Generación de datos binarios correlacionados: una comparación de dos métodos

Hachuel, Leticia; Wojdyla, Daniel; Boggio, Gabriela; Cuesta, Cristina; Servy, Elsa
Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística.

Introducción: Una forma de evaluar el comportamiento de estimadores y tests bajo condiciones que pongan en duda sus propiedades asintóticas es a través de estudios por simulación.
Las conclusiones de dichos estudios se ven reafirmadas en la medida que resulten consistentes a través de diferentes modelos de generación de datos.
En el marco del proyecto "Estudio del comportamiento de estadísticas para medidas repetidas en muestras pequeñas bajo escenarios múltiples" *, este trabajo tiene por objeto la comparacion de algoritmos de generación para luego llevar adelante estudios de Montecarlo relacionados con el comportamiento de estadísticas que consideren la falta de independencia de las observaciones.

Objetivo específico: comparar los algoritmos de generacion de datos binarios correlacionados presentados por:
1) Park, Park y Shin (1996)
2) Servy, Hachuel y Wojdyla (1997-1998)

Características de los algoritmos:

Algoritmo de Park, Park y Shin

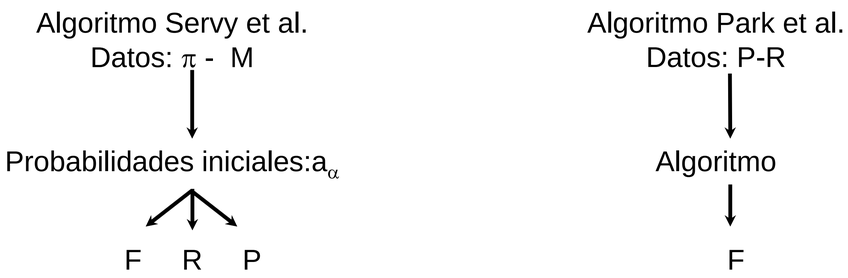
Para generar un vector aleatorio de k variables binarias se considera un conjunto de k variables aleatorias Poisson Y_1, ..., Y_k que son, a la vez, sumas parciales de variables Poisson independientes, $X_1(\beta_1)$, ..., $X_\tau(\beta_\tau)$.
Los valores esperados y la estructura de las correlaciones de las variables binarias Z_i, i=1,...,k definidas por $Z_i=I_{\{0\}}(Y_i)$ se obtienen controlando el esquema de aparición simultánea de $X_1(\beta_1)$, ..., $X_\tau(\beta_\tau)$ y las magnitudes de β_1, ..., β_τ. El algoritmo describe cómo determinar τ (número de variables X's), β_1, ..., β_τ y el esquema de sumas parciales a fin de generar un vector aleatorio binario k-dimensional $(Z_1, ..., Z_k)'$ con un vector de medias pre-especificado $P=(p_1, ..., p_k)'$ y matriz de correlación también pre-especificada R.

Algoritmo de Servy, Hachuel y Wojdyla

Cada vector k-dimensional de variables binarias se genera por los k primeros pasos de una cadena de Markov. La primera respuesta se elige de acuerdo a una distribución de probabilidades $\{a_{u(1)}, u(1) \varepsilon R; R=\{0,1\}\}$. Las respuestas siguientes se simulan según una cadena de Markov, cuyas probabilidades iniciales son $\{a_{u(1)}, u(1) \varepsilon R\}$ y matriz de transición $M=((p_{\alpha\beta}))$ $\alpha,\beta \varepsilon R$ donde $p_{\alpha\beta}$ designa a la probabilidad condicional de que $u(v) = \beta$ dado que $u(v-1) = \alpha$ para v = 2,..k.
Para formar otro conglomerado, es decir otro vector de respuestas binarias correlacionadas, se inicia otra cadena similar independiente de la anterior.
Los parámetros iniciales de este procedimiento son:
π_1: probabilidad de obtener la respuesta 1 en la población y
$M=((p_{\alpha\beta}))$ $\alpha,\beta \varepsilon R$ con R={0,1}.

Descripción de la relación entre los parámetros de ambos algoritmos

Sea k=3; $P=(p_1,p_2,p_3)$, $R=(\rho_{12};\rho_{13};\rho_{23})$ y $F=(p_{000};p_{001};p_{010};p_{100};p_{011};p_{101};p_{110};p_{111})$.



Compatibilización de los parámetros de ambos algoritmos

i) Aplicar el algoritmo Servy et al.
ii) Elegir aquellos casos que originan que los valores de R sean positivos. Calcular F y P.
iii) Considerar estos valores de P y R como datos iniciales para el algoritmo Park et al.

Evaluación de los modelos

Una vez elegidos los casos compatibles en términos de los parámetros, se compara la consistencia de ambos algoritmos generando muestras y calculando las estimaciones de las probabilidades marginales y de las correlaciones entre pares de posiciones.

Para ello se procede a:
i) Simular muestras de tamaño n=30, 50, 70 y 100 con ambos algoritmos y calcular las estimaciones de P y R.
ii) Repetir el procedimiento 1000 veces y calcular el promedio y los desvíos estándares de las estimaciones.
iii) Comparar los resultados obtenidos para los dos algoritmos.

Resultados:

Se eligieron dos escenarios paramétricos para la evaluación de los modelos.

Escenario	π_1	M	$P=(p_1,p_2,p_3)$	$R=(\rho_{12};\rho_{13};\rho_{23})$
1	0.6	0.6 0.4 0.4 0.6	(0.74,0.40, 0.48)	(0.17,0.04, 0.20)
2	0.6	0.7 0.3 0.3 0.7	(0.69,0.58, 0.53)	(0.37,0.15, 0.40)

Promedios y desvíos est. de las estimaciones de las componentes de P y R para el Escenario 1

	Park et al			Servy et al		
n	p_1=0.74	p_2=0.55	p_3=0.51	p_1=0.74	p_2=0.55	p_3=0.51
30	0.74 (0.08)	0.55 (0.09)	0.51 (0.09)	0.74 (0.08)	0.55 (0.09)	0.51 (0.09)
50	0.74 (0.06)	0.55 (0.07)	0.51 (0.07)	0.74 (0.06)	0.54 (0.07)	0.51 (0.07)
70	0.74 (0.05)	0.55 (0.06)	0.51 (0.06)	0.74 (0.05)	0.55 (0.06)	0.51 (0.06)
100	0.74 (0.04)	0.55 (0.05)	0.51 (0.05)	0.74 (0.05)	0.55 (0.05)	0.51 (0.05)
n	ρ_{12}=0.17	ρ_{13}=0.04	ρ_{23}=0.20	ρ_{12}=0.17	ρ_{13}=0.04	ρ_{23}=0.20
30	0.19 (0.19)	0.04 (0.18)	0.20 (0.17)	0.18 (0.18)	0.04 (0.18)	0.19 (0.18)
50	0.17 (0.14)	0.03 (0.14)	0.20 (0.14)	0.17 (0.14)	0.03 (0.14)	0.21 (0.14)
70	0.18 (0.12)	0.04 (0.12)	0.20 (0.12)	0.18 (0.12)	0.03 (0.12)	0.20 (0.12)
100	0.18 (0.10)	0.04 (0.10)	0.20 (0.10)	0.18 (0.10)	0.04 (0.10)	0.20 (0.10)

Promedios y desvíos est. de las estimaciones de las componentes de P y R para el Escenario 2

	Park et al			Servy et al		
n	p_1=0.69	p_2=0.58	p_3=0.53	p_1=0.69	p_2=0.58	p_3=0.53
30	0.69 (0.09)	0.57 (0.09)	0.53 (0.09)	0.69 (0.08)	0.58 (0.09)	0.53 (0.09)
50	0.69 (0.07)	0.58 (0.07)	0.53 (0.07)	0.69 (0.07)	0.58 (0.07)	0.53 (0.07)
70	0.69 (0.05)	0.58 (0.06)	0.53 (0.06)	0.69 (0.05)	0.58 (0.06)	0.53 (0.06)
100	0.69 (0.04)	0.58 (0.05)	0.53 (0.05)	0.69 (0.05)	0.57 (0.05)	0.53 (0.05)
n	ρ_{12}=0.37	ρ_{13}=0.15	ρ_{23}=0.40	ρ_{12}=0.37	ρ_{13}=0.15	ρ_{23}=0.40
30	0.37 (0.18)	0.14 (0.20)	0.40 (0.17)	0.37 (0.18)	0.15 (0.18)	0.39 (0.17)
50	0.38 (0.14)	0.14 (0.14)	0.39 (0.13)	0.36 (0.13)	0.15 (0.15)	0.40 (0.13)
70	0.38 (0.11)	0.15 (0.12)	0.40 (0.11)	0.38 (0.11)	0.15 (0.12)	0.39 (0.11)
100	0.37 (0.10)	0.15 (0.10)	0.39 (0.09)	0.37 (0.09)	0.15 (0.10)	0.39 (0.09)

Discusión:

En la evaluación de cuán estrechamente las muestras generadas por los modelos son capaces de estimar los parámetros especificados, los resultados parciales obtenidos han mostrado una notable concordancia en la estimación de las correlaciones intra-conglomerados y en los valores de las probabilidades marginales .

Sin embargo los escasos escenarios considerados impiden al momento concluir enfáticamente al respecto y se hace necesario completar los estudios ampliando las carácterísticas de los mismos.

* Proyecto de Investigación de la Secretaría de Ciencia y Tecnología de la UNR – Año 2001 (PID 2001).