Hachuel, Leticia
Boggio, Gabriela
Wojdyla, Daniel
Cuesta, Cristina
Servy, Elsa
Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística.
Consejo de Investigación de la Universidad Nacional de Rosario.

ESTUDIO DEL COMPORTAMIENTO DE ESTADÍSTICAS PARA DATOS BINARIOS CORRELACIONADOS EN MUESTRAS PEQUEÑAS

1. INTRODUCCIÓN

Frecuentemente el diseño de las investigaciones involucra múltiples mediciones de una variable respuesta. Por ejemplo, en los estudios longitudinales 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 es la falta de independencia entre las observaciones.

Dentro de este contexto, el estudio de datos binarios correlacionados es un área de interés creciente y un análisis frecuente consiste en describir la relación entre una respuesta dicotómica y covariables a través de la aplicación de métodos de regresión.

Prentice (1988) proporcionó una revisión exhaustiva de los métodos desarrollados para el análisis de regresión de datos binarios correlacionados, donde se destaca el enfoque de la ecuación de estimación generalizada (GEE) sugerido por Liang y Zeger (1986) y Zeger, Liang y Albert (1988). 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) derivaron estadísticas de tipo Wald y score generalizadas 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 ponen en duda las propiedades asintóticas es a través de estudios por simulación. Es importante señalar que las conclusiones de este tipo de estudios se reafirman en la medida que resulten consistentes a través de diferentes modelos de generación de datos. Por tal razón, este trabajo tiene por objeto la comparación de algoritmos de generación de datos binarios correlacionados y la evaluación del comportamiento de estadísticas que consideran la falta de independencia de las observaciones mediante estudios de Montecarlo.

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

2. ALGORITMOS DE GENERACIÓN DE DATOS BINARIOS CORRELACIONADOS

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.

La asignación de diferentes valores a los parámetros de estos modelos dan lugar a 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 modelos de generación de datos ya que las conclusiones de los estudios empíricos toman fuerza cuando son consistentes a través de datos obtenidos por ellos.

Existen diferentes enfoques de generación de variables binarias correlacionadas entre los que se pueden mencionar el de Bahadur (1961), Emrich y Piedmonte (1991) y Lee (1993). Los algoritmos elegidos para este trabajo son el de Park, Park y Shin (1996)y el de Servy, Hachuel y Wojdyla (1997, 1998), los cuales parten de la especificación de diferentes parámetros que determinan las características de los datos binarios a generar.

El modelo de simulación, presentado en Servy et al. (1996, 1997), fue diseñado originalmente 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. El modelo fija inicialmente el valor de la probabilidad de respuesta igual a 1, π, el tamaño k del conglomerado y especifica la matriz de transición M de la cadena de Markov. 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 diferentes posibles perfiles de respuesta en cada conglomerado, (0,0,...,0); (0,0,...,1);....;(1,1,...,1), las probabilidades marginales de respuesta 1 en la primera, segunda y k-ésima posición $(\pi_1,...,\pi_k)$ y las correlaciones entre las respuestas en pares de posiciones diferentes, ρ_{vj}.

El método de Park et al. (1996) parte de fijar las probabilidades de respuesta 1 en cada posición del conglomerado, $(\pi_1,...,\pi_k)$, y las correlaciones de respuesta entre pares de posiciones, ρ_{vj}. A partir de ciertas relaciones y en base a un algoritmo se generan variables Poisson correlacionadas. Luego se definen variables binarias y_v v=1,.., k que cumplen con las especificaciones iniciales y se obtiene la distribución de probabilidades asociadas a los distintos perfiles de respuesta posibles en los conglomerados.

Para comparar ambos algoritmos se debe determinar en primer lugar qué valores asignar a los distintos parámetros a fin de generar datos con las mismas características. Para ello y siendo $P = (\pi_1, \pi_2, \pi_3)$, $F = (P_{000}, P_{001}, P_{010}, P_{100}, P_{011}, P_{101}, P_{110}, P_{111})$ y $R = (\rho_{12}, \rho_{13}, \rho_{23})$ 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 elige un conjunto de escenarios con distintos valores de (π_1, π_2, π_3) y diferentes esquemas de

correlación desde independencia hasta alta correlación. En ellos, 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=15, 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**.**

3. ESTADÍSTICAS PARA DATOS CORRELACIONADOS

Liang y Zeger (1986) presentaron una generalización de las ecuaciones de estimación para el análisis de medidas repetidas. En ella se tiene en cuenta la correlación intragupo en la estimación de los parámetros de regresión de un modelo para la esperanza marginal y se supone que la distribución marginal de la variable pertenece a la familia exponencial. Dichas ecuaciones se conocen como Ecuaciones de Estimación Generalizadas (GEE).

A partir de los resultados sobre propiedades y distribución asintótica de los estimadores obtenidos por GEE, Rotniztky y Jewell (1990) derivan generalizaciones a los clásicos tests chi-cuadrado para probar hipótesis sobre los parámetros de regresión. En particular, definen el test de score generalizado y el test de Wald generalizado.

Se supone para ello, disponer de una muestra de n conglomerados de tamaño variable k_m (m=1,..,n) para comprobar hipótesis del tipo Ho: $\beta_2 = \beta_{20}$, cuando se considera una partición del vector de dimensión px1 de coeficientes de regresión $\beta' = (\beta_1', \beta_2')$ siendo β_1 un vector (p-q)x1 que contiene las (p-q) primeras componentes de β y β_2 un vector de dimensión qx1.

Sea el estimador obtenido por GEE, $\hat{\beta}_G$, la solución de:

$$S_G = \sum_{m=1}^{n} \left(\frac{\partial \mu_m}{\partial \beta} \right)' V_m^{-1} (Y_m - \mu_m) = 0 \qquad (3.1)$$

donde $V_m = A_m^{1/2} R_m(\alpha) A_m^{1/2}$ con $A_m = \text{diag}\{\text{var}(y_{mv})\}$ m=1,..,n ; v=1,..,k_m. Bajo condiciones débiles de regularidad, Liang y Zeger (1986) demuestran que $n^{1/2}(\hat{\beta}_G - \beta)$ tiene distribución asintótica normal con media cero y variancia asintótica dada por:

$$V_\beta = W_\beta \Omega W_\beta, \qquad (3.2)$$

donde:

$W_\beta = n \left(\sum_{m=1}^{n} D_m' V_m^{-1} D_m \right)^{-1}$, siendo $D_m = \frac{\partial \mu_m}{\partial \beta}$ y

$\Omega = \sum_{m=1}^{n} D_m' V_m^{-1} \text{cov}(Y_m) V_m^{-1} D_m$, siendo $\text{cov}(Y_m)$ la verdadera matriz de covariancia de Y_m, la cual se estima por $\{Y_m - \mu_m(\hat{\beta}_G)\}\{Y_m - \mu_m(\hat{\beta}_G)\}'$.

A partir de estos resultados Rotnitzky y Jewell (1990) definen el test de score generalizado de la siguiente manera:

$$X^2_{SG} = n^{-1}\tilde{S}'_{G_2}\tilde{\Sigma}_2^{-1}\tilde{S}_{G_2}, \tag{3.3}$$

siendo:
$$S_{G_2} = \sum_{m=1}^{n}\left(\frac{\partial \mu_m}{\partial \beta_2}\right)' V_m^{-1}(Y_m - \mu_m), \tag{3.4}$$
$$\Sigma_2 = W_{\beta_2}^{-1} V_{\beta_2} W_{\beta_2}^{-1}$$

V_{β_2} es la sub-matriz principal de dimensión qxq de la matriz de covariancias del vector de estimadores de β, V_β, y W_{β_2} es la sub-matriz principal de dimensión qxq de W_β.

En (3.3), las expresiones están valorizadas en $\tilde{\beta} = (\tilde{\beta}_1, \beta_{20})$, donde $\tilde{\beta}_1$ es solución de:

$\tilde{S}_{G1} = S_{G1}(\tilde{\beta}) = 0$; es decir, $\tilde{S}_{G2} = S_{G2}(\tilde{\beta})$ y $\tilde{\Sigma}_2$ es el estimador de Σ_2 valorizado en $\tilde{\beta}$.

La estadística X^2_{SG} tiene una distribución asintótica chi-cuadrado con q grados de libertad bajo condiciones débiles de regularidad y suponiendo una correcta especificación del modelo para la esperanza marginal.

Los autores definen además el test de Wald generalizado que para la misma hipótesis resulta:

$$X^2_{WG} = (\hat{\beta}_{G_2} - \beta_{20})' \hat{V}_{\beta_2}^{-1} (\hat{\beta}_{G_2} - \beta_{20}), \tag{3.5}$$

donde $\hat{\beta}_{G_2}$ es el vector que contiene las últimas q componentes del vector de estimadores $\hat{\beta}_G$ y $\hat{V}_{\beta_2}$ es la submatriz qxq de $\hat{V}_\beta$ correspondiente a las covariancias de $\hat{\beta}_{G_2}$.

Esta estadística se distribuye bajo Ho con distribución chi-cuadrado con q grados de libertad.

Distintos autores están investigando realizar ajustes a la estadística de Wald para tener en cuenta el número de grupos o conglomerados con el fin de producir estadísticas con mejores propiedades para tamaños de muestras moderadas. Shah, Holt y Folsom (1977) presentaron la siguiente modificación de la estadística de Wald, X^2_{WG}, basada en una transformación semejante a la T^2 de Hotelling:

$$X^2_{WG_C} = \frac{n-c}{nc} X^2_{WG}, \tag{3.6}$$

donde n es la cantidad de conglomerados y c son los grados de libertad del contraste. Esta estadística se distribuye según una Fde Snedecor con c y (n-c) grados de libertad.

En este trabajo se estudia el comportamiento de las estadísticas recién definidas - X^2_{SG}, X^2_{WG} y $X^2_{WG_C}$ - para el caso particular de poner a prueba la hipótesis de nulidad del parámetro de regresión en un modelo logit con una única variable explicativa dicotómica aplicado a datos binarios correlacionados. El modelo logit:

$$\ln\frac{\pi}{1-\pi} = \beta_1 + \beta_2 X \quad ; X = 0,1, \tag{3.7}$$

se ajusta a datos generados por el algoritmo de Servy et al. bajo los diferentes escenarios seleccionados según lo explicitado en la sección 2. Se simulan muestras simples aleatorias de n= 10, 15, 30, 50, 100 y 200 conglomerados de tamaño fijo k=3 y se observan todos los individuos dentro de los mismos. En cada muestra se calculan las estadísticas ya definidas

utilizando tres especificaciones diferentes para la matriz de correlación de trabajo $R_m(\alpha)$: la de independencia, AR(1) por ser una buena aproximación al esquema de dependencia que presentan los datos generados y la real, es decir conformada con los valores de $\{\rho_{vj}\}$ utilizados en el proceso de generación de datos.

Para el estudio del control del error de tipo I del test Ho: $\beta_2 = 0$, las muestras se generan bajo el mismo escenario para X=0 y X=1 con asignación aproximadamente balanceada en ambas categorías de X. En cada muestra se calculan las estadísticas presentadas y se decide el rechazo o no de la Ho a un nivel del 5%. Se repite este procedimiento mil veces para cada tamaño de muestra y se calcula el porcentaje de rechazo real.

4. RESULTADOS

4.1. Comparación de algoritmos

Para comprobar que los algoritmos elegidos generan datos binarios con las mismas características se eligieron tres escenarios paramétricos de acuerdo al procedimiento descripto en la Sección 2, a fin de cubrir situaciones de alta, media y baja correlación intra-grupo.

Los resultados hallados muestran que para cualquier valor de la probabilidad de respuesta, (π_1, π_2, π_3) y de la correlación entre pares de unidades intragrupo -bajo, mediano o alto- ambos algoritmos producen, en promedio, buenas estimaciones de dichos parámetros cualquiera sea el tamaño de la muestra (Tablas 1, 2 y 3).

Tabla 1: Promedios y desvíos estándares de las estimaciones de las componentes de P y R para el Escenario 1

n	Park et al.			Servy et al.		
	π_1=0.50	π_2=0.45	π_3=0.45	π_1=0.50	π_2=0.45	π_3=0.45
15	0.51 (0.13)	0.45 (0.13)	0.45 (0.13)	0.50(0.13)	0.45(0.13)	0.44(0.13)
30	0.50 (0.09)	0.45 (0.09)	0.44 (0.09)	0.50(0.10)	0.45(0.09)	0.44(0.09)
50	0.50 (0.07)	0.45 (0.07)	0.44 (0.07)	0.50(0.07)	0.45(0.07)	0.44(0.07)
70	0.50 (0.06)	0.45 (0.06)	0.45 (0.06)	0.49(0.06)	0.45 (0.06)	0.45 (0.06)
100	0.50 (0.05)	0.45 (0.05)	0.45 (0.05)	0.50 (0.05)	0.45 (0.05)	0.45 (0.05)
n	ρ_{12}=0.10	ρ_{13}=0.01	ρ_{23}=0.10	ρ_{12} =0.10	ρ_{13}=0.01	ρ_{23}=0.10
15	0.11 (0.27)	0.00 (0.27)	0.11 (0.26)	0.09 (0.26)	0.01 (0.27)	0.10 (0.27)
30	0.09 (0.18)	0.00 (0.19)	0.10 (0.19)	0.10 (0.18)	0.02 (0.19)	0.10 (0.18)
50	0.10 (0.14)	0.01 (0.14)	0.10 (0.14)	0.10 (0.14)	0.01 (0.14)	0.10 (0.14)
70	0.10 (0.12)	0.01 (0.13)	0.10 (0.12)	0.10 (0.12)	0.01 (0.12)	0.09 (0.12)
100	0.10 (0.09)	0.01 (0.10)	0.10 (0.10)	0.10 (0.10)	0.01 (0.10)	0.10 (0.10)

Tabla 2: Promedios y desvíos estándares de las estimaciones de las componentes de P y R para el Escenario 2

n	Park et al.			Servy et al.		
	π_1=0.96	π_2=0.78	π_3=0.67	π_1=0.96	π_2=0.78	π_3=0.67
15	0.96 (0.05)	0.78 (0.11)	0.67 (0.12)	0.96 (0.05)	0.77(0.11)	0.67(0.12)
30	0.96 (0.04)	0.78 (0.07)	0.67 (0.08)	0.96 (0.04)	0.78(0.07)	0.67(0.08)
50	0.96 (0.03)	0.78 (0.06)	0.67 (0.07)	0.96 (0.03)	0.78(0.06)	0.67(0.07)
70	0.96 (0.02)	0.78 (0.05)	0.67 (0.06)	0.96 (0.02)	0.78(0.05)	0.67(0.06)
100	0.96 (0.02)	0.78 (0.04)	0.67 (0.05)	0.96 (0.02)	0.78(0.04)	0.66(0.05)
n	ρ_{12}=0.28	ρ_{13}=0.15	ρ_{23}=0.53	ρ_{12} =0.28	ρ_{13}=0.15	ρ_{23}=0.53
15	0.40 (0.27)	0.21 (0.26)	0.52 (0.25)	0.41 (0.28)	0.22 (0.28)	0.54 (0.25)
30	0.33 (0.19)	0.18 (0.19)	0.53 (0.17)	0.33 (0.21)	0.16 (0.20)	0.53 (0.17)
50	0.30 (0.15)	0.15 (0.15)	0.53 (0.13)	0.30 (0.16)	0.16 (0.14)	0.53 (0.13)
70	0.29 (0.13)	0.16 (0.13)	0.53 (0.11)	0.28 (0.14)	0.15 (0.13)	0.53 (0.11)
100	0.33 (0.12)	0.15 (0.11)	0.53 (0.09)	0.28 (0.11)	0.15 (0.10)	0.53 (0.09)

Tabla 3: Promedios y desvíos estándares de las estimaciones de las componentes de P y R para el Escenario 3

n	Park et al.			Servy et al.		
	π_1=0.25	π_2=0.30	π_3=0.34	π_1=0.25	π_2=0.30	π_3=0.34
15	0.25 (0.11)	0.31 (0.12)	0.34 (0.12)	0.25 (0.12)	0.31 (0.12)	0.35 (0.12)
30	0.25 (0.08)	0.30 (0.09)	0.34 (0.09)	0.25 (0.08)	0.31 (0.08)	0.35 (0.09)
50	0.26 (0.06)	0.30 (0.07)	0.34 (0.07)	0.25 (0.06)	0.30 (0.07)	0.34 (0.07)
70	0.25 (0.05)	0.30 (0.06)	0.34 (0.06)	0.25 (0.05)	0.30 (0.06)	0.34 (0.06)
100	0.25 (0.04)	0.30 (0.05)	0.34 (0.05)	0.25 (0.04)	0.30 (0.05)	0.34 (0.05)
n	ρ_{12}=0.76	ρ_{13}=0.59	ρ_{23}=0.77	ρ_{12} =0.76	ρ_{13}=0.59	ρ_{23}=0.77
15	0.76 (0.19)	0.60 (0.23)	0.78 (0.18)	0.75 (0.20)	0.57 (0.23)	0.77 (0.18)
30	0.76 (0.13)	0.58 (0.16)	0.77 (0.13)	0.75 (0.14)	0.58 (0.16)	0.78 (0.12)
50	0.76 (0.10)	0.59 (0.12)	0.77 (0.09)	0.76 (0.11)	0.59 (0.12)	0.77 (0.09)
70	0.76 (0.09)	0.59 (0.10)	0.77 (0.08)	0.76 (0.09)	0.57 (0.10)	0.77 (0.08)
100	0.76 (0.07)	0.59 (0.08)	0.78 (0.07)	0.76 (0.07)	0.57 (0.09)	0.77 (0.07)

De la observación de las tablas se desprende que ambos algoritmos estiman bien las probabilidades de respuesta igual a 1 en las distintas posiciones, π_1, π_2, π_3 , y de forma más precisa a medida que el tamaño de la muestra aumenta. En cuanto a los coeficientes de correlación de a pares, también la estimación es buena pero sólo para correlaciones altas la estimación resulta precisa. Estos resultados son válidos para cualquiera de los dos algoritmos. Esta consistencia de los resultados encontrados habilita la comparación del comportamiento de las estadísticas bajo muestras generadas por uno de los algoritmos ya que a priori no puede atribuirse alguna tendencia al algoritmo empleado. El algoritmo elegido para dicha instancia es el de Servy et al.

4.2. Evaluación de estadísticas

Los siguientes gráficos muestran los resultados hallados para los niveles de significación reales de las estadísticas de score generalizada, Wald generalizada y su

correspondiente corrección, aplicados a datos generados por el algoritmo de Servy et al. en los tres escenarios seleccionados (Figura 1, página 8).

En particular, para cada estadística a través de los distintos escenarios de dependencia alta, media y baja, no se observan claras diferencias según haya sido la especificación de la matriz de correlación de trabajo. Sin embargo, puede apreciarse una convergencia mayor hacia el nivel de significación nominal del 5% al aumentar el tamaño de la muestra, cuando la matriz de correlación de trabajo está fija, en menor grado cuando está aproximada (AR(1)), y con una amplitud mayor en el intervalo de convergencia cuando es la de independencia.

En cuanto a los resultados encontrados para cada estadística para muestras chicas se destaca:

- Comportamiento conservador de la estadística X^2_{SG} , el cual se acentúa cuando la intensidad de la correlación entre pares dentro de los grupos es alta.
- Comportamiento liberal de la estadística X^2_{WG} para correlaciones intra-grupo bajas y medias y sin embargo, un tanto conservador para correlaciones altas.
- Comportamiento liberal, atenuado respecto de la anterior, de la estadística $X^2_{WG_C}$, y por lo tanto más conservador ante correlaciones altas, sobre todo para matriz de correlación de trabajo de independencia o AR(1).

5. DISCUSIÓN

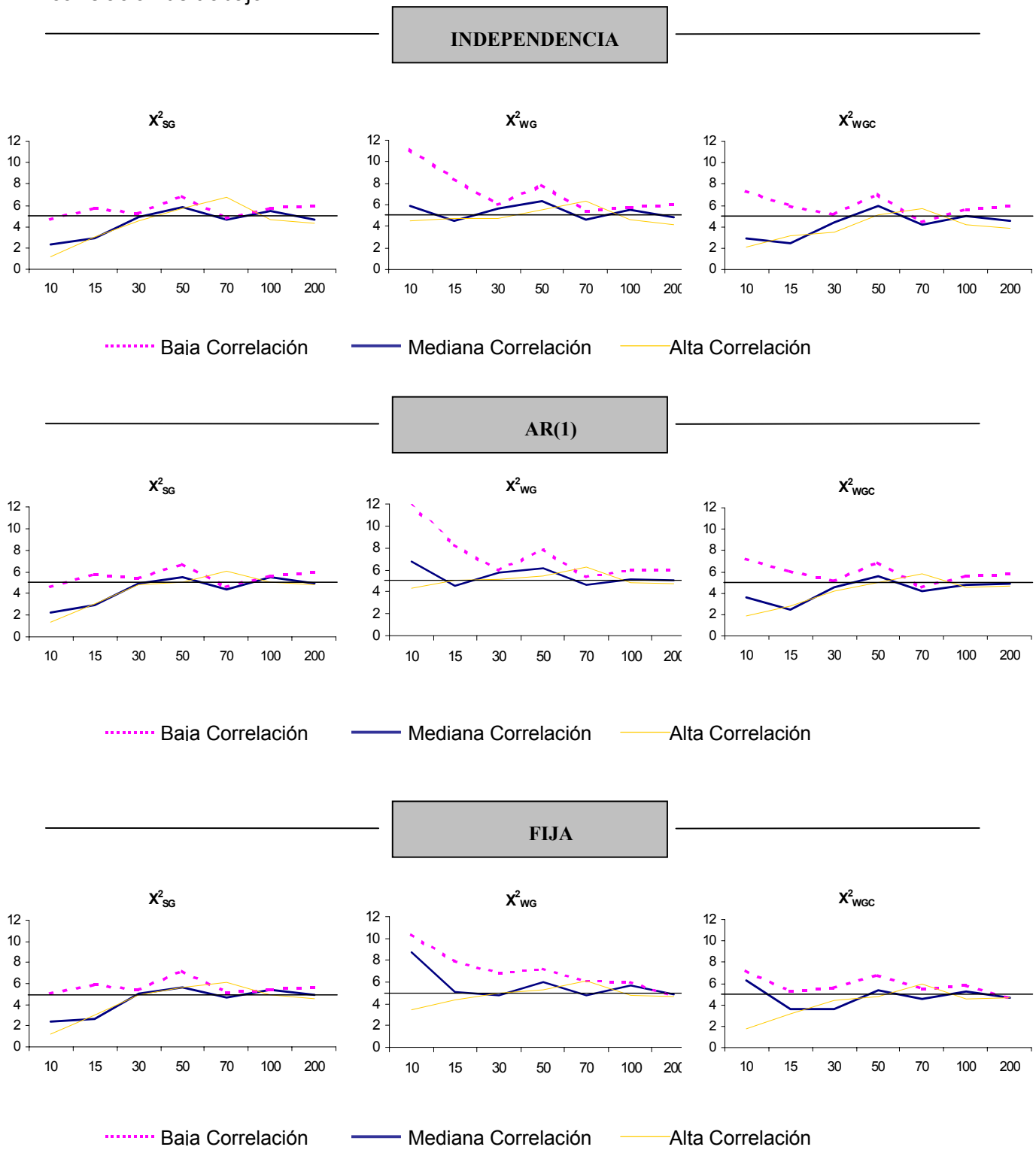
Los resultados obtenidos acerca de los parámetros que caracterizan a los escenarios a partir de los dos algoritmos fueron altamente compatibles. Este hecho motivó la decisión de evaluar, en primera instancia, el comportamiento de las estadísticas a partir de datos generados por el algoritmo de Servy et al.

Respecto a la evaluación de las estadísticas, se confirmó la liberalidad supuesta para la estadística de Wald sobre todo en situaciones de baja correlación. Llamativamente, a medida que la intensidad de la correlación intragrupo es mayor, las tres estadísticas comparadas se tornan conservadoras. De acuedo a la sugerencia generalizada de utilizar estadísticas de comportamiento conservador cuando las muestras son pequeñas, se desprende de este trabajo la recomendación de usar las estadísticas de score generalizado o de Wald corregida, para muestras menores que 30, pero teniendo en cuenta que si la correlación es alta puede extremarse este carácter conservador de las mismas.

Debe tenerse en cuenta que los resultados alcanzados corresponden a una hipótesis particular para un modelo muy simple, por lo que correspondería complejizar el modelo a fin de generalizar las conclusiones.También se considera oportuno repetir la evaluación de las estadísticas con datos generados por el algoritmo de Park et al.

Cabe mencionar que al momento de esta presentación se está ejecutando el estudio de la potencia de los procedimientos evaluados.

Figura 1. Comportamiento de las estadísticas según tamaño de la muestra, de acuerdo a la intensidad de la correlación intragrupo y a las diferentes especificaciones de la matriz de correlación de trabajo.



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.; PIEDMONTE, M. R. “A Method for Generating High-Dimensional Multivariate Binary Variables,” The American Statistician, 49, 302-304. 1991.

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.

PRENTICE, R. L."Correlated binary regression with covariates specific to each binary observation". Biometrics, 44, 1033-1048, 1988 .

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.

SHAH, B. V.; HOLT, M. M.; FOLSOM, R. E. “Inference about regresion models from sample survey data”. Bulletin of the International Statistical Institute, 47 43-57, 1977.

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