QÜESTIIÓ, Vol 11, nº 2, pp. 3-37, 1987

GENERACION DE UN SISTEMA BIVARIANTE CON MARGINALES DADAS Y ESTIMACION DE SU PARAMETRO DE DEPENDENCIA

J. OCAÑA, C.M. CUADRAS

Universitat de Barcelona

En este trabajo se proponen dos posibles estimadores del parámetro de dependencia de una familia de distribuciones bivariantes con marginales dadas y se realiza un estudio de Monte Carlo de sus respectivos sesgo y eficiencia, a fin de determinar cuál de ambos estimadores es preferible. También se propone y se estudia, de forma similar, una posible versión "Jackknife" del mejor de los dos estimadores anteriores. En este estudio se emplean técnicas de reducción de la varianza. Para poder realizar las simulaciones es preciso disponer de un generador eficiente de la familia de distribuciones bivariantes estudiada. Se propone un algoritmo adecuado y se demuestra su validez.

Computer generation of a bivariate system whith specified marginals and dependence parameter estimation

Keywords: Families of bivariate distributions, Estimation of dependence, Variance reduction, Jackknife, Random vector generation.

1. INTRODUCCION

Sean X e Y dos variables aleatorias sobre un espacio de probabilidad (Ω, A, P) con funciones de distribución, ambas continuas, F_1 y F_2 respectivamente, en todo este trabajo supondremos que las distribuciones marginales son absolutamente continuas, sin insistir cada vez en ello. Un problema con interés tanto teórico como aplicado (por ejemplo, en el sentido de ajuste de modelos probabilísticos a datos experimentales) es el de la determinación de sistemas de distribuciones bivariantes $H(x, y)$ tales que tengan por marginales

-Jordi Ocaña i Carles M. Cuadras - Universitat de Barcelona - Dep. d'Estadística - Av. Diagonal, 645 - 08028 Barcelona

-Article rebut el febrer de 1987.

a F_1 y F_2. En general será conveniente que toda familia de distribuciones bivariantes esté caracterizada por un parámetro θ que, de alguna forma, sea una medida de dependencia estocástica.

Dadas las marginales F_1 y F_2, las distribuciones

$$H^+(x,y) = \min\{F_1(x), F_2(y)\}$$

$$H^-(x,y) = \max\{F_1(x) + F_2(y) - 1, 0\} \tag{1}$$

se conocen como cotas de Fréchet. Toda distribución H con marginales F_1 y F_2 verifica

$$H^-(x,y) \leq H(x,y) \leq H^+(x,y) \tag{2}$$

Kimeldorf y Sampson [7] propusieron cinco condiciones para caracterizar lo que sería una adecuada familia de distribuciones bivariantes

$$\{H_\theta : -1 \leq \theta \leq 1\}$$

asociada a un parámetro de dependencia, a saber:

1. $H_1(x,y) = H^+(x,y)$
2. $H_0(x,y) = F_1(x)F_2(y)$,

   es decir, $\theta = 0$ implica independencia estocástica
3. $H_{-1}(x,y) = H^-(x,y)$
4. $H_\theta(x,y)$ es continua en $\theta \in [-1,1]$
5. $H_\theta(x,y)$ es absolutamente continua para todo $\theta, \theta \in (-1,1)$, fijo

Una posible familia de distribuciones bivariantes con marginales dadas es la propuesta por Cuadras y Augé [4]:

$$(3) \qquad F_\theta(x,y) = \begin{cases} F_1(x)^{1-\theta}F_2(y) & \text{si } F_1(x) \geq F_2(y) \\ F_1(x)F_2(y)^{1-\theta} & \text{si } F_1(x) < F_2(y) \end{cases}$$

$$\text{para } \theta \in [0,1]$$

Este sistema se puede extender, de forma natural, al caso de parámetros negativos, resultando

$$(4) \qquad G_\theta(x,y) = \begin{cases} F_1(x)[1-F_1(x)^\theta][1-F_2(y) & \text{si } F(x) \geq 1-F_2(y) \\ F_1(x)\{1-[1-F_2(y)]^{1-\theta}\} & \text{si } F(x) < 1-F_2(y) \end{cases}$$

$$\text{para } \theta \in [-1,0]$$

y el sistema general

$$(5) \qquad H_\theta(x,y) = \begin{cases} F_\theta(x,y) & \text{si } \theta \in [0,1] \\ G_\theta(x,y) & \text{si } \theta \in [-1,0] \end{cases}$$

El sistema de Cuadras–Augé verifica las propiedades 1 a 4 de Kimeldorf y Sampson. No verifica la propiedad 5 puesto que

$$P_\theta[F_1(X) = F_2(Y)] > 0, \text{ para } \theta > 0$$

$$(6) \qquad P_\theta[F_1(X) + F_2(Y) = 1] > 0, \text{ para } \theta < 0$$

Dada la distribución bivariante F con marginales F_1 y F_2, si X e Y son variables aleatorias con distribución conjunta F las variables aleatorias $U = F_1(X)$ y $V = F_2(Y)$ tienen distribución conjunta.

$$(7) \qquad U_F(u,y) = F(F_1^{-1}(u), F_2^{-1}(v))$$

$$\text{para } u,v \in [0,1]$$

con marginales uniformes sobre $[0,1]U_F$ se conoce como la representación uniforme de F. Kimeldorf y Sampson [8] sugirieron las ventajas de su empleo en el estudio de sistemas de distribuciones con marginales dadas.

La representación uniforme del sistema de Cuadras–Augé tiene obviamente la forma

(8a)
$$U_\theta(x,y) = \begin{cases} x^{1-\theta}y & \text{si } 0 \le y \le x \le 1 \\ xy^{1-\theta} & \text{si } 0 < x < y \le 1 \end{cases}$$

para $\theta \in [0,1]$

(8b)
$$U_\theta(y,x) = \begin{cases} x[1-x^\theta](1-y) & \text{si } 0 \le 1-y \le x \le 1 \\ x[1-(1-y)^{1-\theta}] & \text{si } 0 \le x < 1-y \le 1 \end{cases}$$

para $\theta \in [-1,0]$

Si U y V tienen distribución conjunta U_θ, las variables $X = F_1^{-1}(U)$ e $Y = F_2^{-1}(V)$, para las distribuciones contínuas F_1 y F_2, tienen distribución H_θ de Cuadras–Augé (5) con marginales F_1 y F_2 y el mismo parámetro θ.

En [3] se hace un estudio del sentido del parámetro θ como medida de dependencia estocástica. Entre otros resultados, se establecen relaciones entre θ y las medidas de asociación:

(9)
$$\rho = \text{ cov } (X,Y)/(\sigma(X)\sigma(Y))$$

coeficiente de correlación de Pearson

(10)
$$\tau = 4\int_{\mathbf{R^2}} H_\theta(x,y)dH_\theta(x,y) - 1$$

coeficiente de correlación por rangos de Kendall.

(11)
$$\rho_g = 12\int_{\mathbf{R^2}} F_1(x)F_2(y)dH_\theta(x,y) - 3$$

grado de correlación, versión probabilística del coeficiente de correlación por rangos de Spearman.

En [3], para la representación uniforme (8), se establece que

(12) 1. $\rho = 3\theta/(4 - |\theta|)$

(13) 2. $\tau = \theta/(2 - |\theta|)$

(14) 3. $\rho_g = 3\theta/(4 - |\theta|)$

Las relaciones anteriores sugieren de forma inmediata los siguientes posibles estimadores de θ:

(15) 1. $U_R = 4R/(3 + |R|)$

(16) 2. $U_K = 2K/(1 + |K|)$

(17) 3. $U_S = 4S/(3 + |S|)$

indicando R, K y S los coeficientes de correlación muestrales de Pearson, Kendall y Spearman respectivamente.

Sin embargo la relación (12) no es invariante respecto de transformaciones monótonas, a diferencia de las relaciones (13) y (14) que si lo son, de manera que el estimador U_R sólo tiene sentido en el caso de uniformes marginales mientras que U_K y U_S son de aplicabilidad general.

Dada la dificultad de estudiar analíticamente las propiedades de los estimadores U_K y U_S, se ha realizado un estudio de Monte Carlo de su sesgo y eficiencia, a fin de determinar cuál de ambos es preferible o en qué condiciones U_K es superior a U_S o viceversa. La invarianza, respecto de transformaciones monótonas crecientes, de θ, τ y ρ_g (los tres son un índice de concordancia en el sentido establecido en [2] donde entre las propiedades exigidas a un índice de concordancia se cuenta la invarianza respecto de transformaciones monótonas) y de las relaciones que ligan θ con τ y ρ_g, juntamente con las propiedades de la representación uniforme (12), permiten que centremos nuestro estudio en el caso particular de marginales uniformes ya que la validez general de las conclusiones queda asegurada.

La parte 2 de este trabajo se centra en la obtención de un generador para la distribución de Cuadras–Augé. Una vez obtenido éste, se puede abordar el estudio de Monte Carlo propiamente dicho, el cual se describe en la parte 3.

2. UN GENERADOR DEL SISTEMA DE CUADRAS–AUGE

El generador que se propone en esta parte del trabajo, se basa en el siguiente resultado básico:

> Sean U y V dos variables aleatorias estocásticamente independientes con distribución uniforme sobre [O, 1] y sea $\theta \in [-1, 1]$. Definamos las constantes
>
> $$p = |\theta|/(2 - |\theta|)$$
> $$q_1 = q_2 = (1 - |\theta|)/(2 - |\theta|)$$
> $$a = 1/(2 - |\theta|)$$
>
> el vector aleatorio definido como
>
> $$(X, Y) = \begin{cases} (U^a, U^a) & \text{con probabilidad } p \\ (U^a, VU^a) & \text{con probabilidad } q_1 \\ (VU^a, U^a) & \text{con probabilidad } q_2 \end{cases}$$
>
> cuando $\theta \in [0, 1]$, y como
>
> $$(X, Y) = \begin{cases} (U^a, 1 - U^a) & \text{con probabilidad } p \\ (U^a, 1 - VU^a) & \text{con probabilidad } q_1 \\ (VU^a, 1 - U^a) & \text{con probabilidad } q_2 \end{cases}$$
>
> cuando $\theta \in [-1, 0]$
>
> sigue la distribución uniforme de Cuadras–Augé (8)

La validez del resultado anterior se podría comprobar de forma directa. La exposición que sigue demuestra igualmente este resultado, de forma indirecta pero que refleja las ideas y el proceso que han conducido a su obtención. Consideraemos por separado los casos de parámetro no negativo y negativo.

2.1. Caso de Parámetro no negativo

El método más habitual de generación de un par (X, Y) variables aleatorias con distribución conjunta $F(x, y)$ y marginales F_1 y F_2 se basa en la generación de dos variables aleatorias univariantes:

1. En primer lugar se genera X con distribución F_1, por ejemplo mediante $X = F_1^{-1}(U)$ siendo U una variable aleatoria con distribución uniforme sobre [0,1]. Supongamos que se ha obtenido el valor x
2. Si F_x es la distribución de Y condicionada a $X = x$, se genera un segundo valor a partir de $Y = F_x^{-1}(V)$ (o mediante algún otro método distinto de la inversión) con V uniforme sobre [0,1]

El sistema de Cuadras–Augé presenta la dificultad adicional de no ser absolutamente continuo, de manera que es más difícil de manipular. De todas maneras, según la teoría de distribuciones de Schwartz, se puede descomponer U_θ en la suma de dos integrales. En efecto (consideremos de momento solamente el caso de parámetro no negativo para simplificar la notación), la derivada parcial de U_θ respecto de x es

$$\frac{\partial}{\partial x}U_\theta(x,y) = \begin{cases} (1-\theta)x^{-\theta}y & \text{si } x > y \\ y^{1-\theta} & \text{si } x \leq y \end{cases} \tag{18}$$

discontinua sobre $x = y$. Derivándola respecto de y en el sentido de la teoría de distribuciones, tendremos ahora (siendo φ una función de test)

$$< \frac{\partial}{\partial y}\frac{\partial}{\partial x}U_\theta(x,y), \varphi(y) > = - < \frac{\partial}{\partial x}U_\theta(x,y), \frac{\partial}{\partial y}\varphi(y) >$$

$$= -\int_R \frac{\partial}{\partial x}U_\theta(x,y)\frac{\partial}{\partial y}\varphi(y)dy$$

$$= \theta x^{1-\theta} < \delta_{\{x-y\}}, \varphi > + < \{\frac{\partial^2}{\partial x \partial y}U_\theta(x,y)\}, \varphi(y) > \tag{19a}$$

(indicando (h) la distribución representada por la función h, definida en $x < y$, en $x > y$, y no definida para $x = y$) es decir, se obtendrá la densidad generalizada

$$\frac{\partial^2}{\partial x \partial y} = \left\{ \begin{matrix} (1-\theta)x^{-\theta} & \text{si } x > y \\ (1-\theta)y^{-\theta} & \text{si } x < y \end{matrix} \right\} + \theta x^{1-\theta}\delta\{x = y\} \tag{19b}$$

siendo $\delta\{x = y\}$ la distribución delta de Dirac sobre $x = y$, de manera que U_θ se descompone en la suma

$$U_\theta\{x,y\} = pF_\theta(x,y) + qG_\theta(x,y), \quad p + q = 1$$

con

$$pF_\theta(x,y) = \int_0^x \int_0^y f_\theta(u,v)dudv$$

$$f_\theta(x,y) = \begin{cases} (1-\theta)x^{-\theta} & \text{si } x > y \\ (1-\theta)y^{-\theta} & \text{si } x < y \end{cases}$$

$$qG_\theta(x,y) = \int_0^{\min\{x,y\}} \theta x^{1-\theta}dx \tag{20}$$

La probabilidad sobre cualquier boreliano B de $\mathbf{R^2}$ se puede indicar

$$P\ (B) = P((X,Y)^{-1}(B)) = \iint_B f_\theta(x,y)dxdy + \int_{B\cap\{x=y\}} \theta x^{1-\theta}dx \tag{21}$$

y en particular la probabilidad sobre la singularidad $S = \{x = y\}$ es

$$\theta/(2-\theta) = \int_0^1 \theta x^{1-\theta}dx = q \tag{22}$$

Esto sugiere el siguiente método de generación.

1. Con probabilidad $q = \theta/(2-\theta)$ (de singularidad) generemos un par X, Y caso sobre la singularidad S. Esto es equivalente a generar un único valor con distribución.

$$G_\theta(x) = q^{-1}\int_0^x \theta u^{1-\theta}du = x^{2-\theta}, \quad \text{si } x \in [0,1] \tag{23}$$

es decir, mediante $X = V^{1/(2-\theta)}$, siendo V uniforme sobre $[0,1]$ e $Y = X$.

2. Con probabilidad $p = 1-q = 2(1-\theta)/(2-\theta)$ (no singularidad) generemos un par (X,Y) con distribución F_θ

$$F_\theta(x,y) = p^{-1}\int_0^x \int_0^y f_\theta(u,v)dudv$$
$$\text{si } (x,y) \in [0,1]x[0,1]$$

es decir

$$F_\theta(x,y) = \begin{cases} (2(1-\theta))^{-1}[(2-\theta)x^{1-\theta}y - \theta y^{2-\theta}] \\ \qquad\qquad \text{si } 0 \le y \le x \le 1 \\ (2(1-\theta))^{-1}[(2-\theta)xy^{1-\theta} - \theta x^{2-\theta}] \\ \qquad\qquad \text{si } 0 \le x < y \le 1 \end{cases} \tag{24}$$

con marginal para x

$$\begin{aligned} F_{\theta 1}(x) &= \lim_{y->1} F_\theta(x,y) \\ &= (2(1-\theta))^{-1}[(2-\theta)x - \theta x^{2-\theta}] \end{aligned}$$

$$\text{para } x \in [0,1]. \tag{25}$$

La inversión de $F_{\theta 1}$ se tendría que hacer numéricamente, de manera que la generación no sería muy eficiente. En todo caso se tendría que probar un método alternativo, como el rechazo, para generar la variable X. La forma de $F_\theta(y/X=x)$ todavía es menos conveniente.

El hecho de que la densidad f_θ sea simétrica respecto del eje $x=y$ sugiere un perfeccionamiento del método anterior. Obviamente

$$P(X \ge Y/S^*) = P(X < Y/S^*) = 1/2$$

indicando S^* el complemento de S.

Condicionando a los sucesos $[X \ge Y]$ y $[X < Y]$ se tiene

$$\begin{aligned} F_\theta(x,y/X \ge Y) &= 2P([X \le x, Y \le y] \cap [X \ge Y]) \\ &= 2(0.5F_\theta(x,y) + (F_\theta(x,y) - F_\theta(y,y)) \\ &= 2F_\theta(x,y) - F_\theta(y,y) \\ &\text{si } 0 \le y \le x \le 1 \text{(figura 2.1.1.)} \\ F_\theta(x,y/X \ge Y) &= 2(0.5F_\theta(x,x)) = F_\theta x, x) \\ &\text{si } 0 \le x < y \le 1 \text{(figura 2.1.2.)} \end{aligned} \tag{26}$$

de manera que la función de distribución marginal para X es

$$
\begin{aligned}
F_\theta(x/X \geq Y) \quad &= \text{limp}_{y->1} F_\theta(x, y/X \geq Y) \\
&= F_\theta(x, x) \\
&= (2(1-\theta))^{-1}[(2-\theta)x^{2-\theta} - \theta x^{2-\theta}] \\
&= x^{2-\theta} \\
&\text{si } x \in [0,1]
\end{aligned} \tag{27}
$$

con densidad

$$
f_\theta(x/X \geq Y) = (2-\theta)x^{1-\theta} \tag{28}
$$

si $x \in [0,1]$ y cero en caso contrario.

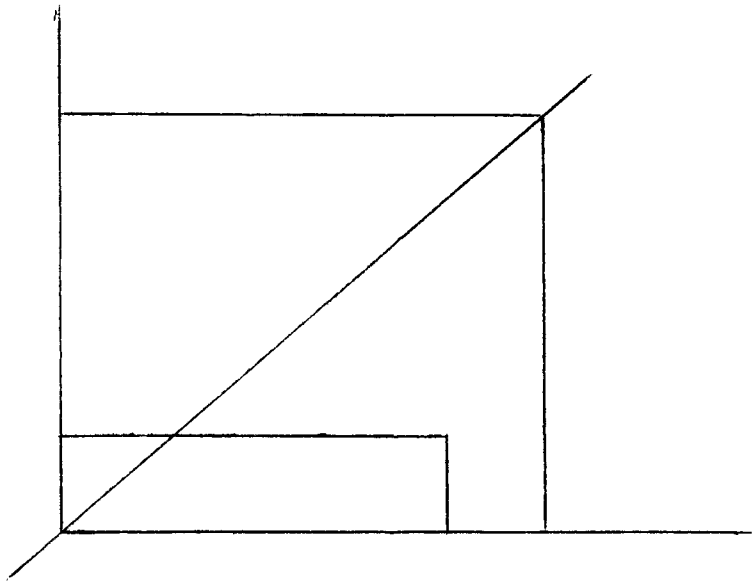

FIGURA 2.1.1. La zona sombreada corresponde a la de masa de probabilidad no nula para la distribución condicionada $F_\theta(x, y|X \geq Y)$.

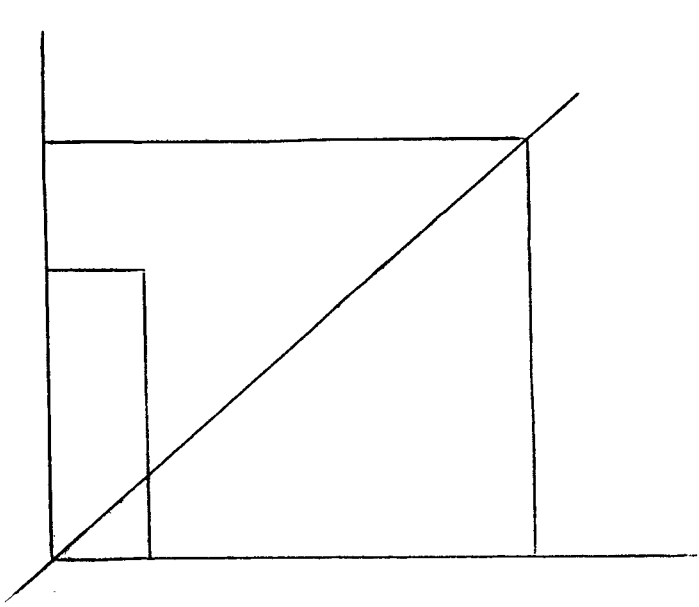

FIGURA 2.1.2. Véase el pié de la Figura 2.1.1.

La densidad conjunta es ahora:

$$f_\theta(x, y/X \geq Y) = \frac{\partial^2}{\partial x \partial y} F_\theta(x, y/X \geq Y)$$

$$(29) \qquad = \begin{cases} (2-\theta)x^{-\theta} & \text{si } 0 \leq y \leq x \leq 1 \\ 0 & \text{en caso contrario} \end{cases}$$

Por lo tanto, la densidad condicionada de Y a $X = x$ es

$$(30) \qquad \begin{aligned} f_\theta(y/X = x, X \geq Y) \quad &= f_\theta(x, y/X \geq Y)/f_\theta(x/X \geq Y) \\ &= x^{-1} \qquad \text{si } y \in [0, x] \end{aligned}$$

es decir , uniforme sobre $[0, x]$.

Similarmente, se obtendría que, condicionando a $X < Y$:

$$F_\theta(y/X < Y) = y^{2-\theta}, \qquad \text{para } y \in [0,1]$$

X condicionada a $Y = y$ es uniforme sobre $[0, y]$.

En resumen, para $\theta \geq 0$, podemos descomponer la función de distribución (8) en

$$U_\theta(x,y) = qG_\theta(x,y) + p_1 F_\theta(x, y/X \geq Y) + p_2 F_\theta(x, y/X < Y) \tag{31}$$

siendo

$$q = \theta/(2-\theta), \quad G_\theta(x,y) = \begin{cases} y^{2-\theta} & \text{si } 0 \leq y \leq x \leq 1 \\ x^{2-\theta} & \text{si } 0 \leq x \leq y \leq 1 \end{cases}$$

$$p_1 = p_2 = (1-\theta)/(2-\theta)$$

$$F_\theta(x, y/X \geq Y) = \begin{cases} (1-\theta)^{-1}[(2-\theta)x^{1-\theta}y - y^{2-\theta}] \\ \qquad \text{si } 0 \leq y \leq x \leq 1 \\ x^{2-\theta} \qquad \text{si } 0 \leq x < y \leq 1 \end{cases}$$

$$F_\theta(x, y/X < Y) = \begin{cases} y^{2-\theta} \qquad \text{si } 0 \leq x \leq y < 1 \\ (1-\theta)^{-1}[(2-\theta)xy^{1-\theta} - x^{2-\theta}] \\ \qquad \text{si } 0 \leq y < x \leq 1 \end{cases} \tag{32}$$

Simplemente sumando la expresión (25) se puede comprobar que esta descomposición es correcta y que por lo tanto lo es el método de generación, que se puede indicar como:

En primer lugar se hará una decisión aleatoria, con probabilidad $\theta/(2-\theta)$, para decidir si se produce la singularidad sobre $\{x = y\}$.

En caso afirmativo invertiremos la distribución $G_\theta(x) = x^{2-\theta}$ para $x \in [0,1]$, es decir, dado que $W = G_\theta(X)$, generaremos una variable W con distribución uniforme sobre $[0,1]$ y

$$\begin{aligned} X &= W^{1/(2-\theta)} \\ Y &= X \end{aligned}$$

En caso contrario (lo cual ocurrirá con probabilidad $2(1-\theta)/(2-\theta)$ se hará una decisión aleatoria, con probabilidades $p_1 = p_2 = 0.5$, para decidir entre las regiones $\{x \geq y\}$ y $\{x < y\}$.

Si $x \geq y$, se invertirá la distribución marginal (27), es decir, como $W = X^{2-\theta}$, se generará una variable aleatoria uniforme W sobre $[0,1]$ y

$$X = W^{1/(2-\theta)}$$

y, según (30) la variable Y como una uniforme entre 0 y X, es decir

$$Y = XV$$

siendo V una variable aleatoria con distribución uniforme sobre $[0,1]$.

Similarmente, si $x < y$, se generará

$$\begin{aligned} Y &= W^{1/(2-\theta)} \\ X &= YV \end{aligned}$$

2.2. Caso de Parametro negativo

El método propuesto en el apartado anterior se puede extender de forma casi inmediata al caso en que el parámetro es negativo. El algoritmo se basa en la misma idea [4] que permitió extender la distribución (3) y (8a) al caso de parámetro negativo: si (T, Z) se distribuye según $U_\theta(t, z)$ para cierto valor del parámetro $\theta > 0$, las variables

$$\begin{aligned} X &= T \\ Y &= 1 - Z \end{aligned}$$

tendrán distribución conjunta

$$\begin{aligned} V_\theta(x,y) &= x - U_\theta(x, 1-y) \\ &= [\min\{x, 1-y\}]^\theta [x(1-y)]^{1-\theta} \end{aligned}$$

$$\text{para } (x,y) \in [0,1]x[0,1] \tag{33}$$

Cambiando θ por $-\theta$ se obtiene la distribución (8b).

Para generar un par de variables (X, Y) según la distribución (8b), para cierto valor del parámetro de dependencia $\theta < 0$, bastará generar un par (X, Z) según el algoritmo de la sección 2.1 y para un valor del parámetro $-\theta$. Posteriormente se substituirá $Y = 1 - Z$

2.3. Caso General

Combinando los generadores de los apartados 2.1 y 2.2 se tendrá un método general para generar valores según la representación uniforme (8) del sistema de Cuadras–Augé.

Según lo indicado en la parte 1 de este trabajo, si (U, V) se genera mediante el método indicado en 2.1 y 2.2 el par $(X, Y) = (F_1^{-1}(U), F_2^{-1}(V)$ se distribuirá según la forma general del sistema (5), para cualesquiera marginales F_1, F_2 y con el mismo parámetro θ

3. ESTUDIO DE LOS ESTIMADORES DEL PARAMETRO DE ASOCIACION

3.1. Estudio Preliminar

Para estudiar el sesgo y la eficiencia de los estimadores (16) y (17) (por curiosidad se incluyó también en ocasiones el estimador (15), a pesar de que su validez solamente está asegurada para el caso de marginales uniformes), se planteó la siguiente serie de experimentos de Monte Carlo.

> Se estudiaron los casos correspondientes a los posibles valores del parámetro θ = 0.8,-0.6, -0.4, -0.2, 0.0, 0.2, 0.4, 0.6, 0.8 (posteriormente el estudio se amplió a valores pequeños, alrededor de 0.0: -0.1, -0.05, -0.01, -0.005, 0.005, 0.01, 0.05, 0.1). Para valores del parámetro $\theta = 1$ o $\theta = -1$, el sesgo y la varianza de los estimadores es obviamente nulo, como se comprobó en simulaciones previas realizadas para verificar que los programas funcionaban correctamente.

Para cada uno de estos valores de los parámetros se estudiaron los estimadores para sucesivos tamaños muestrales n = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 (y en ocasiones 200)

Para cada una de las posibles combinaciones de valores del parámetro y del tamaño muestral, se generaron 10000 muestras. Cada una de ellas permitió obtener un valor $b_i(\theta_n) = u_i(\theta, n) - \theta$, indicando $u_i(\theta, n)$ la i-ésima estimación, i = 1 ..., 10000, para el parámetro θ y el tamaño muestral n.

Se estimó el sesgo mediante

$$b(\theta, n) = 10000^{-1} \sum_i b_i(\theta, n) \tag{34}$$

y la varianza del estimador mediante

$$s^2(\theta, n) = 10000^{-1} \sum_i (s_i(\theta, n) - m_s)^2 \tag{35}$$

$$m_s = 10000^{-1} \sum_i u_i(\theta, n)$$

Para cada combinación de θ y n dada, cada una de las 10000 muestras fue iniciada a partir de una semilla aleatoria distinta, generada independientemente por un generador de números aleatorios auxiliar, distinto del empleado para generar las muestras. De una combinación a otra del parámetro θ y el tamaño n, la muestra número i,$i = 1, \ldots, 10000$, siempre fue iniciada a partir de la misma semilla aleatoria. En otras palabras, se planteó un diseño de secuencias aleatorias comunes. Esto se hizo para reducir la varianza en el sentido de que los resultados de cada una de las (θ, n) fuesen máximamente comparables. Las semillas aleatorias iniciales se tomaron, en el mismo orden en que allí están indicadas, de la tabla 6.8.1 de [1]. En teoría constituyen un conjunto de valores adecuado especialmente para el generador congruencial de números aleatorios empleado, tomado de este mismo texto.

A medida que se obtenía cada una de las muestras, sus valores se ordenaban respecto de la primera componente

$$(X_{(1)}, Y_1), \ldots, (X_{(n)}, Y_{(n)})$$

con

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}$$

Esto motivó que el estimador basado en el coeficiente de correlación por rangos de Kendall se calculase a partir del estadístico

$$K' = \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \mathbf{I}_{[y(i)<y(k)]} \tag{36}$$

indicando $\mathbf{I}_A$ el indicador del suceso A

A partir de (16) y dado que el coeficiente de correlación muestral de Kendall se puede indicar

$$K = 4K'/n(n-1) - 1 \tag{37}$$

se obtiene fácilmente la expresión para el estimador (16)

$$U_k = 2(4K' - n(n-1))/(n(n-1) + |4K' - n(n-1)|) \tag{38}$$

Igualmente, a partir de (17) y dado que el coeficiente de correlación muestral por rangos de Spearman se puede indicar (véase por ejemplo [10])

$$S = 1 - 6D/(n^3 - n) \tag{39}$$

siendo

$$D = \sum (R_i - S_i)^2 = \sum (i - S_1)^2 \tag{40}$$

donde R_i, S_i son los rangos de X_i, Y_i respectivamente, se obtiene la expresión para (17)

$$U_s = 4(n^3 - n - 6D)/\{3(n^3 - n) + |n^3 - n - 6D|\} \tag{41}$$

que fue la finalmente untilizada en el estudio de Monte Carlo.

Las figuras 3.1.1 y 3.1.2 indican el sesgo y la varianza de cada uno de los estimadores, para algunos valores del parámetro θ fijos, en función del tamaño muestral. Los autores disponen de listados con los valores numéricos obtenidos en las simulaciones, con información más completa que las gráficas.

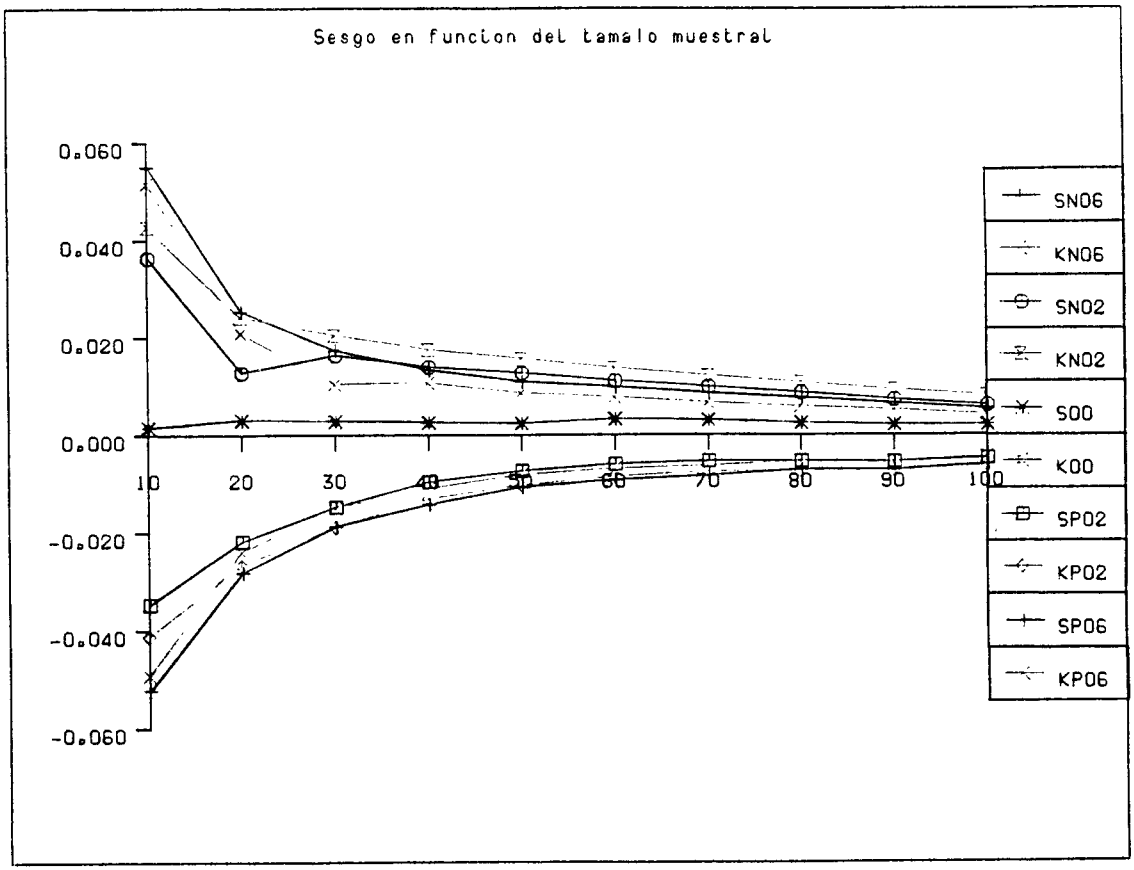

FIGURA 3.1.1. SN06 indica el etimador de Spearman para parámetro negativo $\theta = -0.6$, KN06 indica el estimador de Kendall para $\theta = -0.6$, SP06 el estimador de Spearman para $\theta = +0.6$, etc.

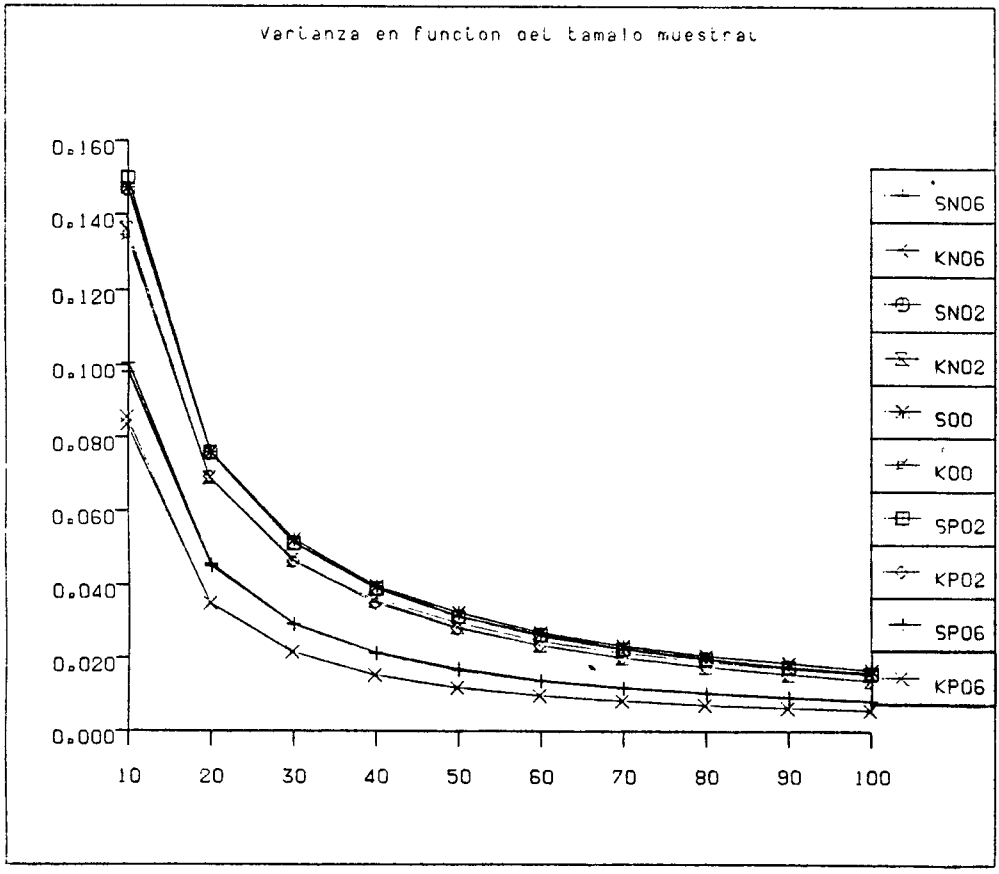

FIGURA 3.1.2. SN06, etc. tienen el mismo significado que en la figura 3.1.1.

TABLA 3.1.1.

Sesgo y función del tamaño muestral

estimador	Kend.	Spear.	Kend.	Spear.	Kend.	Spear.	Kend.	Spear.	Kend.	Spear.
valor parametro	0.6	0.6	0.2	0.2	0.0	0.0	-0.2	-0.2	-0.6	-0.6
n										
10	-0.0493	-0.0521	-0.0413	-0.0346	0.0016	0.0017	0.0427	0.0366	0.0515	0.0552
20	-0.0240	-0.0280	-0.0267	-0.0216	0.0032	0.0032	0.0242	0.0129	0.0209	0.0256
30	-0.0150	-0.0185	-0.0189	-0.0145	0.0027	0.0031	0.0205	0.0165	0.0104	0.0175
40	-0.0110	-0.0140	-0.0131	-0.0094	0.0025	0.0028	0.0175	0.0141	0.0107	0.0135
50	-0.0081	-0.0105	-0.0103	-0.0071	0.0024	0.0026	0.0156	0.0128	0.0086	0.0111
60	-0.0069	-0.0090	-0.0085	-0.0058	0.0032	0.0035	0.0137	0.0112	0.0077	0.0100
70	-0.0064	-0.0082	-0.0075	-0.0052	0.0030	0.0032	0.0121	0.0100	0.0066	0.0087
80	-0.0054	-0.0069	-0.0074	-0.0052	0.0024	0.0026	0.0107	0.0088	0.0058	0.0077
90	-0.0055	-0.0070	-0.0072	-0.0053	0.0021	0.0023	0.0092	0.0074	0.0051	0.0066
100	-0.0046	-0.0060	-0.0065	-0.0047	0.0019	0.0022	0.0080	0.0062	0.0041	0.0055

TABLA 3.1.2.

Varianza en función del tamaño muestral

estimador	Kend.	Spear.	Kend.	Spear.	Kend.	Spear.	Kend.	Spear.	Kend.	Spear.
valor parametro	0.6	0.6	0.2	0.2	0.0	0.0	-0.2	-0.2	-0.6	-0.6
n	KP06	SP06	KP02	SP02	K00	S00	KN02	SN02	KN06	SN06
10	0.0833	0.0978	0.1375	0.1498	0.1344	0.1472	0.1361	0.1469	0.0855	0.1003
20	0.0349	0.0450	0.0691	0.0756	0.0686	0.0757	0.0689	0.0754	0.0349	0.0455
30	0.0214	0.0290	0.0464	0.0512	0.0474	0.0522	0.0467	0.0513	0.0214	0.0289
40	0.0151	0.0210	0.0348	0.0386	0.0361	0.0394	0.0354	0.0389	0.0150	0.0209
50	0.0118	0.0166	0.0278	0.0309	0.0293	0.0320	0.0282	0.0310	0.0116	0.0164
60	0.0096	0.0135	0.0232	0.0258	0.0243	0.0265	0.0232	0.0256	0.0095	0.0134
70	0.0081	0.0115	0.0198	0.0220	0.0211	0.0228	0.0198	0.0219	0.0080	0.0114
80	0.0070	0.0100	0.0174	0.0193	0.0187	0.0201	0.0173	0.0192	0.0070	0.0101
90	0.0062	0.0089	0.0154	0.0171	0.0168	0.0181	0.0153	0.0169	0.0062	0.0089
100	0.0055	0.0079	0.0139	0.0155	0.0153	0.0164	0.0136	0.0152	0.0056	0.0080

Aunque las figuras parecen indicar la existencia de un patrón bastante claro, que se podría enunciar como sigue:

1. Ambos estimadores son asintóticamente insesgados. De hecho, esta propiedad se puede deducir rigurosamente, a partir del hecho de que, por ser ambos estimadores funciones continuas de estadísticos consistentes, son consistentes. Por el hecho de ser consistentes y acotados se concluye que son asintóticamente insesgados.
2. Para valores grandes de $|\theta|$ el estimador U_k es menos sesgado que U_s. La situación se invierte para valores pequeños de $|\theta|$.
3. Ambos estimadores son insesgados para $\theta = 0$ (y para $|\theta| = 1$).
4. Para ambos estimadores, cuando $\theta > 0$ el sesgo es negativo, cuando $\theta < 0$ el sesgo es positivo. Además se verifica que

$$\text{sesgo } (\theta) = - \text{ sesgo } (-\theta)$$

5. El estimador basado en el estadístico de Kendall es en todo momento más eficiente que el estimador basado en el estedístico de Spearman. Este hecho queda reflejado por las gráficas de la varianza y por los valores del error cuadrático, que no ha sido representado (si bien se discute en el apartado siguiente)

hay que constatar el carácter altamente irregular de las gráficas del sesgo. Esto es explicable dado que, a pesar de que el tamaño muestral de 10000 estimaciones generadas puede parecer importante, debido a que la magnitud estimada (el sesgo) toma valores muy pequeños y debido a la magnitud de la varianza, el intervalo de confianza resultante es comparativamente muy amplio. (nos referimos al intervalo de confianza para una combinación de n y θ dada

$$b(\theta, n) \pm z_\alpha s(\theta, n)/(m-1)^{1/2}$$

$$(m = 10000)$$

no a un intervalo simultáneo, difícil de determinar de forma exacta puesto que, debido al diseño experimental de secuencias aleatorias comunes, las variables $b(\theta, n)$ son dependientes. Se podría utilizar un intervalo aproximado, por ejemplo basado en la desigualdad de Bonferoni, pero resultaría todavía más amplio. (Lo cual hace pensar que se debería replantear la validez de estudios similares, como [11], basado en la generación de 1000 estimaciones por ninguna técnica de reducción de la varianza)

Razones de disponibilidad de tiempo de ordenador desaconsejaban aumentar el número de estimaciones. Es por ello que se replanteó el diseño experimental en el sentido de emplear técnicas de reducción de la varianza más sofisticadas (si

bien posteriormente fue también posible aumentar los números de estimaciones, tal como se indica en el apartado 3.3).

3.2. Reducción de la varianza

El problema planteado es el de estimar una función paramétrica, en este caso el sesgo $\beta = E(U) - \theta$, a partir de una muestra $u_1, \ldots, u_m$ (una muestra de valores de un estimador, cada uno de ellos calculado a partir de una muestra $((x_1, y_2)\ldots, (x_n, y_n))$ de valores de la variable (X, Y)).

En el diseño descrito en el apartado anterior, las sucesivas observaciones U_i eran independientes. Si, de alguna forma se modifica el diseño experimental de manera que las observaciones U_{2i-1} y U_{2i} sean dependientes, con covarianza negativa, mientras que los pares $(U_{2i-1}, U_{2i}$, para $i = 1, \ldots, m/2$ (supondremos m par), sean estocásticamente independientes (o al menos estén incorrelacionados), la varianza del estimador (insesgado) del sesgo

$$B = m^{-1} \sum_{i=1}^{m} U_i - \theta \tag{42}$$

será

$$\text{var } (B) = m^{-1} \text{ var } (U) + 2m^{-2} \sum_{i=1}^{m/2} \text{cov } (U_{2i-1}, U_{2i}) \tag{43}$$

menor que la varianza asociada al muestreo aleatorio simple, m^{-1} var (U).

Esta es la idea del método de redución de la varianza basado en el empleo de "secuencias antitéticas". Es remarcable que este método, a diferencia de lo que suele suceder con otras técnicas de reducción de la varianza, prácticamente no complica el análisis estadístico subsiguiente. Por ejemplo, haciendo

$$T_i = (U_{2i-1} + U_{2i})/2, \quad \text{para } i = 1, \ldots, m/2$$

se puede definir el nuevo estimador

$$C = (m/2)^{-1} \sum_{i=1}^{m/2} T_i \tag{44}$$

también insesgado (obviamente es $B = C$), asociado al intervalo de confianza

$$C \pm t(m/2 - 1, \alpha) S/(m/2 - 1)^{1/2}$$

con

$$S = \{(m/2)^{-1} \sum_{i=1}^{m/2} (T_i - C)^2\}^{-1/2} \tag{45}$$

En general, la disminución del tamaño muestral (pasa a ser $m/2$) quedará sobradamente compensada por la reducción de varianza conseguida.

La forma usual de implementación del método se fundamenta en el siguiente teorema [14]:

> Sea V una variable aleatoria con distribución uniforme sobre $[0, 1]$. Para funciones de distribución marginales F_1 y F_2 dadas, ambas con varianza finita, la distribución conjunta de mínima correlación (en el sentido de $\rho < 0$, $|\rho|$ máximo) es la distribución de
>
> $$(F_1^{-1}(V), F_2^{-1}(1 - V))$$
>
> Igualmente, la máxima correlación se alcanza para
>
> $$(F_1^{-1}(V), F_2^{-1}(V))$$

En realidad el resultado anterior no es nada más que otra forma de enunciar el resultado clásico que establece que la máxima y mínima correlación se alcanzan, respectivamente, para las cotas de Fréchet H^+ y H^-, pero en términos de transformaciones sobre variables con distribución uniforme, más acorde con la metodología de generación de variables aleatorias.

Cuando los sucesivos valores generados se obtienen directamente por inversión a partir de un solo valor uniforme, queda plenamente justificada la generación de cada valor U_{2i-1} a partir del valor uniforme v_i y la de U_{2i} a partir del valor antitético $1 - v_i$ para $i = 1, \ldots, m/2$.

En la mayoría de simulaciones, cada valor muestral finalmente obtenido no es función de un solo valor uniforme sino de un cierto número de ellos. El número de valores uniformes puede ser a su vez aleatorio. Por ejemplo en las simulaciones descritas en el apartado anterior, cada estimación U_i se obtiene

a partir de la generación de n pares (x_j, y_j) cada uno de los cuales precisaba a su vez de la generación de dos o tres valores aleatorios. Resultados en parte teóricos (véase por ejemplo [1]) y en parte basados en la experiencia, sugieren que el empleo de la secuencia de números aleatorios.

$$(v_{i1}, v_{i2}, \ldots, v_{ir} \ldots)$$

para generar la observación $2i - 1$, y de la secuencia

$$(1 - v_{i1}, 1 - v_{i2}, \ldots, 1 - v_{ir} \ldots)$$

para generar la observación $2i$, proporciona en general resultados adecuados siempre que se den ciertos requerimientos, del tipo de que, en algún sentido, el valor finalmente obtenido sea una función aproximadamente monótona de la secuencia aleatoria, por ejemplo que sea una función monótona de la suma de los valores aleatorios generados. La obtención de la secuencia antitética es sumamente fácil si se utilizan generadores congruenciales de números aleatorios, de la forma

$$Z_{j+1} = aZ_j + b \bmod M$$

$$v_{j+1} = Z_{j+1}/M$$

siendo a, b, M y Z_j enteros positivos. Bastará iniciar la secuencia número i a partir del valor entero inicial (semilla aleatoria) Z_0 y la secuencia antitética a partir de la semilla $M - Z_0$ (véase por ejemplo [9]).

Las consideraciones anteriores no son válidas en el presente estudio debido a que la variable de salida de cada réplica es una medida de dependencia o asociación. Empleando la técnica descrita en los párrafos precedentes, se genera correlación negativa entre estadísticos como

$$\sum_{j=1}^{n} X_{2i-1.j} \quad \text{y} \quad \sum_{j=1}^{n} X_{2i.j}$$

$$\sum_{j=1}^{n} X_{2i-1.j} \quad \text{y} \quad \sum_{j=1}^{n} Y_{2i.j}$$

pero no entre las correspondientes estimaciones de θ o del sesgo.

Para inducir correlación negativa en el sentido indicado en los párrafos anteriores, se debe modificar la técnica de generación de los pares (X, Y) del siguiente modo:

Se emplearán tres secuencias independientes de números aleatorios enteros Z^1, Z^2, Z^3 en lugar de una sola secuencia. La primera, Z^1, permitirá determinar la singularidad o la región de $[0,1]$ x $[0,1]$ a la que condicionar ($X = Y, X \geq Y$ o bien $X + Y = 1, X + Y \geq 1$ o $X + Y < 1$). Z^2 y Z^3 se emplearán para la generación de X e Y.

Para conseguir una completa sincronización entre pares de réplicas, los tres valores aleatorios se generarán al principio. De este modo cada par (X, Y) requerirá la generación de un número constante de tres valores aleatorios, aunque sea a costa de una velocidad de generación ligeramente menor (según los algoritmos descritos en 2.1 y 2.2 cuando se producía la singularidad sólo era necesaria la generación de dos valores aleatorios, uno para decidir sobre la ocurrencia de esta singularidad y otro para generar $X = Y$)

Asimismo, el diseño experimental se debe modificar en el siguiente sentido:

La réplica $2i$, $i = 1 \ldots, m/2 (m = 10000)$ se iniciará a partir de las semillas Z_0^1, Z_0^2, Z_0^3 para cada una de las tres secuencias aleatorias citadas anteriormente.

La réplica $2i + 1$ se iniciará a partir de las semillas

$$M - Z^1$$

$$Z^2$$

$$M - Z^3$$

es decir, las sencuencias 1 y 3 serán antitéticas pero la 2 no lo será.

Las modificaciones anteriores del diseño experimental inducen correlación negativa entre pares de observaciones de cada estimador y por lo tanto reducción de la varianza. En efecto:

En primer lugar, la sincronización en el generador de (X, Y) asegura una correspondencia uno a uno entre los sucesivos valores aleatorios generados, en el sentido de que, para cada una de las muestras $(X_1, Y_1) \ldots, (X_n, Y_n)$ la semilla aleatoria número j, Z_j^1 sirve para decidir si se ha producido la singularidad o a qué región se condiciona, la semilla Z_j^2 permite generar X_j y la semilla Z_j^3 permite generar Y_j

Tanto el estimador basado en la correlación de Kendall como el basado en la correlación de Spearman dependen únicamente de los rangos de las observaciones (X, Y). A su vez X e Y son funciones monótonas de los valores uniformes, indiquémoslos como W_1 y W_2 respectivamente, utilizados para generarlos. La decisión entre singularidad y cada una de las zonas de $[0,1] \times [0,1]$ depende también de la magnitud de otro valor V generado a partir de una distribución uniforme.

Para completar la demostración basta considerar todos los casos que se pueden presentar, en cuanto a la magnitud de los valores aleatorios V, W_1 y W_2 comparando los valores X e Y que finalmente se obtendrían para la muestra $2i-1$ y para su antitética $2i$. El estudio detallado de todas estas posibilidades alargaría excesivamente este trabajo.

Consideremos solamente las primeras posibilidades:

Caso 1: parámetro positivo

Caso 1.1: en la réplica $2i-1$, en la generación de un par (X, Y) determinado supongamos que:

V es pequeño, en el sentido que se elige la región $X \geq Y$,

W_1 es pequeño, en el sentido de que genera un valor de Y inferior a la medida de esta variable aleatoria.

W_2 es pequeño, en el sentido de que genera un valor de Y inferior a la mediana de esta variable aleatoria.

Entonces, en la réplica $2i$ se generará:

V grande, se elegirá la región $X < Y$,

W_1 pequeño, se generará un valor de X pequeño,

W_2 grande, se generará un valor de Y grande.

Caso 1.2: en la réplica $2i-1$, en la generación de un par (X, Y) determinado supongamos que:

V es pequeño, se elige la región $X \geq Y$

W_1 es pequeño, genera un valor de X inferior a la mediana de esta variable aleatoria,

W_2 es grande, genera un valor de Y superior a la mediana de esta variable aleatoria.

Entonces, en la réplica $2i$ se generará:

V grande, se elegirá la región $X < Y$,

W_1 pequeño, se generará un valor de X pequeño,

W_2 pequeño, se generará un valor de Y pequeño,

etc.

De manera que las réplicas con mayoría de pares (X, Y) concordantes, es decir, con valores altos de los estadísticos (36) o (40), tendrán réplicas antitéticas con los correspondientes valores bajos de estos estadísticos y viceversa.

3.3. Resultados de las Simulaciones

Bajo el diseño experimental anterior, se relizaron una serie de simulaciones para los mismos valores de los parámetros y de tamaños muestrales que los descritos en el apartado 3.1, con la salvedad de que, para cada combinación de un valor del parámetro θ y de un tamaño muestral n no se generó una única serie de $m = 10000$ estimaciones sino, como mínimo, tres de ellas. Es decir, las estimaciones del sesgo se basaron en tamaños reales de $m = 30000$ observaciones del estimador ($m/2 = 15000$ efectivas, recuérdese lo dicho en el apartado anterior). Para valores del parámetro entre 0,2 y 0.005 se hicieron cinco series, de manera que el tamaño fue de 50000 estimaciones. Para $\theta = 0$ el número de series fue de siete. Se aumentó de esta forma el número de observaciones debido a que para valores pequeños del parámetro los intervalos de confianza resultaban comparativamente más grandes.

La reducción de varianza fué del orden del 70%. Los intervalos de confianza alrededor de las estimaciones del sesgo son ahora mucho más estrechos, tal como se puede comprobar a partir de las tablas y listados adjuntos.

Dado que se comprobó que realmente

$$\beta(\theta, n) = -\beta(-\theta, n)$$

(indicando $\beta(\theta, n)$ el sesgo para el tamaño muestral n y el parámetro θ) solamente se indican los resultados correspondientes a valores no negativos del parámetro.

La figura 3.3.1 indica el sesgo en función del valor del parámetro, para algunos tamaños muestrales n y para los dos estimadores de θ en estudio. La figura 3.3.2 indica el sesgo en función del tamaño muestral, para algunos valores del parámetro. La tabla 3.3.1 muestra algunos valores del sesgo, la varianza y del error cuadrático medio para ambos estimadores, así como de sus estadísticos "jackknife" (véase la sección siguiente).

Como se puede apreciar, las indicaciones hechas en el apartado 3.1 quedan plenamente confirmadas. Como conclusión final se puede afirmar que el estimador U_K basado en el estadístico de Kendall es preferible a U_S. Ambos son sesgados, sin que su sesgo sea muy distinto, y cada uno de ellos es menos sesgado que el otro en, aproximadamente, la mitad del espacio paramétrico. En todo momento es más eficiente U_K. El error cuadrático es también siempre menor para este estimador.

Es posible un perfeccionamiento parcial sobre U_K tal como se discute en la siguiente sección.

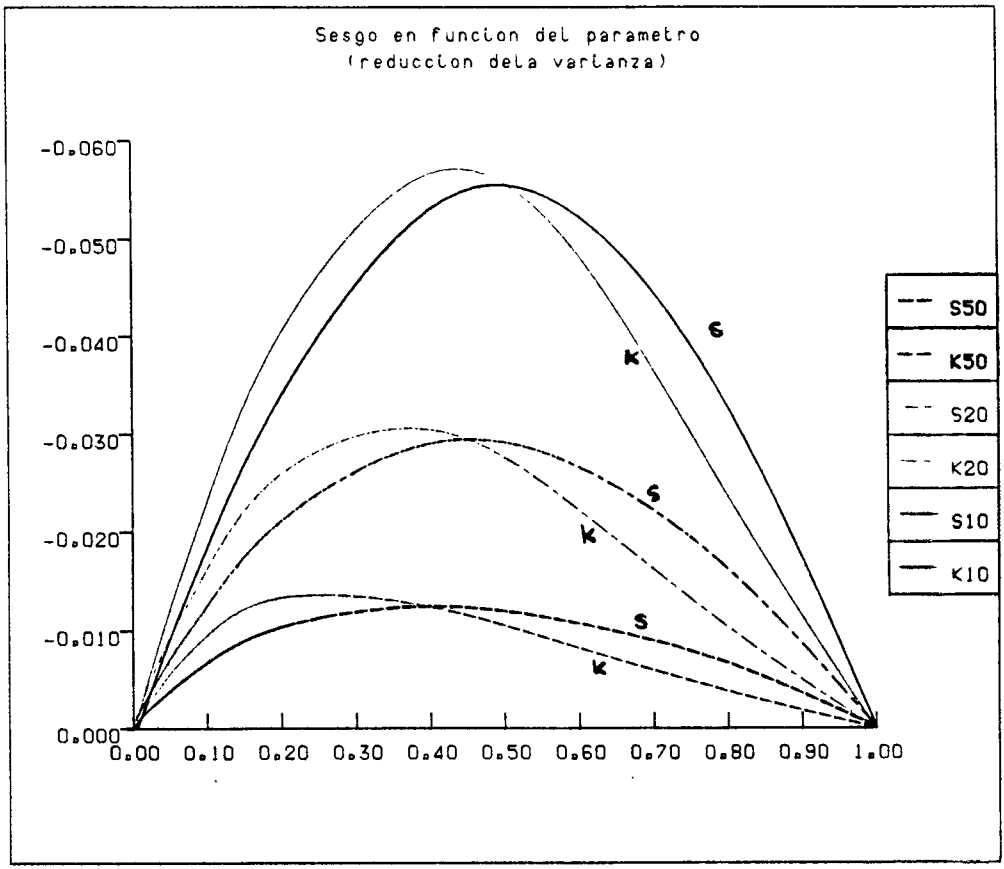

FIGURA 3.3.1. S10 indica la gráfica correspondiente al estimador basado en el estadístico de Spearman para $n = 10$, K10 corresponde al estimador basado en el estadístico de Kendall para $n = 10$, S20 corresponde al estimador de Spearman para $n = 20$, etc.

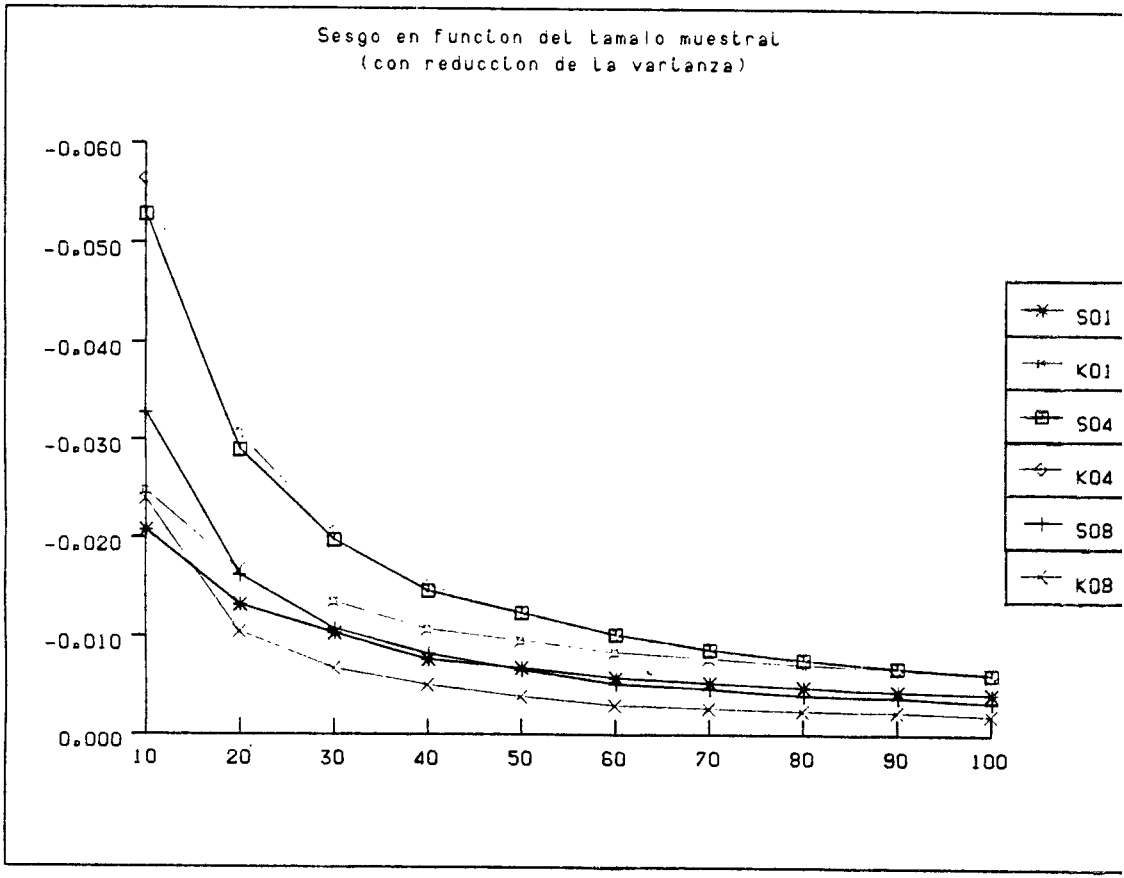

FIGURA 3.3.2. S01 indica la gráfica correspondiente al estimador de Spearman para parámetro $\theta = 0.1$, K01 corresponde al estimador de Kendall para $\theta = 0.1$, etc.

TABLA 3.3.1.

Sesgo. varianza y error cuadrático medio (MSE) de los dos estimadores para $\theta = 0.0$ (0.2) 0.8 y tamaños muestrales $n = 10, 20, 50$. Se indican al mismo tiempo los valores de varianza y error cuadrático verdaderos, propios de los estimadores, y los valores con reducción de la varianza, adecuados para construir intervalos de confianza más precisos para el sesgo.

ESTIMADOR DE SPEARMAN

	ESTIMADOR ORIGINAL					JACKKNIFE					
		REDUCCION DE V.		VERDAD.	VALOR.						
TAMAÑO M.	SESGO	VAR	MSE	VAR	MSE	SESGO	VAR	MSE	VAR	MSE	PARAMETRO
10	-0.00126	0.05133	0.05134	0.14489	0.14496	-0.00097	0.07017	0.07018	0.19637	0.19646	0.0
20	-0.00234	0.02447	0.02448	0.07563	0.07566	-0.00262	0.02980	0.02981	0.09161	0.09165	
50	-0.00161	0.00984	0.00984	0.03130	0.03130	-0.00174	0.01095	0.01095	0.03480	0.03480	
10	-0.03482	0.05609	0.05731	0.14491	0.14623	-0.01132	0.07442	0.07457	0.19162	0.19184	0.2
20	-0.02293	0.02750	0.02804	0.07549	0.07605	-0.00738	0.03257	0.03263	0.08900	0.08910	
50	-0.01121	0.01090	0.01102	0.03014	0.03024	-0.00258	0.01171	0.01172	0.03234	0.03234	
10	-0.05112	0.05109	0.05372	0.12943	0.13218	-0.00966	0.06352	0.06364	0.16006	0.16022	0.4
20	-0.02940	0.02433	0.02520	0.06418	0.06504	-0.00340	0.02669	0.02671	0.07015	0.07017	
50	-0.01239	0.00927	0.00942	0.02406	0.02419	-0.00059	0.00942	0.00942	0.02437	0.02437	
10	-0.04936	0.03593	0.03838	0.09576	0.09814	-0.00160	0.04046	0.04048	0.10502	0.10506	0.6
20	-0.02603	0.01597	0.01665	0.04356	0.04416	0.00009	0.01625	0.01625	0.04359	0.04360	
50	-0.01075	0.00594	0.00605	0.01596	0.01605	-0.00031	0.00595	0.00595	0.01579	0.01580	
10	-0.03330	0.02093	0.02204	0.04965	0.05070	0.00132	0.02013	0.02014	0.04607	0.04608	0.8
20	-0.01750	0.00874	0.00905	0.02105	0.02131	-0.00065	0.00854	0.00854	0.01982	0.01983	
50	-0.00729	0.00316	0.00321	0.00769	0.00773	-0.00076	0.00322	0.00322	0.00751	0.00751	

ESTIMADOR DE KENDALL

	ESTIMADOR ORIGINAL					JACKKNIFE					
		REDUCCION DE V.		VERDAD.	VALOR.						
TAMAÑO M.	SESGO	VAR	MSE	VAR	MSE	SESGO	VAR	MSE	VAR	MSE	PARAMETRO
10	-0.00069	0.04592	0.04593	0.13185	0.13191	0.00051	0.06576	0.06577	0.18618	0.18625	0.0
20	-0.00212	0.02203	0.02204	0.06833	0.06833	-0.00242	0.02797	0.02798	0.08549	0.08552	
50	-0.00160	0.00899	0.00900	0.02863	0.02863	-0.00182	0.01036	0.01036	0.03288	0.03288	
10	-0.04060	0.05163	0.05329	0.13332	0.13503	-0.01614	0.07201	0.07228	0.18328	0.18360	0.2
20	-0.02775	0.02551	0.02629	0.06888	0.06966	-0.01059	0.03124	0.03137	0.08310	0.08325	
50	-0.01404	0.01006	0.01026	0.02729	0.02745	-0.00338	0.01093	0.01094	0.02940	0.02940	
10	-0.05498	0.04669	0.04973	0.11758	0.12069	-0.01009	0.05950	0.05962	0.14668	0.14681	0.4
20	-0.03078	0.02134	0.02230	0.05559	0.05652	-0.00219	0.02278	0.02279	0.05858	0.05859	
50	-0.01229	0.00764	0.00779	0.01963	0.01976	0.00014	0.00741	0.00741	0.01891	0.01892	
10	-0.04574	0.02918	0.03128	0.08132	0.08336	0.00538	0.03080	0.03084	0.08258	0.08265	0.6
20	-0.02180	0.01149	0.01196	0.03351	0.03392	0.00291	0.01034	0.01036	0.02999	0.03001	
50	-0.00814	0.00381	0.00387	0.01135	0.01140	0.00029	0.00364	0.00364	0.01069	0.01069	
10	-0.02443	0.01418	0.01478	0.03564	0.03620	0.00819	0.01084	0.01091	0.02644	0.02653	0.8
20	-0.01137	0.00518	0.00530	0.01367	0.01378	0.00102	0.00447	0.00447	0.01162	0.01163	
50	-0.00439	0.00164	0.00166	0.00480	0.00481	-0.00038	0.00172	0.00172	0.00453	0.00454	

3.4. UN ESTIMADOR "JACKKNIFE"

Sea U_n un estimador sesgado, basado en n observaciones. Indiquemos como U_{-i} al mismo estimador excepto el hecho de que se ha eliminado la observación número i. Los estadísticos

$$J_i(U_n) = nU_n - (n-1)U_i, i = 1, \ldots, n \tag{46}$$

reciben el nombre de pseudovalores y el estadístico

$$\begin{aligned} J(U_n) &= n^{-1} \sum_{i=1}^{n} J_i(U_n) \\ &= nU_n - \left((n-1)/n\right) \sum_{i=1}^{n} U_{-i} \end{aligned} \tag{47}$$

se denomina el "jackknife" de U_n. Este concepto fue introducido por Quenouille [12] como método general de reducción del sesgo de un estimador. Su empleo es común en todo tipo de situaciones en las que se desea reducir el sesgo de un estimador, por ejemplo en Análisis discriminante, en la estimación de la probabilidad de clasificación errónea [13] o en problemas de predicción en medicina [6].

Si U_n es un estimador asintóticamente insesgado, se puede demostrar que ciertamente se obtiene un estimador menos sesgado, bajo condiciones bastante generales (véase por ejemplo [5]). Normalmente si,

$$E_\theta(Un) = \theta + 0(1/n) \tag{48}$$

entonces

$$E_\theta(J(U_n)) = \theta + 0(1/n^2) \tag{49}$$

El mayor inconveniente de este método es que en ocasiones aumenta la varianza de forma apreciable.

Como posible alternativa a los dos estimadores estudiados en la sección anterior, se propone y estudia un estimador jackknife basado en el estimador U_k que finalmente se propuso como el más indicado. En principio se espera que $J(U_k)$ sea manos sesgado qu U_k y, si es menos eficiente, al menos su varianza no excede la de U_s.

Como se recordará, el estimador U_k se basa en el estadístico (36) que se podría indicar, suponiendo $i, j = 1, \ldots, n$

$$K' = \sum_{\{(i,j) \ : \ i<j\}} \mathbf{I}_{[y(i)<y(j]} \tag{50}$$

suponiendo que la muesta ha sido ordenada, de manera que

$$X_(1) < X_(2) < \ldots < X_(n)$$

Si K_{-i} designa el estadístico K' habiendo eliminado la observación i tendremos que

$$\begin{aligned} K_i &= \sum_{\{(j,k) \ : \ j<k, j\neq i, j\neq k\}} \mathbf{I}_{[y(j) < y(k)]} \\ &= K' - \sum_{\{(j,k) \ : \ j=i \ o \ k=i\}} \mathbf{I}_{[y(j)<y(k)]} \\ &= \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} c(i,j,k) \quad \mathbf{I}_{[y(j)<y(k)]} \end{aligned} \tag{51}$$

siendo

$$c(i,j,k) = \begin{cases} 0 \text{ si } i=j \ o \ i=k \\ 1 \text{ en caso contrario} \end{cases} \tag{52}$$

Esta última expresión hace que $J(U_k)$ sea particularmente sencillo de evaluar. En la única interación necesaria para calcular K' se puede evaluar al mismo tiempo cada uno de los estadísticos K_{-i}. Se utilizó este algoritmo en el programa mediante el que se realizaron las simulaciones.

Cada uno de los pseudovalores de U_k se puede obtener de

$$U_{-i} = 2(4K_{-1} - (n-1)(n-2))/((n-1)(n-2) + |4K_{-1} + (n-1)(n-2)|) \tag{53}$$

y de ahí la estimación Jackknife directamente a partir de (47)

Las figuras 3.4.1 y 3.4.2, para los mismos tamaños muestrales indicados en las figuras 3.3, indican conjuntamente el sesgo y la varianza de los estimadores

U_k y $J(U_k)$. Como se puede apreciar, el estimador Jackknife es una posibilidad interesante puesto que su sesgo es claramente menor que el del estimador U_k mientras que la eficiencia no es mucho menor (de hecho en parte del espacio paramétrico el estimador Jackknife es más eficiente) y en el peor de los casos su eficiencia es del mismo orden que la de U_s.

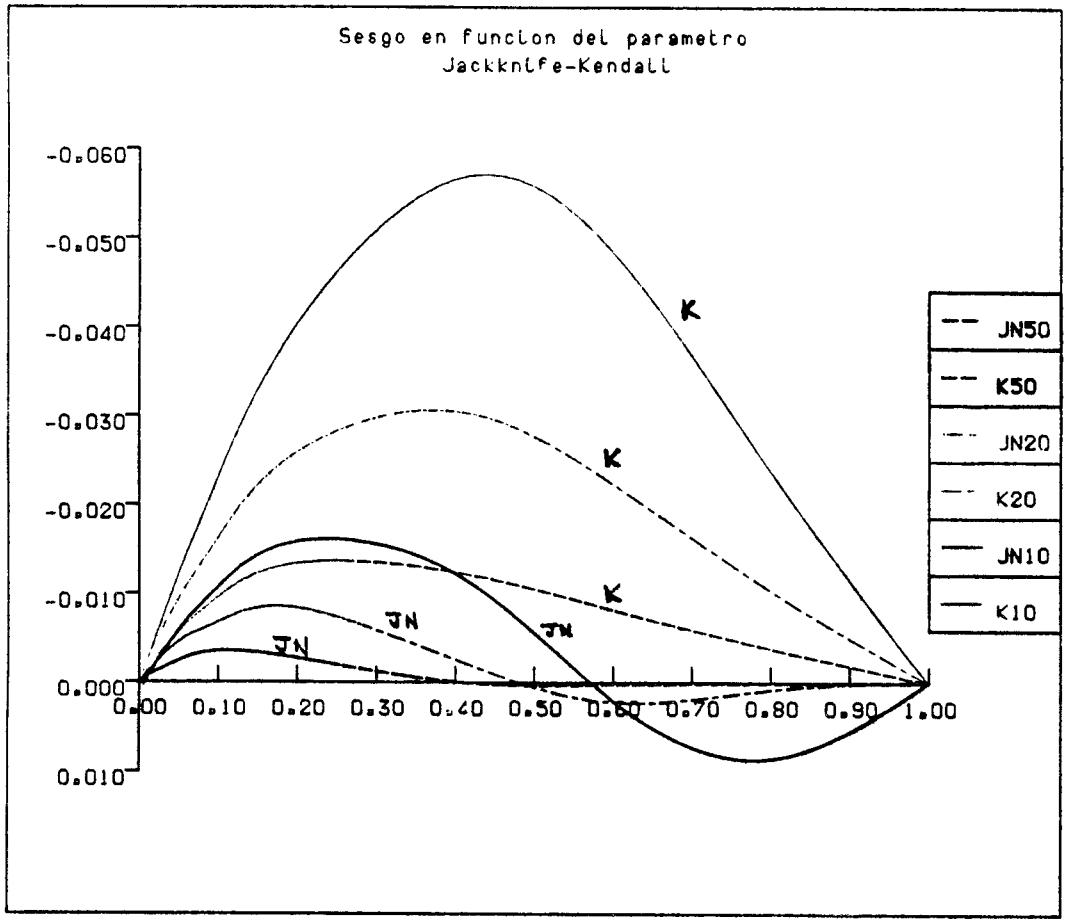

FIGURA 3.4.1. JN10 indica el estimador Jackknife para un tamaño muestral $n = 10$, K10 indica el correspondiente estimador de Kendall, etc.

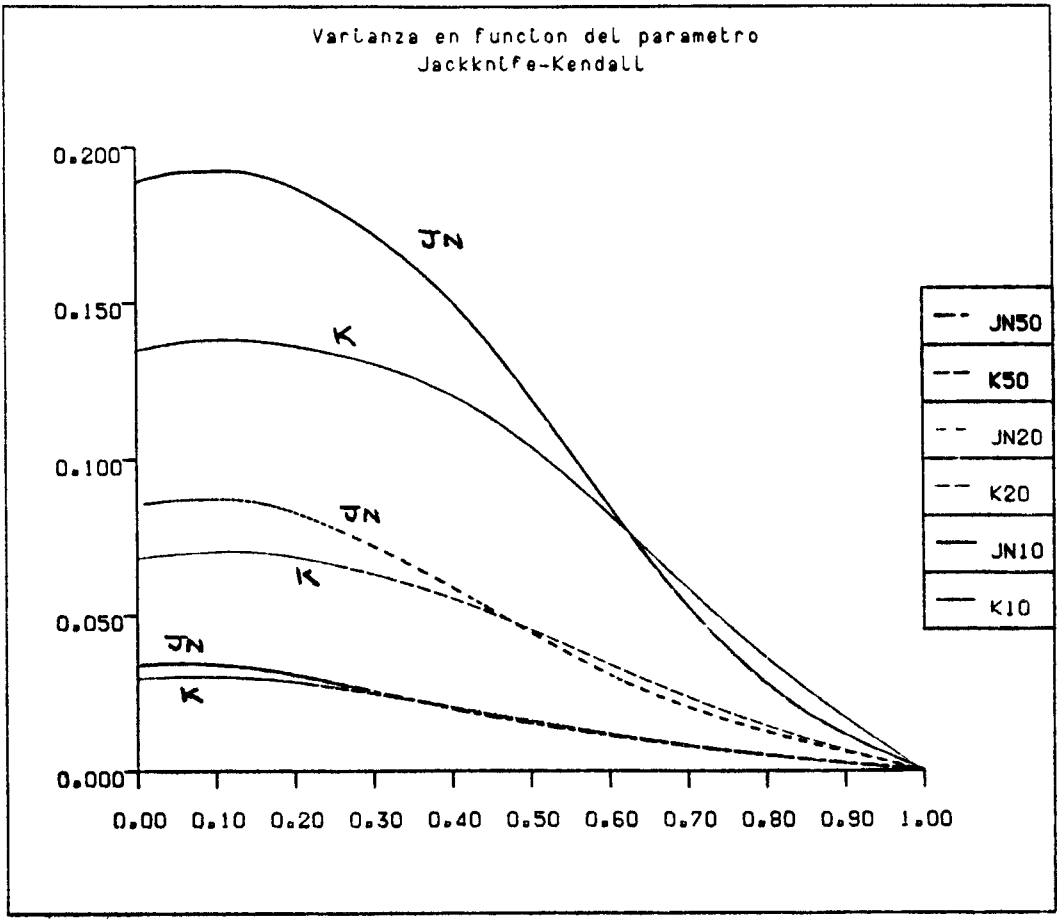

FIGURA 3.4.2. Los símbolos JN10, K10 etc. tienen el mismo sentido que en la figura anterior.

4. CONCLUSIONES E INDICACIONES FINALES

La forma en que, en la sección 2, se ha obtenido un generador aleatorio para la distribución bivariante estudiada, sugiere que el método basado en la descomposición de la función de distribución en una parte singular y una parte absolutamente continua es de validez general para poder realizar estudios de Monte Carlo sobre distribuciones que no sean ni discretas ni absolutamente continuas.

En cuanto a los posibles estimadores de la dependencia estudiados, parece claro que en todo momento es preferible el basado en el estadístico de Kendall en lugar del basado en el estadístico de Spearman. Este resultado, explicable y de validez bastante general, será el objeto de una futura publicación.

A pesar de que la mayoría de trabajos de Estadística en los que se emplea la simulación carecen de consideraciones sobre la precisión estadística de los resultados (con la esperanza de que ya se ha realizado un número "muy grande" de réplicas), de la discusión de las secciones anteriores se desprende la conveniencia de tenerla en cuenta y de, en ocasiones, emplear diseños experimentales (técnicas de reducción de la varianza) distintos del asociado al muestreo aleatorio simple.

En caso de solicitud, los autores facilitarán los programas elaborados en este trabajo, a saber:

> Subrutinas para la generación aleatoria del sistema bivariante de Cuadras–Augé, redactadas en FORTRAN 77.
>
> Muestreo y estimación del sesgo, eficiencia y error cuadrático de los estimadores estudiados. Estos programas están redactados en Pascal (PASCAL VS).

También hay listados disponibles con información más detallada sobre los resultados de las simulaciones que la indicada en las tablas y figuras anteriores.

5. BIBLIOGRAFIA

[1] **Bratley, P., Fox, B. L., Schrage, L. E.** "A guide to simulation". Springer, 1983.

[2] **Consonni, G.** "On measure of concordance". ISI 44th Session. Contrib. Papers, 2: 550-553, 1983.

[3] **Cuadras, C. M.** "Sobre medidas de dependencia estocástica invariantes". "Homenatge a Francesc d'A. Sales." Facultat de Matemàtiques - Universitat de Barcelona: 28-47, 1985.

[4] **Cuadras, C. M., Augé, J.** " A continuous general multivariate distribution and its properties". Commun. Stadist. - Theor. Meth., A 10: 339-353, 1981.

[5] **Efron, B.** "The Jackknife, the Bootstrap an other Resampling Plans". CBMS – NSF Series. SIAM 1982.

[6] **Gong, G.** "Cross-validation, the jackknife, and the bootstrap". "Excess error estimation in forward logistic regression". JASA, 81: 109-113, 1986.

[7] **Kimeldorf, G., Sampson, A.** "One–parameter families of bivariate distributions with fixed marginals". Commun. Statist. 4: 293-301, 1975 a.

[8] **Kimeldorf, G., Sampson, A.** "Uniform representations of bivariate distributions". Sommun. Statist. 4: 617-627, 1975.

[9] **Kleijnen, J. P. C.** "Statistical Techniques in Simulation". Marcel Dekker, 1974.

[10] **Lehmann, E. L.** "Nonparametrics. Statistical Methods Based on Ranks". Holden–Day/Mc Graw–Hill, 1975.

[11] **Looney, S. W.** "A comparison of estimators of a common correlation coefficient". Commun. Statis. Simula., 15: 531-543, 1986.

[12] **Quenouille, M.** "Notes on bias in estimation". Biometrika, 43: 353-360, 1956.

[13] **Rao, P. S. R. S. , Dorvlo, A. S.** "The jackknife procedure for the probabilities of misclassification". Commun. Statist. - Simula., 14: 779-790, 1986.

[14] **Whitt, W.** "Bivariate distributions with given marginals". Ann. Stat., 31: 188-190, 1976.