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

GENERACIÓN DE DATOS BINARIOS CORRELACIONADOS: UNA COMPARACIÓN DE DOS MÉTODOS

1. INTRODUCCIÓN

Frecuentemente el diseño de las investigaciones involucra múltiples mediciones de una variable respuesta. Un ejemplo lo constituyen los estudios longitudinales, en los cuales las mediciones repetidas se obtienen para cada individuo a través del tiempo. En otras aplicaciones, la respuesta de cada unidad experimental se mide bajo múltiples condiciones en lugar de en diferentes momentos. Asimismo, en estudios observacionales una estructura compleja de muestreo genera una situación similar al medir la misma variable en diferentes individuos agrupados en conglomerados. En todos los casos, la característica común de estos ejemplos es la falta de independencia entre las observaciones.

Dentro de este contexto, el análisis de datos binarios correlacionados es un área de interés creciente. Una aplicación típica de este tipo de datos consiste en estudiar la relación entre la medida de respuesta dicotómica y covariables a través de la aplicación de métodos de regresión.

Prentice (1988) proporcionó una revisión comprensiva de los métodos desarrollados para el análisis de regresión de datos binarios correlacionados cuando las covariables están asociadas con cada respuesta binaria y donde se destaca el enfoque de la ecuación de estimación generalizada (GEE) sugerido por Liang y Zeger (1986) y Zeger, Liang y Albert (1988). El enfoque GEE, sólo requiere especificar correctamente la esperanza marginal de las respuestas a fin de obtener estimadores de los coeficientes de regresión consistentes y asintóticamente normales, junto con un estimador consistente de la matriz de covariancias. Aunque las propiedades asintóticas de los estimadores GEE son aceptadas, sus propiedades para tamaños de muestra pequeños no son muy conocidas.

Rotnitzky y Jewell (1990) han derivado tests de tipo Wald y score generalizados para evaluar hipótesis sobre los parámetros de regresión estimados por GEE, también con propiedades a nivel asintótico.

Una forma de evaluar el comportamiento de estimadores y tests bajo condiciones que pongan en duda las propiedades asintóticas es a través de estudios por simulación. Por ejemplo para investigar las propiedades de los estimadores de regresión en análisis que usan el enfoque GEE con tamaños de muestras pequeños, resulta apropiado simular datos con probabilidades de respuesta marginales conocidas y correlaciones de a pares especificadas.

Una revisión de la literatura de generación de variables binarias indica que, mientras que existen algoritmos para generar datos correlacionados con ciertas distribuciones continuas, son escasos y de más reciente aparición los algoritmos para datos dicotómicos

con propiedades deseadas particulares. Es importante señalar que las conclusiones de los estudios por simulación se reafirman en la medida que resulten consistentes a través de diferentes modelos de generación de datos. Por tal razón y en el marco del proyecto "Estudio del comportamiento de estadísticas para medidas repetidas en muestras pequeñas bajo escenarios múltiples"*, la primera etapa de esta investigación tiene por objetivo la comparación 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.

En la sección siguiente se presentan algunos de los enfoques adoptados para la generación de datos binarios.

2. ANTECEDENTES

Entre los diferentes enfoques de generación de variables binarias correlacionadas se encuentra el de Bahadur (1961), quien sugirió un modelo paramétrico que se expresa como una función de masa conjunta de $(Z_1, ..., Z_k)$ siendo pi la esperanza marginal de una variable binaria Z_i, i=1,..., k y $\rho_{ij}=corr(Z_i, Z_j)$. Debido a que la distribución de Badahur es difícil de manejar especialmente para valores de k grandes, este enfoque no es adecuado para generar vectores aleatorios binarios correlacionados de gran dimensión.

Otro algoritmo para simular variables binarias correlacionadas fue propuesto por Emrich y Piedmonte (1991). Sea $\phi(x_1, x_2; r)$ la función de distribución acumulada para una normal bivariada con coeficiente de correlación r. El algoritmo primero resuelve las siguientes ecuaciones:

$$\phi[z(p_i), z(p_j); r_{ij}]= \rho_{ij}(p_i\, q_i\, p_j\, q_j)^{1/2} + p_i\, p_j \qquad (2.1)$$

para r_{ij} (i=1, ..., k-1; j=i+1, ..., k), donde $q_i=1-p_i$ y z(p) es el cuantil p-ésimo de la distribución normal estándar. El próximo paso es generar un vector aleatorio normal k-dimensional $Y=(Y_1, ... ,Y_k)'$ con media cero, variancias unitarias y matriz de correlación $R=[r_{ij}]$. Las variables binarias deseadas se obtienen fijando $Z_i=1$ si $Y_i \leq z(p_i)$ y $Z_i=0$ en otro caso. Una desventaja de este algoritmo es que se deben resolver las ecuaciones no lineales (2.1), lo cual requiere integraciones numéricas en la evaluación de ϕ. Posteriormente, Lee (1993) presentó un método para generar una secuencia aleatoria binaria que es apropiado para cualquier distribución marginal y cualquier conjunto de razones de odds consistentes. Este método también requiere la resolución de un gran número de ecuaciones no lineales.

Gange (1995) propuso un procedimiento iterativo para generar variables categóricas dependientes utilizando el algoritmo de ajuste proporcional. Primeramente la dependencia de las variables categóricas se formula mediante tablas de contingencia; a continuación se construye la probabilidad conjunta ajustando un modelo loglineal a las tablas de contingencia. Finalmente se generan las variables categóricas dependientes comparando números aleatorios uniformes simulados con la probabilidad conjunta. Este procedimiento tiene la desventaja de requerir, por lo tanto, otro procedimiento de ajuste iterativo.

Otros métodos de generación de datos son los desarrollados por Park et al (1996) y Servy et al (1997, 1998), los cuales se describen con mayor detalle a continuación, ya que son los elegidos para llevar a cabo los objetivos de este trabajo.

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

3. ALGORITMO DE PARK, C. G.; PARK, T.; SHIN, D. W.

Este algoritmo permite generar un vector aleatorio $(Z_1, Z_2, ..., Z_k)'$ de variables binarias tales que $E(Z_i)=p_i$ y corr $(Z_i, Z_j)=\rho_{ij} \geq 0$, $i \neq j$. Estas probabilidades y los parámetros de correlación pueden eventualmente expresarse en función de covariables. Para facilitar la presentación se considera en primer término k=2. Sea $X(\alpha)$ una variable aleatoria Poisson con media $\alpha \geq 0$. Se consideran dos variables aleatorias definidas por

$$Y_1=X_1(\alpha_{11} - \alpha_{12}) + X_3 (\alpha_{12})$$

$$Y_2=X_2(\alpha_{22} - \alpha_{12}) + X_3 (\alpha_{12}) \qquad (3.1)$$

donde α_{11}, α_{22} y α_{12} son constantes no negativas y X_i , i=1,..3 variables aleatorias Poisson mutuamente independientes.

Obviamente, Y_1 y Y_2 siguen distribuciones de Poisson con medias α_{11} y α_{22}, respectivamente. Dado que ellas comparten un término común, X_3 (α_{12}), las dos variables Poisson Y_1 e Y_2 están correlacionadas no negativamente. Sea $Z_i = I_{\{0\}}(Y_i)$ i=1,2, donde I_A es la función indicadora de un conjunto A tal que $I_A(y)= 0$ si $y \notin A$. Debido a que Y_1 e Y_2 están correlacionadas no negativamente, de igual manera lo están Z_1 y Z_2 y es claro que el coeficiente de correlación ρ_{12} entre Z_1 y Z_2 es creciente en α_{12}.

Dado que:

$E(Z_1Z_2) = P(X_1 = X_2 = X_3 = 0) = \exp(-(\alpha_{11} + \alpha_{22} - \alpha_{12})) = p_1p_2e^{\alpha_{12}}$ y

$E(Z_i) = E(Z_i^2) = P(X_i = X_3 = 0) = e^{-\alpha_{ii}} = p_i$, i=1,2 resulta que:

$var(Z_i) = p_iq_i$, donde $q_i = 1-p_i$, i=1,2 y

$cov(Z_1, Z_2) = p_1p_2(e^{\alpha_{12}} -1) = \rho_{12}(p_1q_1p_2q_2)^{1/2}$.

Entonces, Z_1 y Z_2 tienen coeficiente de correlación:

$$\rho_{12} = p_1p_2(e^{\alpha_{12}} -1)/(p_1q_1p_2q_2)^{1/2} \qquad (3.2)$$

y a partir de (3.2) se obtiene la expresión deseada para α_{12} :

$$\alpha_{12} = \log(1 + \rho_{12}\{q_1p_1^{-1}q_2p_2^{-1}\}^{1/2}) \quad . \qquad (3.3)$$

Por lo tanto las constantes α_{11}, α_{22} y α_{12} se pueden elegir fijando $E(Z_1)=p_1$, $E(Z_2)=p_2$ y $corr(Z_1,Z_2)= \rho_{12}$.

El caso particular de independencia entre Z_1 y Z_2 se presenta para $\alpha_{12} = 0$.

Como $E(Z_1Z_2) = P(Z_1 = Z_2 = 1) \leq P(Z_i = 1) = p_i$, i=1,2, se tiene entonces que:

$cov(Z_1, Z_2) \leq p_1q_2$ y $cov(Z_1, Z_2) \leq p_2q_1$. De aquí resulta que, como se describe en Emrich y Piedmonte (1991), ρ_{12} no varía libremente en el intervalo [-1,1]. Esto es, se verifica que

$$\rho_{12} \leq (p_2q_1/(p_1q_2))^{1/2} \text{ y } \quad \rho_{12} \leq (p_1q_2/(p_2q_1))^{1/2} \qquad (3.4)$$

Esta desigualdad conduce a la relación $\alpha_{12} \leq \log(p_i^{-1}) = \alpha_{ii}$, i=1, 2. Por lo tanto, las relaciones expresadas en (3.1) generan vectores de variables binarias bivariadas correlacionadas no negativamente.

En el caso general $k \geq 2$, se considera un conjunto de k variables aleatorias Poisson

Y_1, ..., Y_k que son, a la vez, sumas parciales de variables Poisson independientes, $X_1(\beta_1)$, ..., $X_\tau(\beta_\tau)$ para ciertos enteros no negativos τ y números reales no negativos β_1, ..., β_τ. Algunas de las variables $X_1(\beta_1)$, ..., $X_\tau(\beta_\tau)$ pueden aparecer simultáneamente en la conformación de diferentes variables Y-es. Los valores esperados y la estructura de las correlaciones de las variables binarias Z_i, i=1,...,k definidas por $Z_1 = I_{\{0\}}(Y_1)$, ..., $Z_k = I_{\{0\}}(Y_k)$ se pueden encontrar apropiadamente 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 que se presenta a continuación describe cómo determinar τ, 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 vector de medias especificado

$(p_1, ..., p_k)'$ y matriz de correlación también especificada $R=[\rho_{ij}]$ con $\rho_{ij} \geq 0$.

Los pasos del algoritmo son:

Paso 0: Calcular α_{ij} según (3.3) para $1 \leq i, j \leq k$. Sea l=0.

Paso 1: Sea l=l+1, $T_l = \{\alpha_{ij} : \alpha_{ij} > 0,\ 1 \leq i, j \leq k\}$ y sea $\beta_l = \alpha_{rs}$ el menor elemento en el conjunto T_l. Si resulta $\alpha_{rr} = 0$ o $\alpha_{ss} = 0$, entonces se debe suspender el procedimiento. En caso contrario, se elige un conjunto indexado S_l definido:

$S_l = \{s, r\} \bigcup \{\{i, j\} / \alpha_{ij} > 0, i \leq k, j \leq k\}$, es decir para todos los subconjuntos {i,j} pertenecientes a S_l resulta $\alpha_{ij} > 0$.

Paso 2: Para todos los subconjuntos $\{i, j\} \in S_l$ se reemplaza α_{ij} por $\alpha_{ij}-\beta_l$. Si resultan todos los $\alpha_{ij} = 0$, entonces se pasa al Paso 3. En caso contrario, se va al Paso 1.

Paso 3: Sea τ=l. Para i=1, 2, ..., k se construyen las variables Y_i de la siguiente forma: $Y_i = \sum X_l(\beta_l)\, I_{S_l}(i)$. A partir de ellas se definen finalmente las variables binarias objetivo z_i correlacionadas según $Z_i = I_{\{0\}}(Y_i)$, con i=1,..,k.

Debe notarse que en el Paso 1 se elige el β_l mínimo para asegurarse que todas las variables aleatorias Poisson tengan medias no negativas. Es posible que α_{rs} y S_l en el Paso 1 no puedan ser determinados de manera única. En tal caso se debe elegir α_{rs} y S_l arbitrariamente.

Se puede incluso computar luego la distribución de probabilidad de los vectores de respuestas binarias $(Z_1, ..., Z_k)$, F, la cual se puede utilizar para simular la extracción de una muestra de conglomerados o de vectores de respuestas binarias.

4. ALGORITMO DE SERVY, E.; HACHUEL, L.; WOJDYLA, D.

Este modelo de simulación, presentado en Servy, Hachuel y Wojdyla (1997, 1998), fue diseñado para generar muestras de conglomerados cuyos elementos son pares de valores de dos variables categóricas, de forma tal que finalmente la muestra puede presentarse bajo la forma de una tabla de contingencia bivariada. Los conglomerados se generan por los k pasos de una cadena de Markov. Si se ignora una de las variables, dichos conglomerados se transforman en univariados y si esa variable es binaria, el algoritmo se puede utilizar para generar muestras de conglomerados univariados o de vectores de respuestas binarias correlacionadas. A continuación se describe brevemente el modelo.

En forma general, en este esquema de generación de datos, el primer individuo de un conglomerado de tamaño k recibe una respuesta de acuerdo con las probabilidades {a(1),.., a(r)}. La variable u(1), que toma valores en R = {1,..,r} con distribución $\{a_{u(1)}, u(1) \varepsilon R\}$, representa la respuesta del primer individuo del conglomerado. Los individuos siguientes se simulan según una cadena de Marcov, 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) = β dado que u(v-1) = α para v = 2,..k.

Cada conglomerado es una cadena de Markov de longitud k. Para formar otro conglomerado, se inicia otra cadena similar independiente de la anterior. En cualquier conglomerado, la respuesta del v-ésimo individuo está simbolizada por la variable u(v). Las variables {u(v)} pueden tener diferentes distribuciones, cuando v varía de 1 a k. Pero el campo de variación de todas ellas es R. La probabilidad de un conglomerado genérico de tamaño k es,

$$P_{k,u(1),\,u(2),..,\,u(k)} = a_{u(1)}\, p_{u(1)u(2)}\, p_{u(2)u(3)},.., \, p_{u(k-1)\,u(k)} \qquad (4.1)$$

Otras probabilidades de interés son las probabilidades de que el primer individuo del conglomerado posea el valor identificado por α y el v-ésimo posea el identificado con β :

$$((P_{k,\,u(1)\,=\,\alpha,\,u(v)\,=\,\beta}))_{\alpha,\,\beta\,\varepsilon\,RC} = D\,(a_\alpha)\,M^{(v-1)} \qquad (4.2)$$

donde $D(a_\alpha)$ es la matriz diagonal cuyos elementos no nulos son las probabilidades iniciales y el segundo factor es la potencia (v-1)-ésima de la matriz de transición M.

Además,

$$P_{k,u(v)=\alpha} = \sum_{u(1),..,u(k)\neq u(v)}^{rc} P_{k,u(1),...,u(v)=\alpha,..,u(k)} = \sum_{\beta\in RC} P_{k,u(1)=\beta,u(v)=\alpha} = \sum_{\beta\in RC} a_\beta p_{\beta\alpha}^{(v-1)} \qquad (4.3)$$

Las probabilidades $P_{k,\,u(1),..,\,u(k)}$ son fácilmente estimables, porque la muestra de conglomerados es de tipo simple al azar. Sirven para calcular las probabilidades $\{\pi_\alpha\, \alpha \in R\}$, es decir las probabilidades de que al extraer un individuo al azar de la población, el mismo posea la característica indicada α.

Así, la probabilidad de obtener un individuo con características correspondientes a α cuando la población está formada por conglomerados de tamaño k es:

$$\pi_\alpha = \sum_{u(1),..,u(k)} P_{k,u(1),..,u(k)} Pr(\alpha/u(1),..,u(k)) \qquad (4.4)$$

donde Pr(α /u(1),.., u(k)) es la probabilidad de escoger un individuo en la categoría identificada por α cuando se extrajo el conglomerado identificado por (u(1),., u(k)). Siendo la extracción de tipo aleatorio simple, ella coincide con el cociente entre m(α / u(1),.., u(k)), que representa el número de individuos de tipo α en el conglomerado, y k, el tamaño del mismo.

$$\text{O sea, } k\,\pi_\alpha = \sum_{u(1),..,u(k)} P_{k,u(1),..,u(k)}\, m(\alpha/u(1),..,u(k)) \;\; \alpha\in R \qquad (4.5)$$

que alternativamente, se puede escribir:

$$k\,\pi_\alpha = \sum_{v=1}^{k} P_{k,u(v)=\alpha} = a_\alpha + a_1 \sum_{t=1}^{(k-1)} p_{1\alpha}^{(t)} + a_2 \sum_{t=1}^{(k-1)} p_{2\alpha}^{(t)} + ... + a_{(r-1)} \sum_{t=1}^{(k-1)} p_{(r-1)\alpha}^{(t)} \qquad (4.6)$$

El sistema (4.6) es crucial en las simulaciones. Para generar datos que representen una población con parámetros $\{\pi_\alpha \; \alpha \in R\}$ prefijados y que estén relacionados según un proceso de Marcov dado, cuya matriz de transición es M, en principio arbitraria, se comienza despejando las incógnitas $\{a_\alpha : \alpha \varepsilon R\}$ en las ecuaciones (4.6).

Luego, utilizando estas soluciones como probabilidades iniciales y la matriz de transición dada se generan cadenas independientes de un número fijo de pasos (k), que constituyen la muestra de conglomerados deseados.

No todas las soluciones del sistema son satisfactorias como probabilidades iniciales. La estructura del sistema garantiza que la suma de las probabilidades iniciales es igual a 1 pero no garantiza que las soluciones sean positivas. Esto significa que la "arbitrariedad" de $((p_{\alpha\beta}))$, cuando los $\pi's$ están prefijados está acotada por la existencia de las soluciones.

En particular, para generar conglomerados o vectores conformados por variables binarias correlacionadas resulta: R={0,1}, $\{\pi_\alpha \; \alpha = 0,1\}$ y la matriz de transición $M=((p_{\alpha\beta}))$ $\alpha,\beta \varepsilon R$ de dimensión 2x2.

5. COMPARACIÓN DE LOS ALGORITMOS

En los estudios de simulación los datos están generados por modelos probabilísticos que describen de la manera más fiel posible las poblaciones concretas a las cuales se pueden aplicar los procedimientos bajo estudio.

Variando los parámetros de estos modelos se pueden definir diferentes escenarios en los cuales los procedimientos de interés pueden llevarse a cabo. Pero los escenarios posibles que se pueden crear dependen a su vez de los modelos. De allí que es importante comparar diferentes modelos de generación de datos ya que las conclusiones de los estudios empíricos toman fuerza cuando son consistentes a través de datos obtenidos por diferentes modelos.

El objetivo general de este proyecto es comparar el comportamiento de estadísticas a través de estudios de simulación que utilicen diferentes algoritmos de generación de datos binarios correlacionados. Por lo tanto en una primera etapa se trata de comparar los algoritmos recién presentados en términos de si a partir de ellos se reproducen más o menos fielmente los parámetros especificados.

Ambos algoritmos parten de la especificación de diferentes parámetros que determinan las características de los datos binarios generados.

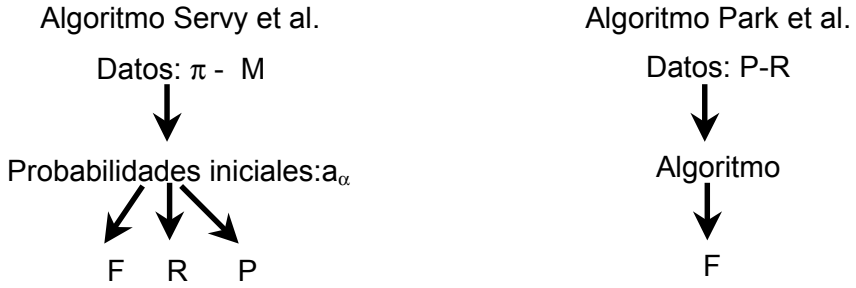
El enfoque de Servy et al. fija el valor de la probabilidad marginal de respuesta $\alpha = 1$, π_1, y especifica la matriz de transición M. A partir de estos valores, se determina el vector de probabilidades iniciales resolviendo un sistema de ecuaciones. Luego es posible determinar las probabilidades de los perfiles de respuesta en cada conglomerado o vector de orden k -secuencias de 0 y 1's-, las probabilidades marginales de respuesta 1 en la primera, segunda y k-ésima posición y las correlaciones entre las respuestas en pares de posiciones diferentes.

El método de Park et al parte de fijar las probabilidades de respuesta 1 en cada posición del conglomerado (p_i = E(z_i=1), i=1...k) y las correlaciones de respuesta entre pares de posiciones (ρ_{ij}). A partir de ciertas relaciones y en base al algoritmo ya presentado se generan variables Poisson correlacionadas. Luego se definen variables Z_i i=1,..,k indicadoras que cumplen con las especificaciones iniciales y la distribución de

probabilidades asociadas a los distintos perfiles de respuesta posibles en los conglomerados.

El siguiente esquema, considerando k=3, simplifica esta descripción.

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



Para comparar ambos algoritmos se debe por lo tanto determinar en primer lugar qué valores asignar a los distintos parámetros a fin de generar datos con las mismas características iniciales. Para ello se siguieron los siguientes pasos:

i) Aplicar el algoritmo Servy et al.

ii) Elegir los 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.

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. A tal fin se decidió:

i) Simular muestras de tamaño n=30, 50, 70 y 100 con ambos algoritmos y calcular las estimaciones de P y R. ($\hat{P}$ y $\hat{R}$).

ii) Repetir el procedimiento 1000 veces y calcular el promedio y desvío estándar de las estimaciones de P y R.

iii) Comparar los resultados obtenidos para los dos algoritmos.

6. RESULTADOS

Para la comparación de los algoritmos elegidos, en una primera instancia, se eligieron dos escenarios paramétricos. Dichos escenarios corresponden a valores de $P=(p_1,p_2,p_3)$ y $R=(\rho_{12};\rho_{13};\rho_{23})$ obtenidos aplicando el algoritmo de Servy et al para diferentes valores de la probabilidad de respuesta 1, π_1, y matriz de transición M según el siguiente detalle (Tabla 1).

Tabla1: Parámetros que determinan los escenarios paramétricos

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)

Las tablas siguientes presentan la estimación promedio de $\hat{P}$ y $\hat{R}$, junto a su desvío estándar, obtenidas por ambos algoritmos a través de 1000 muestras de tamaño n=30, 50, 70 y 100 generadas a partir de los escenarios descritos en la tabla anterior.

Tabla 2a: Promedios y desvíos estándares 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)

Tabla 2b: Promedios y desvíos estándares 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)

Los resultados hallados evidencian que las estimaciones tanto de P como de R resultan muy buenas en ambos algoritmos.

7. DISCUSIÓN

En este trabajo se evalúan dos algoritmos de generación de datos binarios correlacionados de diferentes características. En el procedimiento de Servy et al, cada elemento del conglomerado se genera de acuerdo al algoritmo y el número de individuos dentro del conglomerado (k) puede variar sin que se alteren los restantes parámetros del modelo. En cambio en el procedimiento de Park et al, el número de individuos en los conglomerados determina el número de variables Poisson a generar mediante el algoritmo y determina también el número de parámetros iniciales. Por lo tanto ambos algoritmos presentan diferente grado de flexibilidad.

En cuanto a 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 características de los mismos.

Bibliografía

BAHADUR, R. R. "A Representation of the Joint Distribution of Responses to n Dichotomous Items", in Studies in Item Analysis and Prediction (Stanford Mathematical Studies in the Social Sciences VI), ed. H. Solomon, Stanford, CA: Stanford University Press. 1961.

EMRICH, L. J. Y PIEDMONTE, M. R. "A Method for Generating High-Dimensional Multivariate Binary Variables," *The American Statistician*, 49, 302-304. 1991.

HACHUEL, L. "Estadísticas tipo score para modelos logit con diseños muestrales complejos". Tesis para acceder al grado de Magister en Estadística Aplicada. Universidad Nacional de Córdoba. 2001.

LEE, A. J. "Generating random binary deviates having fixed marginal distributions and specified degrees of association". *The American Statistician* 47, 209-215, 1993.

LIANG, K. Y.; ZEGER, S. L. "Longitudinal data analysis using generalized linear models". *Biometrika* 73, 13-22, 1986.

PARK C. G.; PARK T.; SHIN, D. W. "A Simple Method for Generating Correlated Binary Variates". *The American Statistician* 50, 306-310, 1996.

ROTNITZKY, A.; JEWELL, N. "Hipothesis testing of regression in semiparametric generalized linear models for cluster correlated data". *Biometrika* 77, 485-497, 1990.

SERVY, E.; HACHUEL, L.; WOJDYLA, D. "Análisis de tablas de contingencia para muestras de diseño complejo". *Cuadernos IITAE*. Escuela de Estadística. UNR, 1998.

SERVY, E.; HACHUEL, L.; WOJDYLA, D. "A simulation study for analizing the perfomance of tests of independence under cluster sampling". *Bulletin of the International Statistical Institute.51st Session Istanbul*. Book 2: 411, 1997.

ZEGER, S. L.; LIANG, K. Y.; ALBERT, P. S. “Models for Longitudinal Data: A Generalized Estimating Equation Approach”. *Biometrics* 44, 1049-1060, 1988.