conservamos $X^{(p+1)}$, como el jacobiano de la transformación efectuada es 1 en valor absoluto resultará la siguiente densidad conjunta:

$$(2p+1)!\Big(\prod_{j=1}^{p} f(X^{(p+1)} + e_j) f(X^{(p+1)} - h_j)\Big) f(X^{(p+1)}),$$

con $0 < e_1 < e_2 < \cdots < e_p$, $0 < h_1 < h_2 < \cdots < h_p$.

En nuestro caso (distribución uniforme en [a,b]) tenemos que $f(x) = 1$ si $x \in [a,b]$, $f(x) = 0$ si $x \notin [a,b]$. Es claro que sin pérdida de generalidad podemos suponer $[a,b] = [0,1]$ (el caso general se reduce a éste mediante un cambio lineal de variable) lo cual nos permitirá simplificar la notación. Así, la densidad conjunta de e_i, h_i $(i = 1, \ldots, p)$ y $X^{(p+1)}$ será

$$\begin{cases} (2p+1)! & \text{si } 0 < e_1 < \cdots < e_p < 1 - X^{(p+1)},\ 0 < h_1 < \cdots < h_p < X^{(p+1)}, \\ 0 & \text{en caso contrario.} \end{cases}$$

La densidad conjunta de e_i, h_i $(i = 1, \ldots, p)$ obtenida marginando $X^{(p+1)}$ será entonces:

$$g(e_1, \ldots, e_p, h_1, \ldots, h_p) = \begin{cases} (2p+1)! \int_{h_p}^{1-e_p} dx = (2p+1)!(1 - e_p - h_p) \\ \quad \text{si } 0 < e_1 < \cdots < e_p\ ,\ 0 < h_1 < \cdots < h_p \\ \quad \text{y } e_p + h_p < 1, \\ 0 \text{ en caso contrario.} \end{cases}$$

Esta será nuestra densidad fundamental en lo sucesivo.

Sea

$$A_1 = \{(e_1, \ldots, e_p, h_1, \ldots, h_p) : e_1 < \cdots < e_{p-1} < h_1 < \cdots < h_{p-1} < e_p < h_p\}.$$

La probabilidad de este suceso es:

$$\begin{aligned} \mathbf{P}(A_1) &= \int_{A_1} \cdots \int g(e_1, \ldots, e_p, h_1, \ldots, h_p) de_1 \ldots de_p\, dh_1 \ldots dh_p \\ &= (2p+1)! \int_0^{1/2} de_p \int_{e_p}^{1-e_p} (1 - e_p - h_p) dh_p \int_0^{e_p} dh_{p-1} \int_0^{h_{p-1}} dh_{p-2} \cdots \\ &\qquad \cdots \int_0^{h_2} dh_1 \int_0^{h_1} de_{p-1} \int_0^{e_{p-1}} de_{p-2} \cdots \int_0^{e_2} de_1 \end{aligned}$$

$$= (2p+1)! \int_0^{1/2} de_p \int_{e_p}^{1-e_p} \frac{e_p^{2p-2}}{(2p-2)!} dh_p = (2p+1)! \int_0^{1/2} \frac{1}{2}(1-2e_p)^2 \frac{e_p^{2p-2}}{(2p-2)!} de_p$$

Haciendo el cambio de variable $u = 2e_p$ resulta:

$$\mathbf{P}(A_1) = \frac{1}{2^{2p-1}} \int_0^1 \frac{(2p-1)!}{2!(2p-2)!} (1-u)^2 u^{2p-2} du = \frac{1}{2^{2p-1}}.$$

El cálculo anterior muestra también que si en las primeras $2p-3$ desigualdades que definen el suceso A_1 se permutan las variables $e_1, \ldots, e_{p-1}, h_1, \ldots, h_{p-1}$ respetando el orden $e_1 < e_2 < \cdots < e_{p-1}$, $h_1 < h_2 < \cdots < h_{p-1}$, se obtienen sucesos con la misma probabilidad. El número de estos sucesos es $\binom{2p-2}{p-1}$. Más en general definamos

$$\begin{aligned} A_r = \{(e_1, \ldots, e_p, h_1, \ldots, h_p) : e_1 < \cdots < e_{p-1} < h_1 < \cdots \\ \cdots < h_{p-r} < e_p < h_p < h_{p-r+1} < \cdots < h_p\}. \end{aligned}$$

Un cálculo similar al realizado muestra que

$$\mathbf{P}(A_r) = \frac{1}{2^{2p-r}}, \quad 1 \leq r \leq p.$$

Además, permutando las variables $e_1, \ldots, e_{p-1}, h_1, \ldots, h_{p-r}$ en las primeras $2p - r - 2$ desigualdades que definen el suceso A_r (aunque sin alterar el orden $e_1 < \cdots < e_{p-1}$, $h_1 < \cdots < h_{p-r}$) se obtienen sucesos con igual probabilidad que A_r.

Ahora bien, cada uno de estos sucesos se pueden representar en el plano Oxy mediante una poligonal que parte del origen y llega al punto $(2p, 0)$.

Para esto definamos s_i como $+1$ o -1 según que la variable i-sima en la cadena de desigualdades que define el suceso sea del tipo e o del tipo h. Sea además $S_0 = 0$ y para $1 \leq j \leq 2p$, sea $S_j = s_1 + \cdots + s_j$. La poligonal asociada al suceso es la que tiene como i-simo vértice el punto $P_i = (i, S_i)$.

Para cada uno de estos sucesos el número de índices i tales que $e_i < h_i$ es igual a la mitad del número de segmentos que la poligonal asociada tiene en el semiplano $y > 0$. En efecto, $e_i < h_i$ si y sólo si entre las primeras $2i-1$ variables (en las desigualdades que definen el suceso) hay al menos i del tipo e. Esto a su vez es equivalente a la desigualdad $S_{2i-1} > 0$, la cual se verifica si y sólo si $P_{2i-2}P_{2i-1}$ y $P_{2i-1}P_{2i}$ están en el semiplano $y \geq 0$, etc.

Observemos que con esta representación los sucesos del tipo A_r quedan caracterizados como aquellos cuyas poligonales cortan la recta $x + y = 2p$ por primera vez en el punto $(2p - r, r)$. Dado uno de estos sucesos con poligonal $P_0P_1 \ldots P_{2p}$ existen exactamente 2^r trayectorias posibles para un paseo al azar simétrico entre 0 y $2p$ cuyos primeros $2p - 2r$ segmentos coincidan con $P_0P_1 \ldots P_{2p-2r}$. La probabilidad del suceso, a saber $2^{-(2p-r)}$, es igual a la suma de las probabilidades de las trayectorias mencionadas ($2^r 2^{-2p} = 2^{-2p+r}$) todas las cuales además tienen el mismo número de segmentos en el semiplano $y \geq 0$ que la poligonal $P_0P_1 \ldots P_{2p}$. Consideraciones análogas pueden hacerse para los sucesos simétricos de los del tipo A_r, que se obtienen cambiando e_i por h_i en las desigualdades: les corresponden poligonales que cortan por primera vez la recta $x - y = 2p$ en el punto $(2p - r, -r)$, etc. Los razonamientos anteriores nos permiten afirmar que el problema de calcular $\mathbf{P}(M = k)$ es equivalente al de calcular la probabilidad de que en un paseo al azar simétrico entre 0 y $2p$ haya $2k$ segmentos en el semiplano superior. Pero esta probabilidad se sabe que es $a_{2k,2p} = \binom{2k}{k}\binom{2p-2k}{p-k}2^{-2p}$ (ver [1], p. 82) con lo cual se concluye la prueba. $\square$

Caso de un número par de observaciones

Sea ahora $X^{(1)} < X^{(2)} < \cdots < X^{(2p+2)}$ la muestra ordenada y sea $Z = (X^{(p+1)} + X^{(p+2)})/2$. Definamos $e_i = X^{(p+2+i)} - Z$, $h_i = Z - X^{(p+1-i)}$, $i = 1, \ldots, p$. Sea $M = \#\{i : e_i < h_i\}$. Entonces $\mathbf{P}(M = k) = a_{2k,2p}$.

Otra posibilidad es definir $e_i = X^{(2p+2)} - X^{(2p+2-i)}$, $h_i = X^{(i+1)} - X^{(1)}$. También así se tiene $\mathbf{P}(M = k) = a_{2k,2p}$.

3. Algunas posibles aplicaciones

El teorema A puede utilizarse para construir pruebas de hipótesis, siendo la hipótesis nula que la distribución es uniforme en algún intervalo (no especificado).

Como un segundo ejemplo supongamos que X_1, X_2,. . . ,X_{2p+1} es una muestra aleatoria simple de una distribución, y queremos probar la hipótesis de que proviene de una exponencial de la forma $F(x) = 1 - e^{-m(x-a)}$ (para $x \geq a$), con m conocido y a desconocido. Sin pérdida de generalidad supongamos que $m = 1$ y efectuemos el cambio de variables $Y_i = e^{-X_i}$. Por lo tanto,

$$\mathbf{P}(Y_i \leq z) = \mathbf{P}(X_i \geq -\log z) = e^{-(-\log z - a)} = ze^a.$$

O sea que si la hipótesis es verdadera las Y_i se distribuyen uniformemente entre 0 y e^{-a}. El teorema A provee una distribución exacta que puede usarse para probar la uniformidad de las Y_i.

No es díficil imaginar otras aplicaciones, sin embargo sería necesario un estudio más detenido para evaluar su eficacia.

Referencias

[1] FELLER, W. *An Introduction to Probability Theory and its Applications*, Wiley, 1967.

[2] KNUTH, D. E. *The Art of Computer Programming, vol. 3*, Addison Wesley, 1973.

[3] PEREZ, G. *Sobre la Simetría de una Densidad*, Cursos, Seminarios y Tesis del PEAM, Universidad del Zulia, Maracaibo, 1979.

[4] SARHAN, A. E., GREENBERG, B. G., *Contributions to Order Statistics*, Wiley, 1962.