Revista Colombiana de Estadística
Nºs 19-20, junio-diciembre 1989

TEORIA DE RACHAS

Jimmy A. Corzo S.

Profesor Asistente
Departamento de Matemáticas y Estadística
Universidad Nacional de Colombia

Resumen. Una sucesión de uno o más símbolos idénticos seguidos o precedidos por uno o más símbolos diferentes o por ningún símbolo se llama una racha Gibbons (1971). En (1.9) se da una definición formal de racha.

En este artículo se define un proceso estocástico cuyas variables aleatorias indican el número de rachas hasta cada elemento en un arreglo de un número fijo de símbolos de dos clases. Este proceso se le llama aquí Proceso de Rachas. Para cada variable aleatoria se da la distribución exacta, se calcula su valor esperado y su varianza y se da la covarianza entre cualquier par de variables del proceso. Finalmente se demuestra que con una adecuada normalización, el proceso de rachas es una submartingala.

1. Definiciones, notación y relación entre Rachas y Rangos.

Sean X_i, $i = 1, \ldots, m, m+1, \ldots, N$, $N = m+n$, $m, n = 1, 2, \ldots$

variables aleatorias (v.a's) independientes definidas sobre un espacio de probabilidad (Ω',A',P'), con función de distribución continua $F(x) = P'(X \leq x)$. Entonces el vector $X = (X_1,\ldots X_m,X_{m+1},\ldots,X_N)$ esta definido sobre el espacio producto (Ω,A,P), donde $\Omega = \Omega'\times\ldots\times\Omega'$, $A = A'\otimes\ldots\otimes A'$, $P = P'\otimes\ldots\otimes P'$.

La transición al correspondiente vector de estadísticas de orden $X = (X_{1;N},\ldots,X_{N;N})$ implica la aplicación

$$(1.1) \qquad \vec{X}: (\Omega,A,P) \longrightarrow (\mathbb{R}^N_<,\mathcal{B}_{N,<},\vec{P})$$

donde $\mathbb{R}^N_< = \{ (x_{1,N},\ldots,x_{N;N}) \in \mathbb{R}^N: x_{1,N}<\ldots<x_{N;N} \}$, $\mathcal{B}_{N,<}$ es la σ-álgebra de los conjuntos de Borel en $\mathbb{R}^N_<$ y $\vec{P}$ es la medida inducida por $\vec{X}$ en $\mathcal{B}_{N,<}$.

A través del vector $\varepsilon = (1,\ldots,1,0,\ldots,0)$ quedan representadas las primeras m v.a's con m unos y las segundas n v.a's con n ceros. Sea $C_{m,n}$ el conjunto de todos los arreglos distinguibles de los elementos de ε. Entonces con la aplicación $\vec{X}$ está dado un segundo vector aleatorio.

$$(1.2) \qquad \eta : (\Omega,A,P) \longrightarrow (C_{m,n}, \mathcal{P}(C_{m,n}),\tilde{P})$$

cuya j-ésima componente indica si la j-ésima estadística de orden $X_{j;N}$ corresponde a una de las primeras m v.a's ($\eta_j = 1$) o a una de las segundas m v.a's ($\eta_j = 0$).

Se definen el rango R_i de la i-ésima observación y el antirango D_j de la j-ésima estadística de orden a través del las ecuaciones

$$(1.3) \qquad X_i = X_{R_i;N}, \qquad i = 1,\ldots,N$$

(1.4) $$X_{\mathcal{D}_j} = X_{j;N} \qquad j = 1,\ldots,N.$$

De las definiciones (1.3) y (1.4) se concluye la identidad

(1.5) $$\mathcal{D}_{R_i} = i, \qquad i = 1,\ldots,N.$$

Si $\vec{X} = \vec{x}$, $\vec{x} \in R^N$ y $\eta = c$, $c \in C_{m,n}$ entonces $X = x$, $x \in \mathbb{R}^N$ y $x_i = x_{R_i;N}$. Por lo tanto el par $(\vec{X},\eta)$ es una estadística suficiente para la familia de posibles distribuciones de X.

El evento $(\mathcal{D}_j = k)$ significa que X_k es la j-ésima v.a. más pequeña del vector X. El evento

(1.6) $$A_j = (1 \leqslant \mathcal{D}_j \leqslant m)$$

significa que $X_{j;N}$ es una de las primeras m v.a's. y su complemento

(1.7) $$A_j^c = (m+1 \leqslant \mathcal{D}_j \leqslant N)$$

indica que $X_{j;N}$ es una de las últimas n v.a's. Entonces para la j-ésima componente del vector η se cumple lo siguiente:

(1.8) $$\eta(\omega) = \begin{cases} 1 & \omega \in A_j \\ 0 & \omega \in A_j^c \end{cases}.$$

A la representación de las estadísticas de orden por medio de unos y ceros en las componentes del vector η se le denomina **dicotomización** y a la sucesión $\eta_1,\ldots,\eta_N$, se le llama la **muestra dicotomizada.**

La muestra dicotomizada tiene la siguiente estructura

$$(1.9)\qquad \eta_1 = \ldots = \eta_{L_1} \neq \eta_{L_1+1} = \ldots = \eta_{L_1+L_2} \neq \ldots \neq \eta_{L_1+L_2+\ldots+L_{r_N-1}+1} = \ldots = \eta_{L_1+\ldots+L_{r_N}} .$$

en la cual hay $r_N \geq 2$ grupos de unos y ceros. Cada uno de estos grupos se denomina una racha y el número de elementos en la j-ésima racha L_j, $j = 1,\ldots,r_N$ se llama longitud de la racha. r_N es el número total de rachas en la muestra dicotomizada. Obviamente r_N y L_j, $j = 1,\ldots,r_N$ son v.a's.

OBSERVACION 1. Esta definición de rachas y longitud de las rachas es una adaptación de la definición de empates (ties) dada por Hájek (1967), p.119.

OBSERVACION 2. Nótese que una racha queda completamente caracterizada cuando se especifica su posición dentro de la muestra dicotomizada, su longitud, y la clase de símbolos que representa.

A continuación se definen las v.a's contadores. Sean $I_1(\omega) = 1$ para todo $\omega \in \Omega$ e

$$(1.10)\qquad I_j(\omega) = \begin{cases} 1 & \omega \in \{A_{j-1} \cap A_j^c\} \cup \{A_{j-1}^c \cap A_j\} \\ 0 & \omega \in \{A_{j-1} \cap A_j\} \cup \{A_{j-1}^c \cap A_j^c\}, \end{cases} \qquad j = 2,\ldots N.$$

De (1.8) y (1.10) se deduce que

$$(1.11)\qquad I_j(\omega) = \begin{cases} 1 & \eta_{j-1}(\omega) \neq \eta_j(\omega) \\ 0 & \eta_{j-1}(\omega) = \eta_j(\omega) . \end{cases}$$

Se define el número de rachas hasta el j-ésimo elemento de la muestra dicotomizada por medio de la v.a.

$$
\begin{aligned}
r_j(\omega) &= \sum_{k=1}^{j} I_k(\omega) \qquad j = 1,\ldots,N \\
&= 1 + \sum_{k=2}^{j} I_k(\omega) \qquad j = 2,\ldots,N.
\end{aligned}
\tag{1.12}
$$

OBSERVACION 3. En particular si las primeras m v.a's. del vector X son una muestra de una población con distribución F_1 y las segundas n son una muestra de una población con distribución F_2, entonces para $j = N$, r_N es el número total de rachas en la muestra combinada dicotomizada y corresponde a la estadística de prueba propuesta por Wald y Wolfowitz (1940) para la prueba de la hipótesis de que las dos muestras provienen de la misma población, contra la alternativa general de que éstas provienen de poblaciones distintas.

OBSERVACION 4. La sucesión $\{r_j,\ j = 1,\ldots,N\}$ es una generalización de la estadística propuesta por Wald y Wolfowitz (1940) puesto que genera no solamente el número total de rachas en la muestra dicotomizada sino también el número de rachas hasta cada elemento de ésta.

OBSERVACION 5. Gibbons (1971) presenta una definición similar a (1.12) para el número total de rachas r_N en la muestra combinada dicotomizada, pero no define los contadores I_k, $k = 1,\ldots,N$ a través de los antirangos.

En seguida se establece una relación entre rachas y rangos. Para esto es necesario definir las v.a's. A_j, η_j, I_j y r_j $j = 1,\ldots,N$ con los subíndices aleatorios R_i es decir definir

las v.a's. A_{R_i}, η_{R_i}, I_{R_i} y ν_{R_i}, $i = 1,\ldots,N$.

De la identidad (1.5) se obtiene

$$A_{R_i} = \{1,\ldots,m\}$$

$$A^c_{R_i} = \{m+1,\ldots,N\} \qquad i = 1,\ldots,N$$

y por lo tanto $$\eta_{R_i} = \begin{cases} 1 & i = 1,\ldots,m \\ 0 & i = m+1,\ldots,N \end{cases},$$

lo cual a su vez implica

$$\varepsilon = (\eta_{R_1},\ldots,\eta_{R_N}).$$

En analogía con (1.9) se pueden definir las v.a's. contadoras con subindices aleatorios de la siguiente manera:

$$I_{R_i} = \begin{cases} 1 & \eta_{R_i-1} \neq \eta_{R_i} \\ 0 & \eta_{R_i-1} = \eta_{R_i} \end{cases}$$

con la condición de que $I_{R_i} = 1$ cuando $X_i = \text{mín}\{X_1,\ldots,X_N\}$.

Entonces de acuerdo con (1.12) y (1.13) se puede definir el número de rachas hasta la v.a. X_i así

$$\nu_{R_i} = \sum_{\{k:R_k \leq R_i\}} I_{R_i} \qquad i = 1,\ldots,N. \tag{1.14}$$

En particular si $R_i = N$ es porque $X_i = \text{máx}\{X_1,\ldots,X_N\}$ y por esto se puede interpretar a ν_N como el número de rachas hasta la v.a. $\text{máx}\{X_1,\ldots,X_N\}$. En general nótese que ν_{R_i} se

puede interpretar como el número de rachas hasta la v.a. que se encuentra en la posición R_i en la sucesión $X_1,\ldots,X_N$.

2. Distribución y momentos de las Rachas.

La deducción de la distribución de r_j para cada j fijo, $j=1,\ldots,N$ está basada prácticamente en el mismo argumento que utilizaron Wald y Wolfowitz (1940) para la deducción de la distribución del número total de rachas r_N.

Sean j fijo con $j = 1,\ldots,N$ y K el número de unos que sobran en la muestra dicotomizada $\eta_1,\ldots,\eta_N$ después de eliminar sus últimos $N-j$ elementos. Entonces K es una v.a. con valores $1,\ldots,\nu$, donde $\nu = \min\{m,j\}$ y tiene distribución hipergeométrica es decir:

$$P(K=k) = \frac{\binom{m}{k}\binom{n}{j-k}}{\binom{N}{j}} \qquad \begin{matrix} k = 1,\ldots,\nu \\ j = 1,\ldots,N-1 \end{matrix} .$$

Para j fijo, r_j es el número total de rachas en la subsucesión $r_1,\ldots,r_j$ que queda después de eliminar los últimos $N-j$ elementos de $\eta_1,\ldots,\eta_N$. Entonces de la distribución del número total de rachas calculadas por Wald y Wolfowitz (1940) se puede calcular la siguiente probabilidad condicional:

$$P(r_j = 2h/K=k) = \frac{2\binom{k-1}{h-1}\binom{j-k-1}{h-1}}{\binom{j}{k}} , \qquad h=1,\ldots,m$$

y

$$P(r_j = 2h+1/K=k) = \frac{\binom{k-1}{n}\binom{j-k-1}{h-1}+\binom{k-1}{h-1}\binom{j-k-1}{h}}{\binom{j}{k}} \qquad h=0,..,m-1$$

Usando la fórmula de Bayes y estas dos probabilidades condicionales se concluye que

$$(2.1) \qquad P(r_j=2h) = \sum_{k=1}^{\nu} P(r_j=2h/K=k)P(K=k)$$

$$= \sum_{k=1}^{\nu} \frac{2\binom{k-1}{h-1}\binom{j-k-1}{h-1}\binom{m}{k}\binom{n}{j-k}}{\binom{j}{k}\binom{N}{j}}$$

$$(2.2) \qquad P(r_j=2h+1) = \sum_{k=1}^{\nu} \frac{\binom{k-1}{h}\binom{j-k-1}{h-1}+\binom{k-1}{h-1}\binom{j-k-1}{h}\binom{m}{k}\binom{n}{j-k}}{\binom{j}{k}\binom{N}{j}}$$

En particular cuando $j = N$ se tiene $\nu = m$ y en consecuencia $P(K = m) = 1$. Es decir que en (2.1) y (2.2) solo queda el sumando para $k = m$ que es la distribución del número total de rachas en la muestra dicotomizada.

De las fórmulas (2.1) y (2.2) se pueden calcular directamente los momentos y las covarianzas entre las r_j, $j = 1,...,N$ pero esto resulta engorroso. Una manera mucho más simple de hacerlo es a través de la definición (1.12) y su relación con (1.10).

Después de algunos elementales pero laboriosos cálculos que se hacen a través de la distribución de los antirrangos, se obtienen las siguientes fórmulas para el valor esperado, la varianza y Covarianza, Corzo (1990).

$$(2.3) \qquad E(r_j) = 1 + (j-1)\frac{2mn}{N(N-1)} \qquad j = 1,...N.$$

$$(2.4)\quad \mathrm{Var}(r_j) = \mathrm{Var}\left(\sum_{k=2}^{j} I_k\right) = \sum_{k=2}^{j} \mathrm{Var}(I_k) + \sum_{2\le k_1\ne k_2<j} \mathrm{Cov}(I_{k_1}, I_{k_2})$$

$$= \begin{cases} 0 & j = N = 2 \\ \frac{2}{9}, & j = 2,3, \quad N = 3 \\ \frac{2mn}{N(N-1)}\left\{(j-1)\left[1-\frac{2mn}{N(N-1)}\right] + (j-2)\left[1-\frac{4mn}{N(N-1)}\right] + \right. & \\ \left. 2(j-2)(j-3)\left[\frac{(m-1)(n-1)}{(N-2)(N-3)} - \frac{mn}{N(N-1)}\right]\right. , & 2 \le j \le N \\ & N \ge 4 \end{cases}$$

$$\mathrm{Conv}(r_i, r_j) = E(r_i r_j) - E(r_i)E(r_j)$$

$$= (4i-5)\frac{mn}{N(N-1)} + \left\{(i-2)^2 + (i-1)(j-i-1)\right\}\frac{4mn(m-1)(n-1)}{N(N-1)(N-2)(N-3)}$$

$$- (i-1)(j-1)\frac{4m^2n^2}{N^2(N-1)^2} \qquad 2 \le i \le j \le N, \quad N \ge 4 .$$

Tomando $j = N$ en las fórmulas (2.3) y (2.4) se obtienen las fórmulas conocidas para el valor esperado y la varianza del número total de rachas en la muestra dicotomizada obtenidos por Wald y Wolfowitz en (1940).

3. Estructura de Submartingala del Proceso de Rachas

En este parágrafo se adapta el proceso de rachas $\{(r_j,\ j=1,\ldots,N),\ N \ge 2\}$ a una sucesión de σ-álgebras con respecto a la cual éste, adecuadamente normalizado, es una submartingala.

Para poner en evidencia la obvia dependencia de N de las v.a's. η_j, I_j, π_j, $j = 1,\ldots,N$ definidas en (1.8), (1.10) y (1.12) se utilizará la notación $\eta_{N,j}$, $I_{N,j}$, $\pi_{N,j}$, $j = 1,\ldots,N$ respectivamente.

Sean $T_{N,1} = \{\phi,\Omega\}$, $T_{N,j} = \sigma(\eta_{N,1},\ldots,\eta_{N,j})$, $j = 2,\ldots,N$, $\sigma(I_{N,1},\ldots,I_{N,j})$ y $\sigma(\pi_{N,1},\ldots,\pi_{N,j})$, $j = 1,\ldots,N$, las σ-álgebras generadas por las v.a's. indicadas entre los paréntesis respectivamente.

Entonces de (1.8), (1.11) y (1.12) se deduce que

$$(3.1) \qquad T_{N,j} = \sigma(I_{N,1},\ldots,I_{N,j}) = \sigma(\pi_{N,1},\ldots,\pi_{N,j}) \qquad j=1,\ldots,N.$$

y por lo tanto $\pi_{N,j}$ es medible respecto a $T_{N,j}$ para todo $j = 1,\ldots,N$.

Además por definición de $T_{N,j}$, $j = 1,\ldots,N$, es válida la relación

$$(3.2). \qquad T_{N,1} \subseteq T_{N,2} \subseteq \cdots \subseteq T_{N,N} \subseteq A.$$

Por otra parte de (1.12) $\pi_{N,j}$ se puede escribir también como $\pi_{N,j} = \pi_{N,j-1} + I_{N,j}$. Entonces por la medibilidad de $\pi_{N,j}$ respecto a $T_{N,j}$, $j = 1,\ldots,N$ y como $I_{N,j} \geq 0$ para todo $j = 2,\ldots,N$ se concluye que

$$(3.3) \qquad E(\pi_{N,j}/T_{N,j-1}) = \pi_{N,j-1} + E(I_{N,j}/T_{N,j-1})$$

$$\geq \pi_{N,j-1} \qquad j = 2,\ldots,N, \quad \text{casi seguro (c.s.)}$$

En consecuencia, si se demuestra que una normalización

de $r_{N,j}$ digamos $r^*_{N,j}$ $j = 1,\ldots,N$ tiene valor esperado finito para todo N, entonces la sucesión $\{(r^*_{N,j}, T_{N,j}, j = 1,\ldots,N), N \geq 2\}$ es una sucesión de submartingalas.

A continuación se define la sucesión normalizada $r^*_{N,j}$. Sean

$$S_1^2 = S_2^2 = 1 \tag{3.4}$$

$$S_N^2 = \mathrm{Var}(r_N) = \frac{2mn(2mn-m-n)}{N^2(N-1)} \qquad N \geq 3$$

y

$$r^*_{N,j} = \frac{r_{N,j}}{\sqrt{N}\, S}, \qquad j = 1,\ldots,N \quad N \geq 2 \tag{3.5}$$

De (3.4) y (3.5) sigue

$$E(r_{2,1}) = \frac{1}{2}$$

$$E(r_{2,2}) = \frac{1}{\sqrt{2}}\,(1 + E(I_2)) \tag{3.6}$$

$$E(r_{N,1}) = \frac{1}{\sqrt{N}\, S_N}, \qquad N \geq 3$$

y como $0 \leq r_{N,1} \leq \ldots \leq r_{N,N}$ también vale

$$E(r_{N,j}) = \frac{1}{\sqrt{N}\, S_N}\left(1 + (j-1)\,\frac{2mn}{N(N-1)}\right) \tag{3.7}$$

$$\leq \frac{1}{\sqrt{N}\, S_N} + \frac{2}{N\sqrt{N}\, S_N} \qquad j = 1,\ldots,N, \quad N \geq 2$$

Sean $N = m_N + n_N$, m_N, $n_N = 1,2,\ldots$, y $\lim\inf_N \frac{m_N}{n_N} = \alpha$ una cons-

tante positiva. Entonces para S_N^2 como en (3.4) vale lo siguiente:

$$(3.8)\quad \lim_{m_N\to\infty} \frac{1}{\sqrt{N}\, S_N} = \lim_{m_N\to\infty} \left(\frac{(1-\alpha)(1+\alpha-\frac{1}{n_N}}{2m_N n_N (2\alpha - \frac{1}{n_N}\alpha - \frac{1}{n_N})} \right)^{\frac{1}{2}} = 0 \ .$$

Además

$$\frac{2m_N n_N}{N\sqrt{N}\, S_N} = \frac{\sqrt{2m_N n_N}}{n_N(1+\alpha)} \left(\frac{(1-\alpha)(1+\alpha-\frac{1}{n_N}}{2\alpha - \frac{1}{n_N}\alpha - \frac{1}{n_N}} \right)^{\frac{1}{2}}$$

$$= \left(\frac{2\alpha\,(1+\alpha-\frac{1}{m_N}}{(1+\alpha)(2\alpha - \frac{1}{m_N}\alpha^2 - \frac{\alpha}{m_N}} \right)^{\frac{1}{2}}$$

de aquí se deduce que

$$(3.9)\quad \lim_{m_N\to\infty} \frac{2m_N n_N}{N\sqrt{N}\, S_N} = 1$$

Por lo tanto de (3.6) - (3.9)

$$(3.10)\quad E(r^*_{N,j}) < \infty\ , \quad j = 1,\ldots,N, \quad N \geqslant 2$$

De (3.3) y (3.10) y como $r^*_{N,j}$ es medible con respecto a $T_{N,j}$ para todo $j = 1,\ldots,N$, $N \geqslant 2$ se concluye que $(r^*_{N,j},\ T_{N,j}$, $j = 1,\ldots,N$ es una submartingala. A la sucesión doble $\{(r^*_{N,j}\ ,\ T_{N,j},\ j = 1,\ldots,N),\ N \geqslant 2\}$ se le llama una sucesión de submartingalas.

Por (3.6)-(3.9) y puesto que $r^*_{N,j} \geqslant 0$, $j = 1,\ldots,N$ se tiene también

(3.11) $$\lim_{m_N \to \infty} E(r^*_{N,j}) < \infty$$

y consecuentemente, por el teorema de convergencia de submartingalas (Doob (1959), teorema VII 4.1) existe una v.a. Z tal que

(3.12) $$\lim_{m_N \to \infty} r^*_{N,j} = Z \qquad \text{(c.s.)}$$

y además

$$E(Z) \leq \lim_{m_N \to \infty} E(r^*_{N,j}) < \infty.$$

Conclusiones.

La sucesión $\{r_j,\ j = 1,\ldots,N\}$ no solamente es una generalización de la estadística propuesta por Wald y Wolfowitz (1940) sino que es por sí misma una contribución a la teoría de rachas, puesto que a través de esta queda representada cada una de las variables aleatorias $X_1,\ldots,X_N$ por el número de rachas hasta cada una de ellas como se puede concluir de (1.14). Es decir, dentro de la clase de las estadísticas en rachas, el vector $(r_1,\ldots,r_N)$ desempeña un papel equivalente al vector de rangos $(R_1,\ldots,R_N)$ dentro de las estadísticas basadas en rangos.

En 1954 Mood demostró que para muestras de población Normal, la prueba propuesta por Wald y Wolfowitz en 1940, comparada con las conocidas propuestas t-Student y F-Fisher para alternativas de localización y escala respectivamente, tiene eficiencia asintótica relativa igual a cero (véase Mood (1954)). Este resultado no es sorprendente si se observa que la prueba propuesta por Wald y Wolfowitz (1940) esta basada solamente en el número total de rachas r_N (es decir el número de rachas hasta

el $\max\{X_1,\ldots,X_n\}$ sin tener en cuenta la información contenida en las restantes N-1 estadísticas de rachas $r_1,\ldots,r_{N-1}$.

*

BIBLIOGRAFIA

Corzo, J. Ortiz, J. (1983), *Una Prueba de Dispersión Basada en Secuencias*. Revista de Estadística Nº 8, pp.34-38.

Corzo, J. (1990), *Verallgemeinerte Runtests fur Lage- und Skalen alternativen*. Tesis Doctoral. Univ. Dortmund RFA.

Doob, J.L. (1959), *Stochastic Processes*, John Wiley & Sons, New York.

Gibbons, J. (1971), *Nonparametric Statistical Inference*. Marcel Dekker, New York.

Hajek, J. Sidak, Z. (1976), *Theory of Rank Tests*. Academic Press, New York, Academia Publishing House of the ozechoslovak Academy of Sciences,

Mood, A. (1954), *On the Asymptotic Efficiency of Certain Nonparametric Two-Sample Tests*. Ann. Math. Stat. 25, pp.514-522.

Wald & Wolfowitz (1940), *On a test wether two Samples are from the same Population*. Ann. Math. Stat. 11, 147-162.

* *