Revista Colombiana de Estadística

Volumen 33. Número 2 - diciembre - 2010 **ISSN 0120 - 1751**

Departamento de Estadística
Universidad Nacional de Colombia
Bogotá - Colombia

Revista Colombiana de Estadística

http://www.estadistica.unal.edu.co/revista
http://www.matematicas.unal.edu.co/revcoles
http://www.emis.de/journals/RCE/
revcoles_fcbog@unal.edu.co

Indexada en: Scopus, Science Citation Index Expanded (SCIE), Web of Science (WoS), SciELO Colombia, Current Index to Statistics, Mathematical Reviews (MathSci), Zentralblatt Für Mathematik, Redalyc, Latindex, Publindex (A_1)

La *Revista Colombiana de Estadística* es una publicación semestral del Departamento de Estadística de la Universidad Nacional de Colombia, sede Bogotá, orientada a difundir conocimientos, resultados, aplicaciones e historia de la estadística. La Revista contempla también la publicación de trabajos sobre la enseñanza de la estadística.

Se invita a los editores de publicaciones periódicas similares a establecer convenios de canje o intercambio.

Dirección Postal:
Revista Colombiana de Estadística
© Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Carrera 30 No. 45-03
Bogotá – Colombia
Tel: 57-1-3165000 ext. 13231
Fax: 57-1-3165327

Adquisiciones:
Punto de venta, Facultad de Ciencias, Bogotá.
Suscripciones:
revcoles_fcbog@unal.edu.co
Solicitud de artículos:
Se pueden solicitar al Editor por correo físico o electrónico; los más recientes se pueden obtener en formato PDF desde la página Web.

Edición en LaTeX: Patricia Chávez R.
Impresión: proCEditor, Tel. 57-1-5602317, Bogotá.

Revista Colombiana de Estadística	Bogotá	Vol. 33	Nº 2
ISSN 0120 - 1751	COLOMBIA	diciembre-2010	Págs. 167-339

Contenido

Editorial

Beatriz Piedad Urdinola[a]

Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Naciones Unidas, a través de su Consejo Económico y Social, decretó el pasado 20 de octubre de 2010 como el Día Mundial de la Estadística. Esta celebración busca "concientizar sobre las muchas aportaciones de las estadísticas oficiales basadas en los valores básicos de servicio, integridad y profesionalidad". Y, entre otras razones, se escogió el presente año para inaugurar esta celebración, por ser un año censal en varios de los países miembros de Naciones Unidas. Es decir, durante este año se realiza el operativo oficial más extenso y dispendioso al que puede enfrentarse un estadístico, pero también con mayores resultados y gratificaciones, como son los censos nacionales de población.

Esta editorial está dedicada al espíritu de esta celebración, en la medida en que exalta la importancia de la producción estadística oficial y los valores éticos que se requieren en la recolección, elaboración y análisis de estas cifras. En particular, los operativos censales sirven de ejemplo tanto en la complejidad que demanda el quehacer estadístico, desde las oficinas oficiales de producción estadística, como en la integridad y el profesionalismo que se requieren y que cubren muchos más ámbitos que los de la recolección y presentación de cifras oficiales.

Los censos nacionales de población representan uno de los retos más interesantes para cualquier oficina de estadísticas oficiales. Desde el momento de su preparación logística, financiera y técnica hasta la presentación de resultados oficiales y estudios poscensales, se requieren años de preparación en todas y cada una de estas áreas, si se quieren producir datos de calidad, así como altísimos niveles de compromiso y profesionalismo, que no pueden faltar en ninguno de los eslabones de su producción. Son muchos más los ejemplos de censos exitosos, que de aquellos infructíferos, y para exaltar la labor de los estadísticos y otros profesionales vinculados a estos procesos, Naciones Unidas ha decretado esta fecha de celebración.

Sin embargo, son también varios los censos en diferentes latitudes que han fallado en su integridad y profesionalismo. Seltzer (2005) nos presenta una triste revisión de casos de mediciones censales dejados en manos perversas, quizás en muchos casos por la falta de ética y profesionalismo de algunos colegas, que han violado derechos humanos. Basta con mencionar el reconocido caso del censo de Estados Unidos, durante las guerras mundiales, cuya oficina oficial de censos violó

[a]Editora de la Revista Colombiana de Estadística, Profesora asociada.
E-mail: bpurdinolac@bt.unal.edu.co

la confidencialidad estadística y filtró información censal que generó olas de desplazamiento forzado a los japoneses en tierras americanas. Casos similares se han presentado en China, países nórdicos, Ruanda y Francia, entre otros.

Colombia -satisfactoriamente, y hasta el momento- no ha reportado casos de este tipo, y tengo confianza en que el DANE seguirá manteniendo la calidad excepcional en el manejo de la confidencialidad estadística en sus censos y encuestas de hogares, en una sociedad con los problemas sociales, económicos y políticos como la colombiana, que así lo requiere. Quiero, entonces, destacar y felicitar a todos los estadísticos y profesionales que comparten el oficio de producción de estadísticas oficiales en Colombia que han logrado con su esfuerzo llevar a excelente término esta dispendiosa labor de la producción de estadísticas oficiales, y ojalá vengan muchos más años caracterizados por el profesionalismo y la integridad ética en esta importante labor.

Referencias

Seltzer, W. (2005), 'On the use of Population Data Systems to Target Vulnerable Population Subgroups for Human Rights Abuses', *Coyuntura Social* (32). Fedesarrollo, Bogotá.

Revista Colombiana de Estadística
Diciembre 2010, volumen 33, no. 2, pp. 167 a 190

Verificación y monitoreo de la aleatoriedad de los juegos de números de d dígitos

Testing and Monitoring the Randomness of d Digit Number Game

Humberto Gutiérrez-Pulido[a], Juan Manuel García[b]

Departamento de Matemáticas, CUCEI, Universidad de Guadalajara, Guadalajara, México

Resumen

El interés de este trabajo se centra en el problema de probar la aleatoriedad de los resultados de los juegos de números de d dígitos. Es usual que este problema se aborde con pruebas aproximadas del tipo χ^2 y otras pruebas de independencia de resultados sucesivos. Pero estas pruebas, tienen entre otras limitantes, el hecho de que requieren muestras grandes. Como alternativa, en este trabajo se detalla una prueba bayesiana basada en el modelo multinomial. Además, para monitorear los resultados de este tipo de juego de azar y detectar en forma oportuna patrones y resultados no aleatorios, se propone la utilización de una carta de control geométrica. Se hace un breve estudio Monte Carlo para comprender mejor las características de la carta propuesta. Como caso práctico se analizan los resultados de 500 sorteos de la lotería mexicana Tris y se detectan problemas de falta de aleatoriedad, tanto con la prueba bayesiana como con la carta de control.

Palabras clave: carta de control, distribución geométrica, distribución multinomial, distribución Dirichlet, lotería, prueba ji-cuadrada.

Abstract

this work is centered on testing the randomness of d-digit number game. It is usual that this problem is studied by the χ^2 test and other tests for independence of successive draws. However, these tests require of large sample sizes. As an alternative, it is proposed a Bayesian methodology based on the multinomial model. This methodology does not depend on asymptotic results. Besides, for monitoring the results of this type of game, it is proposed a geometric control chart. Monte Carlo study is carried out to analyze this chart. As a practical case, the data of 500 draws of mexican lottery Tris were analyzed, and problems of lack of randomness are detected.

Key words: Control chart, Dirichlet distribution, Geometric distribution, Lottery Chi-square test, Multinomial distribution.

[a]Profesor titular. E-mail: humberto.gutierrez@cucei.udg.mx
[b]Profesor titular. E-mail: jmgarciat@prodigy.net.mx

1. Introducción

La fascinación y afición por los juegos y apuestas de azar tienen profundo arraigo en la historia de la humanidad, y están documentadas desde los primitivos juegos de dados en la antigüedad hasta los actuales juegos de póker y apuestas en línea (ver Schwartz 2003, 2006). El gran interés por los juegos de azar motivó, sobre todo a partir del siglo XVII, el desarrollo teórico del cálculo de probabilidades. En la actualidad, con el incremento del tiempo de ocio y la aparición de "casinos electrónicos", los juegos de azar son una práctica para mucha gente, lo que se debe reconocer como un riesgo, porque hay personas que alcanzan una adicción patológica a este tipo de juegos. A esta adicción se le conoce como ludopatía, y ha llevado a que en diferentes épocas y países, los juegos y apuestas de azar hayan sido prohibidos. Sin embargo, la tendencia ha sido a liberalizar y reglamentar lo que hoy sin duda es una industria mundial en la que participan los propios gobiernos, con la aparición de las loterías nacionales.

En este contexto, uno de los mayores retos es darle confianza y certeza al jugador y a la sociedad en cuanto a que los resultados de los sorteos son aleatorios. Hay dos formas básicas con las que se ha buscado dar tal confianza y certeza; la primera, controlando el proceso, y para ello se buscan mejores equipos para realizar los sorteos, mayor uniformidad entre las esferas utilizadas, elección aleatoria de urnas, sorteos en vivo, etc. La otra forma, es mostrar evidencias estadísticas de que los resultados se pueden considerar aleatorios. Sobre esto último, en el Reino Unido, la Comisión Nacional de la Lotería (National Lottery Commission, NLC) reguladora de la Lotería Nacional, tiene un programa de investigación para comprobar si hay elementos de no aleatoriedad en los diferentes juegos de la Lotería Nacional del Reino Unido (ver Royal Statistical Society 2000, 2002 y University of Salford 2004, 2005*a*, 2005*b*).

Adicionalmente, en la literatura especializada en estadística hay diferentes trabajos que abordan problemas de la aleatoriedad de los juegos de azar y aspectos relacionados. Unos enfocados a estudiar juegos de números de d dígitos, en los que el número ganador se determina por la extracción al azar de d esferas, usualmente cada una proveniente de una urna que contiene diez esferas numeradas con los dígitos: $0, 1, 2, \ldots, 9$. Otros trabajos se enfocan a los juegos de lotería del tipo k de N, en los que la combinación ganadora se determina por la extracción al azar de k esferas sin remplazo de una urna que contiene esferas similares numeradas del 1 al N. Sobre este tipo de juegos, en Joe (1987) se estudian las posibles distribuciones de un conjunto de k-tuplas, dadas las frecuencias marginales de los números del 1 al N, y se analizan datos de la lotería Lotto 6/49 de Canadá. En Stern & Cover (1989) se estima la distribución de máxima entropía de los boletos comprados en la lotería Lotto 6/49 de Canadá, buscando determinar números impopulares para el público. En Joe (1990) se cuestiona la estrategia de comprar un boleto con una k-tupla de números impopulares. Por su parte, Johnson & Klotz (1993) desarrollan métodos numéricos para calcular las estimaciones de máxima verosimilitud de las probabilidades para los diferentes números de una lotería, y dan el estadístico de razón de verosimilitud generalizado $-2\ln()$ para probar que los números son equiprobables; estos métodos los aplican a un conjunto de datos de

la lotería Lotto América. En Joe (1993) se obtiene la modificación del estadístico de prueba ji-cuadrada para probar la uniformidad de los números del 1 al N, dado que los números se extraen sin remplazo, y se analizan datos de la lotería Lotto 6/49 de Canadá. Finkelstein (1995) estima la distribución de frecuencia de los números a los que se les apuesta en la lotería de California, y prueba la hipótesis de que los números pequeños son más populares. En Genest, Lockhart & Stephens (2002) se discute el uso del estadístico χ^2 para bondad de ajuste, para probar la equiprobabilidad de la ocurrencia de números de una lotería, el cual no sigue una distribución χ^2 simple porque los números se extraen sin remplazo, y muestran que su distribución asintótica es la de una suma ponderada de variables aleatorias χ^2; este resultado lo utilizan para comprobar la uniformidad de números ganadores en la lotería Lotto 6/49 de Canadá. Koning & Vermaat (2002) calculan las probabilidades de ganar premios y el pago esperado en la Lotto holandesa; también prueban la hipótesis de que los números y colores se extraen aleatoriamente. Percy (2006) describe una metodología bayesiana para probar la aleatoriedad de los sorteos de lotería, y considera el caso de datos de la Lotería Nacional del Reino Unido. En Coronel-Brizio, Hernández-Montoya, Rapallo & Scalas (2008) se abordan también las loterías del tipo k/N mediante la obtención de la media y varianza de la variable aleatoria que representa a los números extraídos; también usan la distribución hipergeométrica, y sus resultados los aplican a la lotería mexicana Melate y a la lotería italiana. De esta revisión se desprende que los estudios estadísticos para los juegos de lotería del tipo k de N son variados; y algunos han propuesto metodologías para verificar la aleatoriedad de los resultados, y otros han estudiado la forma en que la gente elige los números a los que les apuestan.

En cuanto a los juegos de números de d dígitos, en Clotfelter & Cook (2008) se usan datos del juego de números de la lotería de Maryland para probar la hipótesis de que los jugadores de lotería creen que la probabilidad de un evento disminuye cuando ese evento ha ocurrido recientemente ("la falacia del jugador"), y lo comprueban mostrando cómo las apuestas por los números que han ganado recientemente disminuye en los sorteos siguientes. En Teo & Leong (2002) se estudia el diseño de un mecanismo de control para el manejo de las apuestas en un juego de números de cuatro dígitos, esto es, para decidir si las apuestas deben ser aceptadas o rechazadas por el operador que ofrece el juego. En términos generales, el tema de juegos de números de d dígitos ha recibido mucha menos atención en la literatura especializada; sin embargo el tipo de herramientas que se utilizan para probar la aleatoriedad de estos tipo de sorteos presentan algunos problemas. Por ejemplo, usualmente para verificar la igualdad de la frecuencia de los dígitos o grupos de ellos se aplican pruebas aproximadas, como la tradicional prueba χ^2, que entre otros inconvenientes requiere tamaños de muestra grandes para lograr buenas aproximaciones (Cai & Krishnamoorthy 2006); otro problema es que para probar la independencia se aplican usualmente pruebas que precisan el supuesto de normalidad.

El objetivo del presente artículo es proponer alternativas de solución a los dos problemas anteriores. Para el primero, en la sección 2 se propone utilizar el procedimiento bayesiano basado en la distribución multinomial que no requiere resultados asintóticos. Para el segundo problema, en la sección 3 se propone el uso

de una carta geométrica como alternativa para estar monitoreando en tiempo real la aleatoriedad de los resultados de un juego de d dígitos. Adicionalmente, en la sección 4 se presentan los resultados de un estudio de simulación mediante el cual se analizan las propiedades de la carta de control propuesta. En la sección 5 se aplican los procedimientos propuestos a un conjunto de resultados del sorteo Tris, un juego popular en México, promovido por una institución pública. Mediante ambos procedimientos se detectan problemas significativos en la aleatoriedad de dicho sorteo. Por último, en la sección 6 se hace la discusión y se plantean las conclusiones del trabajo.

2. Igualdad de frecuencia de los dígitos

Una primera forma de probar la aleatoriedad de los resultados de los juegos de números de d dígitos es ver si hay igualdad de frecuencia de los diferentes dígitos, ya sea considerando cada urna por separado o todas a la vez. Esto se puede expresar de manera general, considerando n ensayos independientes (n sorteos), donde cada esfera tiene probabilidad p_i de ser extraída, con $p_i > 0$ y $\sum_{i=1}^{m} p_i = 1$, y si X_i es el número de veces que se extrae la esfera i, entonces el vector $X = (X_1, \ldots, X_m)$ sigue una distribución multinomial con parámetro n y $\mathbf{p}$, donde $\mathbf{p} = (p_1, \ldots, p_m)$. La densidad de probabilidades multinomial está dada por

$$f(x_1, \ldots, x_m; n, \mathbf{p}) = \frac{n!}{x_1! \cdots x_m!} \prod_{i=1}^{m} p_i^{x_i} \tag{1}$$

con $n = \sum_{i=1}^{m} x_i$. Es de interés probar la siguiente hipótesis

$$\begin{aligned} &H_0 : p_1 = p_{1,0}, \ldots, p_{m-1} = p_{m-1,0} \\ &H_a : p_i \neq p_{i,0} \text{ para algún } i = 1, 2, \ldots, m-1 \end{aligned} \tag{2}$$

donde en el caso del juego de números de d dígitos, $m = 10$ y $p_{i,0} = 0.1$.

2.1. Algunas pruebas basadas en la distribución ji-cuadrada

En Cai & Krishnamoorthy (2006) se analizan varias pruebas para (2) donde se usa de alguna forma la distribución χ^2. La primer prueba, que es la usual y que originalmente desarrolló Pearson en 1990, se basa en el estadístico dado por:

$$\chi_0^2 = \sum_{i=1}^{m} \frac{(X_i - np_{i,0})^2}{np_{i,0}} \tag{3}$$

Bajo H_0 este estadístico sigue en forma aproximada una distribución χ^2_{m-1} con $m-1$ grados de libertad (Cai & Krishnamoorthy 2006), por lo que la hipótesis nula en (2) se rechaza, con un nivel de significancia de α, si $\chi_0^2 \geq \chi^2_{\alpha,m-1}$.

Por su parte, la prueba de la razón de verosimilitud se basa en en el estadístico

$$R = 2 \sum_{i=1}^{m} X_i \ln \left(\frac{X_i}{np_{i,0}} \right) \tag{4}$$

que sigue en forma aproximada una distribución χ^2_{m-1} con $m-1$ grados de libertad (Cai & Krishnamoorthy 2006). Nass (1959) propuso otra aproximación a la distribución de (3), con

$$c\chi_0^2 \sim \chi_v^2 \tag{5}$$

donde $c = 2E(\chi_0^2)/Var(\chi_0^2)$ y $v = cE(\chi_0^2)$. Estas constantes se obtienen (ver Cai & Krishnamoorthy 2006), a partir de

$$E(\chi_0^2) = m-1 \quad \text{y} \quad Var(\chi_0^2) = 2(m-1) - (m^2+2m-2)/n + \sum\nolimits_{i=1}^{m} (np_{i,0})^{-1} \tag{6}$$

Una variante propuesta por Nass (1959), de especial interés en este trabajo, es cuando se tiene igual probabilidad en las celdas de la distribución multinomial, es decir, cuando $p_{i,0} = 1/m$ para todo i. En este caso se hace una corrección por continuidad para que el estadístico (3) tome la forma siguiente:

$$\chi_c^2 = \frac{\sum_{i=1}^{m} X_i^2 - 1}{(n/m)} - n \tag{7}$$

con lo que la varianza de χ_c^2 en (6) toma la forma $2(m-1)(n-1)/n$.

Una problemática de estas pruebas para validar la aleatoriedad de los juegos de azar de d dígitos es que son aproximadas, y requieren valores grandes en el número de sorteos, lo que restringe su aplicación. En Cai & Krishnamoorthy (2006) se muestra que las pruebas aproximadas pueden tener un inadecuado comportamiento para tamaños de muestra pequeños y para $p_{i,0}$ pequeños, que serían las situaciones de mayor interés en el caso de juegos de número de d dígitos, ya que para empezar $p_{i,0} = 0.1$.

Finalmente es importante agregar otras variantes sobre las que existen muchos trabajos en la literatura, que usan la aproximación χ^2 para probar la hipótesis en (2) por el método de intervalos de confianza. Estas variantes consisten en estimar intervalos de confianza simultáneos para las m proporciones p_i del modelo multinomial (ver por ejemplo Sison & Glaz 1995, Hou, Chiang & Tai 2003). El primer trabajo sobre el particular es el de Quesenberry & Hurst (1964), y una mejora de este, bastante fácil de implementar, es la propuesta por Goodman (1965), que señala que un intervalo simultáneo aproximado, con una confianza de $100(1-\alpha)$ para cada una de las p_i se obtiene con:

$$\frac{B + 2x_i \pm \{B\,[B + 4x_i(n-x_i)/n]\}^{1/2}}{2(n+B)} \tag{8}$$

donde B es el $100 \times (1-\alpha/m)$ percentil superior de la distribución χ^2 con un grado de libertad.

2.2. Alternativa bayesiana

Dado el modelo multinomial (1), la verosimilitud correspondiente está dada por

$$L(p_1, \ldots, p_m; n, X) \propto \prod_{i=1}^{m} p_i^{x_i} \tag{9}$$

A priori, para el modelo multinomial se puede utilizar su conjugada (Bernardo & Smith 1994), que es la distribución Dirichlet (m, α):

$$\pi(p_1, \ldots, p_m) = \frac{1}{B(\alpha)} \prod_{i=1}^{m} p_i^{\alpha_i - 1} \tag{10}$$

donde $p_i > 0, \sum_{i=1}^{m} p_i = 1$, $\alpha_i > 0$, $\alpha = (\alpha_1, \ldots, \alpha_m)$ y

$$B(\alpha) = \frac{\prod_{i=1}^{m} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{m} \alpha_i\right)}$$

Con $\alpha_i = 1$ para todo i, se logra una distribución a priori no informativa. Una propiedad interesante es que si $\mathbf{p}$ tiene una distribución Dirichlet (m, α), entonces la distribución marginal de p_i es una distribución Beta $(\alpha_i, \alpha_0 - \alpha_i)$, con $\alpha_0 = \sum_{i=1}^{m} \alpha_i$. Además la distribución marginal de cualquier sub-vector de una Dirichlet es también Dirichlet; por ejemplo la distribución marginal de $(p_i, p_j, 1 - p_i - p_j)$ es Dirichlet $(\alpha_i, \alpha_j, \alpha_0 - \alpha_i - \alpha_j)$.

De (9) y (10) se concluye que la distribución posterior está dada por

$$\begin{aligned} \pi(p_1, \ldots, p_m \mid X) &\propto \prod_{i=1}^{m} p_i^{x_i} \prod_{i=1}^{m} p_i^{\alpha_i - 1} \\ &\propto \prod_{i=1}^{m} p_i^{\alpha_i + x_i - 1} \end{aligned} \tag{11}$$

que, salvo por una constante, corresponde a una distribución Dirichlet (m, α^*), con $\alpha^* = (\alpha_1 + x_1, \ldots, \alpha_m + x_m)$. De esta manera, con base en (11) se puede hacer inferencia para $\mathbf{p} = (p_1, \ldots, p_m)$, lo que permite explorar la hipótesis (2) o el caso particular de interés en los juegos de número de d dígitos, donde $m = 10$ y $p_{i,0} = 0.1$. Para explorar la distribución posterior resulta conveniente obtenerla por simulación; por ejemplo en el software libre R, que se puede bajar de *http://cran.r-project.org* junto con el paquete MCMCPack, que al cargarlo puede generar la distribución posterior con la función *MCmultinomdirichlet* $\mathbf{x}$, α^*, mc, donde $\mathbf{x} = (\mathbf{x_1}, \ldots, \mathbf{x_m})$ y mc es el número de realizaciones que se quiere de $\mathbf{p} \mid \mathbf{X}$. También puede ser de utilidad la función *ddirichlet* $(\mathbf{p}, \alpha)$.

Alternativamente a la simulación se puede recurrir a las expresiones analíticas. Por ejemplo, de acuerdo con lo que se dijo antes, la distribución marginal posterior de cualquiera de las p_i es una Beta $(\alpha_i + x_i, \alpha_0^* - (\alpha_i + x_i))$, con $\alpha_0^* = \sum_{i=1}^{m} (\alpha_i + x_i)$. Así, para cualquier p_i se tiene que

$$\pi(p_i \mid X) = \frac{\Gamma(\alpha_0^*)}{\Gamma(\alpha_i + x_i)\,\Gamma(\alpha_0^* - (\alpha_i + x_i))} p_i^{\alpha_i + x_i - 1} (1 - p_i)^{\alpha_0^* - (\alpha_i + x_i) - 1} \tag{12}$$

Para explorar esta distribución se puede graficar para alguna i y ver qué tantas posibilidades se le da a $p_i = 0.1$, ya sea calculando un intervalo de probabilidad para p_i con los cuantiles $q_{\gamma 2}$ y $q_{1-\gamma 2}$ de la distribución $\pi(p_i \mid X)$, o bien calculando $\Pr(p_i > 0.1 \mid X)$ o $\Pr(p_i < 0.1 \mid X)$.

El procedimiento descrito también se puede ajustar fácilmente para plantear la aleatoriedad de grupos de dígitos. Basta hacer los ajustes en el modelo multinomial correspondiente. El procedimiento descrito no requiere aproximaciones asintóticas, por lo que se puede aplicar en cualquier momento, además de ser más informativo al obtener la distribución posterior para $\mathbf{p}$ o las p_i de interés.

3. Independencia y monitoreo de resultados

Puede pasar que con base en las pruebas de la sección anterior se declare que la frecuencia con la que ocurre cada dígito es similar; sin embargo esto no es suficiente para asegurar la aleatoriedad de los resultados, ya que pudieron darse rachas de sorteos que favorecieran a ciertos dígitos, pero que al combinar toda la información y resumir las frecuencias, estas rachas queden ocultas. Por ello se requiere también verificar la independencia entre los resultados sucesivos.

La forma tradicional de verificar esta independencia es ver los resultados de una urna a través de los diferentes sorteos como una serie de tiempo, y calcular la autocorrelación para medir la correlación entre los valores de la serie (ver Royal Statistical Society 2002). Sea x_j una serie de n valores ($j = 1, 2, \ldots, n$), el coeficiente de autocorrelación simple r_k con retardo k, típicamente se estima con

$$r_k = \frac{\sum_{t=1}^{n-k} (x_t - x)(x_{t+k} - \overline{x})}{\sum_{t=1}^{n} (x_t - \overline{x})^2} \tag{13}$$

donde $\overline{x}$ es la media aritmética de la variable x en la serie. Es usual calcular r_k para varios valores de k, usualmente pequeños. Una prueba típica de independencia de las series es la de Box & Pierce (1970), que se basa en el siguiente estadístico:

$$Q(r) = n \sum_{k=1}^{t} r_k^2 \tag{14}$$

que asintóticamente tiene una distribución χ^2 con t grados de libertad, por lo que se rechaza la independencia de la serie si $Q(r) > \chi^2_{1-\alpha}$. En Ljung & Box (1978) se propone una modificación a la prueba de Box-Pierce, que tiene un mejor desempeño cuando se satisface el supuesto de normalidad. La propuesta se basa en el estadístico

$$L(r) = n(n+2) \sum_{k=1}^{t} (n-k)^{-1} r_k^2 \tag{15}$$

que para n grande se distribuye como χ^2_t. El que estas pruebas requieran normalidad, muestras grandes y el hecho de aportar poca evidencia gráfica, son limitantes importantes para su aplicación práctica en juegos de lotería.

Una alternativa a las pruebas de independencia sería monitorear los resultados de los sorteos en tiempo real y no un año o dos después, cuando todo es historia. Es reconocido que las cartas de control son una de las mejores formas de monitorear procesos repetitivos (Gutiérrez & de la Vara 2009), por lo que dado que cada sorteo

de un juego de azar es una repetición en la realización de un proceso, se propone monitorear la aleatoriedad de los juegos de azar mediante cartas de control.

Una forma práctica e intuitiva de monitorear los resultados de un juego de números de d dígitos es viendo el dígito que resultó en el más reciente sorteo y analizar hace cuántos sorteos (ensayos) había aparecido tal dígito; de esta manera, si Y es el número de ensayos o sorteos que transcurrieron para que apareciera nuevamente tal dígito, entonces bajo el supuesto de aleatoriedad, Y tiene una distribución geométrica con parámetro $p = 0.1$, por lo que

$$f(y \mid p) = \Pr(Y = y \mid p) = (1-p)^{y-1}p \tag{16}$$

con $y = 1, 2, \ldots$

La distribución acumulada está dada por

$$F(y \mid p) = \Pr(Y < y \mid p) = p\sum_{i=1}^{y}(1-p)^{i-1} = 1 - (1-p)^{y} \tag{17}$$

El límite de control superior (LCS) e inferior (LCI) indican dónde se espera esté Y bajo el supuesto de aleatoriedad. Si se fija una probabilidad de error tipo I igual a γ, entonces estos límites se obtienen con $\Pr(Y > LCS \mid p) = \gamma 2$ y $\Pr(Y < LCI \mid p) = \gamma 2$. Sin embargo, con la distribución geométrica con $p = 0,1$ se obtiene que $P(Y = 1) = 0,1$, por lo que dado que usualmente los valores de γ son pequeños (el valor más comúnmente usado en control de procesos es $\gamma = 0,0027$), entonces no se debe usar LCI, por lo que $\Pr(Y > LCS \mid p) = \gamma$. De aquí que

$$\begin{aligned} \Pr(Y &> LCS \mid p) = 1 - \Pr(Y \leq LCS \mid p) \\ &= 1 - F(LCS + 1 \mid p) \\ &= (1-p)^{LCS+1} = \gamma \end{aligned}$$

Despejando a LCS de esta última igualdad, se obtiene que el valor aproximado para el LCS está dado por

$$LCS = \frac{\ln(\gamma)}{\ln(1-p)} - 1 \tag{18}$$

Nótese que para la aplicación específica de la carta es necesario partir de que $p = 0.1$; si no fuera así, entonces la incertidumbre sobre p la podría aportar una distribución Beta, y hacer inferencia con la distribución posterior correspondiente. En Gutiérrez (2006) se presenta una versión bayesiana para una carta de control bajo el supuesto anterior. La versión frecuentista de la carta geométrica y algunos detalles de su caracterización se pueden consultar en Yang, Xie, Kuralmani & Tsui (2002).

Si $\gamma = 0.0027$, entonces el límite de control superior LCS se puede tomar igual a 56, ya que $\Pr(Y > 56 \mid p = 0,1) = 0,002739$. Además de los límites de control es usual establecer otras zonas que ayudan a detectar otros patrones no aleatorios en el comportamiento del proceso (ver Gutiérrez & de la Vara 2009). Sin embargo,

en el caso de la carta de control geométrica no es posible utilizar las zonas típicas debido a la forma de la distribución geométrica, que tiene una forma exponencial donde acumula las probabilidades más altas a los números enteros más pequeños.

Como alternativa se propone generar cinco zonas con una cobertura aproximada de 0,20 en la probabilidad para los correspondientes valores de Y. Estas zonas se detallan en la tabla 1, y se han obtenido con el apoyo de la distribución acumulada (17). Como en el juego de números de d dígitos lo que se quiere garantizar es la aleatoriedad, entonces esto será equivalente a observar un proceso (juego) bajo control estadístico en una carta geométrica, en donde la proporción de puntos en las diferentes zonas a lo largo de la carta de control sea consistente con la probabilidad indicada en la tabla 1. La separación de las zonas de la tabla 1 puede darse con el valor medio que las separa, así $l_A = 2,5$, $l_B = 5,5$, $l_C = 9,5$, $l_D = 16,5$.

Tabla 1: Zonas en la carta de control geométrica.

Zona	Valores de Y	Probabilidad
A	$\{1, 2\}$	$0,19000$
B	$\{3, 4, 5\}$	$0,21951$
C	$\{6, 7, 8, 9\}$	$0,20307$
D	$\{10, \ldots, 16\}$	$0,20212$
E	$\{Y \geq 17\}$	$0,18530$

De esta forma, para monitorear el comportamiento de Y en un juego de número de d dígitos, la observación se puede centrar en cada una de las urnas con las que usualmente se generan los dígitos del juego. Monitoreando en forma separada cada urna y llevando un registro (carta) particular para cada dígito -para ello cada vez que se realice un sorteo se ve qué dígito fue extraído de la urna bajo análisis-, se analiza cuántos sorteos fueron necesarios para que volviera a aparecer tal dígito, con lo que en cada sorteo se tiene el valor de Y. Para aplicar la carta desde el primer sorteo, se puede suponer que en el sorteo cero salieron todos los dígitos.

Si se viera un punto fuera del LCS, se deberá ver qué pasó, particularmente qué dígito i es, porque esto identificará la esfera que ha tardado demasiados sorteos en salir, es decir, esto sugerirá que $p_i < 0,1$. Al considerar que $\sum_{i=1}^{10} p_i = 1$, y si $p_i < 0,1$ para algún i, entonces esto implica en contraste que $p_j > 0,1$ para una o más $j = 1, \ldots 10$, con $j \neq i$. Esto, trasladado en la carta de control, se reflejará en que mientras los valores de Y para la esfera i, Y_i, tenderán a ser más grandes, los valores de Y_j tenderán a ser más pequeños. Ambas cosas se detectarán en las zonas de la carta de control. Esta estructura de dependencia plantea la necesidad de hacer un monitoreo simultáneo e individualizado para todas las esferas. Simultáneo en el sentido que se tenga claro que a quien se está monitoreando es a Y (número de ensayos o sorteos que transcurrieron para que saliera nuevamente el dígito que acaba de aparecer), pero individualizado porque se propone que el valor de Y se registre en la celda (carta) que le corresponde a cada dígito.

De esta manera en cada carta, para cada dígito, se debe vigilar visualmente que aproximadamente la quinta parte de los valores de Y caigan en cada una de

las zonas de la carta. Si hay sospechas de que esto no está ocurriendo para un dígito, porque por ejemplo hay una racha larga de puntos donde los valores de Y tienden a caer en las últimas zonas de la correspondiente carta de control; esto también se reflejará en forma opuesta en por lo menos otra zona de otro dígito, donde se tendrán muchos puntos en las primeras zonas de la carta de control. Identificado esto, para verificar formalmente las posibles rachas de puntos que siguen un patrón no aleatorio se tienen dos posibilidades. En la primera se parte de un modelo multinomial para el dígito i con los cinco sucesos dados por las zonas de la carta, es decir, un modelo multinomial con $\mathbf{p} = (p_A, p_B, p_C, p_D, p_E)$, donde se observan x_i veces la esfera i; entonces la distribución marginal posterior para cualquiera de las p_k (con $k = A, \ldots, E$), se obtiene a partir de (12), para precisar mejor que quede lo indicado como la distribución a priori no informativa, $\alpha_k = 1$. Específicamente la distribución es Beta$(1 + w_k, 5 + x_i - (1 + w_k))$:

$$\pi(p_k \mid x_i, w_k) = \frac{\Gamma(5 + x_i)}{\Gamma(1 + w_k)\,\Gamma(5 + x_i - (1 + w_k))} p_k^{1+w_k-1} (1 - p_k)^{5+x_i-(1+w_k)-1} \quad (19)$$

donde w_k es el número de puntos que cayeron en la zona k de la carta de control para el dígito i. A partir de esta distribución marginal se puede calcular un intervalo de probabilidad para ver si éste incluye al correspondiente valor de p_k dado en la tabla 1. Adicionalmente se pueden combinar dos o más de las zonas de la tabla 1, y hacer los análisis correspondientes mediante (19).

El otro criterio para verificar rachas de valores de Y que no sigan un patrón aleatorio es la aplicación de la distribución binomial. Supóngase que hay una racha sospechosa donde h de t puntos consecutivos cayeron en la zona k de la carta geométrica, considerando que la zona k puede ser alguna de las indicadas en la tabla 1, o bien una combinación de algunas zonas de esta tabla u otras que reúnan algún otro criterio. Entonces es claro que bajo el supuesto de aleatoriedad, la probabilidad de que tal racha ocurra está dada por la distribución binomial (t, p_k), donde p_k es la probabilidad de la zona k construida como se ha indicado. Por tanto

$$\Pr(h \mid t) = \binom{t}{h} p_k^h (1 - p_k)^{t-h} \quad (20)$$

Por ejemplo, supóngase que en la zona A de la carta de control geométrica para algún dígito caen tres puntos consecutivos, lo que significa que tal dígito salió como resultado en tres sorteos consecutivos o casi consecutivos. Para calcular la probabilidad de esto, se tiene que $h = 3, t = 3$ y $p_k = 0,19$; y de acuerdo con (20), la probabilidad de tal racha es igual a $0,006859$. Esta probabilidad es suficientemente baja como para pasar por alto la situación; por tanto lo que se sugiere hacer es evaluar de cerca lo que está pasando con el proceso de aleatorización, como por ejemplo se podrían simular sorteos para corroborar la sospecha con más datos. En la sección 5, donde se analizan datos de un caso práctico, se aplica (20) para varias rachas sospechosas.

Para analizar el funcionamiento de la carta propuesta se hace el siguiente estudio, y además más adelante se verá el caso práctico del sorteo Tris.

4. Estudio Monte Carlo

Para analizar y visualizar el funcionamiento de la carta geométrica propuesta en la sección previa se hizo un estudio Monte Carlo. Las rutinas fueron programadas con S-Plus. Se buscó simular lo que ocurre en una urna que contiene diez esferas numeradas con los dígitos, y en cada sorteo se extrae una. Se asignó una probabilidad de extracción p_i para cada esfera y se simularon 300 mil repeticiones del sorteo. Además de cuantificar la proporción de veces que resultó cada esfera, se registró el número de sorteos que tuvieron que pasar para que volviera a salir el mismo resultado, es decir se cuantificó Y, junto con la proporción de veces, respecto al total de sorteos en que el valor de Y fue superior al límite de control superior ($LCS = 56$). Además, para cada esfera se registró la proporción de veces en que el valor de Y cayó en cada una de las zonas de la carta para cada dígito (tabla 1). En la tabla 2 se reportan los resultados para dos casos. En el primero se supuso que dos esferas, la 3 y 4, tenían probabilidades de ser extraídas de $p_3 = 0,07$ y $p_4 = 0,13$, y a las restantes se les asignó una probabilidad de $0,1$. Se observa que la proporción en la que cada esfera fue el resultado del sorteo fue muy similar a su correspondiente probabilidad; por ejemplo, el $7,0\,\%$ de las ocasiones la esfera 3 fue el resultado del sorteo, y la cuatro un $13,0\,\%$ de las veces. Por otro lado, en cuanto a los valores de Y en las zonas A a E para cada esfera, se ve que en el caso de la esfera 3, las proporciones para las primeras tres zonas fue de $0,138$, $0,168$ y $0,172$, que son visiblemente menores a las esperadas, de alrededor de $0,20$; en contraste, en la última zona, la E, tiene una proporción de $0,313$, contra una esperada de 0.185. En otras palabras la esfera tres tardaba demasiados sorteos en salir. En cuanto a la esfera 4, los resultados son opuestos: más veces cayó Y en las primeras zonas de la carta y pocas en las últimas; por ejemplo, en la zona E la proporción fue de $0,107$, que es casi la mitad de la esperada bajo aleatoriedad $(0,185)$. En los restantes dígitos los resultados en las diferentes zonas son muy similares a los esperados bajo aleatoriedad.

Para el segundo caso de la tabla 2, se asignó a la esfera 2 una probabilidad de ser extraída de $p_2 = 0,145$, y este aumento respecto a $0,10$ se quitó de forma equitativa al resto de las esferas, con lo que sus probabilidades de extracción fueron de $0,095$. Se observa que la proporción con la que salió cada esfera es muy similar a la probabilidad asignada, por ejemplo, un $14,5\,\%$ de las ocasiones la esfera 2 fue el resultado del sorteo, mientras que las proporciones de los valores de Y para tal esfera que caen en las zonas A y B son de $0,271$ y $0,270$ respectivamente, ambas muy por arriba de $0,20$. Estos aumentos contrastan con los datos observados para la zona E, a la que le corresponden sólo el $8,0\,\%$ de los valores de Y para la esfera 2. Respecto del resto de las esferas, se nota una ligera disminución de la proporción de valores de Y para la zona A.

Con los resultados de las simulaciones de la tabla 2 se ilustra que un incremento en el porcentaje de extracciones de un dígito en particular se puede detectar, además de en el porcentaje mismo, en la carta de control mediante un incremento de la proporción de valores de Y que caen en las primeras zonas de la carta y una baja proporción de valores de Y en las últimas zonas, principalmente en la zona E, mientras que cuando se tienen proporciones de extracción menores a $0,1$, se

Tabla 2: Resultados de estudio de simulación para dos casos seleccionados.

		Resultado del sorteo	Puntos fuera del LCS	Proporción de puntos en cada zona A	B	C	D	E
Prob. teórica $p_i = 0,1$		$0,100$	$0,00027$	$0,190$	$0,220$	$0,203$	$0,202$	$0,185$
Resultados para los dígitos, con $p_3 = 0,07$ $p_4 = 0,13$, y los restantes igual a $0,1$	1	$0,100$	$0,00027$	$0,188$	$0,221$	$0,203$	$0,206$	$0,183$
	2	$0,100$	$0,00024$	$0,192$	$0,219$	$0,203$	$0,201$	$0,186$
	3	$0,070$	$0,00119$	$0,138$	$0,168$	$0,172$	$0,209$	$0,313$
	4	$0,130$	$0,00006$	$0,245$	$0,255$	$0,214$	$0,179$	$0,107$
	5	$0,100$	$0,00025$	$0,191$	$0,217$	$0,204$	$0,202$	$0,185$
	6	$0,100$	$0,00032$	$0,191$	$0,220$	$0,200$	$0,202$	$0,186$
	7	$0,100$	$0,00035$	$0,193$	$0,218$	$0,203$	$0,203$	$0,183$
	8	$0,099$	$0,00028$	$0,191$	$0,221$	$0,201$	$0,196$	$0,191$
	9	$0,101$	$0,00026$	$0,190$	$0,219$	$0,206$	$0,202$	$0,183$
	0	$0,100$	$0,00032$	$0,189$	$0,219$	$0,199$	$0,208$	$0,185$
Resultados para los dígitos, con $p_2 = 0,145$ y los restantes igual a $0,095$	1	$0,095$	$0,00035$	$0,182$	$0,211$	$0,197$	$0,206$	$0,204$
	2	$0,145$	$0,00001$	$0,271$	$0,270$	$0,216$	$0,163$	$0,080$
	3	$0,095$	$0,00033$	$0,184$	$0,211$	$0,198$	$0,206$	$0,201$
	4	$0,095$	$0,00034$	$0,180$	$0,216$	$0,199$	$0,202$	$0,203$
	5	$0,095$	$0,00037$	$0,181$	$0,215$	$0,197$	$0,206$	$0,201$
	6	$0,095$	$0,00038$	$0,183$	$0,215$	$0,196$	$0,208$	$0,198$
	7	$0,095$	$0,00037$	$0,184$	$0,208$	$0,203$	$0,204$	$0,201$
	8	$0,094$	$0,00036$	$0,179$	$0,215$	$0,197$	$0,203$	$0,206$
	9	$0,095$	$0,00034$	$0,182$	$0,214$	$0,196$	$0,203$	$0,204$
	0	$0,095$	$0,00038$	$0,197$	$0,207$	$0,202$	$0,207$	$0,202$

detecta en la carta, con puntos fuera del LCS, mayores proporciones de valores de Y que caen en las zonas D y E, y una menor proporción para las primeras zonas de la carta. Las sospechas deben ser validadas con las pruebas descritas en la subsección 2.2.

En relación con el uso del LCS para detectar cambios, primero es importante puntualizar la probabilidad del error tipo I, que consiste en declarar que ha habido un cambio cuando en realidad no fue así, por lo que el correspondiente punto fuera del LCS será una falsa alarma. En la tabla 2, se ha estimado la proporción de falsas alarmas debido a puntos fuera del LCS. Se ha hecho respecto al total de sorteos, porque al final de cuentas se está monitoreando sólo un proceso, y los resultados se reportan de forma separada para cada dígito como una estrategia para separar fuentes de variabilidad atribuible y así lograr un mejor monitoreo del proceso. Ante esto, el LCS se ha establecido con $\gamma = 0,002739$, como es usual en cartas de control, de tal forma que en condiciones de control estadístico (aleatoriedad) se esperaría que entre todos los dígitos sólo se tuviera esa proporción de falsas alarmas. Si a $0,002739$ se le divide entre diez, entonces se esperaría que cada dígito aporte $0,00027$ puntos fuera del LCS. En la tabla 2 se puede ver que en las esferas que tuvieron $p_i = 0,1$, su proporción de puntos fuera está muy cerca de lo esperado, mientras que para las esferas en que $p_i < 0,1$ se tienen más puntos fuera, porque tarda más la esfera en salir como resultado; por ejemplo, para la esfera 3 en el caso 1, se tiene que la proporción de puntos fuera fue de $0,00119$, en cambio, si $p_i > 0,1$, la proporción de puntos fuera es mucho menor, como se ve en la esfera 4 del caso 1, donde tal proporción fue de $0,00006$. Por tal razón el tipo de

problemas de falta de aleatoriedad que se detectan con el LCS son aquellos donde $p_i < 0,1$, y no así cuando $p_i > 0,1$, esto último debido a que la carta geométrica no tiene LCI.

Con lo que se ha dicho respecto al error tipo I, queda claro que la probabilidad de este error γ, que es con la que se calcula el LCS, se mantendrá respecto al total de sorteos bajo el supuesto de aleatoriedad, a pesar de que se lleve el registro por separado de los diferentes dígitos. Además, dadas las limitaciones del LCS para detectar cambios, es necesario el uso de las cartas para cada dígito junto con las zonas propuestas en la tabla 1. Esto incrementa la potencia para detectar problemas mediante la identificación de rachas o patrones con sospecha de falta de aleatoriedad. Pero para mantener en un nivel bajo la posibilidad de declarar cambios cuando no los ha habido, error tipo I global, entonces cuando haya sospechas de una racha no aleatoria en una carta de control para un dígito, esto deberá ser corroborado con criterios probabilísticos, como los sugeridos en la sección anterior, con las expresiones (12), (19) y (20).

5. Aplicación al sorteo Tris de México

El sorteo Tris es un juego de d dígitos de la empresa Pronósticos para la Asistencia Pública del gobierno mexicano. Inicialmente el participante seleccionaba tres dígitos, posteriormente fue cambiado para que la selección fuera de cuatro dígitos, y en el año 2007 se transformó en un juego de cinco dígitos, con lo cual el participante forma un número entre 00000 y 99999. El sorteo se realiza en forma pública, y para extraer las esferas que definen el número premiado se utiliza un dispositivo electromecánico (urna). Los números ganadores del sorteo Tris están disponibles en la página *www.pronosticos.gob.mx*. Para este trabajo se tomaron los resultados del dígito que formaba las unidades de millar, a la que se identifica como la urna 4, para un total de 500 sorteos Tris entre el 20/07/1996 (sorteo 2913) y el 28/03/1997 (sorteo 3412). La frecuencia resultante para cada uno de los dígitos en estos 500 sorteos en la urna 4 se muestra en la tabla 3.

Tabla 3: Resultados de 500 sorteos Tris para la urna 4 (unidades de millar).

Dígito	1	2	3	4	5	6	7	8	9	0
Frecuencia	48	56	45	52	49	57	56	43	60	34
Proporción	0,096	0,112	0,09	0,104	0,098	0,114	0,112	0,086	0,12	0,068

Es evidente que bajo la hipótesis de selección aleatoria de las esferas, cada dígito tiene una probabilidad teórica de ser seleccionado de 0,1. Sin embargo, a primera vista los resultados de la tabla 3 parecen contradecir esta hipótesis, debido particularmente a los dígitos cero y nueve. Este último salió casi el doble de veces que el primero. Para confirmar estas sospechas se procede a probar la hipótesis (2) con $m = 10$ y $p_{i,0} = 0,1$, mediante las diferentes pruebas aproximadas descritas en la subsección 2.1. Para empezar, la prueba usual basada en el estadístico (3),

donde se tiene que

$$\chi_0^2 = \sum_{i=1}^{10} \frac{(x_i - 50)^2}{50} = 11,2 \tag{21}$$

como el valor crítico es $\chi^2_{0,05,9} = 16,919$, no se rechaza H_0. Este no rechazo se ve también con el valor-p para (21), que es de $0,2622$. Así, esta prueba no aporta evidencia significativa en contra de la aleatoriedad de los resultados sintetizados en la tabla 3. Algo similar ocurre con la prueba de la razón de verosimilitud del estadístico (4), ya que el valor de $R = 11,703$, lleva a no rechazar H_0, puesto que el valor-p correspondiente es igual a $0,2306$. Por su parte, con la prueba de Nass basada en el estadístico (7), se tiene que $\chi_c^2 = 11,18$, y dado que los grados de libertad para la distribución χ^2 de referencia son $v = 9,02$, entonces esta prueba tampoco rechaza H_0. En cuanto a los intervalos simultáneos para las diez p_i, basados en (8), con $\alpha = 0,05$, todos incluyen al $0,1$; por ejemplo los intervalos para p_1, p_9 y p_{10} están dados, respectivamente, por $[0,065, 0,139]$, $[0,085, 0,167]$ y $[0,043, 0,107]$, con p_{10} la probabilidad para el dígito cero. Como todos los intervalos incluyen a $0,1$, no se rechaza (2) con esta metodología. Es de señalar que el intervalo para el dígito cero es el que estuvo más cercano de no incluir al $0,1$.

Para aplicar el procedimiento bayesiano descrito en la subsección 2.2, se tiene que $n = \sum_{i=1}^{10} x_i = 500$ y $\alpha_0^* = 5 + 500$. Un primer asunto de interés es explorar la distribución conjunta posterior de la probabilidad del dígito cero, p_{10}, y la del nueve, p_9. Para ello se hizo una corrida de dos mil simulaciones de la distribución posterior de $\mathbf{p}$, utilizando el MCMCPack de R, como se indicó en la subsección 2.2. En la parte superior de la figura 1 se ve una muestra de la distribución conjunta posterior de (p_{10}, p_9), de donde se aprecia la poca posibilidad que se le da a que $p_{10} \geq 01$; de hecho, de los dos mil puntos sólo 14 están a la derecha de la línea vertical $p_{10} = 0,1$, lo que en proporción significa $0,007$. Estas evidencias refuerzan el hecho de que $p_{10} < 0,1$. Por su parte, la posibilidad de que $p_9 > 0,1$ es un poco mayor, ya que de los dos mil puntos 165 están abajo de la línea horizontal $p_9 = 0,1$, lo que significa una proporción de $0,0825$.

Una conclusión similar a la anterior se llega al observar la parte baja de la figura 1 que muestra la gráfica de contornos con diez cortes de la distribución posterior de $\mathbf{p}_m = (p_{10}, p_9, 1 - p_{10} - p_9)$, que de acuerdo con lo establecido en la subsección 2.2, es Dirichlet$(\mathbf{p}_m, \alpha_m)$, con $\alpha_m = (1 + 34, 1 + 60, 510 - 2 - 34 - 60)$. En esta figura se ve la poca posibilidad que se le da a la región $p_{10} \geq 0,1$.

De la misma manera se puede obtener la distribución posterior marginal (12) para cualquiera de los dígitos, que según lo establecido en la subsección 2.2 es una distribución Beta$(1 + x_i, 510 - 1 - x_i)$. Como es de particular interés hacer inferencia sobre los dígitos nueve y cero, de la tabla 3 se ve que éstos tuvieron una frecuencia de $x_9 = 60$ y $x_{10} = 34$, respectivamente. En la figura 2 se muestra la distribución posterior para estos dos casos, de donde es claro que en el caso del dígito cero se debe rechazar la idea de que $p_{10} = 0,1$, como requiere el supuesto de aleatoriedad. Esto se respalda si se calculan intervalos de probabilidad para p_{10} con los cuantiles $q_{\gamma 2}$ y $q_{1-\gamma 2}$ de la distribución $\pi(p_{10} \mid X)$. Tomando $\gamma = 0,05$, se obtiene que estos intervalos están dados por $[0,0929,\ 0,1491]$ y $[0,0484,\ 0,0921]$, para p_9 y p_{10}, respectivamente.

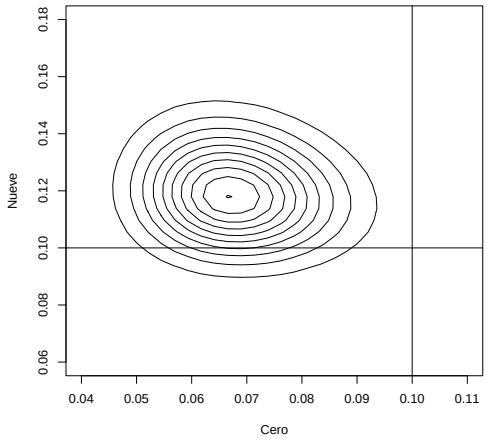


FIGURA 1: Distribución posterior conjunta para los dígitos cero y nueve.

Como en el caso del dígito nueve, el valor $p_9 = 0,1$ está dentro del correspondiente intervalo, en este caso no se alcanza a rechazar la hipótesis de aleatoriedad, cosa que si ocurre en el caso del dígito cero, ya que $p_{10} = 0,1$ está fuera del correspondiente intervalo de probabilidad. De hecho la probabilidad de ver este evento o uno peor respecto a aleatoriedad, se puede obtener calculando el área en la cola de la distribución marginal posterior, es decir calculando $\Pr(p_{10} > 0,1 \mid X)$ que es igual a $0,0056$. Esto muestra que se tiene una fuerte evidencia contra la aleatoriedad del sorteo Tris en el período analizando. Mientras que en relación al dígito nueve se tiene que $\Pr(p_9 < 0,1 \mid X) = 0,0807$, lo que da una cierta evidencia en contra de la aleatoriedad, aunque no concluyente como ya se había puntualizado en la figura 1.

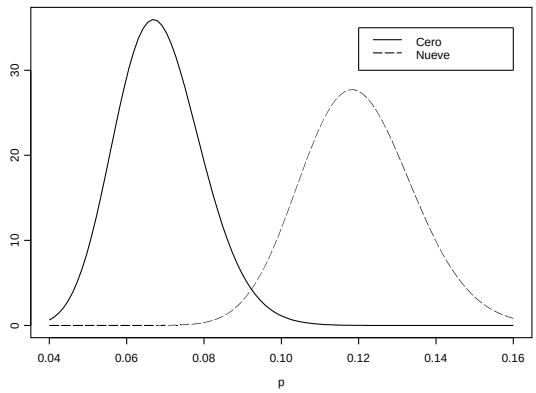


FIGURA 2: Distribución marginal posterior para los dígitos cero y nueve.

Si se quisiera un estimador puntual para p_9 y p_{10}, en lugar de los estimadores de intervalo que se han dado antes, se puede calcular la moda de las correspondientes distribuciones marginales posteriores. Estas modas están dadas por $\widehat{p}_9 = 0,1181$ y $\widehat{p}_{10} = 0,0669$.

Por lo anterior, contrario a las pruebas tradicionales basadas en aproximar diferentes estadísticos con la distribución χ^2, el procedimiento bayesiano sí detecta que hay problemas en la aleatoriedad de los resultados del sorteo Tris, como intuitivamente se esperaba a partir de la tabla 2, con la ventaja adicional de que se aporta evidencia de la distribución posterior, conjunta o marginal, para los p_i con problemas.

Para probar la independencia de los resultados del sorteo Tris, con los datos de la urna del millar de los 500 sorteos se calculó el coeficiente de autocorrelación r_k, para $k = 1, \ldots, 24$, y se obtuvo que el estadístico Box-Pierce de la expresión (14) toma el siguiente valor:

$$Q = 500 \sum_{i=1}^{24} r_i^2 = 36,74$$

que tiene un valor-$p = 0,0464$, por lo que con una significancia de $\alpha = 0,05$ se rechaza la independencia de los resultados de la serie. Al aplicar la prueba Ljung-Box con el estadístico (14), se obtiene que $L(r) = 38,00$, que tiene un valor-$p = 0,0346704$, por lo que también se rechaza la independencia de la serie, aunque la evidencia gráfica de la falta de independencia, no reportada en este trabajo, para toda la serie es poco clara.

Con el fin de aportar una mejor evidencia gráfica y monitorear el comportamiento a través del tiempo de los resultados del sorteo Tris, primero,c on base en cada dato de la serie original se calculó Y, es decir se identificó qué dígito fue el resultado y se contó cuántos sorteos fueron necesarios para que volviera a aparecer tal dígito. Para aplicar la carta desde el primer sorteo, se supuso que en el sorteo cero salieron todos los dígitos. En la tabla 4 se muestra un extracto de los valores de Y.

Tabla 4: Muestra de los valores de Y para los primeros 20 sorteo 5 de Tris.

Sorteo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Dígito	9	9	5	2	3	5	7	1	8	9	9	0	2	5	3	0	2	4	6	1
Y	1	1	3	4	5	3	7	8	9	8	1	12	9	8	10	4	4	18	19	12

Con los valores de Y para cada dígito se obtuvieron las correspondientes cartas de control geométricas. Por ejemplo en la figura 3 se muestran las cartas de control para los dígitos nueve y cero, que ya se vio tienen problemas de cumplimiento del supuesto de aleatoriedad, por lo que es de interés ver cómo lo reflejan las cartas de control. En el caso del dígito cero, el punto 26 con valor de $Y = 60$ está fuera del LCS, lo que indica que tuvieron que pasar 60 sorteos para que volviera a salir la esfera del dígito cero en la urna del millar, lo que no es congruente bajo el supuesto de aleatoriedad en el sorteo. Además, visualmente parece haber demasiados puntos en la parte alta de la zona E de la carta. En particular parece sospechoso el que

9 de 34 valores de Y sean mayores que 22. Para corroborar esta sospecha (racha) mediante la aplicación de (20), se tiene que $h = 9$, $t = 34$, y como la zona bajo sospecha en la carta está dada por aquellos puntos de Y que son mayores que 22, entonces para calcular la probabilidad de que un punto caiga en esta zona se puede usar la distribución geométrica acumulada $F(y)$ de la expresión (17), con $p = 0,1$; por tanto $p_k = \Pr(Y \geq 23) = 1 - F(23 \mid p = 0,1) = 0,0886$, entonces aplicando (20) se obtiene que la probabilidad de la racha referida es igual a $0,00174$, lo que es una probabilidad muy baja. Esto confirma la idea de que $p_{10} < 0,1$ para el dígito cero. Aquí la carta de control hubiese resultado un buen instrumento de monitoreo del sorteo, ya que hubiera avisado que con frecuencia pasaban demasiados sorteos en donde la esfera cero no salía como resultado.

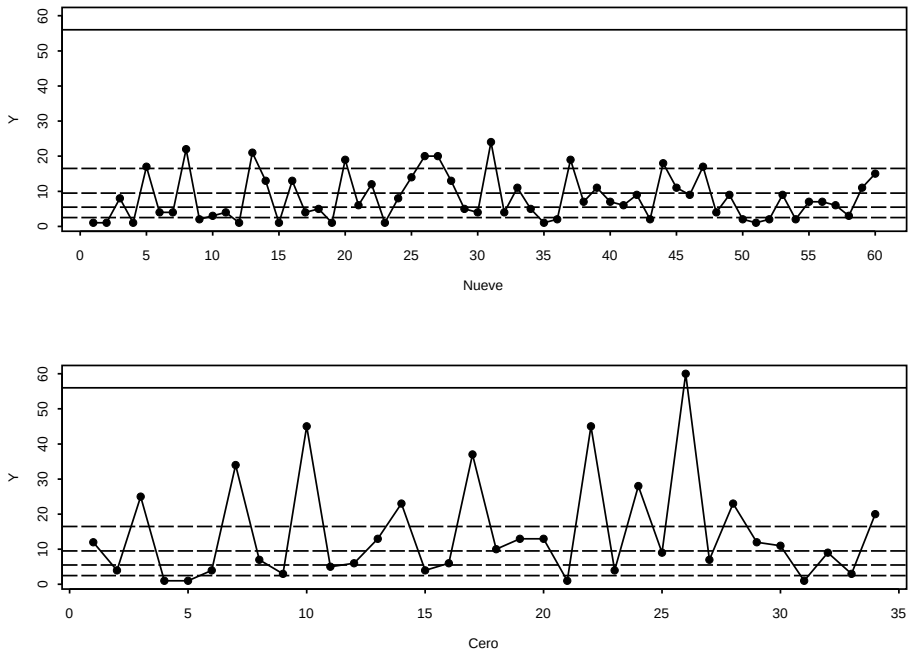


FIGURA 3: Cartas geométricas para los dígitos cero y nueve del sorteo Tris.

En la parte baja de la figura 3 se muestra la carta de control para el dígito nueve, que presentó una situación opuesta a la del dígito cero, ya que los valores de Y tienden a ser más pequeños; por ejemplo, hubo nueve veces en que $Y = 1$, lo que implica que en nueve ocasiones hubo dos sorteos consecutivos donde se tuvo como resultado al nueve. Además, de las 60 veces que salió el número 9, en todas $Y < 25$, y como de acuerdo con (17) se tiene que $P(Y < 25) = F(25 \mid p = 0,1) = 0,9282$, al aplicar (20) con $h = 60$, $t = 60$ y $p_k = 0,9282$, se tiene que la probabilidad de que en los sesenta sorteos ocurra que $Y < 25$ es igual a $0,011$, lo que es una baja probabilidad. Lo anterior se suma a la sospecha de que en el período analizado, el

número nueve salía con más frecuencia de lo esperado. Esto confirma la utilidad de la carta, y en el caso del dígito nueve se tiene que $p_9 > 0,1$.

Igualmente se analizó cuántos valores de Y cayeron en cada zona de la correspondiente carta de control para cada dígito. Los resultados de esto se muestran en la tabla 5. Como por diseño la frecuencia con la que los valores de Y caen en cada zona para cada dígito es más o menos similar, resaltan los casos de las esferas 3 y 8 debido a que las primeras zonas de sus respectivas cartas de control incluyen una cantidad pequeña de puntos. Para evaluar con mayor detalle esto, se utiliza (19). Si se parte del supuesto de que $p = 0,1$ y si se junta la información de las zonas A y B, se tiene que $p_{AB} = p_A + p_B = 0,41$ (ver tabla 1); entonces en el caso del dígito 3 se tendrá que $x_i = 45$ y $w_k = 5 + 6$, y por lo que la distribución posterior de p_{AB} es Beta con parámetros $(12, 37)$. Al calcular con esta distribución un intervalo de probabilidad para p_{AB} con una cobertura de $0,95$, se obtiene que el intervalo es $[0,1364, 0,3731]$, que no incluye a $p_A + p_B = 0,41$, lo que es una evidencia contra el supuesto de que $p_3 = 0,1$ para la esfera 3; esto se ratifica si se calcula $\Pr(p_{AB} > 0.41 \mid x_i, w_k) = 0,0068$, lo que sugiere que a través del tiempo Y para la esfera 3 no siguió una distribución geométrica con $p_3 = 0,1$, y que más bien parece que $p_3 < 0,1$ por lo menos por un período largo, lo que es una evidencia más contra la aleatoriedad del sorteo Tris.

En la parte superior de la figura 4 se muestra la carta de control para el dígito 3, y la gráfica sugiere dos periodos: el primero con $p_3 > 0,1$, que va del punto 1 hasta aproximadamente el 20, donde los valores de Y tienden a que Y tenga valores pequeños; por ejemplo del punto 14 al 18, se da una racha donde 4 de 5 puntos consecutivos caen en la zona A de la carta; luego aplicando (20) con $h = 4$, $t = 5$ y $p_k = 0,19$, se tiene que bajo el supuesto de aleatoriedad la probabilidad de tal racha es de apenas $0,005$. Por el contrario, en la segunda parte de la carta parece que los valores de Y tienden a ser más grandes; por ejemplo, en esta segunda parte se da una racha desde el punto 19 hasta el 45 donde ningún valor de Y cayó en la zona A, lo que sugiere que en esta segunda parte de la carta más bien $p_3 < 0,1$. Al mezclarse estos dos periodos hacen que la frecuencia total con la que resultó la esfera 3 no suscite ninguna sospecha de problemas de aleatoriedad. Lo que ha ocurrido con la esfera 3 muestra la utilidad de la carta de control geométrica propuesta para monitorear la aleatoriedad de este tipo de juegos.

En la carta de control para el dígito 8, que no se presenta aquí y que es otro caso donde hay pocos valores de Y en las zonas A y B, ocurre algo relativamente similar a lo del dígito 3: en los primeros 19 puntos de la carta, 18 caen fuera de las zonas A y B, lo que bajo el supuesto de aleatoriedad tiene una probabilidad muy baja de ocurrir de apenas $0,0006$. En cambio, en la segunda parte de la carta los valores de Y fluctúan de la manera esperada.

En la parte inferior de la figura 4 se muestra la carta de control para el dígito 6, que presenta un caso opuesto al dígito 3, ya que hay pocos valores de Y en las zonas D y E, lo que indica que la esfera 6 tendía a salir demasiado pronto, por lo que pareciera que $p_6 > 0,1$ en algunos periodos. Por ejemplo del punto 15 al 38 se presentó una racha donde de 24 puntos sólo uno cayó en la zona E, lo que tiene una probabilidad de ocurrir de $0,04$. Aunque no es una evidencia contundente,

sí apoya un tanto que $p_6 > 0,1$ durante ese periodo. En apoyo a esto se destaca visualmente que tres puntos consecutivos de Y cayeron en la zona A (del 10 al 12), y la probabilidad de que esto ocurra bajo el supuesto de aleatoriedad es de $(p_A)^3 = (0,19)^3 = 0,00686$, que se puede considerar baja.

Tabla 5: Frecuencia de los valores de Y por zona de la carta para el sorteo Tris.

	Zona						Puntos fuera
Dígito	A	B	C	D	E	Total	del LCS
1	9	10	10	8	11	48	0
2	11	10	16	11	8	56	0
3	5	6	13	11	10	45	0
4	13	11	11	7	10	52	0
5	8	9	12	13	7	49	0
6	10	14	16	8	9	57	0
7	12	13	9	17	5	56	0
8	5	6	9	15	8	43	0
9	15	12	13	10	10	60	0
0	4	7	6	7	10	34	1

En suma, más allá del posible incremento del error tipo I al rechazar la hipótesis de aleatoriedad en el proceso de extracción de las esferas de la urna del millar en el sorteo Tris al evaluar la aleatoriedad de diferentes rachas, lo que importa mostrar con las cartas de control de la figura 4 es que para probar la aleatoriedad de juegos de d dígitos no basta probar la igualdad de la frecuencia con la que los diferentes dígitos aparecen como resultado, como se hizo en el caso del sorteo Tris para el dígito 0 y 9 a través de las figuras 1 a 3, sino además es fundamental monitorear la rapidez con la que los diferentes dígitos aparecen, y así poder identificar rachas no aleatorias con apoyo de las cinco zonas de la carta geométrica para cada dígito, o que permitirá actuar con mayor oportunidad y de manera más preventiva. Pero como ya se ha dicho, es importante evaluar formalmente la aleatoriedad de las rachas sospechosas, para lo cual se han propuesto las opciones que representan las expresiones (19) y (20).

6. Discusión y conclusiones

Verificar la aleatoriedad y legalidad de los resultados de los juegos de azar es un asunto que toma cada día mayor importancia social, dada la existencia de estos juegos y sorteos tanto en formato tradicional como en formato electrónico o en internet. Por ello se justifica profundizar en metodologías estadísticas que ayuden a verificar tal aleatoriedad. En la revisión bibliográfica hecha para este trabajo se destaca que los juegos de números del tipo k de N han recibido más atención por parte de los estadísticos, aunque varias de las pruebas propuestas en la literatura están basadas en el estadístico χ^2, que requiere tamaños de muestra grandes para lograr buenas aproximaciones, por lo que sería deseable repensar procedimientos que no dependen de aproximaciones asintóticas, como las opciones

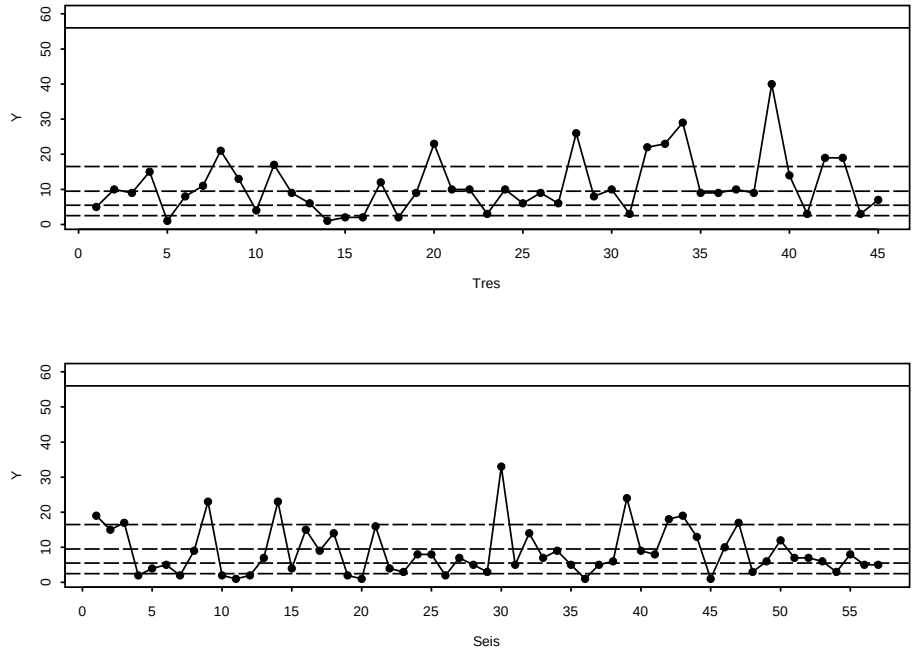


FIGURA 4: Cartas geométricas para los dígitos 3 y 6 del sorteo Tris.

bayesianas. Además es importante pensar en herramientas para monitorear en forma cotidiana este tipo de sorteos.

Por su parte, los juegos de números de d dígitos, preocupación central de este trabajo, han recibido menos atención, pero también requieren mejores formas de verificar su aleatoriedad, ya que la prueba χ^2, que es el procedimiento estándar para verificar la igualdad en la frecuencia con la que aparece como resultado cada dígito, depende de aproximaciones asintóticas y precisa tamaños de muestra grandes. Por ello, como alternativa se propuso en este artículo utilizar el procedimiento bayesiano basado en un modelo multinomial, lo que permite hacer inferencia con base en la distribución posterior $\pi(\mathbf{p} \mid X)$, que es fácilmente obtenida por simulación. Además se pueden obtener distribuciones posteriores marginales de una o varias proporciones que se quieran analizar con mayor detalle. Esta alternativa, aparte de no requerir aproximaciones asintóticas, es un procedimiento sencillo e intuitivo que permite generar inferencia sobre el valor de p_i.

El caso práctico analizado en este trabajo sobre los resultados del sorteo Tris de México ha mostrado claramente lo anterior, ya que a pesar de que la frecuencia de los diferentes dígitos parecía no ser igual (ver tabla 3), ninguna de las pruebas tradicionales de la subsección 2.2 fue capaz de evidenciar la falta de aleatoriedad, cosa que sí se logra con base en $\pi(\mathbf{p} \mid X)$ (ver figuras 1 y 2).

Para verificar la aleatoriedad de un juego de números de d dígitos, también es necesario verificar la aleatoriedad a través del tiempo. En la literatura se encontró que esto se hace con base en pruebas de independencia entre ensayos sucesivos, y que un procedimiento utilizado se basa en un estadístico χ^2 obtenido a partir de los coeficiente de autocorrelación (ver Royal Statistical Society 2002) y Box & Pierce (1970). Sin embargo, estas pruebas dependen de supuestos distribucionales y de resultados asintóticos que usualmente requieren tamaño de muestra grande. Además, cuando estas pruebas detectan falta de independencia aportan pocos elementos para diagnosticar el origen del problema. Por ello, y como alternativa se ha propuesto en este trabajo el uso de la carta de control geométrica con cinco zonas para monitorear la aleatoriedad de los juegos de número de d dígitos, ya que esto ratifica que no sólo es importante verificar la igualdad de la frecuencia de los dígitos, sino analizar la rapidez con la que aparece en los diferentes dígitos. La carta de control es intuitiva, no requiere aproximaciones asintóticas ni supuestos distribucionales adicionales; además permite actuar oportunamente cuando se detecten resultados o patrones que evidencien la falta de aleatoriedad.

De esta manera, el hecho de monitorear la variable Y, que es igual al número de sorteos para que vuelva a aparecer como resultado un mismo dígito, y llevar un registro separado para cada dígito permite detectar evidencias del desempeño del proceso, e identificar específicamente cuál es la esfera (dígito) que está teniendo problemas. Para facilitar la interpretación de la carta geométrica e incrementar la potencia para detectar rachas de puntos que se desvíen de la aleatoriedad, se propuso dividir la carta en cinco zonas con una cobertura aproximada de $0,20$ cada una (ver tabla 1).

La aplicación de la carta geométrica propuesta a los datos del sorteo Tris aportó evidencia adicional contra la falta de aleatoriedad de los dígitos 0 y 9. Además se detectaron otros problemas para otro par de dígitos (el 3 y el 6), que aunque en sus frecuencias totales no parecía que tuvieran problemas, la velocidad con la que aparecieron mostró rachas, donde en un periodo aparecían más rápido y en el otro fue más lento. Esto se validó con las pruebas estadísticas propuestas en las expresiones (19) y (20).

De esta manera, la carta geométrica propuesta resulta un instrumento útil para monitorear la aleatoriedad de juegos de azar del tipo d dígitos, tanto por sus propiedades teóricas como por su sencillez y sentido práctico.

En cierto sentido, los problemas detectados en la aleatoriedad de los resultados del sorteo Tris sorprenden porque la empresa responsable de este sorteo es una empresa pública que goza de un buen prestigio en México. Pero es claro que requiere mejores formas de asegurar la aleatoriedad de sus sorteos, como también lo concluye Coronel-Brizio et al. (2008), al detectar problemas de aleatoriedad en el sorteo Melate de esta misma empresa, que es un juego del tipo k/N.

[Recibido: julio de 2009 — Aceptado: septiembre de 2010]

Referencias

Bernardo, J. M. & Smith, A. F. M. (1994), *Bayesian Theory*, John Wiley, Chichester.

Box, G. E. P. & Pierce, D. A. (1970), 'Distribution of Residua Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models', *Journal of the American Statistical Association* **65**, 1509–1526.

Cai, Y. & Krishnamoorthy, K. (2006), 'Exact Size and Power Properties of Five Tests for Multinomial Proportions', *Communications in Statistics-Simulation and Computation* **35**, 149–160.

Clotfelter, C. T. & Cook, P. J. (2008), 'The "Gambler's Fallacy" in Lottery Play', *Management Science* **39**(12), 1521–1525.

Coronel-Brizio, H., Hernández-Montoya, A., Rapallo, F. & Scalas, E. (2008), 'Statistical Auditing and Randomness Test of Lotto k/N-Type Games', *Physica A* **387**(25), 6385–6390.

Finkelstein, M. (1995), 'Estimating the Frequency Distribution of the Numbers bet on the California Lottery', *Applied Mathematics and Computation* **69**(2-3), 195–207.

Genest, C., Lockhart, R. A. & Stephens, M. A. (2002), 'χ^2 and the lottery', *Journal of the Royal Statistical Society: Series D (The Statistician)* **51**(2), 243–257.

Goodman, L. A. (1965), 'On Simultaneous Confidence Intervals for Multinomial Proportions', *Technometrics* **7**, 247–254.

Gutiérrez, H. (2006), 'Cartas de control bayesianas para atributos y el tamaño de subgrupo grande en la carta p', *Revista Colombiana de Estadistica* **29**(2), 163–180.

Gutiérrez, H. & de la Vara, R. (2009), *Control estadístico de calidad y seis sigma*, segunda edn, McGraw-Hill, México.

Hou, C. D., Chiang, J. & Tai, J. J. (2003), 'A Family of Simultaneous Confidence Intervals for Multinomial Proportions', *Computational Statistics & Data Analysis* **43**, 29–45.

Joe, H. (1987), 'An Ordering of Dependence for Distribution of k-tuples, with Applications to Lotto Games', *The Canadian Journal of Statistics* **15**(3), 227–238.

Joe, H. (1990), 'A Winning Strategy for Lotto Games?', *The Canadian Journal of Statistics* **18**(3), 233–244.

Joe, H. (1993), 'Tests of Uniformity for Sets of Lotto Numbers', *Statistics & Probability Letters* **16**(3), 181–188.

Johnson, R. & Klotz, J. (1993), 'Estimating hot Numbers and Testing Uniformity for the Lotto', *Journal of the American Statistical Association* **88**(422), 662–668.

Koning, R. H. & Vermaat, M. B. (2002), A Probabilistic Analysis of the Dutch Lotto, Technical report, University of Groningen. Research report.

Ljung, G. M. & Box, G. E. P. (1978), 'On a Measure of Lack of Fit in Time Series Models', *Biometrika* **65**(2), 297–303.

Nass, C. A. G. (1959), 'The 2-test for Small Expectations in Contingency Tables, with Special Feference to Accidents and Absenteeism', *Biometrika* **46**, 365–385.

Percy, D. F. (2006), Bayesian Methods for Testing the Randomness of Lottery Draws, Technical report, University of Salford, Centre for Operational Research and Applied Statistics. Research report.

Quesenberry, C. P. & Hurst, D. C. (1964), 'Large-Sample Simultaneous Confidence Intervals for Multinomial Proportions', *Technometrics* **6**, 191–195.

Royal Statistical Society (2000), Reports on the Randomness of U. K. Lottery Games, Technical report. Obtenidos en septiembre de 2007.
*http://www.natlotcomm.gov.uk

Royal Statistical Society (2002), Reports on the Randomness of U. K. Lottery Games, Technical report. Obtenidos en septiembre de 2007.
*http://www.natlotcomm.gov.uk

Schwartz, D. G. (2003), *Suburban Xanadu: The Casino Resort on the Las Vegas Strip and Beyond*, Routledge, New York.

Schwartz, D. G. (2006), *Roll the Bones: The History of Gambling*, Gotham Books, New York.

Sison, C. P. & Glaz, J. (1995), 'Simultaneous Confidence Intervals and Sample size Determination for Multinomial Proportions', *Journal of the American Statistical Association* **90**(429), 366–369.

Stern, H. & Cover, T. M. (1989), 'Maximum Entropy and the Lottery', *Journal of the American Statistical Association* **84**, 980–985.

Teo, C. P. & Leong, S. M. (2002), 'Managing Risk in a Four-Digit Number Game', *SIAM REVIEW* **44**(4), 601–615.

University of Salford (2004), Randomness of the Lotto Draws: Summary of Findings, Technical report, Centre for the Study of Gambling. Obtenido en septiembre de 2007.
*http://www.natlotcomm.gov.uk

University of Salford (2005*a*), Randomness of the Lotto Lucky Dip, Technical report, Centre for the Study of Gambling. Obtenido en septiembre de 2007. *http://www.natlotcomm.gov.uk

University of Salford (2005*b*), Randomness of thunderball draws, Technical report, Centre for the Study of Gambling. Obtenido en septiembre de 2007. *http://www.natlotcomm.gov.uk

Yang, Z., Xie, M., Kuralmani, V. & Tsui, K. L. (2002), 'On the Performance of Geometric Charts with Estimated Control Limits', *Journal of Quality Technology* **34**(4), 448–459.

Revista Colombiana de Estadística
Diciembre 2010, volumen 33, no. 2, pp. 191 a 218

Estimación probabilística del cambio climático en Venezuela mediante un enfoque bayesiano

Probabilistic Estimation of Climate Change in Venezuela using a Bayesian approach

Alexis Durán[1,a], Lelys Guenni[2,b]

[1]Departamento de Estadística, Facultad de Ciencias, Universidad Experimental Ezequiel Zamora, San Carlos, Venezuela

[2]Departamento de Cómputo Científico y Estadística, División de Ciencias Físicas y Matemáticas, Universidad Simón Bolívar, Caracas, Venezuela

Resumen

El problema del cambio climático es uno de los grandes problemas ambientales que enfrenta la humanidad, ya que ligeras variaciones en las variables climáticas pueden traer graves consecuencias en las actividades económicas y el bienestar humano en general. Hoy en día los modelos de circulación general (MCG) de la atmósfera son la principal herramienta para estudiar los cambios climáticos. El Ministerio del Ambiente y de los Recursos Naturales (MARN) lideró en el año 2005 la Primera Comunicación Nacional en Cambio Climático de Venezuela, utilizando salidas de 16 MCGs a escala global (resolución de $5^\circ \times 5^\circ$), cuyas proyecciones estiman incrementos para la temperatura y disminución en la precipitación para los próximos años. Cada MCG arroja diferentes resultados generando incertidumbre en la señal del cambio climático futuro. Este trabajo utiliza un enfoque Bayesiano y una extensión del método *Reliability Ensemble Average* (REA) (Tebaldi, Smith, Nychka & Mearns 2005), combinando las salidas (presente y futura) de precipitación y temperatura de los 16 MCG con observaciones de las condiciones climáticas actuales, con el fin de determinar las distribuciones de probabilidad del cambio climático futuro para estas dos variables climáticas en nueve regiones de Venezuela. Para el estudio se toman en cuenta dos criterios: *sesgo*, el cual considera la diferencia entre las salidas de los modelos y el clima actual, y *convergencia*, que cuantifica las diferencias en los cambios simulados por los múltiples modelos del clima futuro. El principal resultado obtenido del trabajo es que aún existe considerable incertidumbre en las proyecciones de los MCG, ya que estos no incluyen todos los aspectos sobre el funcionamiento del sistema climático. También se pudo establecer que mientras menor sea la variabilidad natural de la variable climática, más

[a]Profesor instructor. E-mail: duranalexis@yahoo.com
[b]Profesor titular. E-mail: lbravo@cesma.usb.ve

efectiva será su proyección.

Palabras clave: estimación Bayes, inferencia posterior, modelo probabilístico.

Abstract

The changing climate is one of the main environmental problems facing humanity, since slight variations in the climate variables might have terrible consequences in the economic activities and human well-being. Nowadays atmospheric Global Circulation Models (GCMs) are the main tools to study changing climate. The Ministry of Environment and Natural Resources (MENR) led in 2005 the First Communication in Climate Change of Venezuela, using the outputs of 16 GCMs at a global scale (resolution of $5° \times 5°$) whose projections estimate increasing temperature and diminishing precipitation in the coming years. Each GCM gives different results, generating uncertainty in the future changing climate signal. This work uses a Bayesian approach and an extension of the *Reliability Ensemble Average* (REA) (Tebaldi et al. 2005) method, combining the outputs (present and future) of precipitation and temperature of the 16 GCMs with observations of present climate conditions, to determine the probability distributions of future changing climate change for these two climate variables in 9 regions in Venezuela. For this study, two criteria are used: *bias*, which considers the difference between the model outputs and the present climate; and *convergence*, which quantifies the differences among the simulated changes of future climate by multiple models. The main result of this work is that a large amount of uncertainty still exists in the GCMs projections, since they as yet do not include all aspects of the climate system functioning. It was also concluded that the lower the natural variability in the climate variable, the more effective is its projection.

Key words: Bayes estimation, Probabilistic model, Posterior inference.

1. Introducción

La importancia que tiene el estudio del cambio climático para Venezuela viene dada por la circunstancia de ser un país cuya economía es altamente dependiente de la producción y exportación de petróleo (la actividad petrolera aportó el 16,8 % del PIB y representó el 87,2 % de las exportaciones de bienes en 2004). Además Venezuela es un país de gran diversidad biológica, ecosistemas frágiles, poseedor de costas bajas y territorios insulares vulnerables, que no cuenta con la capacidad suficiente para la debida atención de contingencias derivadas de la ocurrencia de fenómenos meteorológicos extremos.

La capacidad actual del país para enfrentar la variabilidad climática natural no es muy alta, ya que existen debilidades en las áreas de medición sistemática de los elementos climáticos, escasez de personal especializado en el área de aplicaciones prácticas de la información climática y una débil integración interinstitucional para la organización de las actividades productivas en función de aprovechar al máximo las potencialidades del clima y reducir los riesgos asociados a éste. De

no modificarse tal situación, el nivel de vulnerabilidad o sensibilidad de Venezuela ante el cambio climático se incrementará aún más. Actualmente, en los grandes centros de investigación climática a nivel mundial se hacen proyecciones del cambio climático con el fin de buscar las medidas de adaptación más viables a este gran problema. Estas proyecciones son discutidas ampliamente en el último reporte del Panel Intergubernamental sobre el Cambio Climático (IPCC 2007).

Hoy en día, los modelos de circulación general (MCG) de la atmósfera son las principales herramientas que existen para hacer proyecciones sobre el cambio climático a nivel mundial (Benioff, Guill & Lee 1996). Los MCG son representaciones numéricas tridimensionales, que se emplean para simular el comportamiento del sistema climático global a gran escala (incluyendo la atmósfera, los océanos, la biosfera, la criosfera y la superficie terrestre), pero poseen una resolución espacial insuficiente o inadecuada para aplicar sus resultados como base de los estudios de impacto climático a nivel local o regional. Uno de los mayores inconvenientes con los modelos MCG radican en que para algunas regiones se obtienen resultados diferentes que generalizan incertidumbre en la señal del cambio climático futuro, y por consiguiente, es difícil hacer proyecciones confiables acerca de estas condiciones climáticas futuras (Vera, Silvestri, Liebmann & González 2006).

En la primera comunicación nacional sobre cambio climático (CNCC) para Venezuela (MARN 2005), se realizó un estudio sobre algunas de las implicaciones ambientales más generales que se derivan del cambio de los patrones espacio-temporales de la precipitación y la temperatura, lo que repercute en aspectos tales como el tipo climático de grandes áreas del país, la disponibilidad hídrica desde el punto de vista climático y el confort humano y animal. Para este estudio se analizó el comportamiento de 16 modelos acoplados de circulación general atmósfera-océano (MACGAO) utilizados en el tercer reporte del IPCC (IPCC-TAR), que fueron incluidos en la herramienta MAGICC/SCENGEN (Model for the Assessment of Greenhouse Induced Climate Change / Scenario Generator), desarrollado por la Climate Research Unit (CRU, University of East Anglia, UK) (Hulme, Wigley, Barrow, Raper, Centella, Smith & Chipanshi 2000), bajo cinco escenarios de emisión de gases del efecto invernadero (IS92A, SRESA1, SRESA2, SRESB1, SRESB2) (IPCC 2001) y tres niveles de sensibilidad climática, a fin de considerar con amplitud las diferentes incertidumbres asociadas al proceso de simulación del clima futuro. Basándose en la idea de que el clima en el siglo XXI debería continuar la tendencia mostrada en el siglo XX, incluyendo las características del impacto del fenómeno El Niño en el país, y aceptando que se espera en el futuro un incremento en la frecuencia de ocurrencia de este evento, se observó que los modelos que mejor se adaptan a estas condiciones fueron el modelo UKTR (desarrollado por el United Kingdom Meteorological Office, Inglaterra) y el modelo CCC-EQ (desarrollado por el Canadian Center for Climate Modelling and Analysis, Canadá) (MARN 2005), en donde las proyecciones obtenidas revelan un importante incremento en la temperatura y una disminución de las precipitaciones (tendencia general).

Esta investigación desarrollada previamente en Durán (2008), analiza los cambios de precipitación y temperatura producidos por todos los MCG utilizados en la primera CNCC para Venezuela mediante un enfoque bayesiano, que puede ser

considerado una extensión del método *Reliability Ensemble Average* (REA) propuesto por Nychka & Tebaldi (2002), el cual consiste en derivar la distribución de probabilidad a posteriori de las proyecciones presente y futura de estas variables climáticas a escala regional en puntos de grilla de tamaño $5° \times 5°$. Para lograrlo, se combinan las salidas de los distintos MCG y se hace inferencia de la distribución de probabilidad a posteriori de las proyecciones presente y futura utilizando métodos MCMC (Markov Chain Monte Carlo), evaluando la incertidumbre asociada a cada uno de los MCG incluidos en el estudio. El método toma en cuenta dos criterios: el *sesgo*, dado por la diferencia entre las salidas del modelo para el clima presente (1960-1989) y el promedio (ponderado por las precisiones a posteriori) del clima actual, y la *convergencia*, dada por la diferencia entre las salidas del modelo del clima futuro (años 2025, 2050 y 2100) y el promedio (ponderado por las precisiones a posteriori) del clima futuro. Mientras menor sea el valor de estas dos cantidades, mayor será la efectividad del modelo en cuestión para proyectar el cambio climático futuro. Para este estudio se contó con las salidas del MAGICC/SCENGEN de 16 MCG bajo una sensibilidad climática intermedia y un escenario SRESA2 (recomendado y suministrado por el MARN) que describe un mundo muy heterogéneo donde la tasa de crecimiento demográfico está siempre en aumento durante todo el siglo XXI; el desarrollo económico está orientado regionalmente (poca globalización) y tanto el crecimiento económico per cápita como el cambio tecnológico son muy lentos y fragmentados. Por otra parte, se obtuvieron los datos de precipitación y temperatura observados a partir de la Red Regional Digital de Datos Hidrometeorológicos para América del Sur, América Central y el Caribe (R-Hydronet[1]) del año 1960 a 1989, los cuales fueron recalculados a 9 grillas de $5° \times 5°$ que abarcan todo el territorio nacional, como se muestra en la figura 1.

La estructura del trabajo es como sigue: en la sección 2 se introduce la metodología del *Reliability Ensemble Average* (REA) utilizada para evaluar la efectividad o fiabilidad de los MCG incluidos en el análisis. Allí se explica cómo se combinan las salidas de todos los modelos climáticos mediante un enfoque bayesiano, lo cual permite estimar las distribuciones a posteriori de los cambios proyectados en las variables temperatura y precipitación. En la sección 3 se resumen los datos de precipitación y temperatura utilizados para evaluar las salidas de los modelos climáticos durante el siglo XX y se describen los modelos climáticos utilizados en este estudio. En la sección 4 se presentan los resultados de los modelos probabilísticos propuestos para representar las salidas de los modelos climáticos y los datos observados, así como también los resultados del método REA en la selección de los modelos más eficientes en proyectar los cambios de temperatura y precipitación. Finalmente se discuten los resultados obtenidos y se presentan las conclusiones más importantes de este trabajo.

[1] http://www.r-hydronet.sr.unh.edu

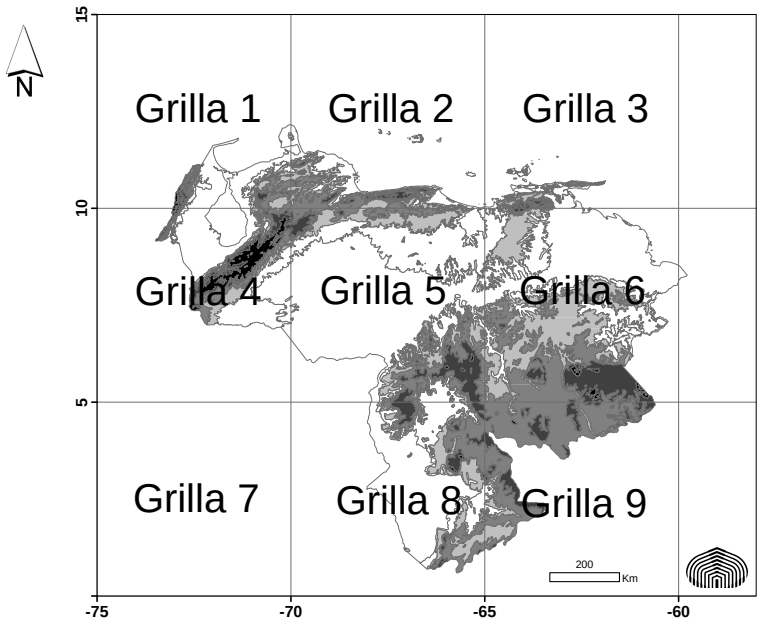


FIGURA 1: Mapa de Venezuela dividido en nueve grillas de acuerdo con las salidas de los MCG.

2. Combinación de simulaciones de múltiples modelos climáticos

El enfoque bayesiano utilizado en esta investigación para estudiar los cambios futuros de la temperatura y la precipitación, puede considerarse como una extensión del método *Reliability Ensemble Average* (REA), el cual fue publicado por primera vez en 2003 por Giorgi & Mearns (2002), y recientemente utilizado por Tebaldi et al. (2005). Este método combina todas las salidas disponibles de los MCG con los datos observados de las variables climáticas para simular las distribuciones a posteriori de los cambios climáticos utilizando métodos MCMC (Markov Chain Monte Carlo); en particular se utiliza el muestreador de Gibbs o *Gibbs sampler* (Gelfand & Smith 1990). El método brinda dos criterios que permiten evaluar la efectividad o fiabilidad de los MCG incluidos en la investigación, estos son: el sesgo y la convergencia. Ambos términos se describen a continuación.

2.1. Definiciones de sesgo y convergencia

2.1.1. Sesgo

La idea se basa en suponer que un modelo climático es más confiable si es capaz de proyectar correctamente los cambios climáticos observados en el siglo XX, es

decir, este criterio permite evaluar la incertidumbre que existe en los MCG para proyectar el clima presente. La manera de calcular el sesgo es consiguiendo una media (ponderada por las precisiones a posteriori) entre las salidas de los MCG del clima presente y los datos observados de la variable climática de estudio, para luego calcular la diferencia entre las salidas de cada modelo y esta media ponderada. En forma matemática se puede escribir de la siguiente manera:

$$\widetilde{\mu} = \frac{X_0 * \tau_0 + \sum_{i=1}^{n} X_i * \tau_i}{\tau_0 + \sum_{i=1}^{n} \tau_i} \tag{1}$$

$$\text{Sesgo}_i = X_i - \widetilde{\mu}$$

El subíndice i varía de 1 a n, donde $n = 16$ corresponde al total de modelos climáticos utilizados, X_i es el promedio de la variable climática del clima presente para el modelo i con τ_i como precisión, X_0 es el promedio de los datos observados con τ_0 como precisión y $\widetilde{\mu}$ es la media (ponderada) del clima presente. Esta notación utiliza la precisión como el inverso de la varianza.

2.1.2. Convergencia

En este caso se piensa que un modelo es más confiable si su capacidad para proyectar futuros cambios climáticos tiene gran similitud con las señales y magnitudes de los cambios proyectados por otros modelos; en otras palabras, la convergencia permite evaluar la incertidumbre entre los MCG para proyectar el clima futuro. La manera de calcular este criterio es similar a la del sesgo, con la diferencia de que no se toman en cuenta los datos observados, y en lugar de calcular los promedios de las proyecciones de los MCG del clima presente se toman los valores futuros Y_i. En forma matemática, la convergencia para el modelo i se puede expresar así:

$$\widetilde{\nu} = \frac{\sum_{i=1}^{n} Y_i * \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{2}$$

$$\text{Convergencia}_i = Y_i - \widetilde{\nu}$$

De nuevo el subíndice i varía de 1 a n, con $n = 16$ que corresponde al número total de modelos, Y_i son las proyecciones promedio del clima futuro para el modelo i con λ_i como precisión y $\widetilde{\nu}$ es la media (ponderada) del clima futuro.

2.2. Modelo para el caso independiente

En esta parte del estudio no se asume ningún tipo de relación entre el clima presente y futuro. En este caso se consideran X_{ij} y Y_{ij} la precipitación o temperatura simulada presente y futura, respectivamente, por los distintos MCG. El subíndice i va de 1 a 16, correspondiente a cada uno de los modelos, y j va de 1 hasta 9, correspondiente a las grillas en las cuales se ha dividido Venezuela (ver figura 1). El análisis se lleva a cabo mensualmente, con valores de precipitación y temperatura promediados por trimestre (diciembre-enero-febrero, DEF; marzo-abril-mayo, MAM; junio-julio-agosto, JJA, y septiembre-octubre-noviembre, SON).

2.2.1. Verosimilitud de los datos

Dado que el estudio descriptivo sugiere que los datos observados siguen una distribución normal, se utiliza esta distribución para calcular la verosimilitud de los datos. En consecuencia se tiene:

$$X_{ij} \sim N\left(\mu_j, \tau_{ij}^{-1}\right) \tag{3}$$

$$Y_{ij} \sim N\left(\nu_j, (\theta\tau_{ij})^{-1}\right) \tag{4}$$

μ_j y ν_j representan el promedio de la variable climática en el presente y futuro, respectivamente, para una determinada región (grilla j) y trimestre; por tanto, el parámetro de interés será $\Delta_j = \nu_j - \mu_j$, que mide el cambio climático esperado en la grilla j, por lo que un valor positivo indicaría un aumento promedio de la variable climática en los próximos años y viceversa para los valores negativos. El parámetro τ_{ij} es la precisión del modelo i en la grilla j (inverso de la varianza) para la variable climática proyectada en el presente y $\theta\tau_{ij}$ es la precisión del modelo i en la grilla j para la variable climática proyectada en el futuro. Nótese que la diferencia entre estas precisiones es el factor θ. En cuanto a la verosimilitud de los datos climáticos observados se asume:

$$X_0 \sim N\left(\mu_j, \tau_0^{-1}\right) \tag{5}$$

Obsérvese que las distribuciones de X_0 y X_{ij} tienen en común el parámetro μ_j lo cual es aceptable, porque ambas son modelos probabilísticos del clima presente, pero las precisiones sí son diferentes, ya que para X_0 la precisión depende de la variabilidad natural, región y longitud de las observaciones, mientras que para X_{ij} las precisiones τ_{ij} dependen de las aproximaciones numéricas y dr las parametrizaciones utilizadas para modelar los parámetros climáticos, así como también de las resoluciones espaciales y temporales de los modelos empleados en este estudio.

Los parámetros de los modelos definidos en (3) y (4) son estimados utilizando el paradigma bayesiano. Para ello es necesario definir las distribuciones de probabilidad a priori para el vector de parámetros $(\mu_j, \nu_j, \tau_{ij}, \theta)$.

2.2.2. Distribuciones a priori de los parámetros

Para las distribuciones a priori de los parámetros $(\mu_j, \nu_j, \tau_{ij}, \theta)$ de los modelos definidos en (3) y (4), se tomaron distribuciones previas no informativas como sigue:

- Se asume una distribución Gamma para τ_{ij} con parámetros a, b conocidos. Esta es la suposición usual para la precisión de un modelo con distribución normal, ya es una distribución conjugada para este modelo y la verosimilitud tiene la misma forma general que esta distribución cuando se asume como una función de la precisión. De esta manera se facilitan los cálculos y la inferencia a posteriori, ya que la distribución a posteriori para τ_{ij} será también una distribución Gamma al condicionarla en el resto de los parámetros.

- La distribución a priori para θ al igual que para τ_{ij} se asume Gamma con parámetros c, d conocidos. Para utilizar distribuciones a priori difusas se asume $a = b = c = d = 0,0001$. Esto implica que la media de la distribución será igual a 1 y la varianza será igual a 10.000, por lo que se supone muy poco conocimiento a priori sobre los parámetros del modelo, y se le da un mayor énfasis a la información que proveerán los datos para la inferencia a posteriori sobre los parámetros.

- Para la distribución a priori de μ_j y ν_j se seleccionó una distribución uniforme restringida a valores finitos, pero con un amplio rango que cubra los posibles valores sobre estos parámetros. Nuevamente en este caso la escogencia de una distribución uniforme facilita los cálculos y asegura que las distribuciones a posteriori para μ_j y ν_j sean distribuciones normales al condicionarlas en el resto de los parámetros.

Mayor información sobre las diferentes especificaciones de las distribuciones a priori puede ser consultada en Migon & Gamerman (1999).

2.2.3. Distribuciones a posteriori

Aplicando el teorema de Bayes y asumiendo independencia entre los parámetros de interés a priori, la distribución posterior del vector de parámetros $(\mu_j, \nu_j, \tau_{ij}, \theta)$ condicional en los datos $X_0, \mathbf{X}, \mathbf{Y}$ queda proporcional a:

$$\begin{aligned} p(\mu_j, \nu_j, \tau_{ij}, \theta \mid X_0, \mathbf{X}, \mathbf{Y}) \propto \prod_{j=1}^{m}\prod_{i=1}^{n} \tau_{ij}^{a-1} \exp(b\tau_{ij})\tau_{ij}\theta^{\frac{1}{2}} \\ \exp\left\{-\frac{\tau_{ij}}{2}\left[(X_{ij}-\mu_j)^2 + \theta(Y_{ij}-\nu_j)^2\right]\right\} \\ \times \theta^{c-1}\exp\left\{-\left[d\theta + \frac{\tau_0}{2}(X_0-\mu_j)^2\right]\right\} \quad (6) \end{aligned}$$

donde $m = 9$ y $n = 16$. Como se puede observar, la distribución a posteriori no forma parte de ninguna familia de distribuciones paramétricas, por lo que no es posible hacer inferencia de estas distribución en forma analítica. Este también es el caso para las distribuciones marginales a posteriori de los parámetros (Tebaldi et al. 2005), las cuales no pueden ser obtenidas mediante integrales de forma cerrada. Por tanto se utilizan los métodos MCMC para hacer inferencia, generando muestras de la distribución a posteriori para todos parámetros desconocidos. La técnica por utilizar en esta investigación, como se dijo anteriormente, es el muestreador de Gibbs (Gelfand & Smith 1990). Para su implementación es necesario derivar las distribuciones condicionales completas de todos los parámetros.

Por simplicidad se consideran las ecuaciones independientes de cada grilla j, ya que las distribuciones son las mismas para cada una de ellas y lo que cambia son los datos para ser utilizados en cada grilla. De aquí en adelante utilizamos un solo subíndice para X_{ij}, Y_{ij} y τ_{ij}, esto es, X_i, Y_i y τ_i. Se considera que $\mathbf{X} = (X_1, \ldots, X_n)$ y $\mathbf{Y} = (Y_1, \ldots, Y_n)$ son los datos simulados del clima presente y futuro para todos los modelos en alguna grilla particular j. También se elimina el subíndice j para μ_j

y ν_j y utilizamos μ y ν. La distribución a posteriori de todo el vector de parámetros para cualquier grilla j tiene la forma:

$$\begin{array}{l} p(\mu,\nu,\tau_i,\theta \mid X_0,\mathbf{X},\mathbf{Y}) \propto \tau_i^{na}\theta^{c+\frac{n}{2}-1} \\ \exp\left\{\sum_{i=1}^{n}\left[b\tau_i - \frac{\tau_i}{2}((X_i-\mu)^2+\theta(Y_i-\nu)^2)\right] - \left[d\theta + \frac{\tau_0}{2}(X_0-\mu)^2\right]\right\} \end{array} \tag{7}$$

2.2.4. Distribuciones condicionales a posteriori

- Promedio de las proyecciones del clima presente (μ):

$$p(\mu \mid \nu,\tau_i,\theta,X_0,\mathbf{X},\mathbf{Y}) \propto N\left(\widetilde{\mu},(\sum_{i=0}^{n}\tau_i)^{-1}\right) \tag{8}$$

  donde

$$\widetilde{\mu} = \frac{\sum_{i=0}^{n}\tau_i X_i}{\sum_{i=0}^{n}\tau_i}$$

- Promedio de las proyecciones del clima futuro (ν):

$$p(\nu \mid \mu,\tau_i,\theta,X_0,\mathbf{X},\mathbf{Y}) \propto N\left(\widetilde{\nu},(\theta\sum_{i=1}^{n}\tau_i)^{-1}\right) \tag{9}$$

  donde

$$\widetilde{\nu} = \frac{\sum_{i=1}^{n}\tau_i Y_i}{\sum_{i=1}^{n}\tau_i}$$

- Distribución condicional para θ

$$p(\theta \mid \mu,\nu,\tau_i,X_0,\mathbf{X},\mathbf{Y}) \propto \Gamma\left(\frac{n}{2}+c,\frac{1}{2}\sum_{i=1}^{n}\tau_i(Y_i-\nu)^2+d\right) \tag{10}$$

- Distribución condicional para la precisión del modelo i (τ_i):

$$p(\tau_i \mid \mu,\nu,\theta,X_0,\mathbf{X},\mathbf{Y}) \propto \Gamma\left(a+1,\frac{1}{2}\left[(X_i-\mu)^2+\theta(Y_i-\nu)^2\right]+b\right) \tag{11}$$

2.3. Modelo para el caso dependiente

En este caso se propone un modelo en el que existe una relación lineal entre clima presente y futuro. Al igual que en el caso anterior, las ecuaciones son independientes de la grilla por ser las mismas en cada una de ellas variando solamente los datos para ser utilizados para la grilla j. Se emplea además la misma nomenclatura que la empleada anteriormente.

2.3.1. Verosimilitud de los datos

El modelo propuesto en este caso para el clima presente X_i y el clima futuro Y_i del modelo i en cualquier grilla j es el siguiente:

$$X_i = \mu + \eta_i \tag{12}$$

$$Y_i = \nu + \beta(X_i - \mu) + \frac{\varepsilon_i}{\sqrt{\theta}} \tag{13}$$

donde η_i y ε_i son aproximadamente normales con media 0 y precisión τ_i. En este modelo se establece una relación lineal entre $(Y_i - \nu)$ y $(X_i - \mu)$, que está definida por el parámetro β. Esta relación puede ser positiva (valores positivos de β) o negativa (valores negativos de β). El valor de β también interviene en aspectos como: la correlación entre X_i y Y_i, la señal del cambio climático producida por el modelo i y en la cuantificación de $(X_i - \mu)$ dentro del cálculo del sesgo. La verosimilitud de X_i es igual a la del caso independiente, mientras que la verosimilitud de Y_i queda de la forma:

$$Y_i \mid \nu, \theta, \mu, \beta, X_i, \tau_i \propto \sqrt{\theta\tau_i}\exp\left\{-\frac{1}{2}\theta\tau_i(Y_i - \nu - \beta(X_i - \mu))^2\right\} \tag{14}$$

Nótese que si β toma el valor cero, el modelo se reduce al del caso independiente. Como en el caso anterior, el parámetro de interés es $\Delta_j = \nu_j - \mu_j$, que es el cambio climático esperado en la grilla j.

2.3.2. Distribución a priori de los parámetros

Todos los parámetros de este caso tienen la misma distribución a priori que en el caso independiente, y para el nuevo parámetro β se asume una previa no informativa uniforme en el intervalo $(-1, 1)$.

2.3.3. Distribución posterior

Al aplicar el teorema de Bayes, la distribución posterior del vector de parámetros $(\mu, \nu, \tau_i, \beta, \theta)$ queda proporcional a:

$$p(\mu, \nu, \tau_i, \beta, \theta \mid X_0, \mathbf{X}, \mathbf{Y}) \propto \tau_i^{na+1-1}\theta^{c+\frac{n}{2}} \exp\Big\{\sum_{i=1}^{n}\Big[-b\tau_i - \frac{\tau_i}{2}((X_i - \mu)^2 + \theta(Y_i - \nu - \beta(X_i - \mu))^2)\Big] - \Big[d\theta + \frac{\tau_0}{2}(X_0 - \mu)^2\Big]\Big\} \tag{15}$$

En forma similar al caso independiente, se puede observar que tanto la distribución posterior como las distribuciones marginales de los parámetros de interés no forman parte de alguna familia de distribuciones paramétricas, por lo que no es posible hacer inferencia de estas distribuciones en forma analítica. Por consiguiente, también se utiliza el muestreador de Gibbs para hacer inferencia sobre todos los parámetros desconocidos. Para ello se necesitan las distribuciones condicionales completas de los parámetros, las cuales se dan a continuación.

2.3.4. Distribuciones condicionales a posteriori

- Promedio de las proyecciones del clima presente (μ):

$$p(\mu \mid \nu, \tau_i, \theta, X_0, \mathbf{X}, \mathbf{Y}) \propto N\left(\widetilde{\mu}, \left(\sum_{i=0}^{n} \tau_i + \theta\beta^2 \sum_{i=1}^{n} \tau_i\right)^{-1}\right) \tag{16}$$

donde

$$\widetilde{\mu} = \frac{\sum_{i=0}^{n} \tau_i X_i - \theta\beta \sum_{i=1}^{n} \tau_i (Y_i - \nu - \beta X_i)}{\sum_{i=0}^{n} \tau_i + \theta\beta^2 \sum_{i=1}^{n} \tau_i}$$

Como se observa en la distribución (16), la precisión y media del clima presente dependen de los siguientes factores: la variabilidad natural y la efectividad de los MCG en proyectar el clima presente y futuro. La variabilidad natural depende de las características climáticas de la zona o grilla, reflejada en la precisión de los datos observados (τ_0). La efectividad de las proyecciones del clima presente y futuro depende de las aproximaciones numéricas y de las parametrizaciones de los procesos físicos utilizados en los modelos climáticos. Esto se refleja en las precisiones de cada uno de los modelos (τ_i). Además, para el clima futuro, la efectividad depende de β, que como se dijo anteriormente mide la asociación lineal entre clima presente y futuro.

- Promedio de las proyecciones del clima futuro (ν):

$$p(\nu \mid \mu, \tau_i, \theta, X_0, \mathbf{X}, \mathbf{Y}) \propto N\left(\widetilde{\nu}, \left(\theta \sum_{i=1}^{n} \tau_i\right)^{-1}\right) \tag{17}$$

donde

$$\widetilde{\nu} = \frac{\sum_{i=1}^{n} \tau_i (Y_i - \beta(X_i - \mu))}{\sum_{i=1}^{n} \tau_i}$$

- Distribución condicional para θ

$$p(\theta \mid \mu, \nu, \tau_i, X_0, \mathbf{X}, \mathbf{Y}) \propto \Gamma\left(\frac{n}{2} + c, \frac{1}{2} \sum_{i=1}^{n} \tau_i (Y_i - \nu - \beta(X_i - \mu))^2 + d\right) \tag{18}$$

- Distribución condicional para las precisiones del modelo i (τ_i):

$$p(\tau_i \mid \mu, \nu, \theta, X_0, \mathbf{X}, \mathbf{Y}) \propto \Gamma\left(a + 1, \frac{1}{2}\left[(X_i - \mu)^2 + \theta(Y_i - \nu - \beta(X_i - \mu))^2\right] + b\right) \tag{19}$$

Toda la inferencia bayesiana de este trabajo está implementada en el software R (R Development Core Team 2007), utilizando los datos disponibles descritos anteriormente, mediante el método MCMC (Markov Chain Monte Carlo)[2]. Este método fue utilizado por Tebaldi et al. (2005) para la cuantificación de la incertidumbre en las proyecciones de cambio climático a nivel regional.

[2]http://www.cgd.ucar.edu/ nychka/REA/

3. Datos climáticos para Venezuela en años recientes

3.1. Datos históricos

En esta parte del trabajo se describen las principales características de la precipitación y temperatura para Venezuela, a partir de los datos observados extraídos de la Red Regional Digital de Datos Hidrometeorológicos para América del Sur, América Central y el Caribe (R-Hydronet) para el periodo 1960 a 1989, los cuales fueron promediados por trimestre y recalculados para nueve grillas de $5° \times 5°$ que abarcan todo el territorio nacional (ver figura 1). Este análisis previo ayuda a comprender de una mejor manera los resultados que se muestran más adelante.

En general, de acuerdo con el régimen estacional de la precipitación en Venezuela, se pueden considerar cuatro trimestres que son: diciembre-enero-febrero (DEF, época seca), marzo-abril-mayo (MAM, transición de época seca a lluvia), junio-julio-agosto (JJA, época de lluvia) y septiembre-octubre-noviembre (SON, transición de época de lluvia a seca). Con base en esta partición del año, se presentan los siguientes análisis.

3.1.1. Temperatura

En la tabla 1 se resume la información de los datos observados de temperatura trimestral promedio para cada grilla rectangular definida en la figura 1.

Tabla 1: Promedio de la temperatura ($\pm$ 1 desviación estándar) en ℃ para Venezuela durante el periodo 1960-1989.

	Trimestre				
Grilla	DEF	MAM	JJA	SON	MEDIA
1	25,42 $\pm$ 0,46	26,37 $\pm$ 0,44	26,93 $\pm$ 0,32	26,18 $\pm$ 0,47	26,23 $\pm$ 0,42
2	24,60 $\pm$ 0,44	25,85 $\pm$ 0,49	26,03 $\pm$ 0,33	26,10 $\pm$ 0,33	25,64 $\pm$ 0,40
3	24,61 $\pm$ 0,47	25,85 $\pm$ 0,55	26,02 $\pm$ 0,45	26,03 $\pm$ 0,42	25,63 $\pm$ 0,47
4	23,96 $\pm$ 0,46	24,47 $\pm$ 0,52	23,78 $\pm$ 0,32	23,76 $\pm$ 0,31	23,99 $\pm$ 0,40
5	26,45 $\pm$ 0,44	27,21 $\pm$ 0,69	25,18 $\pm$ 0,35	25,94 $\pm$ 0,32	26,20 $\pm$ 0,45
6	24,71 $\pm$ 0,42	25,75 $\pm$ 0,57	24,93 $\pm$ 0,33	25,58 $\pm$ 0,45	25,24 $\pm$ 0,44
7	25,48 $\pm$ 0,40	25,27 $\pm$ 0,39	23,97 $\pm$ 0,34	24,89 $\pm$ 0,29	24,90 $\pm$ 0.36
8	26,12 $\pm$ 0,43	25,96 $\pm$ 0,48	24,72 $\pm$ 0,32	25,67 $\pm$ 0,33	25,62 $\pm$ 0,39
9	25,38 $\pm$ 0,44	25,37 $\pm$ 0,59	24,34 $\pm$ 0,33	25,39 $\pm$ 0,45	25,12 $\pm$ 0,45
MEDIA	25,19 $\pm$ 0,44	25,79 $\pm$ 0,52	25,10 $\pm$ 0,34	25,50 $\pm$ 0,37	25,40 $\pm$ 0,42

A partir de la tabla 1, se observa que los trimestres de transición son los que en promedio registran valores de temperatura más altos, seguidos por la época seca y por último la época de lluvia, en donde las grillas con mayor temperatura promedio fueron las grillas 1 y 5 con $26,23$ ℃ y $26,20$ ℃ respectivamente. La grilla con los valores más bajos de temperatura fue la 4, con $23,99$ ℃, la cual abarca gran parte de la zona andina. En cuanto a la dispersión, se puede ver que el trimestre MAM posee la desviación estándar más alta, seguido por el trimestre DEF. Esto hace pensar que para estos meses se encontrarán las precisiones más bajas para los valores simulados.

Para determinar la normalidad de los datos se utiliza la prueba estadística de Wilks-Shapiro. De acuerdo con el p-valor (la probabilidad asociada al valor del estadístico calculado) y usando un nivel de significancia del 5 % y 1 %, se deduce que todos los datos observados (X_0) para la temperatura en cada una de las grillas se distribuyen normalmente.

3.1.2. Precipitación

En la tabla 2 se resumen las principales características de los datos observados de precipitación para Venezuela.

Tabla 2: Promedio de la precipitación ($\pm$ 1 desviación estándar) en mm para Venezuela durante el periodo 1960-1989.

	Trimestre				
Grilla	DEF	MAM	JJA	SON	MEDIA
1	28,98 $\pm$ 10,62	85,29 $\pm$ 14,00	91,18 $\pm$ 19,93	166,31 $\pm$ 24,20	92,94 $\pm$ 22,37
2	41,47 $\pm$ 14,11	48,94 $\pm$ 13,80	108,05 $\pm$ 12,02	97,73 $\pm$ 19,58	74,05 $\pm$ 23,34
3	76,97 $\pm$ 16,31	54,81 $\pm$ 16,26	172,38 $\pm$ 19,55	156,50 $\pm$ 23,02	115,17 $\pm$ 19,23
4	56,44 $\pm$ 10,38	167,56 $\pm$ 25,43	206,78 $\pm$ 17,68	206,88 $\pm$ 22,30	159,41 $\pm$ 13,22
5	23,86 $\pm$ 6,43	122,37 $\pm$ 29,22	308,89 $\pm$ 23,80	156,82 $\pm$ 12,72	152,99 $\pm$ 16,66
6	75,54 $\pm$ 15,10	109,53 $\pm$ 29,11	269,80 $\pm$ 25,44	154,65 $\pm$ 17,38	152,38 $\pm$ 16,81
7	107,91 $\pm$ 19,94	269,60 $\pm$ 26,89	311,92 $\pm$ 18,92	233,78 $\pm$ 19,38	230,80 $\pm$ 10,70
8	150,02 $\pm$ 17,16	278,40 $\pm$ 30,99	366,94 $\pm$ 19,94	225,22 $\pm$ 15,25	255,14 $\pm$ 8,69
9	99,90 $\pm$ 13,00	222,33 $\pm$ 31,05	326,68 $\pm$ 25,73	165,50 $\pm$ 17,90	203,60 $\pm$ 11,42
MEDIA	73,45 $\pm$ 13,67	150,98 $\pm$ 24,08	240,29 $\pm$ 20,33	173,71 $\pm$ 19,08	159,61 $\pm$ 19,29

En la tabla 2 se observan los promedios de precipitación mensual por trimestre. Por ejemplo, la grilla 1 registró 28,98 mm de precipitación mensual promedio durante el trimestre de sequía. Cabe aclarar que esta interpretación se mantendrá en todos los análisis de precipitación que se van a presentar en esta investigación, a menos que se indique lo contrario. Ahora bien, en esta tabla se observan precipitaciones más bajas para los meses de sequía y precipitaciones más altas para los meses de lluvia. Este comportamiento coincide con el comportamiento esperado del régimen estacional de la precipitación. En cuanto a los trimestres de transición (MAM y SON), la precipitación se mantuvo con un valor promedio que se ubica entre el valor de la época de lluvia y el valor de la época de sequía, siendo un poco mayor en SON para las grillas 1 a 6.

Al comparar las tablas 1 y 2, se nota una mayor variabilidad natural en la precipitación. Esto hace pensar que de las dos variables climáticas estudiadas, la precipitación es la más difícil de proyectar, por lo que se espera una menor eficiencia por parte de los modelos globales en proyectar esta variable. El trimestre MAM en las dos primeras grillas es el que posiblemente refleje más esta deficiencia, por presentar la mayor dispersión.

De acuerdo con la prueba estadística de Wilks-Shapiro, se puede decir que en la mayoría de los casos, los datos de precipitación mensual tienen una distribución normal a un nivel de significancia del 5 % y 1 %, excepto en los trimestres de sequía (grilla 1) y de transición a la época de lluvias en la zona costera del país (grillas 1,

2 y 3). Los casos de no normalidad pueden ser consecuencia de la alta variabilidad interanual que ha presentado la precipitación en los últimos años. Una de las causas puede ser atribuida al efecto del fenómeno El Niño-Oscilación del Sur (ENSO por sus siglas en inglés).

3.2. Proyecciones de la temperatura y la precipitación mediante el uso de modelos climáticos globales

Los modelos acoplados de circulación general atmósfera-océano (MACGAO) son herramientas tecnológicas que requieren un enorme poder computacional que no poseen los países en vías de desarrollo (caso Venezuela), que ante la Convención Marco de las Naciones Unidas para el Cambio Climático (CMNUCC) se denominan *países No Anexo I*. El programa de apoyo a las comunicaciones nacionales coordinó el desarrollo de una metodología simplificada que se materializa en el MAGICC/SCENGEN. Las proyecciones climáticas utilizadas en este trabajo provienen de 16 modelos MACGAO, que fueron empleados por el MARN en la Primera Comunicación Nacional sobre Cambio Climático de Venezuela, bajo una sensitividad climática intermedia y un escenario SRES-A2, por tener las características más adecuadas a las condiciones de nuestro país. Estas proyecciones fueron obtenidas del Banco Nacional de Datos de la Dirección de Hidrología y Meteorología del MARN y constan de promedios trimestrales para un lapso de 30 años centrados en los años 2025 (2015-2039), 2050 (2035-2064) y 2100 (2085-2114) con una cobertura de nueve grillas de $5^\circ \times 5^\circ$ que abarcan todo el territorio nacional, como se observa en la figura 1. Una lista de los modelos utilizados se presenta en la tabla 3 con una breve descripción de cada uno de ellos.

4. Resultados

En esta sección se presentan los principales resultados obtenidos a partir de 5.000 simulaciones de las distribuciones a posteriori de los modelos propuestos en la sección 2 para la temperatura y la precipitación, los cuales fueron obtenidos utilizando los métodos MCMC descritos anteriormente. Dicho análisis está dividido por trimestres, como se indicó en la sección 2.1, y por caso (independiente y dependiente). Se hacen recomendaciones sobre los modelos más adecuados para proyectar las señales del cambio climático en Venezuela, con base en el sesgo y en la convergencia. Se indican también las zonas o grillas del país en donde es mayor la deficiencia de estas proyecciones, entre otros análisis.

4.1. Temperatura

4.1.1. Caso independiente

En la figura 2 se muestran los diagramas de caja para la variable $\Delta = \nu - \mu$ para 30 años de datos centrados en el año 2100. En esta figura se aprecia que todos los promedios de este parámetro son positivos y mantienen su magnitud

Tabla 3: Modelos del MAGICC/SCENGEN considerados por el MARN (2005)

Nombre	Procedencia	Descripción
HadCM2	Inglaterra	En vertical tiene 19 capas atmosféricas y está acoplado a un modelo oceánico de 20 capas.
UKTR	Inglaterra	En vertical tiene 11 capas atmosféricas y está acoplado a un modelo oceánico de 17 capas.
CSIRO-TR	Australia	En vertical tiene 9 capas atmosféricas y está acoplado a un modelo oceánico de 21 capas.
ECHAM4	Alemania	En vertical tiene 19 capas atmosféricas y está acoplado a un modelo oceánico OPYC3.
UKHI-EQ	Inglaterra	En vertical tiene 11 capas atmosféricas y está acoplado a un modelo oceánico de capa mezclada.
CSIRO2-EQ	Australia	En vertical tiene 9 capas atmosféricas y está acoplado a un modelo oceánico de capa mezclada.
ECHAM3	Alemania	En vertical tiene 19 capas atmosféricas y está acoplado a un modelo oceánico geostrófico con 11 capas.
UIUC-EQ	Estados Unidos	En la vertical tiene 11 capas atmosféricas y fue añadido a un modelo oceánico "slab".
ECHAM1	Alemania	En la vertical tiene 19 capas atmosféricas y está acoplado a un modelo oceánico geostrófico con 11 capas.
CSIRO1-EQ	Australia	En la vertical tiene 9 capas atmosféricas y está acoplado a un modelo oceánico de capa mezclada.
CCC-EQ	Canadá	En la vertical tiene 10 capas atmosféricas y está acoplado a un modelo oceánico de capa mezclada.
GFDL-TR	Estados Unidos	En la vertical tiene 9 capas atmosféricas y está acoplado a un modelo oceánico de 12 capas.
BMRC-EQ	Australia	En la vertical tiene 9 capas atmosféricas y está acoplado a un modelo oceánico de capa mezclada.
CGCM1-TR	Canadá	En la vertical tiene 10 capas atmosféricas y está acoplado a un modelo oceánico de 29 capas.
NCAR-DOE	Estados Unidos	En la vertical tiene 9 capas atmosféricas y está acoplado a un modelo oceánico de 20 capas.
CCRS/NIES	Japón	En la vertical tiene 20 capas atmosféricas y está acoplado a un modelo oceánico de 17 capas.

en cada uno de los trimestres, indicando para los próximos años un incremento en la temperatura en todo el país para todos los trimestres del año. Dicho esto y tomando en cuenta los promedios de temperatura mostrados en la tabla 1, se deduce que los trimestres de transición seguirán siendo los más calientes y la época de lluvia la más fría. Por otra parte, la grilla 9 posiblemente sea la más afectada por el cambio climático, por mostrar altos valores de Δ desde diciembre hasta agosto. Los incrementos varían aproximadamente entre 0.2 a $0,75$ ℃ en el año 2025, entre $0,7$ y $1,45$ ℃ en 2050 y entre $1,5$ y $3,5$ ℃ para 2100 con respecto al período base (año 2000), es decir, que el incremento cada 25 años va en aumento. Esto puede ser consecuencia de los muchos cambios ambientales que se experimentan en el país y en el mundo, con una mayor demanda de agua, alimentos, recursos naturales y una

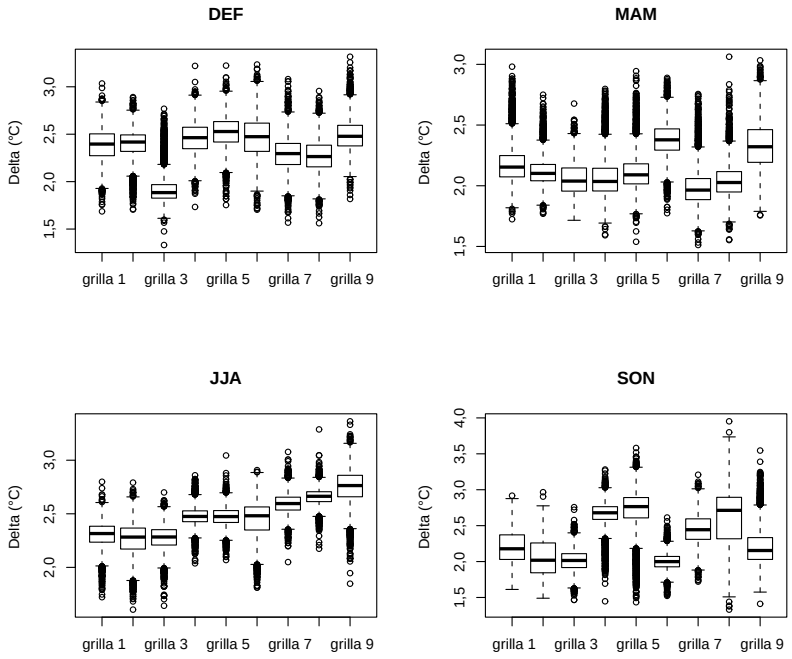


FIGURA 2: Diagrama de caja de los Deltas (Δ) simulados para la temperatura en el caso independiente centrado en el año 2100.

mayor quema de combustibles fósiles, lo que trae consigo un mayor aumento de la concentración de gases de efecto de invernadero, produciéndose así un incremento en la temperatura de la superficie terrestre.

Otros resultados relevantes fueron los sesgos (1) y las convergencias (2) definidos en la sección 2. Estos criterios se analizaron a partir de las gráficas de los promedios de las 5.000 simulaciones realizadas, con sus respectivos intervalos de probabilidad del 90 % (gráfico no incluido). Estos resultados fueron analizados para cada una de las grillas, y se pudo apreciar que estos criterios tienen comportamientos similares en todas las grillas, con la diferencia de que la convergencia es ligeramente más dispersa. Luego de este análisis se llegó a la conclusión de que no existe una diferencia considerable de eficiencia entre los distintos modelos para proyectar el clima presente y futuro. Ahora bien, al observar las magnitudes del sesgo y la convergencia, la grilla 4 presenta las proyecciones con la más baja confiabilidad, ya que posee los valores más altos de estos dos criterios, mientras que la grilla 6, seguida de la 9, fueron las más confiables. Las grillas con las proyecciones más confiables poseen la mayor variabilidad natural (ver tabla 1), resultado que es contrario a lo esperado, ya que se prevé una proyección climática más confiable donde hay menor variabilidad natural.

En la tabla 4 se muestran los modelos que se consideran como los más adecuados para realizar las proyecciones de temperatura. Los criterios que se utilizaron para la selección toman en cuenta las precisiones a posteriori τ_i más altas, y los valores de sesgo y convergencia más bajos, todo lo cual indica una menor incertidumbre en las proyecciones climáticas.

Tabla 4: Modelos adecuados para realizar las proyecciones de temperatura, caso independiente.

Trimestre	Modelo	Grilla
DEF	GFDL-TR	Todas
MAM	CSIRO1-EQ	Todas
JJA	ECHAM3	1 - 2 - 3
	GFDL-TR	4 - 5 - 6 - 7 - 8 - 9
SON	CSIRO-TR	1 - 2 - 3 - 5 - 6 - 8 - 9
	ECHAM3	4 - 7

4.1.2. Caso dependiente

En la figura 3 se presentan los diagramas de caja para los Deltas (Δ) obtenidos a partir de 5.000 simulaciones de la distribución posterior definida en la ecuación (14) para el caso dependiente. Estos valores muestran un comportamiento similar al de la figura 2 en el sentido de que ambos casos proyectan un incremento en la temperatura en todos los trimestres del año, con la diferencia de que en este caso se observa una dispersión más alta. Esta diferencia puede ser ocasionada por la presencia del parámetro adicional β considerado en el modelo del caso dependiente. La presencia de β aumenta la incertidumbre en el modelo, y en consecuencia las simulaciones tienen precisiones más bajas. La tabla 5 muestra los valores promedio de este parámetro a partir de las simulaciones realizadas para cada grilla para valores centrados en el año 2100.

Tabla 5: Estadísticos resumen de 5.000 simulaciones de β por grilla, centrados en el año 2100.

Grilla	Percentil 5 %	Media	Mediana	Percentil %95
1	7,65302	9,61837	9,68360	11,36844
2	8,24154	10,15864	10,23301	11,88699
3	6,55086	8,55633	8,51311	10,68685
4	9,19478	10,38972	10,43891	11,39611
5	8,80463	10,06858	10,08505	11,27856
6	8,34223	9,83922	9,837431	11,34465
7	8,75636	10,75940	10,82612	12,51741
8	8,79315	10,25467	10,29290	11,57691
9	9,37951	10,51142	10,52238	11,58817

Otro resultado relevante observado en la figura 3 es que la grilla 9 es más afectada para todos los trimestres por el cambio de temperatura en los próximos

años, comportamiento que coincide con el caso independiente. Los cambios en grados centígrados oscilan de $0,19$ a $0,8$ para el año 2025; de $0,31$ a $2,03$ para 2050 y de $1,08$ a $4,38$ para 2100 con respecto al año base (2000), lo que significa que son más altos los incrementos para este caso en comparación con el caso independiente. En ambos casos, los cambios o incrementos se van amplificando con el transcurso de los años.

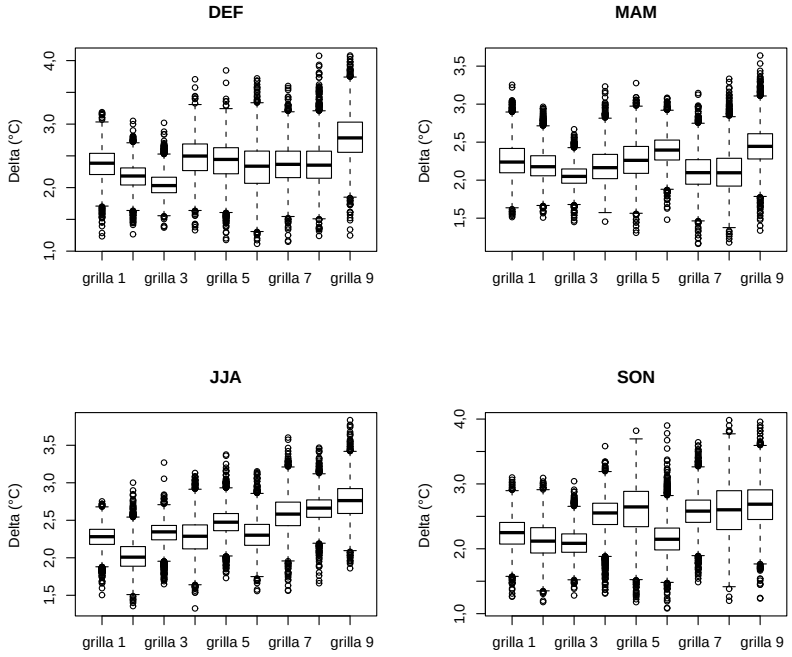


FIGURA 3: Diagrama de caja de los Deltas (Δ) simulados para la temperatura en el caso dependiente centrado en el año 2100.

En general existe el mismo comportamiento de los criterios del método REA (sesgo y convergencia) en los casos independiente y dependiente. Se llega a esta deducción al comparar los resultados de las simulaciones obtenidas. En otras palabras, ambos casos sugieren la misma eficiencia en las proyecciones de la temperatura para Venezuela utilizando los modelos indicados en la tabla 4 para el caso independiente y los de la tabla 6 para el caso dependiente.

En la figura 4 se muestran los diagramas de caja de la precisión a posteriori τ_i (inverso de la varianza) para cada uno de los modelos descritos en la tabla 3 para el caso dependiente. Para ello se procesaron los datos de las 5.000 simulaciones de la distribución posterior del modelo descrito en las ecuaciones (12) y (13). Se utilizaron como observaciones futuras los 30 años de datos centrados en el año 2100. De la figura 4 se concluye por ejemplo, que el modelo 1 (M1:HadCM2) presenta una muy baja precisión en casi todos los trimestres.

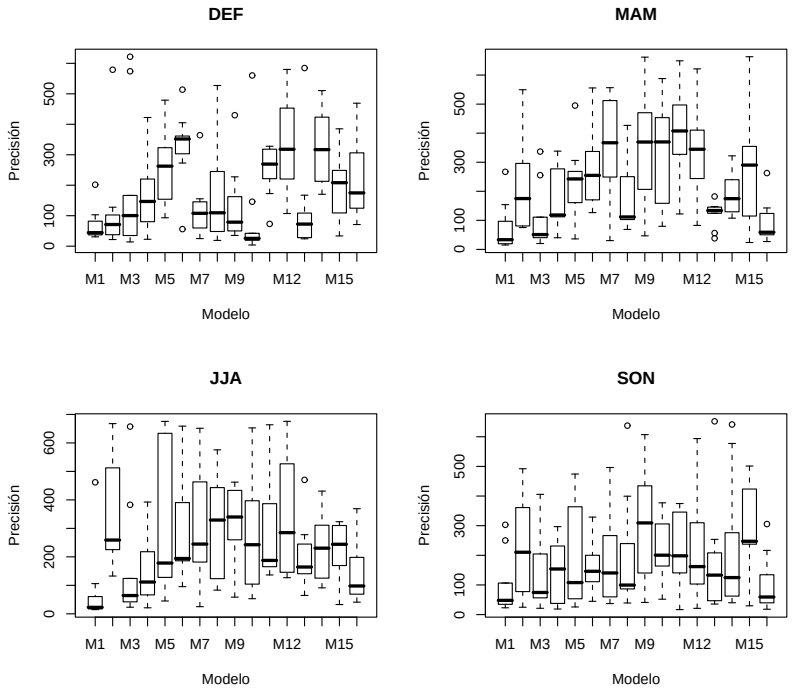


FIGURA 4: Precisión de la temperatura obtenida a partir de las simulaciones a posteriori para los datos centrados en el año 2100, caso dependiente.

En la tabla 6 se indican los modelos más favorables para las proyecciones de cambio de temperatura en el caso dependiente, teniendo en cuenta la precisión τ_i del modelo i y los criterios del método REA.

TABLA 6: Modelos adecuados para realizar las proyecciones de temperatura, caso dependiente.

Trimestre	Modelo	Grilla
DEF	GFDL-TR	Todas menos la 4
	UKTR	4
MAM	CCC-EQ	Todas menos la 7
	UKTR	7
JJA	UIUC-EQ	4 - 5 - 6 - 7 - 8 - 9
	BMRC-EQ	1 - 2 - 3
SON	UKTR	1 - 2 - 3 - 4 - 5 - 6
	CSIRO1-EQ	7 - 8 - 9

4.2. Precipitación

4.2.1. Caso independiente

En la figura 5 se presenta el valor acumulado promedio anual (de 30 años) del parámetro Delta (Δ) (para valores centrados en los años 2025, 2050 y 2100) para la precipitación. En ella se observa un resultado bastante significativo que indica una disminución de la precipitación anual promedio hasta el año 2050, que luego se mantiene aproximadamente constante con 50 mm por año por debajo del valor del año base (2000). Sin embargo, la precipitación trimestral promedio disminuye sostenidamente en los trimestres JJA y SON, y aumenta en los trimestres DEF y MAM.

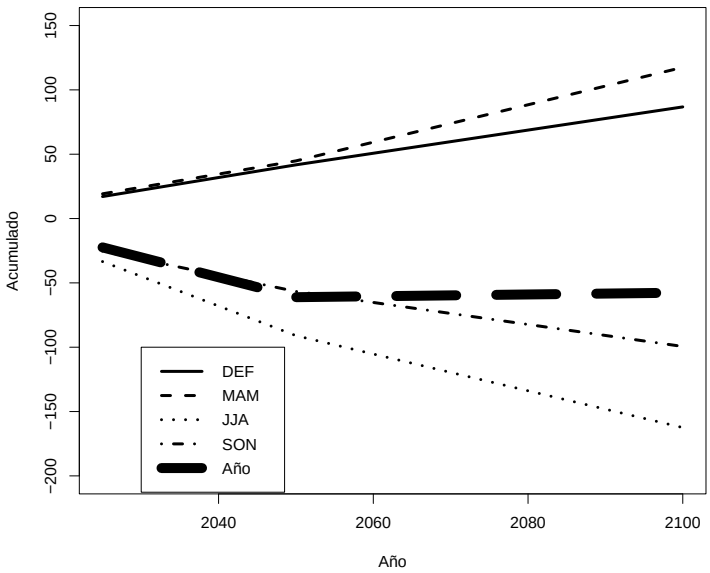


FIGURA 5: Precipitación acumulada promedio anual (de 30 años) del parámetro Delta (Δ) para valores centrados en los años 2025, 2050 y 2100, caso independiente.

La figura 6 muestra los Deltas obtenidos de 5.000 simulaciones de la distribución posterior (ecuación 7) para la precipitación en el caso independiente presentados por trimestre mediante diagramas de cajas. Esta figura corresponde a las simulaciones centradas en el año 2100. En ella se observan incrementos de lluvias para la mayor parte del país en el lapso de diciembre a mayo, en donde las grillas 7 y 8, que corresponden al estado Amazonas, presentan el mayor incremento. Por otra parte, para la segunda mitad del año, la figura refleja disminuciones en las lluvias proyectadas en casi todo el territorio nacional. Aquí las grillas más afectadas por el cambio son las grillas 4, 6 y 8 para el trimestre JJA, y 1, 3 y 6 para el trimestre SON, donde se observan las disminuciones más altas.

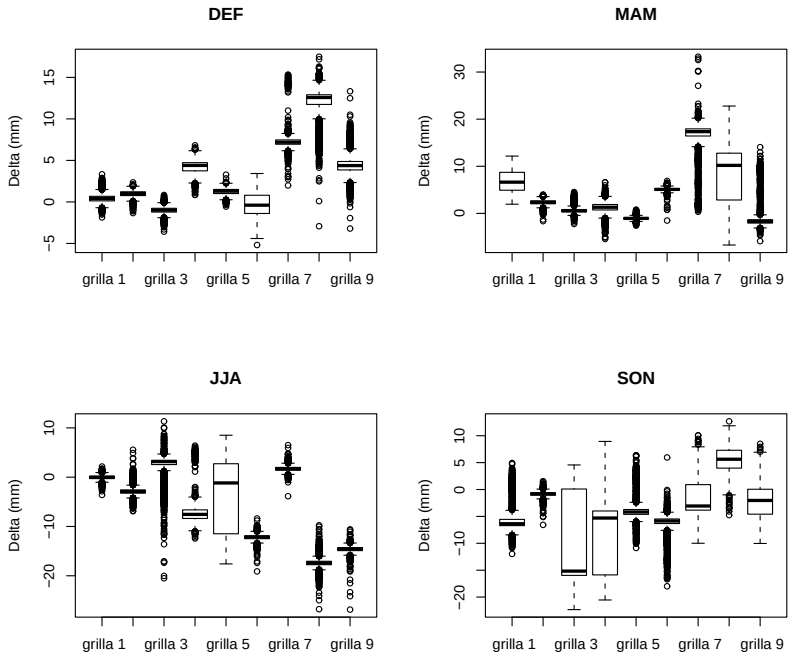


FIGURA 6: Diagrama de caja de los Deltas (Δ) simulados para la precipitación en el caso independiente centrados en el año 2100.

Al igual que con la temperatura, para esta variable climática se calculó el sesgo (ecuación 1) y la convergencia (ecuación 2) y se graficaron los valores promedios para las 5.000 simulaciones realizadas, con sus respectivos intervalos de probabilidad del 90 % (gráfico no incluido). Al analizar estos resultados por grillas se pudo observar que la convergencia va aumentando con el tiempo, lo que indica una pérdida en la eficiencia de los modelos en las proyecciones del clima futuro. En general, se observan valores de sesgo y convergencia altos en la mayoría de las grillas, en comparación con los valores obtenidos para la temperatura. De allí se deduce que existe una menor confiabilidad en las proyecciones del cambio climático para la variable precipitación en comparación con las proyecciones de temperatura. La mayor incertidumbre de estas proyecciones se presenta en los meses que van desde marzo hasta agosto, cuando los valores de sesgo, convergencia y desviación estándar son más altos. Por el contrario, el trimestre DEF posee los valores más bajos de estos criterios, y en consecuencia presenta las proyecciones más confiables para esta variable climática.

En la tabla 8 aparecen los MCG más adecuados para cada trimestre. Los modelos fueron seleccionados en función de los criterios del método REA, es decir, aquellos modelos con menores valores en sesgo y convergencia.

Tabla 7: Modelos adecuados para realizar las proyecciones de precipitación, caso independiente.

Trimestre	Modelo	Grilla
DEF	GFDL-TR	1 - 5
	ECHAM1	2 - 3 - 4
	CGCM1-TR	6 - 7 - 8 - 9
MAM	BMRC-EQ	1 - 2 - 3
	ECHAM4	4 - 5 - 6 - 9
	CCSR/NIES	7 - 8
JJA	ECHAM1	1 - 2
	GFDL-TR	3 - 4
	CCC-EQ	5 - 6
	BMRC-EQ	7 - 8 - 9
SON	CSIRO-TR	1 - 5 - 6 - 9
	ECHAM3	2 - 3
	UKTR	4 - 7 - 8

Como se observa en la tabla anterior, se deben utilizar al menos tres modelos en cada trimestre para realizar proyecciones más efectivas, en donde el modelo BMRC-EQ resultó ser el modelo más recomendado. También se observa que se requiere una mayor cantidad de modelos para la precipitación, comparado con la temperatura, a fin de lograr una mejor eficiencia en las proyecciones. Esto puede ser consecuencia de la baja precisión mostrada en la mayoría de los modelos para esta variable climática.

4.2.2. Caso dependiente

La figura 7 indica los cambios de precipitación para este caso mediante diagramas de cajas para los datos centrados en el año 2100. En términos generales, se observa mucha variación de los resultados con el transcurso del tiempo. Para los datos centrados el año 2050 (figura no incluida), se observa un alto incremento de la precipitación para la época seca sobre la grilla 8, lo que se obtiene nuevamente para las grillas 7 y 8 pero con mayor contundencia para los datos centrados en el año 2100, como se puede observar en la figura 7. En consecuencia, la grilla 8, que corresponde prácticamente al estado Amazonas, es la zona del país que posiblemente sea la más afectada por el cambio de la precipitación en los próximos años, de acuerdo con los resultados obtenidos para el caso dependiente.

En cuanto a los Deltas acumulados a partir de 5.000 simulaciones de la distribución posterior dada por la ecuación (14), se muestra en la figura 8 un pequeño incremento de la precipitación en los primeros años hasta el año 2050; de allí en adelante el incremento se hace mucho más intenso. Este comportamiento se debe al incremento observado en las grillas 7 y 8 para los últimos años proyectados (2085-2114) en los meses MAM, mientras para el resto de las grillas prácticamente no hubo cambios. En cuanto a los otros trimestres se proyecta lo siguiente: para DEF, aumentos considerables hasta 2050; de allí en adelante comienza a bajar

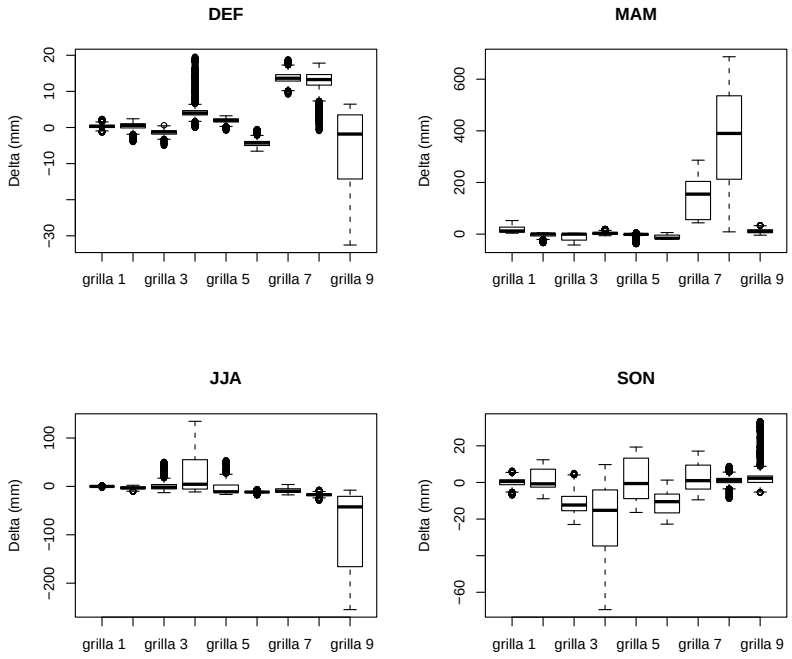


FIGURA 7: Diagrama de caja de los Deltas (Δ) simulados para la precipitación en el caso dependiente centrados en el año 2100.

lentamente pero manteniéndose con un valor acumulado por encima al del año base (2000). Los dos trimestres restantes permanecen casi sin cambios hasta el año 2050, año a partir del cual se observa una disminución en la precipitación.

Al comparar los sesgos y las convergencias entre los casos dependiente e independiente, se observa que estos criterios conservan comportamientos similares en ambos casos, con ligeras variaciones en algunas grillas. Sin embargo, se esperaban valores más altos y con mayor dispersión para el primer caso con respecto al caso independiente, donde los modelos tienen menor precisión a posteriori. Se puede concluir que no existen diferencias considerables en la eficiencia de los MCG entre estos dos casos para proyectar las señales de cambio de precipitación en Venezuela.

En la figura 9 se muestran los diagramas de caja de las precisiones a posteriori para la precipitación a partir de 5.000 simulaciones de la distribución posterior del modelo definido en las ecuaciones (12) y (13). Estos resultados arrojan los valores de precisión más bajos obtenidos en este estudio, por lo que se espera un mayor sesgo y convergencia, y por ende una baja eficiencia en las proyecciones de cambio para esta variable. Este resultado tiene sentido, porque como se dijo anteriormente, la precipitación posee una mayor variabilidad natural que la temperatura, y el caso dependiente añade un nuevo parámetro desconocido que aumenta la incertidumbre en el modelo.

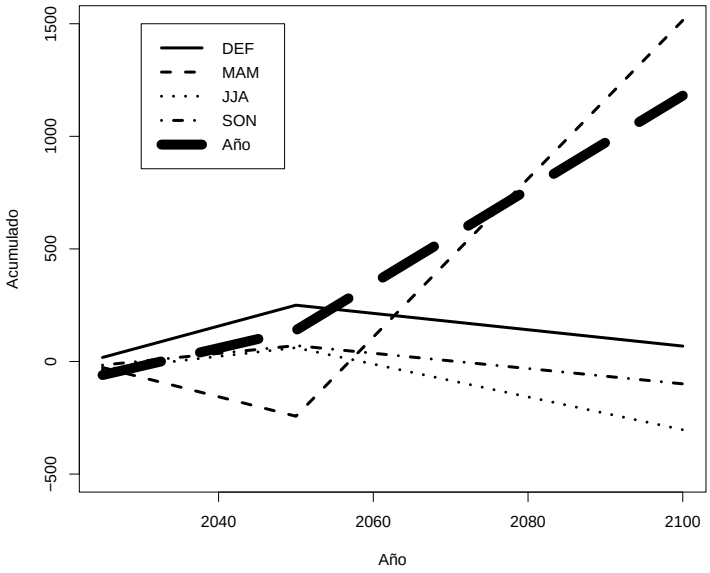


FIGURA 8: Precipitación acumulada promedio anual (de 30 años) del parámetro Δ para valores centrados en el año 2025, 2050 y 2100, caso dependiente.

Al analizar las precisiones τ_i y los criterios del método REA obtenidos para este caso, se recomiendan los modelos que se muestran en la tabla 8. Aquí se puede ver que el CGCM1-TR fue el modelo más adecuado para proyectar las señales de cambio de precipitación, con una mayor frecuencia de valores bajos de sesgo y convergencia.

5. Discusión y conclusiones

El cambio climático es uno de los grandes problemas ambientales que enfrenta la humanidad hoy en día. Pequeñas variaciones en los parámetros climáticos, como la temperatura y la precipitación, pueden acarrear consecuencias muy negativas en las actividades económicas de la sociedad, e incluso afectar considerablemente la salud del ser humano. En la actualidad, el planeta ya está teniendo un calentamiento con muchas consecuencias medibles para los seres vivos, por lo que es de vital importancia realizar proyecciones de cómo será en un futuro este fenómeno con la finalidad de buscar posibles medidas de adaptación y evitar así mayores consecuencias adversas.

Para predecir el cambio climático hay que conocer exactamente cómo funciona el sistema climático, que debido a su complejidad aún hoy no se conoce del todo; sin embargo, existen herramientas para hacer proyecciones de este cambio, siendo la

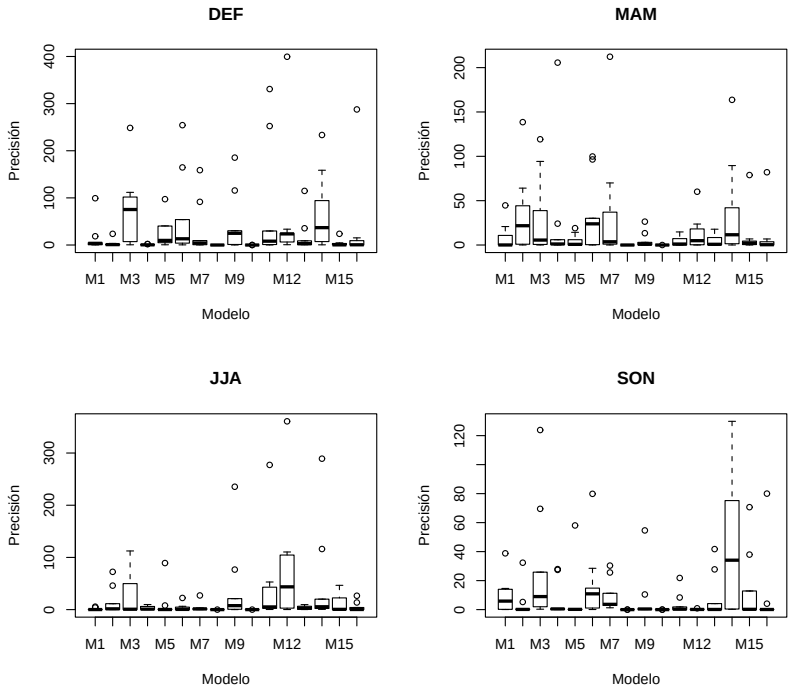


FIGURA 9: Precisión de la precipitación obtenida a partir de las simulaciones a posteriori para los datos centrados en el año 2100, caso dependiente.

TABLA 8: Modelos adecuados para realizar proyecciones de precipitación, caso dependiente

Trimestre	Modelo	Grilla
DEF	ECHAM1	1 - 2 - 5
	CGCM1-TR	3 - 4 - 6 - 7
	HADCM2	8
	BMRC-EQ	9
MAM	CSIRO-TR	1 - 2 - 3
	CSIRO2-EQ	4 - 7 - 8 - 9
	CGCM1-TR	5 - 6
JJA	CSIRO-TR	1 - 2
	GFDL-TR	3 - 4
	CCC-EQ	5 - 6 - 7 - 8 - 9
SON	CGCM1-TR	1 - 5 - 6 - 9
	BMRC-EQ	2
	NCAR-DOE	3
	HADCM2	4 - 7 -

principal de ellas los modelos de circulación general de la atmósfera (MCG). Existe una gran diversidad de modelos climáticos, y cada uno de ellos arroja proyecciones diferentes en la señal del cambio climático. No obstante, existen métodos que permiten estudiar la eficiencia de estos modelos, como es el caso del *Reliability Ensemble Average* (REA), que consiste en combinar las salidas de los modelos (clima presente y futuro) con los datos observados para simular promedios de las variables climáticas proyectadas, y con base en dos criterios, *el sesgo* y *la convergencia*, medir la confiabilidad de las proyecciones del clima presente y futuro, respectivamente.

En Venezuela, al aplicar el método REA para la precipitación y temperatura bajo dos condiciones, el caso independiente (no existe ningún tipo de relación entre el clima presente y futuro), y el caso dependiente (existe relación entre el clima presente y futuro), se obtuvieron los siguientes resultados:

- La mayoría de los modelos estudiados en este trabajo producen de manera general las mismas tendencias en cuanto al incremento de la temperatura en los próximos años; es por ello que se obtuvieron altas precisiones y valores bajos de sesgo y convergencia a partir de las 5.000 simulaciones obtenidas de las distribuciones a posteriori de los parámetros para los modelos propuestos, estimadas utilizando el paradigma bayesiano. Esto mejora la confianza en las proyecciones (mayor efectividad) realizadas por los modelos propuestos por muchos científicos alrededor del mundo. Sin embargo, hay que tener siempre en cuenta que el sistema climático no se conoce a plenitud. En consecuencia existen otros procesos de retrolimentación positiva o negativa que interfieren con el cambio climático y que aún no están incluidos en los modelos disponibles.

- En cuanto a la precipitación, existe una alta incertidumbre en los modelos estudiados por la gran variabilidad que existe en sus proyecciones, reflejadas en las bajas precisiones a posteriori y en los altos valores de sesgo y convergencia para la mayoría de los modelos analizados. Esto hace que las proyecciones actuales sobre la precipitación no sean del todo confiables. Por tanto, podemos decir que para la precipitación, los MCG tienen una menor eficiencia en los cambios proyectados en comparación con la temperatura.

- En el caso dependiente existe una menor eficiencia comparada con el caso independiente, debido a que al establecer una relación entre el clima presente y futuro se añade una nueva variable aleatoria al modelo, aumentando así la complejidad y la incertidumbre, lo cual se vio reflejado en las precisiones a posteriori (más altas para el caso independiente que para el caso dependiente).

- Los modelos más precisos son los que lógicamente deberían tener los menores valores de sesgo y convergencia (en la mayor parte del territorio nacional) por tener un mayor peso en la media ponderada. Sin embargo, en algunos casos no se observó este comportamiento; por tanto, el que un modelo climático tenga una precisión alta, no es garantía de que sea el más eficiente, según el método REA.

- La tendencia observada para la temperatura en ambos casos estudiados, coincide con lo indicado en la Primera Comunicación Nacional de Cambio Climático en Venezuela, donde se indica que esta variable climática aumentará considerablemente en los próximos años. Este incremento es más fuerte a medida que el lapso de tiempo de proyección es más amplio, comportamiento que puede ser explicado por el crecimiento económico mundial y los patrones de combustión de combustibles fósiles, que son los principales responsables de la emisión de gases de efecto de invernadero y del calentamiento global.

- En la precipitación sí hubo diferencias considerables entre el caso dependiente y el caso independiente. El caso dependiente arroja una disminución de la precipitación en los primeros 60 años de proyección, seguido de un incremento considerable para el resto del periodo de estudio. En el caso independiente se observa una disminución sostenida en el tiempo. Tomando en cuenta la opinión de expertos en meteorología y climatología, establecer una relación entre el clima presente y futuro no pareciera funcionar más allá de 60 años para esta variable climática, ya que de acuerdo con las condiciones del país se espera una disminución de la precipitación en los próximos años como lo indica la Primera Comunicación Nacional de Cambio Climático en Venezuela.

- Los modelos elegidos en la Primera Comunicación Nacional de Cambio Climático en Venezuela por medio de un taller de expertos en meteorología y climatología fueron el modelo UKTR y el modelo CCC-EQ, por adaptarse mejor a las condiciones climáticas del país. Sin embargo, los resultados obtenidos en esta investigación indican que estos modelos son bastantes eficaces en sus proyecciones solo en algunos meses y regiones del país, por lo que se recomienda para futuros estudios del cambio climático en Venezuela, tomar en cuenta otras opciones, como por ejemplo el modelo GFDL-TR que para la temperatura mostró ser el más eficiente para el trimestre DEF, entre otros.

Agradecimientos

Agradecemos extensamente a la profesora María Teresa Martelo de la Universidad Central de Venezuela, por proveer los datos del MAGICC/SCENGEN y por sus acertados y útiles comentarios durante la elaboración de este trabajo.

[Recibido: agosto de 2009 — Aceptado: septiembre de 2010]

Referencias

Benioff, R., Guill, S. & Lee, J. (1996), *Vulnerability and Adaptation Assessments: An International Handbook*, Kluwer Academic Publishers, Dordrecht, The Neatherlands.

Durán, A. J. (2008), Enfoque bayesiano para la estimación probabilística de los cambios climáticos en Venezuela, Tesis de maestría, Universidad Simón Bolívar, Baruta, Estado Miranda.

Gelfand, A. E. & Smith, A. F. M. (1990), 'Sampling-Based Approaches to Calculating Marginal Densities', *Journal of the American Statistical Association* **85**(410), 398–409.

Giorgi, F. & Mearns, L. (2002), 'Calculation of Average, Uncertainty Range and Reliability of Regional Climate Changes from AOGCM Simulations via the ŞReliability Ensemble Averaging (REA)Ť Method', *Journal of Climate* **15**, 1141–1158.

Hulme, M., Wigley, T. M. L., Barrow, E. M., Raper, S. C. B., Centella, A., Smith, S. J. & Chipanshi, A. C. (2000), *Using a Climate Scenario Generator for Vulnerability and Adaptation Assessments: MAGICC and SCENGEN version 2.4 Workbook*, Climatic Research Unit, Norwich UK.

IPCC (2001), *Intergovernmental Panel on Climate Change, Working Group II Report, Third Assessment Report (TAR)*, UNEP/WMO, Ginebra, Suiza.

IPCC (2007), *Intergovernmental Panel on Climate Change, Working Group III Report, Fourth Assessment Report (AR-4)*, UNEP/WMO, Ginebra, Suiza.

MARN (2005), *Primera comunicación nacional en cambio climático de Venezuela*, MARN/GEF/PNUD, Caracas, Venezuela.

Migon, H. S. & Gamerman, D. (1999), *Statistical Inference: An integrated Approach*, A Hodder Arnold Publication, New York, United States.

Nychka, D. & Tebaldi, C. (2002), 'Comment on Calculation of Average, Uncertainty Range and Reliability of Regional Climate Changes from AOGCM Simulations via the ŚReliability Ensemble AveragingŠ (REA) method', *Journal of Climate* **16**(5), 883–884.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Tebaldi, C., Smith, R., Nychka, D. & Mearns, L. (2005), 'Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles', *Journal of Climate* **18**(12), 1524–1540.

Vera, C., Silvestri, G., Liebmann, B. & González, P. (2006), 'Climate change scenarios for seasonal precipitation in South America from IPCC-AR4 models', *Geophysical Research Letters* **33**(L13707), doi:10.1029/2006GL025759.

Revista Colombiana de Estadística
Diciembre 2010, volumen 33, no. 2, pp. 219 a 231

Una extensión de la distribución Weibull de dos parámetros

An Extension of the Two-Parameter Weibull Distribution

Juan F. Olivares-Pacheco[1,a], Héctor C. Cornide-Reyes[2,b], Manuel Monasterio[2,c]

[1]Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó, Chile

[2]Departamento de Ingeniería Informática y Ciencias de la Computación, Facultad de Ingeniería, Universidad de Atacama, Copiapó, Chile

Resumen

En este artículo se presenta una extensión de la distribución Weibull de dos parámetros, con el objetivo de flexibilizar el modelo en términos de la kurtosis. Se estudian las propiedades básicas de la nueva densidad obtenida, así como su función de distribución, momentos, coeficientes de asimetría y kurtosis. Se realizan estudios de simulación para algunos casos particulares, ilustrando la utilidad de la extensión considerada.

***Palabras clave*:** coeficiente de asimetría, distribución Weibull, kurtosis, Slash.

Abstract

In this paper, we present an extension of the Two-parameter Weibull distribution to make it even more flexible in terms of its kurtosis coefficient. Properties involving moments and asymmetry and kurtosis indexes are studied. Simulation studies for some cases, illustrating the usefulness of the extension considered, are carried out.

***Key words*:** Asymmetry, Kurtosis, Slash distribution, Weibull distribution.

[a]Instructor. E-mail: jolivares@mat.uda.cl
[b]Profesor asistente. E-mail: hcornide@diicc.uda.cl
[c]Profesor asistente. E-mail: mmonasterio@diicc.uda.cl

1. Introducción

La distribución Weibull ha sido ampliamente utilizada para modelar tiempos de vida de componentes (o sistemas), en organizaciones que desarrollan programas de mantenimiento preventivo en sus máquinas, ya que permite estudiar la tasa de falla de un componente crítico. Esta distribución fue introducida por Weibull (1951) y posteriormente se han desarrollado muchos trabajos que otorgan una mayor flexibilidad al modelo original modificando su estructura paramétrica. Por ejemplo, Tang, Xie & Goh (2003) y Chen (2000) trabajan sobre el parámetro de forma para obtener un mejor ajuste de funciones con tasa de falla con forma de bañera o creciente. Zhang & Xie (2007) generan un nuevo modelo al incorporar un parámetro adicional a la familia de distribuciones Weibull con dos parámetros.

La función de densidad de probabilidad (fdp) de una variable aleatoria Weibull de dos parámetros es de la forma

$$f_X(x \mid \alpha, \beta) = \frac{\beta}{\alpha^\beta} x^{\beta-1} e^{-(x/\alpha)^\beta},\ x > 0 \tag{1}$$

donde $\alpha > 0$ es el parámetro de escala, y $\beta > 0$ es el parámetro de forma. Si X se distribuye en Weibull de parámetros (α, β) lo denotaremos por $X \sim \mathrm{W}(\alpha, \beta)$. A partir de la distribución Weibull, se pueden derivar los siguientes casos particulares; si $\beta = 0,5$, X se distribuye Hiperexponencial; si $\beta = 1$, X se distribuye exponencial de parámetro $\alpha > 0$ y $\beta = 2$, tenemos que X se distribuye en Rayleigh de parámetro $\alpha > 0$. Por otro lado, la función de distribución acumulada (fda) es

$$F_X(x \mid \alpha, \beta) = 1 - e^{-(x/\alpha)^\beta},\ x > 0 \tag{2}$$

y la función de confiabilidad viene dada por

$$R_X(x \mid \alpha, \beta) = 1 - F_X(x \mid \alpha, \beta) = e^{-(x/\alpha)^\beta},\ x > 0 \tag{3}$$

Desde (1) y (3), tenemos que la función de tasa de falla es

$$h_X(x \mid \alpha, \beta) = \frac{f_X(\alpha, \beta)}{R_X(x \mid \alpha, \beta)} = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1},\ x > 0 \tag{4}$$

La función de tasa de falla dada en (4) es decreciente para $\beta < 1$, constante para $\beta = 1$ y creciente para $\beta > 1$.

Actualmente, la preocupación está centrada en desarrollar extensiones a la distribución Weibull para lograr un mejor ajuste a la situación de cada componente en estudio. Sobre este mismo punto, ya existen trabajos similares como, por ejemplo, el de Gómez, Quintana & Torres (2007), que utilizan la familia de distribuciones elípticas para generar una nueva familia de distribuciones denominadas Slash-elípticas, así como un trabajo reciente de Gómez, Olivares-Pacheco & Bolfarine (2009), quienes generan una extensión de la distribución Birnbaum-Saunders a partir del tratamiento de su kurtosis.

En este sentido, la distribución Slash canónica se define como la razón de dos variables aleatorias independientes, a saber; una normal estándar (N(0, 1)) y una uniforme en (0, 1) (U(0, 1)), donde la fdp es dada por

$$p(x) = \begin{cases} \frac{\phi(0)-\phi(x)}{x^2}, & \text{si } x \neq 0 \\ \frac{1}{2}\phi(0), & \text{si } x = 0 \end{cases} \tag{5}$$

donde $\phi(\cdot)$ es la densidad normal estándar. Esta distribución presenta colas más pesadas que la distribución normal, y como consecuencia, es una distribución con mayor kurtosis. Algunas propiedades de esta familia son discutidas en Rogers & Tukey (1972) y en Mosteller & Tukey (1977). Los estimadores de máxima verosimilitud (EMV) para los parámetros de localización y escala son discutidos en Kafadar (1982). Wang & Genton (2006) proponen una versión skew multivariada para la distribución Slash estándar.

Una representación estocástica para la distribución Slash estándar es dada por

$$S = \frac{Z}{U^{1/q}} \tag{6}$$

donde $Z \sim \mathrm{N}(0, 1)$, $U \sim \mathrm{U}(0, 1)$, Z es independiente de U y $q > 0$. Si $q = 1$, entonces se obtiene la distribución Slash canónica, y si $q \to \infty$, se obtiene la distribución normal estándar.

Por tanto, el objetivo de este trabajo, es estudiar la variable aleatoria S en (6), considerando la variable aleatoria $Z \sim \mathrm{W}(\alpha, \beta)$. Se llamará a esta nueva densidad Slash-Weibull de parámetros (α, β, q), denotado por $\mathrm{SW}(\alpha, \beta, q)$. Esta variable aleatoria S presentará una mayor kurtosis que el modelo Weibull original (1); ejemplo de datos que poseen un mayor índice de kurtosis son aquellos relacionados con tiempo de fallas de componentes sometidos a un estrés cíclico (ver Gómez et al. 2009).

Este trabajo está organizado como sigue. En la sección 2, se presenta la densidad de la familia Slash-Weibull, y muestran algunos casos particulares derivados del modelo Slash-Weibull, se obtienen los momentos y se estudian los coeficientes de asimetría y kurtosis del modelo. Esta sección finaliza con un análisis de las funciones de confiabilidad y tasa de falla. En la sección 3, se estudian los aspectos inferenciales del modelo, en particular los estimadores de máxima verosimilitud y se presenta un estudio de simulación relacionado con los parámetros involucrados. Finalmente, la sección 4 es dedicada a las principales conclusiones.

2. La distribución Slash-Weibull

En esta sección se define la densidad de la familia estudiada y se muestran algunas de sus propiedades básicas. Por un lado, la representación estocástica del modelo Slash-Weibull es de la forma

$$W = \frac{X}{U^{1/q}} \tag{7}$$

donde $X \sim \mathrm{W}(\alpha, \beta)$, $U \sim \mathrm{U}(0,1)$, X es independiente de U y $q > 0$. Diremos que W se distribuye de acuerdo en la distribución Slash-Weibull de parámetros (α, β, q), denotaremos la distribución de (7) usando la notación $W \sim \mathrm{SW}(\alpha, \beta, q)$.

2.1. Función de densidad

La siguiente proposición muestra la fdp de la distribución Slash-Weibull, obtenida a partir de la representación estocástica dada en (7).

Proposición 1. *Sea $W \sim SW(\alpha, \beta, q)$. Entonces, la fdp de W es dada por*

$$f_W(w \mid \alpha, \beta, q) = \frac{q\beta}{\alpha^\beta} w^{\beta-1} T_W(w \mid \alpha, \beta, q),\ w > 0 \tag{8}$$

donde $\alpha,\ \beta > 0$, $q > 0$ y $T_W(w \mid \alpha, \beta, q)$ se define como

$$T_W(w \mid \alpha, \beta, q) = \int_0^1 u^{\beta+q-1} e^{-(uw/\alpha)^\beta}\, du$$

Demostración. Desde (7), y usando la independencia de X y U, sea $\varphi : \mathbb{R}^2 \to \mathbb{R}^2$, la transformación $\varphi(x,u) = (x/u^{1/q}, u^{1/q})$ para $u \neq 0$ y con inversa $\varphi^{-1}(w,v) = (wv, v^q)$, entonces el jacobiano de la transformación inversa es $J(w,v) = qv^q$. Por tanto,

$$\begin{aligned} f_W(w \mid \alpha, \beta, q) &= \int_{-\infty}^{\infty} f_X(wv \mid \alpha, \beta) f_U(v) \mid J(w,u) \mid\, dv \\ &= q \int_0^1 v^q f_X(wv \mid \alpha, \beta)\, dv \end{aligned}$$

donde $f_X(\cdot \mid \alpha, \beta)$ es dado por (1). □

La figura 1 muestra la densidad Slash-Weibull para diferentes elecciones de los parámetros α, β y q. En la figura 1 se observa claramente el efecto producido por el parámetro q en el modelo Slash-Weibull, es decir, presenta colas más pesadas que la distribución Weibull, y como consecuencia, el nuevo modelo tiene mayor kurtosis.

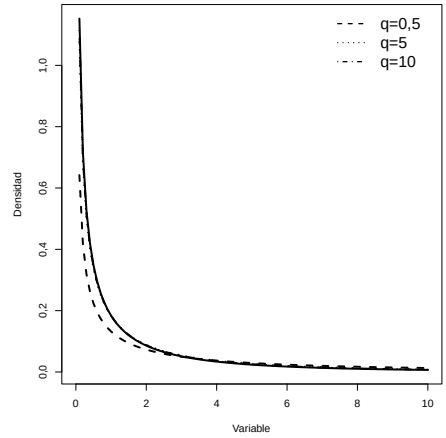
Los siguientes corolarios son consecuencias directas de (8) y se obtienen como extensiones de los casos particulares derivados del modelo Weibull.

Corolario 1. *La variable aleatoria X se distribuye de acuerdo con la distribución Slash-Hiperexponencial y lo denotaremos por $X \sim SH(\alpha, q)$, con fdp dada por*

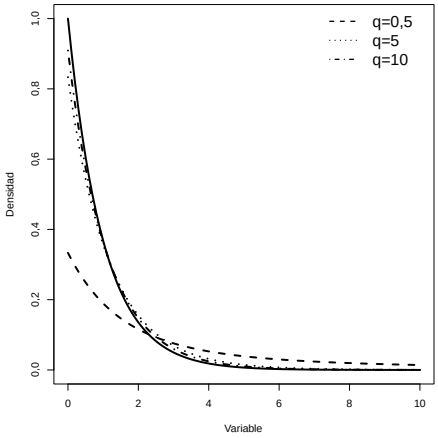
$$f_X(x \mid \alpha, q) = \frac{q}{2} (\alpha x)^{-1/2} T_X(x \mid \alpha, q),\ x > 0 \tag{9}$$

donde $\alpha > 0$, $q > 0$ y $T_W(w \mid \alpha, q)$ se define como

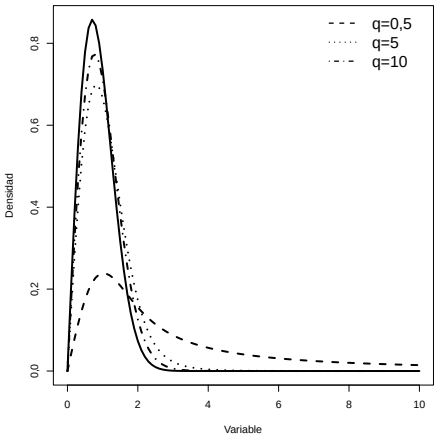
$$T_X(x \mid \alpha, q) = \int_0^1 u^{q-1/2} e^{-(ux/\alpha)^{1/2}}\, du$$



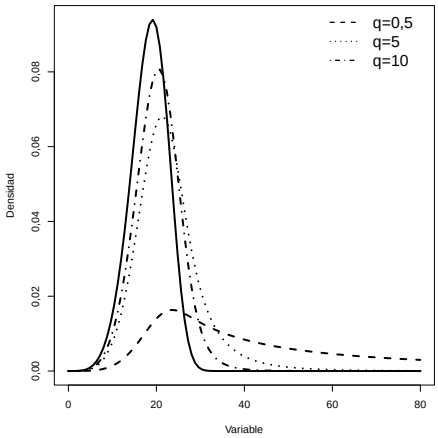
(a) $f_W(w \mid \alpha = 1, \beta = 0,5, q)$



(b) $f_X(x \mid \alpha = 1, \beta = 1, q)$



(c) $f_X(x \mid \alpha = 1, \beta = 2, q)$



(d) $f_X(x \mid \alpha = 20, \beta = 5, q)$

FIGURA 1: FDP $f_W(w \mid \alpha, \beta, q)$ para varias elecciones de (α, β, q). La línea continua corresponde a la densidad Weibull de dos parámetros.

Corolario 2. *La variable aleatoria X se distribuye de acuerdo con la distribución Slash-Exponencial y lo denotaremos por $X \sim SE(\alpha, q)$, con fdp dada por*

$$f_X(x \mid \alpha, q) = \frac{q}{\alpha} T_X(x \mid \alpha, q),\ x > 0 \tag{10}$$

donde $\alpha > 0$, $q > 0$ y $T_W(w \mid \alpha, q)$ se define como

$$T_X(x \mid \alpha, q) = \int_0^1 u^q e^{-(ux/\alpha)}\, du$$

Corolario 3. *La variable aleatoria X se distribuye de acuerdo con la distribución Slash-Rayleigh y lo denotaremos por $X \sim SR(\alpha, q)$, con fdp dada por*

$$f_X(x \mid \alpha, q) = \frac{2q}{\alpha^2} x T_X(x \mid \alpha, q),\ x > 0 \tag{11}$$

donde $\alpha > 0$, $q > 0$ y $T_W(w \mid \alpha, q)$ se define como

$$T_X(x \mid \alpha, q) = \int_0^1 u^{q+1} e^{-(ux/\alpha)^2}\, du$$

Las figuras 1(a), 1(b) y 1(c) muestran las formas de la densidad para diferentes elecciones de los parámetros (α, q), para los modelos Slash-Hiperexponencial, Slash-Exponencial y Slash-Rayleigh, respectivamente.

Además, de la representación estocástica dada en (7) es fácil generar variables aleatorias del modelo Slash-Weibull, a partir de la generación de variables aleatorias Weibull y Uniforme.

2.2. Momentos

Los momentos del modelo Slash-Weibull vienen dados por la siguiente proposición.

Proposición 2. *Sea $W \sim SW(\alpha, \beta, q)$. Entonces, con $r = 1, 2, 3, \ldots$ y $q > r$, tenemos*

$$\mathbb{E}[W^r] = \frac{q}{q-r} \alpha^r \Gamma\left(1 + \frac{r}{\beta}\right) \tag{12}$$

donde $\Gamma(u) = \int_0^\infty t^{u-1} \exp(-t)\, dt$, $u > 0$ (función Gamma).

Demostración. Dado que X y U son independientes, a través de la representación estocástica dada en (7), tenemos

$$\mathbb{E}[W^r] = \mathbb{E}\left[\left(\frac{X}{U^{1/q}}\right)^r\right] = \mathbb{E}\left[U^{-r/q}\right] \mathbb{E}[X^r]$$

ya que $U \sim \mathrm{U}(0,1)$, se sigue que $\mathbb{E}\left[U^{-r/q}\right] = \frac{q}{q-r}$, $q > r$ y por otro lado, como $X \sim \mathrm{W}(\alpha, \beta)$, tenemos $E[X^r] = \alpha^r \Gamma(1 + \frac{r}{\beta})$, (Johnson, Kotz & Balakrishnan 1995). □

Usando la proposición 2, es posible obtener las expresiones para la esperanza y varianza de la variable aleatoria W, las cuales son dadas en el siguiente corolario.

Corolario 4. *Si $W \sim SW(\alpha, \beta, q)$, entonces*

$$\begin{aligned} E[W] &= \frac{q}{q-1} \alpha \Gamma\left(1 + \frac{1}{\beta}\right),\ q > 1 \\ V(W) &= q\alpha^2 \left\{ \frac{1}{q-2} \Gamma\left(1 + \frac{2}{\beta}\right) - \frac{q}{(q-1)^2} \Gamma^2\left(1 + \frac{1}{\beta}\right) \right\},\ q > 2 \end{aligned}$$

2.3. Coeficiente de asimetría y kurtosis

Proposición 3. *Sea $W \sim SW(\alpha, \beta, q)$. Entonces el coeficiente de asimetría es*

$$\sqrt{\beta_1} = \frac{\dfrac{\Gamma(1+3/\beta)}{q^2(q-3)} - \dfrac{3\Gamma(1+1/\beta)\Gamma(1+2/\beta)}{q(q-1)(q-2)} + \dfrac{2\Gamma^3(1+1/\beta)}{(q-1)^3}}{\left[\dfrac{\Gamma(1+2/\beta)}{q(q-2)} - \dfrac{\Gamma^2(1+1/\beta)}{(q-1)^2}\right]^{3/2}}, \; q > 3 \qquad (13)$$

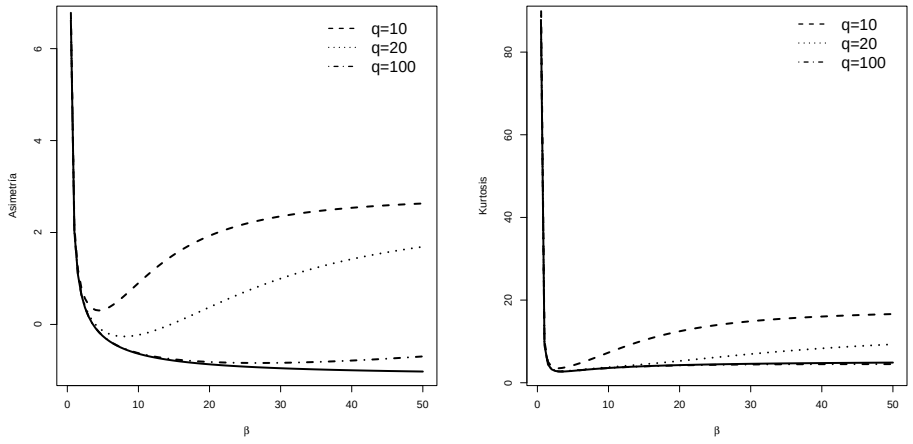
y el coeficiente de kurtosis es

$$\beta_2 = \frac{\frac{\Gamma(1+4/\beta)}{q^3(q-4)} - \frac{4\Gamma(1+1/\beta)\Gamma(1+3/\beta)}{q^2(q-1)(q-3)} + \frac{6\Gamma^2(1+1/\beta)\Gamma(1+2/\beta)}{q(q-1)^2(q-2)} - \frac{3\Gamma^4(1+1/\beta)}{(q-1)^4}}{\left[\frac{\Gamma(1+2/\beta)}{q(q-2)} - \frac{\Gamma^2(1+1/\beta)}{(q-1)^2}\right]^2}, \; q > 4 \qquad (14)$$

Demostración. Usando los coeficientes de asimetría y kurtosis estandarizados, tenemos

$$\sqrt{\beta_1} = \frac{\mu_3 - 3\mu_1\mu_2 + 2\mu_1^3}{(\mu_2 - \mu_1^2)^{3/2}} \quad \text{y} \quad \beta_2 = \frac{\mu_4 - 4\mu_1\mu_3 + 6\mu_1^2\mu_2 - 3\mu_1^4}{(\mu_2 - \mu_1^2)^2}, \qquad (15)$$

donde $\mu_r = \mathbb{E}[W^r]$, con $r = 1, 2, 3, \ldots$ y $q > r$ es definido en (12). □



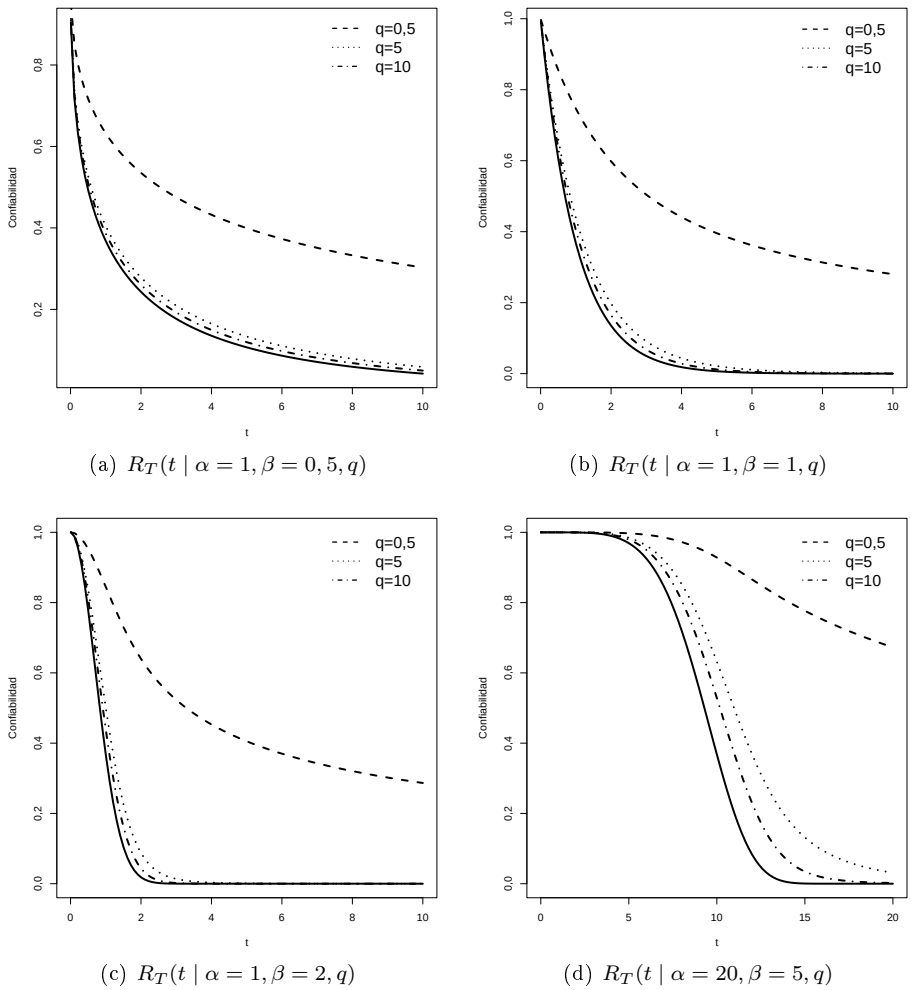
(a) Coeficiente de asimetría. (b) Coeficiente de kurtosis.

FIGURA 2: Coeficiente de asimetría y kurtosis para la distribución Slash-Weibull y Weibull (línea continua), para diferentes valores de q.

La figura 2 presenta $\sqrt{\beta_1}$ y β_2 como función del parámetro de forma. Esta figura muestra el comportamiento de los coeficientes de asimetría y kurtosis para las distribuciones Slash-Weibull y Weibull, para diferentes valores de q. Además, se puede ver que cuando el valor de q se incrementa, los coeficientes de asimetría y kurtosis tienden a los correspondientes coeficientes de la distribución Weibull. A partir de la figura 2(b) se observa claramente el efecto producido por el parámetro q en el modelo Slash-Weibull, es decir, el nuevo modelo tiene mayor kurtosis.

2.4. Función de confiabilidad y tasa de falla

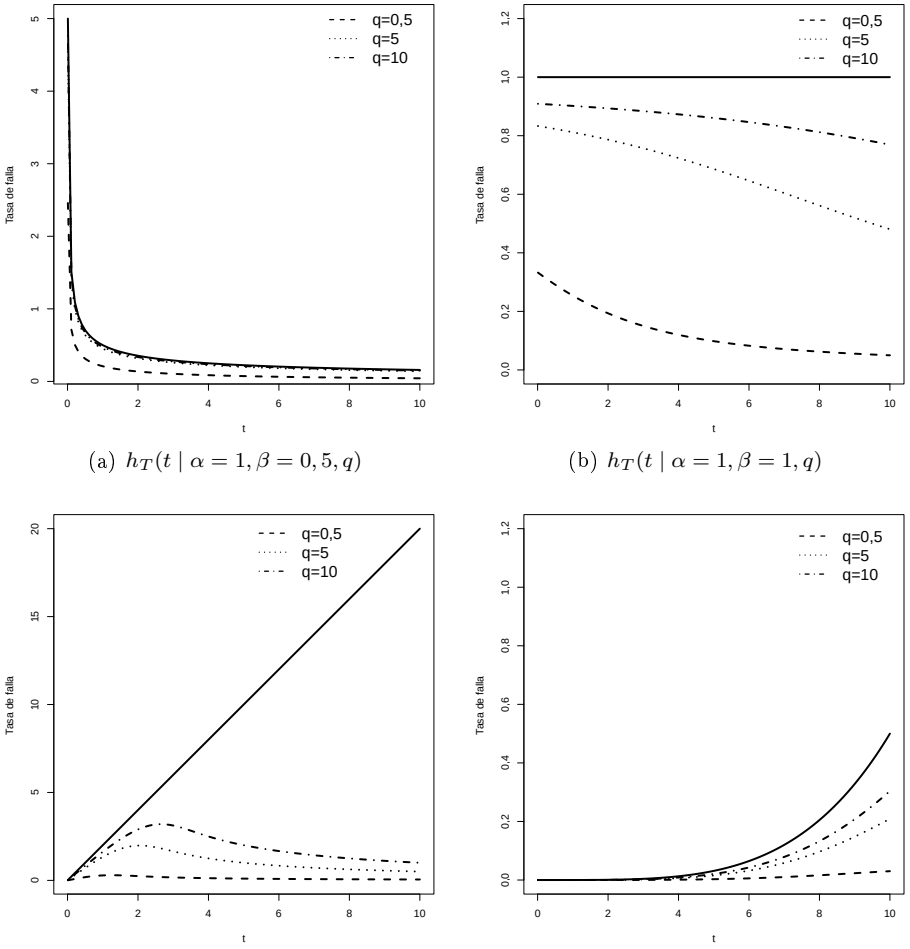
Es esta sección se estudian las funciones de confiabilidad y tasa de falla del modelo Slash-Weibull, es decir, se considera una variable aleatoria $T \sim \mathrm{SW}(\alpha, \beta, q)$. La confiabilidad de un componente (o sistema) en el tiempo t, es definido como $R_T(t \mid \alpha, \beta, q) = \mathbb{P}(T > t) = 1 - F_T(t \mid \alpha, \beta, q)$, donde T es el tiempo de vida del componente y $F_T(t \mid \alpha, \beta, q)$ es la función de distribución acumulada de la variable aleatoria T, R_T es también conocida como la función de confiabilidad de un componente (o sistema). En la figura 3 aparece la forma de la función de confiabilidad $R_T(t \mid \alpha, \beta, q)$ para varias elecciones de (α, β, q).



(a) $R_T(t \mid \alpha = 1, \beta = 0,5, q)$ (b) $R_T(t \mid \alpha = 1, \beta = 1, q)$

(c) $R_T(t \mid \alpha = 1, \beta = 2, q)$ (d) $R_T(t \mid \alpha = 20, \beta = 5, q)$

FIGURA 3: Función de confiabilidad $R_T(t \mid \alpha, \beta, q)$ para varias elecciones de (α, β, q). La línea continua corresponde a la función de confiabilidad de la densidad Weibull de dos parámetros.



(a) $h_T(t \mid \alpha = 1, \beta = 0,5, q)$ (b) $h_T(t \mid \alpha = 1, \beta = 1, q)$

(c) $h_T(t \mid \alpha = 1, \beta = 2, q)$ (d) $h_T(t|\alpha = 20, \beta = 5, q)$

FIGURA 4: Función de tasa de falla $h_T(t \mid \alpha, \beta, q)$ para varias elecciones de (α, β, q). La linea continua corresponde a la función de tasa de falla de la densidad Weibull de dos parámetros.

La figura 4 muestra la función de tasa de falla

$$h_T(t \mid \alpha, \beta, q) = \frac{f_T(t \mid \alpha, \beta, q)}{R_T(t \mid \alpha, \beta, q)}$$

correspondiente a la variable aleatoria $T \sim \text{SW}(\alpha, \beta, q)$, donde $f_T(t \mid \alpha, \beta, q)$ es la fdp de T. De la función de tasa de falla para la variable aleatoria T se desprenden las siguientes propiedades: (i) $h_T(t \mid \alpha, \beta, q) \geq 0$, $\forall\, t$, y (ii) $h_T(0 \mid \alpha, \beta, q) = 0$.

Una de las gráficas más ampliamente conocidas y utilizadas para modelar tiempos de vida de componentes (o sistemas) a través de la distribución Weibull es la curva de la bañera. En la figura 4 se puede observar la función de tasa de falla del

modelo Slash-Weibull, y que a través de este es posible construir la curva de la bañera para distintos valores de α, β y q. Cabe destacar que al ir aumentando el valor de q, el modelo Slash-Weibull se aproxima al comportamiento mostrado por la distribución Weibull de dos parámetros.

En la figura 4(a), con $\beta < 1$, se observa un comportamiento decreciente en la curva, debido a que se considera que los componentes en el inicio de su utilización presentan una alta cantidad de fallas que va disminuyendo en el tiempo. En la figura 4(b), con $\beta = 1$, se observa un comportamiento constante en la curva, debido a que se considera que los componentes pasan por un periodo con tasas de fallas constantes en el tiempo. En las figuras 4(c) y 4(d), con $\beta > 1$, se observa un comportamiento creciente en la curva, debido a que se considera que los componentes al aproximarse al final de su vida útil aumentan sus tasas de fallas en el tiempo.

3. Aspectos inferenciales

A partir de (12) es posible obtener los tres primeros momentos poblacionales y estos pueden ser usados para el cálculo de los estimadores de momentos para los respectivos parámetros (con $q > 3$). Entonces, en esta sección nos enfocaremos en el calculo de los EMV, y se realizará un análisis de simulación para estudiar el comportamiento de las estimaciones.

3.1. Estimación por máxima verosimilitud

La función de log-verosimilitud correspondiente a una muestra aleatoria $X_1, \ldots, X_n$ desde la distribución $\mathrm{SW}(\alpha, \beta, q)$ en (8) puede ser escrita como

$$\ell(\theta) = n\log q + n\log\beta - n\beta\log\alpha + (\beta-1)\sum_{i=1}^{n}\log w_i + \sum_{i=1}^{n}\log T_W(w_i \mid \alpha, \beta, q) \quad (16)$$

donde, $T_W(\cdot \mid \alpha, \beta, q)$ se define en la proposición 1.

Los EMV vienen dados por la maximización de (16). Dada la forma de (16), el sistema obtenido a partir de la obtención de las respectivas derivadas parciales debe ser resuelto numéricamente. La maximización se realiza por medio de la función optim del software R, el método específico es L-BFGS-B (Byrd, Lu, Nocedal & Zhu 1995), ya que permite que cada variable involucrada se pueda acotar. Este método corresponde a una modificación del método quasi-Newton. A continuación se presenta un estudio de simulación, donde el principal objetivo es estudiar el comportamiento de los estimadores de máxima verosimilitud para los parámetros α, β y q.

3.2. Análisis de simulación

En esta sección se estudian los resultados de varios estudios de simulación relacionados con los parámetros α, β y q. El principal objetivo es analizar el

comportamiento de los estimadores de máxima verosimilitud para los parámetros α, β y q.

El estudio es realizado por la generación de 1.000 muestras aleatorias simuladas de la distribución Slash-Weibull para diferentes valores de los parámetros del modelo. Luego del cálculo de los EMV para cada parámetro del modelo, para cada muestra generada, el valor medio y la desviación estándar empírica para las 1.000 estimaciones de cada parámetro son calculadas. Estos resultados se pueden ver en la tabla 1, donde se observa que las estimaciones son bastante estables, y lo más importante, las estimaciones son cercanas a los valores verdaderos para el tamaño de muestra considerado.

TABLA 1: Media y desviaciones estándar simuladas para los EMV de α, β y q.

			$n = 100$		
α	β	q	$\widehat{\alpha}(SD)$	$\widehat{\beta}(SD)$	$\widehat{q}(SD)$
1	0,5	3	0,965(0,219)	0,513(0,043)	2,642(0,335)
		5	1,056(0,223)	0,512(0,039)	4,990(0,122)
		7	1,017(0,204)	0,503(0,041)	7,004(0,019)
1	1	3	0,947(0,118)	1,040(0,121)	2,762(0,411)
		5	0,979(0,118)	1,025(0,076)	4,890(0,588)
		7	0,994(0,114)	1,020(0,088)	6,979(0,264)
1	2	3	0,996(0,081)	2,061(0,239)	3,004(0,449)
		5	0,999(0,059)	2,074(0,198)	4,904(0,564)
		7	1,004(0,056)	2,073(0,187)	6,850(0,577)
10	5	3	10,085(0,581)	5,176(0,856)	3,093(0,419)
		5	9,977(0,430)	5,211(0,637)	4,953(0,574)
		7	10,020(0,240)	5,070(0,573)	6,954(0,540)
			$n = 200$		
α	β	q	$\widehat{\alpha}(SD)$	$\widehat{\beta}(SD)$	$\widehat{q}(SD)$
1	0,5	3	0,938(0,148)	0,508(0,031)	2,659(0,338)
		5	1,010(0,166)	0,502(0,029)	5,016(0,095)
		7	1,010(0,136)	0,504(0,028)	7,002(0,010)
1	1	3	0,987(0,103)	1,011(0,063)	2,882(0,441)
		5	0,988(0,067)	1,030(0,065)	4,756(0,498)
		7	0,996(0,077)	1,003(0,052)	6,943(0,298)
1	2	3	0,998(0,060)	2,025(0,183)	3,010(0,393)
		5	0,978(0,046)	2,055(0,158)	4,851(0,570)
		7	0,994(0,045)	2,016(0,131)	6,838(0,530)
10	5	3	10,043(0,390)	5,058(0,521)	3,063(0,365)
		5	9,964(0,264)	5,097(0,440)	4,927(0,514)
		7	9,971(0,212)	5,069(0,439)	6,860(0,513)

4. Conclusiones

La nueva familia introducida, llamada distribución Slash-Weibull, presenta un coeficiente de kurtosis mayor que la distribución Weibull de dos parámetros considerada; este hecho puede ser útil para el ajuste de un conjunto de datos con una kurtosis mayor que la distribución Weibull ordinaria. El estudio de simulación desarrollado muestra que el modelo Slash-Weibull puede producir un ajuste mucho mejor que el modelo Weibull. Además, también se puede decir desde el estudio de simulación que los EMV presentan un comportamiento bastante bueno en términos del sesgo empírico y el error cuadrático medio.

Agradecimientos

Los autores agradecen a los árbitros y al editor por sus valiosos comentarios. J. F. Olivares-Pacheco agradece a la Comisión Nacional de Ciencia y Tecnología-Conicyt por financiar sus estudios de doctorado en la Pontificia Universidad Católica de Chile. La investigación de H. C. Cornide-Reyes ha sido parcialmente financiada por el Proyecto DIUDA 221171 de la Dirección de Investigación y Postgrado de la Universidad de Atacama, Chile.

[Recibido: marzo de 2009 — Aceptado: octubre de 2010]

Referencias

Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. (1995), 'A Limited Memory Algorithm for Bound Constrained Optimization', *SIAM Journal on Scientific Computing* **16**(5), 1190–1208.

Chen, Z. (2000), 'An New Two-parameter Lifetime Distribution with Bathtub-Shape or Increasing Failure Rate Function', *Statistics and Probability Letters* **49**(2), 155–161.

Gómez, H. W., Olivares-Pacheco, J. F. & Bolfarine, H. (2009), 'An Extension of the Generalized Birnbaum-Saunders Distributions', *Statistics and Probability Letters* **79**(3), 331–338.

Gómez, H. W., Quintana, F. A. & Torres, F. J. (2007), 'A New Family Slash-Distributions with Elliptical Contours', *Statistics and Probability Letters* **77**(7), 717–725.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995), *Continuous Univariate Distributions*, second edn, Wiley, New York.

Kafadar, K. (1982), 'A Biweight Approach to the One-Sample Problem', *Journal of the American Statistical Association* **77**(378), 416–424.

Mosteller, F. & Tukey, J. W. (1977), *Data Analysis and Regression*, Addison-Wesley.

Rogers, W. H. & Tukey, J. W. (1972), 'Understanding Some Long-tailed Symmetrical Distribution', *Statistics Neerlandia* **26**, 211–226.

Tang, Y., Xie, M. & Goh, N. T. (2003), 'Statistical Analysis of a Weibull Extension Model', *Communications in Statistics: Theory and Methods* **32**(5), 913–928.

Wang, J. & Genton, M. G. (2006), 'The Multivariate Skew-Slash Distribution', *Journal of Statistical Planning and Inference* **136**(1), 209–220.

Weibull, W. (1951), 'A Statistical Distribution Function of Wide Applicability', *Journal of Applied Mechanics* **18**, 293–297.

Zhang, T. & Xie, M. (2007), 'Failure Data Analysis with Extended Weibull Distribution', *Communications in Statistics: Simulation and Computation* **36**(3), 579–592.

Revista Colombiana de Estadística
Diciembre 2010, volumen 33, no. 2, pp. 233 a 250

Procedimiento y algoritmo de estimación en modelos multinivel para proporciones

Procedure and Estimation Algorithm in Multilevel Models for Proportions

Ernestina Castells[1,a], Mario M. Ojeda[2,b], Minerva Montero[3,c]

[1]Facultad de Matemática, Universidad Autónoma de Guerrero, Acapulco, México
[2]Facultad de Estadística e Informática, Universidad Veracruzana, Xalapa, México
[3]Departamento de Matemática, Instituto de Cibernética, Matemática y Física, La Habana, Cuba

Resumen

En este artículo se describe un procedimiento para la estimación de parámetros fijos y aleatorios en modelos multinivel para proporciones. El procedimiento de estimación se basa en el método de los mínimos cuadrados generalizados. Una vez que se formula el modelo, se demuestra que es posible aplicar la teoría asintótica de estimación en el marco del modelo lineal general. Se elabora un algoritmo que permite calcular los estimadores propuestos. La aplicación se ilustra con un ejemplo de meta-análisis. Se concluye que el procedimiento presentado puede ser una estrategia favorable en investigaciones aplicadas.

Palabras clave: mínimos cuadrados generalizados iterativos, modelos multinivel, tablas de contingencia.

Abstract

This paper describes a procedure for the estimation of fixed and random parameters in multilevel model for proportions. The estimation procedure is developed using Iterative Generalized Least Squares. Once the model is formulated, we demonstrate that it is possible to apply the asymptotic estimation theory in the framework of the general lineal model. An algorithm to calculate the proposed estimators is elaborated. We illustrate the application using an example of meta-analysis. It is concluded that the proposed procedure can be favorable strategy to do applied research.

Key words: Contingency tables, Iterative generalized least squares, Multilevel models.

[a]Profesora titular. E-mail: ernestinacg@yahoo.com
[b]Profesor titular. E-mail: mojeda@uv.mx
[c]Investigadora auxiliar. E-mail: minerva@icmf.inf.cu

1. Introducción

La flexibilidad de la modelación multinivel ha implicado un papel cada vez más importante de esta técnica dentro de la teoría estadística y las aplicaciones; sin embargo, la inferencia estadística acerca de los modelos multinivel para proporciones es una problemática que presenta dificultades teóricas y computacionales aún sin resolver. Para evitar la carga computacional y la inestabilidad asociada con la compleja integración numérica que implica la estimación de parámetros de modelos multinivel no lineales se han desarrollado varios métodos. Entre los más comunes se encuentran los que se basan en aproximaciones de la integral involucrada en la verosimilitud marginal, los métodos bayesianos basados en algoritmos de Monte Carlo mediante Cadena de Markov (MCMC), los métodos EM estocásticos y los métodos de verosimilitud simulada (Fotouhi 2003). Estos métodos, aunque en principio aplicables con cierto éxito, presentan diversos inconvenientes; una de las más notables desventajas es su alto costo computacional.

En la actualidad, continúa siendo de particular interés explorar nuevos métodos que le brinden al investigador herramientas adecuadas para la estimación de parámetros en modelos multinivel para datos categóricos. El objetivo de este artículo es fundamentar la racionalidad del método propuesto por Montero, Castell & Ojeda (2007). Este enfoque se basa fundamentalmente en la integración de tres estrategias estadísticas: 1) El uso, como base fundamental para elaborar métodos de inferencia para tablas de contingencia, de modelos asociados a la distribución asintótica gaussiana de un vector de transformaciones de las frecuencias observadas; 2) El uso de aproximaciones simplificadoras para la matriz de covarianza asintótica de los errores aleatorios de nivel-1 en el modelo multinivel propuesto y 3) La aplicación de mínimos cuadrados generalizados iterativos para la estimación de parámetros de los modelos gaussianos resultantes.

En la sección 2 se describe brevemente la filosofía del análisis multinivel para datos categóricos. En la sección 3 se formula un modelo multinivel para proporciones y en la sección 4 se demuestra que los estimadores propuestos tienen propiedades que permiten aplicar la teoría de estimación para muestras grandes. En la misma sección se demuestra la consistencia de los estimadores, tanto de los parámetros asociados a efectos fijos como a componentes de la varianza. En la sección 5 se presenta el algoritmo de estimación descrito en el anexo. Finalmente, en la sección 6, la aplicación del método se ilustra con un ejemplo.

2. Motivación para el análisis multinivel de proporciones

Muchos problemas prácticos tratan datos categóricos que conllevan al análisis de un conjunto de tablas de contingencia. Ejemplos importantes se pueden encontrar en problemas de meta-análisis (Glass 1976, Hedges & Olkin 1985), en el análisis de datos de panel (Hsiao 1995, Hamerle & Honning 1995), en los estudios de casos y controles multicentro (Lubin, Blot, Berrino, Flamant, Gillis, Kunze,

Schmäwhl & Visco 1984, Fears & Brown 1986, Breslow & Zhao 1988) y en problemas de estimación en áreas pequeñas (Rudas 1986). En todos estos casos, los individuos se seleccionan de diferentes grupos y es posible reconocer que las observaciones pertenecientes a un mismo grupo son más parecidas entre sí que las que se encuentran en grupos diferentes. Ignorar la estructura de grupos puede provocar serios problemas inferenciales (Snijders & Bosker 1999).

Cuando se analizan datos en un conjunto de tablas de contingencia es posible establecer una estructura jerárquica de datos de dos niveles, en la que los grupos que conforman las tablas se reconocen como las unidades de nivel-2 y los individuos anidados dentro de los grupos se identifican como las unidades de nivel-1. Un enfoque que distingue los diferentes niveles de las variables explicativas y toma en cuenta explícitamente la varianza dentro y entre grupos es la modelación multinivel (Goldstein 1995). En los modelos multinivel, también conocidos como modelos jerárquicos (Brik & Raudenbush 1992) o modelos de coeficientes aleatorios (Longford 1995), algunos parámetros varían de grupo a grupo, lo que permite tratarlos como efectos aleatorios.

Frecuentemente, el interés del estudio en tablas de contingencia se centra en el análisis de un gran número de funciones de las probabilidades de las celdas. Uno de los modelos más utilizados es el de regresión logística multinivel (Efron 1996, Hartzel, Liu & Agresti 2001, Lee & Nelder 2002); sin embargo, la aparición de problemas cada vez más complejos exige que otras funciones (Forthofer & Lehnen 1981) diferentes de las conocidas funciones logit o probit, también sean tomadas en consideración.

La idea básica del enfoque descrito en este artículo es usar funciones de las probabilidades como las variables dependientes en un modelo lineal multinivel (Montero 2006). Este enfoque se basa en el uso de los mínimos cuadrados ponderados o mínimos cuadrados generalizados para datos categóricos. El empleo de esta metodología, presentada por primera vez por Grizzle, Starmer & Koch (1969) se había limitado sólo al caso donde los parámetros del modelo son fijos (Grizzle et al. 1969). En este artículo esta estrategía se extiende al caso multinivel.

La aplicación del método se ilustra a través de un ejemplo de meta-análisis, que puede considerarse como un problema estadístico multinivel, ya que la información dentro de los estudios se combina en presencia de una heterogeneidad potencial entre los estudios.

3. El modelo lineal multinivel para proporciones

En esta sección se describe la estructura jerárquica impuesta por un conjunto de tablas de contingencia y se formula un modelo multinivel para proporciones. Para simplicar la futura discusión de la teoría de estimación que se presenta en la sección 4 el modelo se expresa en términos matriciales.

Sea Y una variable respuesta con R categorías. Sean $X_1, X_2, \ldots, X_t$, un conjunto de variables explicativas. Sea G un factor de estratificación en J grupos (unidades de nivel-2). Sea $\mathcal{A}$ el conjunto formado por las diferentes combinaciones de

los valores de $X_1, X_2, \ldots X_t$. Sean I el cardinal del conjunto $\mathcal{A}$ y $C = \{1, 2, \cdots I\}$. A cada elemento de $\mathcal{A}$ se le hace corresponder biunívocamente un elemento c de $\mathcal{C}$. Todos los individuos (unidades de nivel-1) con la misma combinación de valores $(x_1, x_2, \ldots x_t)$ se tratan como un subgrupo dentro de los grupos determinados por G. De cada subgrupo se seleccionan muestras aleatorias independientes de tamaño n_{ji}. Sea n_{jir} el número de individuos en la muestra del i-ésimo subgrupo del j-ésimo grupo, clasificados en la r-ésima categoría de respuesta ($i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, J$; $r = 1, 2, \ldots R$).

Sea $\pi_{r|ji} = P(Y = r \mid c = i, g = j)$. Se supone que $(n_{ji1}, n_{ji2}, \ldots, n_{jir})$ para $i-j$ fijos sigue una distribución multinomial con probabilidades $(\pi_{1|ji}, \pi_{2|ji}, \ldots, \pi_{R|ji})$ y que las muestras correspondientes a diferentes subgrupos y/o diferentes grupos son independientes. La estructura de datos descrita anteriormente puede resumirse en J tablas de contingencia $I \times R$ como la que se muestra en la tabla 1.

Tabla 1: Tabla de contingencia para el j-ésimo grupo.

	Categorías de la respuesta				
Subgrupo	1	2	$\cdots$	R	Total
1	n_{j11}	n_{j12}	$\cdots$	n_{j1R}	n_{j1}
2	n_{j21}	n_{j22}	$\cdots$	n_{j2R}	n_{j2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	n_{jI1}	n_{jI2}	$\cdots$	n_{jIR}	n_{jI}

El vector de probabilidades, π_j, asociado a la j-ésima tabla se puede escribir como $\boldsymbol{\pi}_j = \left(\boldsymbol{\pi}'_{j1}, \boldsymbol{\pi}'_{j2}, \ldots, \boldsymbol{\pi}'_{jI}\right)'$, donde $\boldsymbol{\pi}_{ji} = \left(\pi_{1|ji}, \pi_{2|ji}, \ldots, \pi_{R|ji}\right)'$ con $\sum_{r=1}^{R} \pi_{r|ji} = 1$, para cada (i, j). Cada conjunto de probabilidades tiene $R - 1$ elementos funcionalmente independientes. Los J vectores de probabilidades en cada tabla de contingencia conforman un único vector $\boldsymbol{\pi} = (\boldsymbol{\pi}'_1, \boldsymbol{\pi}'_2, \ldots, \boldsymbol{\pi}'_J)'$.

Sea $\boldsymbol{F}_m(\boldsymbol{\pi}_j)$, $m = 1, 2, \ldots, a$ un conjunto de funciones de $\boldsymbol{\pi}_j$. Defínase $\boldsymbol{F}(\boldsymbol{\pi}_j) = [\boldsymbol{F}_1(\boldsymbol{\pi}_j), \boldsymbol{F}_2(\boldsymbol{\pi}_j), \ldots, \boldsymbol{F}_a(\boldsymbol{\pi}_j)]'$ el vector de funciones de $\boldsymbol{\pi}_j$ para $j = 1, 2, \ldots, J$ y $a \leq I(R-1)$. Estas funciones expresan la estructura relevante de los datos *dentro* de los grupos. La estructura *entre* grupos se toma en cuenta al definir un único vector $\boldsymbol{F}(\boldsymbol{\pi}) = [\boldsymbol{F}(\boldsymbol{\pi}_1)', \boldsymbol{F}(\boldsymbol{\pi}_2)', \ldots, \boldsymbol{F}(\boldsymbol{\pi}_J)']'$ de funciones de probabilidades de orden $(aJ \times 1)$. Ya que se supone que la estructura *dentro* de los grupos es la misma para todos los grupos, se aplican las mismas funciones a cada uno de ellos.

Diferentes tipos de funciones pueden representarse en una manera relativamente simple usando notación matricial (Grizzle et al. 1969, Forthoper & Koch 1973). En la sección 6 se presenta un ejemplo con la función logit.

Para investigar acerca de la relación de la función de las probabilidades con las variables explicativas consideradas sobre la misma población objeto de estudio, Montero et al. (2007) proponen el siguiente modelo lineal multinivel.

Para cada uno de los J grupos se postula un modelo de nivel-1:

$$\boldsymbol{F}(\boldsymbol{\pi}_j) = \boldsymbol{X}_j \boldsymbol{\beta}_j, \quad j = 1, 2, \ldots, J$$

donde $\boldsymbol{X}_j$ es una matriz de diseño de orden $(a \times t)$ y rango t; $\boldsymbol{\beta}_j = (\beta_{1j}, \beta_{2j}, \ldots, \beta_{tj})'$ es un vector de parámetros aleatorios de orden $(t \times 1)$, donde β_{kj} es el coeficiente de la variable X_k en la la ecuación j.

La variabilidad de los J coeficientes $(\beta_{k1}, \beta_{k2}, \ldots, \beta_{kJ})$ de la k-ésima variable $(k = 1, \ldots, t)$ puede explicarse a través de un conjunto adicional de variables $Z_1, Z_2, \ldots, Z_q$, medidas a nivel de grupo, así:

$$\beta_{kj} = \boldsymbol{Z}'_{kj}\boldsymbol{\Gamma}_k + u_{kj}, \qquad k = 1, \ldots, t, \quad j = 1, 2, \ldots, J \tag{1}$$

donde $\boldsymbol{\Gamma}_k$ es el vector de orden $(q_k \times 1)$ de coeficientes asociados a las variables medidas en el nivel-2, Z_{kj} representa las observaciones de las q_k variables en el nivel-2, en la j-ésima tabla y u_{kj} son los errores aleatorios no observables.

La ecuación en (1) puede describirse de una forma más compacta, de manera que el modelo en el nivel-2 se define como:

$$\boldsymbol{\beta}_j = \boldsymbol{Z}_j\boldsymbol{\Gamma} + \boldsymbol{u}_j, \quad j = 1, 2, \ldots, J$$

donde $\boldsymbol{Z}_j = diag\left(\boldsymbol{Z}'_{1j}, \boldsymbol{Z}'_{2j}, \ldots, \boldsymbol{Z}'_{tj}\right)$ es una matriz diagonal en bloques de orden $(t \times Q)$, con $Q = q_1 + q_2 + \ldots + q_t$ y cuyos elementos son los valores de las variables explicativas en el nivel-2; $\boldsymbol{\Gamma}$ es el vector de efectos fijos de orden $(Q \times 1)$ y $\boldsymbol{u}_j$ es el vector de errores aleatorios de orden $(t \times 1)$. Se asume $E(\boldsymbol{u}_j) = 0$ y $Cov(\boldsymbol{u}_j)$ se denota por $\boldsymbol{\Omega}_{u_j}$. En la matriz de covarianza $\boldsymbol{\Omega}_{u_j}$ se utilizará $\sigma_{u_{kk^*}}$ $(k, k^* = 1, 2, \ldots, t)$ para denotar las componentes de la varianza del nivel-2. Además, $Cov(\boldsymbol{u}_j, \boldsymbol{u}_{j^*}) = 0$, para $(j, j^* = 1, 2, \ldots, J)$.

Una forma conveniente de expresar el modelo es mediante la formulación matricial dado en 2:

$$\boldsymbol{F}(\boldsymbol{\pi}) = \boldsymbol{A}\boldsymbol{\Gamma} + \boldsymbol{X}\boldsymbol{u} \tag{2}$$

donde $\boldsymbol{A} = \left(\boldsymbol{Z}'_1\boldsymbol{X}'_1, \boldsymbol{Z}'_2\boldsymbol{X}'_2, \ldots, \boldsymbol{Z}'_J\boldsymbol{X}'_J\right)'$, con $\boldsymbol{X}_j = \boldsymbol{X}_{j^*}$ para todo $j \neq j^*$, es una matriz de dimensión $(aJ \times Q)$ y $\boldsymbol{X} = diag(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_J)$ es una matriz de dimensión $(aJ \times tJ)$. El vector $\boldsymbol{\Gamma}$ está formado por los Q efectos fijos y $\boldsymbol{u} = (u_1, u_2, \ldots, u_J)'$.

Se supone que los efectos aleatorios de grupos diferentes son mutuamente independientes, $E(\boldsymbol{u}) = \boldsymbol{0}$ y $Cov(\boldsymbol{u}) = \boldsymbol{\Omega}_u$, con $\boldsymbol{\Omega}_u = \left[I_J \otimes \boldsymbol{\Omega}_{u_j}\right]$, donde $\otimes$ representa el producto de Kronecker. De aquí que $E[\boldsymbol{F}(\boldsymbol{\pi})] = \boldsymbol{A}\boldsymbol{\Gamma}$ y $Var[\boldsymbol{F}(\boldsymbol{\pi})] = \boldsymbol{V}_{\boldsymbol{F}(\boldsymbol{\pi})}$, donde: $\boldsymbol{V}_{\boldsymbol{F}(\pi)} = \boldsymbol{X}\boldsymbol{\Omega}_u\boldsymbol{X}'$.

Es importante destacar que, a diferencia del caso en el que los parámetros del modelo son fijos y las inferencias conciernen sólo a los grupos especificados (clases, regiones, etc.), en el modelo multinivel para proporciones el interés de las inferencias se dirige hacia la población de grupos, no sólo a aquellos que pasan a representar la muestra. Los grupos de individuos que conforman las tablas pueden considerarse como unidades anónimas, de la misma forma que lo son las observaciones elementales.

Una vez formulado el modelo multinivel, es posible aplicar la teoría asintótica en el marco del modelo lineal general. En la próxima sección se muestran los resultados teóricos que permiten demostrar la validez de este enfoque.

4. Estimación

Para estimar los efectos fijos y los componentes de la varianza de modelos multinivel del tipo definido en la ecuación (2), se propone aplicar un enfoque basado en los mínimos cuadrados generalizados.

Supóngase que se han hecho observaciones en J grupos diferentes según la estructura de datos definida en la sección 3. Asociado a la muestra del i-ésimo subgrupo de la j-ésima tabla existe un vector de proporciones muestrales $\boldsymbol{p}_{ji} = (p_{ji1}, p_{ji2}, \ldots, p_{jiR})^{'}$, donde $p_{jir} = n_{jir}/n_{ji}$. Sea $\boldsymbol{p}_j = [\boldsymbol{p}'_{j1}, \boldsymbol{p}'_{j2}, \ldots, \boldsymbol{p}'_{jI}]'$ la estimación muestral de $\boldsymbol{\pi}_j$. Cuando las muestras de los I subgrupos son independientes, la matriz de covarianza de $\boldsymbol{p}_j$ está dada por $\boldsymbol{V}_{\pi_j}$, la cual es una matriz diagonal en bloques de dimensión $(IR \times IR)$, con las matrices $\boldsymbol{V}_{\pi_{ji}} = 1/n_{ji}[\boldsymbol{D}_{\pi_{ji}} - \boldsymbol{\pi}_{ji}\boldsymbol{\pi}^{'}_{ji}]$, para $i = 1, 2, \ldots, I$ y $j = 1, 2, \ldots, J$, sobre la diagonal principal, donde $\boldsymbol{D}_{\boldsymbol{\pi}_{ji}}$ es una matriz diagonal de dimensión $(I \times I)$ con elementos del vector $\boldsymbol{\pi}_{ji}$ sobre la diagonal principal.

Los J vectores de proporciones observadas en cada tabla de contingencia conforman un único vector $\boldsymbol{p} = [\boldsymbol{p}'_1, \boldsymbol{p}'_2, \ldots, \boldsymbol{p}'_J]'$ con media $\boldsymbol{\pi}$ y matriz de covarianza $V_{\boldsymbol{\pi}}$, la cual involucra a las matrices de covarianza de $\boldsymbol{p}_j$. Cuando las observaciones de diferentes grupos son independientes, la covarianza entre observaciones de diferentes grupos es cero, por tanto, la matriz de covarianza de $\boldsymbol{p}$ tiene forma de una matriz diagonal en bloques con las matrices $\boldsymbol{V}_{\boldsymbol{\pi}_j}$ sobre la diagonal principal.

Se puede demostrar por el teorema central del límite multivariado (Rao 1973, p. 128) que las propociones muestrales (y las proporciones muestrales condicionadas) tienen distribución asintóticamente normal. Por el método delta (Agresti 2002, p. 577), las funciones (y funciones condicionales) de tales proporciones, son también asintóticamente normales. El teorema 1, a continuación, especifica las distribuciones de tales funciones.

Teorema 1. *Sea $\boldsymbol{F}(\boldsymbol{p})$ la estimación muestral de $\boldsymbol{F}(\boldsymbol{\pi})$. Se supone que $\boldsymbol{F}$ tiene derivadas parciales continuas de segundo orden en una región abierta que contiene a $\boldsymbol{\pi}$. Sea $\boldsymbol{H}$ la matriz jacobiana de la función $\boldsymbol{F}(\boldsymbol{\pi})$, evaluada correspondientemente, la cual se supone no nula.*

Entonces:

$$i)\boldsymbol{F}(\boldsymbol{p}) \mid \boldsymbol{u} \xrightarrow{d} N\left(\boldsymbol{A\Gamma} + \boldsymbol{Xu}, \boldsymbol{H}\boldsymbol{V}_{\pi}\boldsymbol{H}'\right)$$

$$ii)\boldsymbol{F}(\boldsymbol{p}) \xrightarrow{d} N\left(\boldsymbol{A\Gamma}, \boldsymbol{X\Omega}_u\boldsymbol{X}^{'} + \boldsymbol{H}\boldsymbol{V}_{\pi}\boldsymbol{H}'\right)$$

La demostración del teorema es consecuencia directa de aplicar el método delta.

En esta fase del análisis es posible postular el siguiente modelo en términos de las proporciones observadas:

$$\boldsymbol{F}(\boldsymbol{p}) = \boldsymbol{A\Gamma} + \boldsymbol{Xu} + \boldsymbol{e} \tag{3}$$

donde $\boldsymbol{e}$ se reconoce como el vector de errores aleatorios en el nivel-1. $\boldsymbol{A}$, $\boldsymbol{\Gamma}$, $\boldsymbol{X}$ y $\boldsymbol{u}$ se definen como en la ecuación en (2).

La parte aleatoria del modelo comprende dos errores: uno para cada nivel. Se supone que $\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{\Omega}_u)$ y $\boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{H}\boldsymbol{V}_\pi\boldsymbol{H}')$, donde las matrices de covarianza para los errores de nivel-2 y los errores de nivel-1 son funciones de los parámetros $\boldsymbol{u}$ y $\boldsymbol{\pi}$, respectivamente. Sin embargo, como sugirió Goldstein (1987), se puede introducir una simplificación y requerir simplemente que las varianzas de los errores en el nivel-1 sean inversamente proporcional es a n_{ji}. Si además de la reparametrización se supone una variación simple aleatoria a través de las tablas, entonces se puede suponer que la varianza entre tablas es la misma para cada uno de los I subgrupos. En este caso $\boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{\Omega}_e)$, donde $\boldsymbol{\Omega}_e$ es una matriz diagonal con los elementos σ_e^2/n_{ji} en la diagonal principal.

El hecho de que hay más de un término de error añade una complicación a los procedimientos de estimación. Existen tres tipos de parámetros que pueden ser estimados: los efectos fijos, los coeficientes aleatorios del primer nivel (que pueden ser estimados o no, porque el modelo general que resulta de sustituir los modelos del nivel-2 en el modelo de nivel-1 no depende de ellos) y los componentes de la varianza y covarianza. A continuación se proponen los estimadores para los efectos fijos y los componentes de la varianza. El desarrollo para los efectos fijos y los componentes de la varianza hace uso de las ideas presentadas en Castells (1985).

4.1. Efectos fijos

Bajo las suposiciones de los errores del modelo en (3), éste se puede escribir como un modelo lineal con parámetros no aleatorios, tal que:

$$\boldsymbol{F}(\boldsymbol{p}) = \boldsymbol{A}\boldsymbol{\Gamma} + \boldsymbol{e}^* \quad \text{donde} \quad \boldsymbol{e}^* = \boldsymbol{X}\boldsymbol{u} + \boldsymbol{e} \tag{4}$$

Dado el modelo en (4) y de acuerdo con (Rao 1973, p. 306), se tiene que el mejor estimador lineal de $\boldsymbol{\Gamma}$ es el estimador mínimo cuadrado generalizado:

$$\widehat{\boldsymbol{\Gamma}} = \left(\boldsymbol{A}'\boldsymbol{V}_\lambda^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}'\boldsymbol{V}_\lambda^{-1}\boldsymbol{F}(\boldsymbol{p})$$

donde $\boldsymbol{V}_\lambda = \boldsymbol{X}\boldsymbol{\Omega}_u\boldsymbol{X}' + \boldsymbol{\Omega}_e$.

La existencia del estimador $\widehat{\boldsymbol{\Gamma}}$ está sujeta a la existencia de las inversas involucradas. Siendo el estimador $\widehat{\boldsymbol{\Gamma}}$ de la forma dada, está claro (Rao 1973, p. 307) que es insesgado para $\boldsymbol{\Gamma}$ y que

$$Var\left(\widehat{\boldsymbol{\Gamma}}\right) = \left(\boldsymbol{A}'\boldsymbol{V}_\lambda^{-1}\boldsymbol{A}\right)^{-1}$$

Sea $\overrightarrow{\mathbf{C}}$ el vector que resulta de escribir las columnas de una matriz cualquiera $\mathbf{C}$, una debajo de las otras. Entonces la matriz $\boldsymbol{V}_\lambda$ depende del parámetro $\boldsymbol{\lambda} = \begin{pmatrix} \overrightarrow{\boldsymbol{\Omega}_u} \\ \overrightarrow{\boldsymbol{\Omega}_e} \end{pmatrix}$ y si éste fuese conocido, $\widehat{\boldsymbol{\Gamma}}$ y su matriz de covarianza serían calculables. Suponer $\boldsymbol{\lambda}$ conocido significaría una restricción para la aplicación práctica de estos estimadores. Una situación mucho más realista sería suponer λ desconocido,

estimar éste y sustituir las expresiones de sus estimadores en $\widehat{\boldsymbol{\Gamma}}$ logrando así un estimador en dos etapas para Γ de la forma:

$$\widehat{\widehat{\boldsymbol{\Gamma}}} = \left(\boldsymbol{A}'\widehat{\boldsymbol{V}}_{\lambda}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}'\widehat{\boldsymbol{V}}_{\lambda}^{-1}\boldsymbol{F}\left(\boldsymbol{p}\right)$$

donde $\widehat{\boldsymbol{V}}_{\lambda}$ es la estimación de $\boldsymbol{V}_{\lambda}$.

4.2. Componentes de la varianza

Un procedimiento para estimar las componentes de la varianza es el siguiente: el vector de errores $\boldsymbol{e}^{*}$ se puede escribir en la forma: $\boldsymbol{e}^{*} = \boldsymbol{\varpi}_1\,\boldsymbol{\xi}_1 + \boldsymbol{\varpi}_2\boldsymbol{\xi}_2$, donde: $\boldsymbol{\varpi}_1 = \boldsymbol{X}, \boldsymbol{\varpi}_2 = \boldsymbol{I}_{aJ}, \xi_1 = \boldsymbol{u}$ y $\xi_2 = \boldsymbol{e}$, con: $E\left(\boldsymbol{\xi}_i\right) = 0\ \forall\ i = 1, 2$; $Cov(\boldsymbol{\xi}_1) = \boldsymbol{\Omega}_u$; $Cov\left(\boldsymbol{\xi}_2\right) = \boldsymbol{\Omega}_e$; $Cov\left(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right) = Cov\left(\boldsymbol{\xi}_2, \boldsymbol{\xi}_1\right) = 0$; por tanto: $Var\left(\boldsymbol{e}^{*}\right) = \boldsymbol{\varpi}_1\boldsymbol{\Omega}_u\boldsymbol{\varpi}_1' + \boldsymbol{\Omega}_e$.

Queda claro que éste es un modelo de componentes de la varianza.

Sea $L\left(\boldsymbol{A}\right)$ el espacio generado por las columnas de la matriz $\boldsymbol{A}$; $\left[L\left(\boldsymbol{A}\right)\right]^{\perp}$ el espacio complemento ortogonal de $L\left(\boldsymbol{A}\right)$; $\boldsymbol{R} = \left[\boldsymbol{I} - \boldsymbol{A}\left(\boldsymbol{A}'\boldsymbol{A}\right)^{-1}\boldsymbol{A}'\right]$ el proyector sobre $\left[L\left(\boldsymbol{A}\right)\right]^{\perp}$, y $r\left(\mathbf{C}\right)$ el rango de una matriz cualquiera $\mathbf{C}$. Entonces $r\left(\boldsymbol{R}\right) = aJ - r\left(\boldsymbol{A}\right) = n$. Sea $\boldsymbol{M}$ una matriz de orden $n \times aJ$ tal que $r\left(\boldsymbol{M}\right) = n$ y $\boldsymbol{MA} = \mathbf{0}$.

Defínase $\widetilde{\boldsymbol{F}} = \boldsymbol{MF}\left(\boldsymbol{p}\right)$; se tiene que

$$E\left(\widetilde{\boldsymbol{F}}\right) = E\left(\boldsymbol{MF}\left(\boldsymbol{p}\right)\right) = \boldsymbol{MA\Gamma} = \mathbf{0}$$

$$Cov\left(\widetilde{\boldsymbol{F}}\right) = Cov\left(\boldsymbol{MF}\left(\boldsymbol{p}\right)\right) = \boldsymbol{MV}_{\lambda}\boldsymbol{M}'$$

Como $\boldsymbol{F}\left(\boldsymbol{p}\right) \xrightarrow{d} N\left(\boldsymbol{AF}, \boldsymbol{V}_{\lambda}\right)$ entonces $\widetilde{\boldsymbol{F}}\left(\boldsymbol{p}\right) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{MV}_{\lambda}\boldsymbol{M}'\right)$

Se tiene que:

$$E\left(\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'}\right) = \overrightarrow{\boldsymbol{MV}_{\lambda}\boldsymbol{M}'} = \left(\overrightarrow{\boldsymbol{M}\boldsymbol{\varpi}_1\boldsymbol{\varpi}_1'\boldsymbol{M}'} \vdots \overrightarrow{\boldsymbol{MM}'}\right)\begin{pmatrix}\overrightarrow{\boldsymbol{\Omega}_u}\\ \overrightarrow{\boldsymbol{\Omega}_e}\end{pmatrix} = \boldsymbol{Z}^{*}\boldsymbol{\lambda}$$

donde $\boldsymbol{z}^{*} = \left(\overrightarrow{\boldsymbol{M}\boldsymbol{\varpi}_1\boldsymbol{\varpi}_1'\boldsymbol{M}'} \vdots \overrightarrow{\boldsymbol{MM}'}\right)$ y $\boldsymbol{\lambda} = \left(\overrightarrow{\boldsymbol{\Omega}_u} \vdots \overrightarrow{\boldsymbol{\Omega}_e}\right)'$

Sea el modelo lineal dado por: $\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'} = \boldsymbol{Z}^{*}\boldsymbol{\lambda} + \boldsymbol{\mu}$, con $\boldsymbol{\mu} = \left(\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'} - \boldsymbol{Z}^{*}\boldsymbol{\lambda}\right)$, $E\left(\boldsymbol{\mu}\right) = \mathbf{0}$ y $Cov\left(\boldsymbol{\mu}\right) = \boldsymbol{V}_{\lambda}^{*}$, donde $\boldsymbol{V}_{\lambda}^{*}$ tiene la expresión:

$$\boldsymbol{V}_{F}^{*} = E\left[\left(\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'}\right)\left(\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'}\right)'\right] - \left[E\left(\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'}\right)E\left(\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'}\right)'\right]$$

pero

$$\left[\left(\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'}\right)\left(\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'}\right)'\right] = \begin{bmatrix} \widetilde{\boldsymbol{f}_1}\widetilde{\boldsymbol{f}_1}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}' & \widetilde{\boldsymbol{f}_1}\widetilde{\boldsymbol{f}_2}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}' & \dots & \widetilde{\boldsymbol{f}_1}\widetilde{\boldsymbol{f}_n}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}' \\ \widetilde{\boldsymbol{f}_2}\widetilde{\boldsymbol{f}_1}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}' & \widetilde{\boldsymbol{f}_2}\widetilde{\boldsymbol{f}_2}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}' & \dots & \widetilde{\boldsymbol{f}_2}\widetilde{\boldsymbol{f}_n}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}' \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\boldsymbol{f}_n}\widetilde{\boldsymbol{f}_1}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}' & \widetilde{\boldsymbol{f}_n}\widetilde{\boldsymbol{f}_2}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}' & \dots & \widetilde{\boldsymbol{f}_n}\widetilde{\boldsymbol{f}_n}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}' \end{bmatrix}$$

donde $\widetilde{\boldsymbol{f}_i}$ es la i-ésima columna de la matriz $\widetilde{\boldsymbol{F}}$.

Calculando $E\left(\widetilde{\boldsymbol{f}_i}\widetilde{\boldsymbol{f}_j}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'\right)$, empleando el momento de orden 4 de una distribución normal multivariada (ver Anderson 1958), se obtiene:

$$\begin{aligned} E\left(\widetilde{\boldsymbol{f}_i}\widetilde{\boldsymbol{f}_j}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'\right) &= \tau_{ij}\boldsymbol{R} + \left[\tau_{i1}\boldsymbol{R}(j), \tau_{i2}\boldsymbol{R}(j), \dots, \tau_{in}\boldsymbol{R}(j)\right] \\ &\quad + \left[\tau_{j1}\boldsymbol{R}(i), \tau_{j2}\boldsymbol{R}(i), \dots, \tau_{jn}\boldsymbol{R}(i)\right] \\ &= \tau_{ij}\boldsymbol{R} + \boldsymbol{R}'(i) \otimes \boldsymbol{R}(j) + \boldsymbol{R}'(j) \otimes \boldsymbol{R}(i) \end{aligned}$$

En la expresión anterior, y con el objetivo de simplificar la formulación matemática, se ha denotado: $\boldsymbol{M}\boldsymbol{V}_\lambda\boldsymbol{M}' = \boldsymbol{R}$ y $\boldsymbol{R}(i)$ una columna cualquiera de $\boldsymbol{R}$.

Se ve fácilmente que:

$$\begin{aligned} E\left[\left(\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'}\right)\left(\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'}\right)'\right] &= \boldsymbol{R} \otimes \boldsymbol{R} + \left[\boldsymbol{R} \otimes \boldsymbol{R}(1), \boldsymbol{R} \otimes \boldsymbol{R}(2), \dots, \boldsymbol{R} \otimes \boldsymbol{R}(n)\right] \\ &\quad + \left[\boldsymbol{R}'(1) \otimes \overrightarrow{\boldsymbol{R}}, \boldsymbol{R}'(2) \otimes \overrightarrow{\boldsymbol{R}}, \dots, \boldsymbol{R}'(n) \otimes \overrightarrow{\boldsymbol{R}}\right] \\ &= \boldsymbol{R} \otimes \boldsymbol{R} + \left[\boldsymbol{R} \otimes \boldsymbol{R}(1), \boldsymbol{R} \otimes \boldsymbol{R}(2), \dots, \boldsymbol{R} \otimes \boldsymbol{R}(n)\right] + \overrightarrow{\boldsymbol{R}'} \otimes \overrightarrow{\boldsymbol{R}} \end{aligned}$$

Por otra parte:

$$\left[E\left(\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'}\right)E\left(\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'}\right)'\right] = \left(\overrightarrow{\boldsymbol{R}}\right)\left(\overrightarrow{\boldsymbol{R}}\right)' = \overrightarrow{\boldsymbol{R}'} \otimes \overrightarrow{\boldsymbol{R}}$$

Entonces:

$$\begin{aligned} Cov\left(\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'}\right) &= \left[\boldsymbol{R} \otimes \boldsymbol{R}\right] + \left[\boldsymbol{R} \otimes \boldsymbol{R}(1), \boldsymbol{R} \otimes \boldsymbol{R}(2), \dots, \boldsymbol{R} \otimes \boldsymbol{R}(n)\right] \\ &= \boldsymbol{V}^*_{\lambda_0} \end{aligned}$$

Minimizando, respecto a λ, la norma $\left\|\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}}'} - \boldsymbol{Z}^*\boldsymbol{\lambda}\right\|_{\mathbf{V}^{*-}_{\lambda_0}}$ con:

$$\begin{aligned} \boldsymbol{V}^*_{\lambda_0} &= \left[\left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}\boldsymbol{M}'\right) \otimes \left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}\boldsymbol{M}'\right)\right] \\ &\quad + \left[\left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}\boldsymbol{M}'\right) \otimes \left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}\boldsymbol{M}'\right)(1) \dots \left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}\boldsymbol{M}'\right) \otimes \left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}\boldsymbol{M}'\right)(n)\right] \end{aligned}$$

donde λ_0 es un valor inicial para $\boldsymbol{\lambda}$. Este valor podría obtenerse de la siguiente manera: se puede buscar el estimador mínimo cuadrado generalizado para $\boldsymbol{\Gamma}$, suponiendo $Cov(\boldsymbol{\xi}_1) = 0$, y luego, examinando los residuos, obtener un valor inicial para λ.

El estimador para los componentes de la varianza tendrá la forma:

$$\widehat{\boldsymbol{\lambda}} = \left(\boldsymbol{Z}^{*'}\boldsymbol{V}_{\lambda_0}^{*-}\boldsymbol{Z}^{*}\right)\boldsymbol{Z}^{*'}\boldsymbol{V}_{\lambda_0}^{*-}\left(\overrightarrow{\widetilde{\boldsymbol{F}}\,\widetilde{\boldsymbol{F}'}}\right) \tag{5}$$

(se ha utilizado el símbolo $\boldsymbol{B}^{-}$ para indicar la inversa generalizada de una matriz $\mathbf{B}$ cualquiera). La expresión de $\widehat{\lambda}$ se da en función de la inversa generalizada porque las inversas involucradas nunca existen.

4.3. Propiedades de los estimadores

En esta sección se prueba que el estimador $\widehat{\boldsymbol{\lambda}}$ en 5 es un estimador del tipo $\boldsymbol{\lambda}_0$-MINQUE [1] propuesto por Kleffe (1976).

Teorema 2. *El estimador $\widehat{\boldsymbol{\lambda}}_{\lambda_0}$ es un estimador del tipo $\boldsymbol{\lambda}_0$-MINQUE.*

Demostración. Sólo hay que considerar el resultado del teorema 3 del anexo (dado por Castells 1985) y probar que $\widehat{\boldsymbol{\lambda}}_{\lambda_0}$ es una solución de la ecuación:

$$S_Z(\boldsymbol{\lambda}_0)\,\lambda = \mathbf{U}(\boldsymbol{\lambda}_0, \mathbf{F}(\mathbf{p}))$$

donde

$$\boldsymbol{U}_i = \boldsymbol{F}(\boldsymbol{p})'\boldsymbol{M}'\left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}\boldsymbol{M}'\right)^{-1}\boldsymbol{M}\widetilde{\mathbf{F}}_i\boldsymbol{M}'\left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}\boldsymbol{M}'\right)^{-1}\boldsymbol{M}\widetilde{\boldsymbol{F}}_j M'$$

con

$$S_M(\boldsymbol{\lambda}_0)_{ij} = tr\left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}\boldsymbol{M}'\right)^{-1}\boldsymbol{M}\widetilde{\mathbf{F}}_i\boldsymbol{M}'\left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}\boldsymbol{M}'\right)^{-1}\boldsymbol{M}\widetilde{\mathbf{F}}_j\boldsymbol{M}', \quad j = 1,2$$

$$\widetilde{\boldsymbol{F}}_1 = \boldsymbol{I}_{aJ}, \qquad \widetilde{\boldsymbol{F}}_2 = \boldsymbol{\omega}_1\boldsymbol{\omega}_1'$$

El elemento i del vector $\left(\boldsymbol{A}'\boldsymbol{V}_{\lambda_0}^{*-}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}'}\right)$ se puede escribir de la siguiente forma:

$$\begin{aligned}
\left(\boldsymbol{A}'\boldsymbol{V}_{\lambda_0}^{*-}\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}'}}\right) &= \left(\overrightarrow{\boldsymbol{M}_i\widetilde{\boldsymbol{F}}_i\boldsymbol{M}'}\right)'\boldsymbol{V}_{\lambda_0}^{*-}\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}'}} \\
&= \left(\overrightarrow{\boldsymbol{M}\widetilde{\boldsymbol{F}}_i\boldsymbol{M}'}\right)'\Theta^{-1}\overrightarrow{\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}'}} \\
&= \frac{1}{2}tr\left[\left(\boldsymbol{M}\boldsymbol{V}_{\boldsymbol{\lambda}_0}\boldsymbol{M}'\right)^{-1}\boldsymbol{M}\widetilde{\boldsymbol{F}}_i\boldsymbol{M}'\left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}\boldsymbol{M}'\right)^{-1}\widetilde{\boldsymbol{F}}\widetilde{\boldsymbol{F}'}\right] \\
&= \frac{1}{2}\widetilde{\mathbf{F}}'\left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}\boldsymbol{M}'\right)^{-1}\boldsymbol{M}\widetilde{\boldsymbol{F}}_i\boldsymbol{M}'\left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}\boldsymbol{M}'\right)^{-1}\widetilde{\boldsymbol{F}}
\end{aligned}$$

[1] En inglés: *minimum norm quadratic unbiased estimation.*

El elemento i-j de la matriz $\left(\boldsymbol{A}'\boldsymbol{V}_{\lambda_0}^{-}\boldsymbol{A}\right)$ viene expresado por:

$$\begin{aligned}\left(\boldsymbol{A}'\boldsymbol{V}_{\lambda_0}^{-}\boldsymbol{A}\right)_{ij} &= \left(\overrightarrow{\boldsymbol{M}\widetilde{\boldsymbol{F}}_i\boldsymbol{M}'}\right)' \boldsymbol{V}_{\boldsymbol{\lambda}_0}^{*-} \left(\overrightarrow{\boldsymbol{M}\widetilde{\boldsymbol{F}}_j\boldsymbol{M}'}\right) \\ &= \frac{1}{2}tr\left[\left(\boldsymbol{M}\widetilde{\boldsymbol{F}}_i\boldsymbol{M}'\right)\left(\boldsymbol{M}\boldsymbol{V}_{\boldsymbol{\lambda}_0}^{*}\boldsymbol{M}'\right)^{-1}\left(\boldsymbol{M}\widetilde{\boldsymbol{F}}_j\boldsymbol{M}'\right)\left(\boldsymbol{M}\boldsymbol{V}_{\boldsymbol{\lambda}_0}^{*}\boldsymbol{M}'\right)^{-1}\right] \\ &= \frac{1}{2}tr\left[\left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}^{*}\boldsymbol{M}'\right)\left(\boldsymbol{M}\widetilde{\boldsymbol{F}}_i\boldsymbol{M}'\right)^{-1}\left(\boldsymbol{M}\boldsymbol{V}_{\lambda_0}^{*}\boldsymbol{M}'\right)^{-1}\left(\boldsymbol{M}\widetilde{\mathbf{F}}_j\boldsymbol{M}'\right)\right]\end{aligned}$$

lo cual prueba que $\widehat{\lambda}_{\lambda_0}$ es un estimador del tipo definido $\boldsymbol{\lambda}_0$-MINQUE. □

5. Validación del algoritmo de estimación

Como parte de la estrategia de estimación, en este artículo se presentan (ver anexo) los pasos básicos de un algoritmo de mínimos cuadrados generalizados iterativos siguiendo el enfoque propuesto. La validez del procedimiento se ha venido explorando a través de varios estudios de simulación para conjuntos de datos balanceados (el tamaño de las muestras es el mismo para todos los subgrupos) y desbalanceados (el tamaño de las muestras de los subgrupo es diferente). En todos los casos se utilizó un modelo de regresión logística donde los logaritmos de la razón de riesgo (*log odds ratios*) se reconocen como efectos aleatorios de una población de grupos. El interés de la investigación se centró en el análisis de las estimaciones de los dos parámetros fijos y la varianza al nivel de grupo.

En el caso balanceado (Montero, Castell & Ojeda 2008) se examinó la influencia de diferentes tamaños de muestras y magnitudes de la varianza de los efectos aleatorios sobre la precisión de las estimaciones. Se consideraron setenta diseños diferentes, dados por las combinaciones de cinco números de tablas de contingencia (10, 25, 50, 75, 100), siete tamaños de muestra (10, 25, 50, 75, 100, 200, 300) de los subgrupos y dos magnitudes de varianza, una grande y otra pequeña. Los resultados indican que las estimaciones de los parámetros fijos son precisas para muestras de tamaño moderadamente pequeñas, pero se necesita un mayor número de observaciones para alcanzar un comportamiento razonable del estimador para la varianza de nivel-2. En general, para alcanzar mayor precisión en las estimaciones es más importante un número grande de individuos por subgrupos ($\geqslant 200$) que un número grande de grupos. La diferencia en precisión de las estimaciones para diseños con 50 o más tablas de contingencia no es sustancial.

En el caso desbalanceado se encontró que en ciertas situaciones, las estimaciones de la varianza producen sesgos grandes y estimaciones negativas. Para mejorar los resultados se propuso corregir el algoritmo aplicando una técnica basada en la descomposición de valores singulares truncados en la solución de los mínimos cuadrados generalizados para estimar los componentes de la varianza. Mediante simulación se mostró la efectividad de la técnica en cuanto a la reducción del sesgo de los estimadores (Montero & Guerra 2005). En el estudio se fijó el número de tablas. Los tamaños muestrales en los subgrupos se generaron a partir de tres distribuciones uniformes diferentes, dando lugar a tres tipos de diseños (ligeramente

desbalanceado, moderadamente desbalanceado y muy desbalanceado). Se consideraron dos magnitudes de varianza de nivel-2, las mismas que en el caso balanceado. Para las especificaciones consideradas los resultados mostraron niveles aceptables de sesgo y precisión. Se observó además que la calidad de las estimaciones no se afectó por el grado de desbalance de los datos.

Para todos los casos se encontró que el método de estimación se comporta mejor cuando la varianza de los efectos aleatorios es pequeña.

6. Ejemplo

Para ilustrar el enfoque discutido en este artículo, se usaron los datos reportados en Turner, Omar, Yang, Goldstein & Thompson (2000), consistentes en 22 ensayos clínicos realizados para investigar el efecto de descontaminación selectiva del tracto digestivo, sobre el riesgo de infección del tracto respiratorio. Se seleccionaron aleatoriamente pacientes de unidades de cuidados intensivos, ya sea para recibir un tratamiento de combinación de antibióticos o para no recibir tratamiento. En la tabla 3 (ver anexo) se presenta la proporción de pacientes infectados en cada tratamiento para cada uno de los ensayos, así como los logaritmos de la razón de riesgo (*log odds ratios*) y sus varianzas.

Los datos pueden considerarse dentro de una estructura jerárquica, donde los subgrupos de pacientes (tratamiento y control) se anidan dentro de los ensayos. De esta manera, los subgrupos de pacientes se consideran como las unidades del nivel-1 y los ensayos, las unidades de nivel-2.

El meta-análisis de los ensayos con respuesta binaria se realizó ajustando un modelo de regresión logística en el que se fijan los efectos de los tratamientos y se permite que los logaritmos de la razón de riesgo (*log odds ratios*) varíen a través de los J ensayos (Turner et al. 2000):

$$logit\left(\pi_{ij}\right) = \beta_{0j} x_{ij} + \sum_{k=1}^{J} \beta_k D_{kij} \quad \text{en el nivel-1 (“dentro” de los ensayos)} \tag{6}$$

$$\begin{aligned} \beta_{0j} &= \theta + u_j \quad \text{en el nivel-2 (“entre” los ensayos)} \\ u_j &\sim N\left(0, \sigma_u^2\right) \end{aligned} \tag{7}$$

donde π_{ij} es la probabilidad de infectarse para los individuos del i-ésimo subgrupo en el j-ésimo ensayo, $x_{ij} = 0/1$ indica su subgrupo control/tratamiento, y el conjunto de $D_{1ij}, \ldots, D_{Jij} = 0/1$, su pertenencia al ensayo.

Sustituyendo 7 en 6 se obtiene el siguiente modelo:

$$logit\left(\pi_{ij}\right) = \left(\theta + u_j\right) x_{ij} + \sum_{k=1}^{J} \beta_k D_{kij} \tag{8}$$

donde θ (*log odds ratio*) representa el efecto promedio de interés.

Las estimaciones de los parámetros del modelo en (8), obtenidas mediante el enfoque de mínimos cuadrados generalizados (GLS[2]) propuesto en este artículo, se comparan con las mostradas por Turner et al. (2000), utilizando otros métodos disponibles en el sistema de programas MLwiN (Goldstein, Rasbash, Plewis, Draper, Browne, Yang, Woodhouse & MJR 1998): el método de cuasi-verosimilitud marginal (MQL), el método de cuasi-verosimilitud penalizada (PQL) y un método Bootstrap paramétrico (ver Goldstein & Rasbash 1996, Kuk 1995, para una discusión detallada de los métodos).

En la tabla 2 se muestran los valores estimados de θ y σ_u^2 y los correspondientes intervalos de confianza Wald. Nótese que la estimación GLS de θ es muy similar a la obtenida por el método Bootstrap de corrección de sesgos. En el caso de σ_u^2, la estimación GLS de σ_u^2 es la más extrema.

Tabla 2: Estimaciones del meta-análisis de los datos de infección del tracto respiratorio.

Método	Log OR θ	(IC 95 %)	Varianza entre ensayos σ_u^2	(IC 95 %)
MQL	$-1{,}43$	$(-1{,}80,\ -1{,}07)$	0,46	(0,08, 0,84)
PQL	$-1{,}49$	$(-1{,}90,\ -1{,}07)$	0,64	(0,14, 1,14)
PQL-Bootstrap	$-1{,}66$	$(-2{,}26,\ -1{,}05)$	0,71	(0,00, 1,13)
GLS	$-1{,}63$	$(-2{,}06,\ -1{,}04)$	1,06	(0,44, 1,69)

Para completar la comparación de enfoques, las estimaciones GLS se comparan con las obtenidas aplicando un método completamente bayesiano, que utiliza el muestreador Gibbs para Monte Carlo mediante Cadena de Markov (MCMC) (Schmidt, Spiegelhalter & Thomas 1995), y con las obtenidas mediante aproximaciones al modelo completamente bayesiano (Abrams & Sansó 1998), propuestas para reducir los cálculos requeridos. Estos métodos se pueden implementar en sistemas como el BUGS (Gilks, Thomas & Spiegelhalter 1994).

La estimación GLS de θ se presenta con el valor más extremo ($-1{,}63$ comparado con el $-1{,}49$ del completamente bayesiano y el $-1{,}50$ del bayesiano aproximado) y la estimación de σ_u^2 exhibe un valor muy cercano al obtenido mediante los otros métodos (1,06 comparado con el 1,09 del completamente bayesiano y 0,96 del bayesiano aproximado).

Como se ha podido observar, el procedimiento propuesto produjo estimaciones de los parámetros, razonablemente cercanas a las obtenidas por otros métodos ya conocidos. De esta forma, se pretende poner de manifiesto las importantes posibilidades que ofrece un método de estimación computacionalmente muy simple de aplicar en comparación con otros métodos basados en estrategias más complejas.

[2]En este artículo los métodos de estimación utilizados se identifican por sus siglas en inglés.

7. Conclusiones

El procedimiento de estimación presentado en este artículo ubica al análisis de un conjunto de tablas de contingencia en una clase de problemas que pueden tratarse utilizando mínimos cuadrados generalizados.

Una de las principales ventajas del enfoque propuesto es que puede usarse en situaciones donde otros métodos imponen la solución de complicadas expresiones matemáticas. Otra ventaja importante es la facilidad que le ofrece al investigador para construir una amplia familia de funciones de particular interés para el análisis multinivel de las proporciones esperadas en un conjunto de tablas de contingencia.

En general, el enfoque propuesto ofrece una herramienta muy útil para modelar una gran variedad de situaciones que ocurren frecuentemente en la práctica, constituyendo así una eficaz alternativa a la modelación multinivel para datos categóricos jerárquicos.

[Recibido: abril de 2009 — Aceptado: octubre de 2010]

Referencias

Abrams, K. & Sansó, B. (1998), 'Approximate Bayesian Inference for Random Effects meta-analysis', *Statistics in Medicine* **17**, 201–218.

Agresti, A. (2002), *Categorical Data Analysis*, Wiley, New York.

Anderson, T. W. (1958), *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

Breslow, N. E. & Zhao, L. P. (1988), 'Logistic Regression for Stratified Case-Control Studies', *Biometrics* **44**, 891–899.

Brik, A. S. & Raudenbush, S. W. (1992), *Hierarchical Linear Models: Applications and Data Analysis Methods*, Sage Publications, California.

Castells, E. (1985), Estimación en un modelo con parámetros aleatorios, Tesis de maestría, Facultad de Matemática, Universidad de La Habana, La Habana.

Efron, B. (1996), 'Empirical Bayes Methods for Combining Likelihoods', *Journal of the American Statistical Association* **96**(434), 538–565.

Fears, T. R. & Brown, C. C. (1986), 'Logistic Regression Methods for Retrospective Case-Control Dtudies using Complex Sampling Procedures', *Biometrics* **42**, 955–960.

Forthofer, R. N. & Lehnen, R. G. (1981), *Public Program Analysis: A New Categorical Data Approach*, Lifetime Learning Publications, Belmont, California.

Forthoper, R. N. & Koch, G. G. (1973), 'An Analysis for Compounded Functions of Categorical Data', *Biometrics* **29**, 143–157.

Fotouhi, A. R. (2003), 'Comparisons of Estimation Procedures for Nonlinear Multilevel Models', *Journal of Statistical Software* **8**(9), 1–39.

Gilks, W. R., Thomas, A. & Spiegelhalter, D. J. (1994), 'A language and Program for Complex Bayesian Modelling', *Statistician* **43**, 169–177.

Glass, G. V. (1976), 'Primary, Secondary and Meta-Analysis of Research', *Educational Researcher* **5**, 3–8.

Goldstein, H. (1987), *Multilevel Models in Educational and Social Research*, Charles Griffin, London.

Goldstein, H. (1995), *Multilevel Statistical Models*, 2 edn, Halsted Press, New York.

Goldstein, H. & Rasbash, J. (1996), 'Improved Approximations for Multilevel Models with Binary Responses', *Journal of the Royal Statistical Society. Series A* (57), 395–407.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G. & MJR, H. (1998), *A user's guide to MLwiN*, Institute of Education.

Grizzle, J. E., Starmer, C. F. & Koch, G. G. (1969), 'Analysis of Categorical Data by Linear Models', *Biometrics* **25**, 489–504.

Hamerle, A. & Honning, G. (1995), Panel analysis for qualitative variables, *in* G. Arminger, C. C. Clogg & M. E. Sobel, eds, 'A Handbook for Statistical Modeling in the Social and Behavioral Sciences', Plenum, New York, pp. 401–451.

Hartzel, J., Liu, I.-M. & Agresti, A. (2001), 'Describing Heterogeneous Effects in Stratified Ordinal Contingency Tables, with Application to Multi-CCenter clinical trials', *Computational Statistics and Data Analysis* **35**, 429–499.

Hedges, L. V. & Olkin, I. (1985), *Statistical Methods for Meta-analysis*, Academic Press, New York.

Hsiao, C. (1995), *Analysis of Panel Data*, Cambridge University Press, New York.

Kleffe, J. (1976), 'A Note on MINQUE for Normal Models', *Mathematische Operationsforschung und Statistik* **7**, 107–114.

Kuk, A. Y. C. (1995), 'Asymtotically Unbiased Estimation in Generalized Linear Models with Random Effects', *Journal of the Royal Statistical Socity* **57**, 395–407.

Lee, Y. & Nelder, J. A. (2002), 'Analysis of Ulcer data Using Hierarchical Generalized Linear Models', *Statistics in Medicine* **21**, 191–202.

Longford, N. (1995), *Random coefficient models: Handbook of Statistiscal Modeling for the Social and Behavioral Sciences*, Plenum Press, New York.

Lubin, J. H., Blot, W. J., Berrino, F., Flamant, R., Gillis, C. R., Kunze, M., Schmäwhl, D. & Visco, G. (1984), 'Patterns of Lung Cancer Risg According to Type of Cigarrette Smoked', *International Journal of Cancer* **33**, 569–576.

Montero, M. (2006), Análisis de tablas de contingencia: un enfoque multinivel, Tesis de doctorado, Facultad de Matemática, Universidad de La Habana, La Habana.

Montero, M., Castell, E. & Ojeda, M. M. (2007), 'Fitting a Multilevel Model to a Sample of Contingency Tables using the GSK Approach', *Revista Investigación Operacional* **28**(3), 204–214.

Montero, M., Castell, E. & Ojeda, M. M. (2008), 'Analysis of a Contingency Tables sample: A Simulation Study', *Revista Ciencias Matemáticas* **24**, 83–92.

Montero, M. & Guerra, V. (2005), 'Estimating Multilevel Models for Categorical Data via Generalized Least Squares', *Revista Colombiana de Estadística.* **21**(8), 63–76.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, segunda edn, John Wiley & Sons, Inc., New York.

Rudas, T. (1986), 'A Monte Carlo Comparison of the Small Sample Behavior of the Pearson, the Likelihood Ratio and the Cressie-Read Statistics', *Journal of Statistical Computation and Simulation* (24), 107–120.

Schmidt, T. C., Spiegelhalter, D. J. & Thomas, A. (1995), 'Bayesian Approaches to Random'effects Meta-analysis: A Comparative Study', *Statistics in Medicine* **14**, 2685–2699.

Snijders, T. A. B. & Bosker, R. (1999), *Introduction to Basic and Advanced Multilevel Modelling*, Sage, London.

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H. & Thompson, S. G. (2000), 'Multilevel models for meta-analysis of clinical trials with binary outcomes', *Statistics in Medicine* **19**, 3417–3432.

Apéndice A.

Teorema 3. *Sean las matrices:*

$$\begin{aligned} \boldsymbol{V}_{\lambda}^{*} &= \boldsymbol{R} \otimes \boldsymbol{R} + [\boldsymbol{R} \otimes \boldsymbol{R}(1), \boldsymbol{R} \otimes \boldsymbol{R}(2), \ldots, \boldsymbol{R} \otimes \boldsymbol{R}(n)] \quad \boldsymbol{y} \\ \Theta &= 2\,(\boldsymbol{R} \otimes \boldsymbol{R}), \end{aligned}$$

y $\mathbf{A}$ *y* $\mathbf{B}$ *dos matrices simétricas cualesquiera de orden* $n \times n$*, entonces se cumple:*

$$\begin{aligned} i)\ \boldsymbol{V}_{\lambda}^{*}\overrightarrow{\boldsymbol{B}} &= \Theta\overrightarrow{\mathbf{B}} \\ ii)\overrightarrow{\boldsymbol{A}}'\boldsymbol{V}_{\lambda}^{*-}\overrightarrow{\boldsymbol{B}} &= \overrightarrow{\boldsymbol{A}}'\Theta^{-1}\overrightarrow{\boldsymbol{B}} \end{aligned} \tag{9}$$

Algoritmo de mínimos cuadrados generalizados iterativos

Tabla 3: Infecciones del tracto respiratorio en los grupos tratamiento y control de 22 ensayos para la descontaminación selectiva del tracto digestivo.

Ensayo	Infección/Total		*Odds ratio*	Log OR	Varianza (Log OR)
	Tratamiento	Control			
1	7/47	25/54	0,20	−1,59	0,24
2	4/38	24/41	0,08	−2,48	0,38
3	20/96	37/95	0,41	−0,89	0,11
4	1/14	11/17	0,04	−3,17	1,33
5	10/48	26/49	0,23	−1,46	0,21
6	2/101	13/84	0,11	−2,20	0,60
7	12/161	38/170	0,28	−1,27	0,12
8	1/28	29/60	0,04	−3,23	1,10
9	1/19	9/20	0,07	−2,69	1,26
10	22/49	44/47	0,06	−2,89	0,44
11	25/162	30/160	0,79	−0,23	0,09
12	31/200	40/185	0,66	−0,41	0,07
13	9/39	10/41	0,93	−0,07	0,28
14	22/193	40/185	0,47	−0,76	0,08
15	0/45	4/46	0,10	−2,27	2,27
16	31/131	60/140	0,41	−0,88	0,07
17	4/75	12/75	0,30	−1,22	0,36
18	31/220	42/225	0,71	−0,34	0,07
19	7/55	26/57	0,17	−1,75	0,23
20	3/91	17/92	0,15	−1,89	0,42
21	14/25	23/23	0,03	−3,62	2,20
22	3/65	6/68	0,50	−0,69	0,53

1. Obtener estimaciones iniciales de los parámetros fijos utilizando un ajuste de mínimos cuadrados generalizados para datos categóricos, asumiendo que los errores aleatorios en el nivel-2 tienen varianza 0.

2. A partir de estas estimaciones formar los residuos "crudos"

$$\widetilde{\boldsymbol{F}}(\boldsymbol{p}) = \boldsymbol{F}(\boldsymbol{p}) - \boldsymbol{A}\widehat{\boldsymbol{\Gamma}}$$

3. Calcular la matriz de productos cruzados: $\boldsymbol{F}(\boldsymbol{p})^{*} = \widetilde{\boldsymbol{F}}(\boldsymbol{p})\,\widetilde{\boldsymbol{F}}(\boldsymbol{p})'$.

4. Formar el vector: $\boldsymbol{F}(\boldsymbol{p})^{**} = \overrightarrow{\boldsymbol{F}(\boldsymbol{p})^{*}}$.

5. Similarmente, construir el vector $\overrightarrow{\boldsymbol{V}_{\lambda}}$.

6. La relación entre los vectores definidos en los pasos 4 y 5 se puede expresar mediante el modelo lineal: $E\left(\boldsymbol{F}\left(\boldsymbol{p}\right)^{**}\right) = \boldsymbol{Z}^{*}\boldsymbol{\lambda}$, donde $\boldsymbol{z}^{*}$ es la matriz de diseño para los parámetros aleatorios (o sea, los elementos de $\boldsymbol{\Omega}_u$ y $\boldsymbol{\Omega}_e$). Luego, en este paso es posible postular el modelo: $\boldsymbol{F}\left(\boldsymbol{p}\right)^{**} = \boldsymbol{Z}^{*}\boldsymbol{\lambda} + \boldsymbol{R}$.

7. Para estimar $\boldsymbol{\lambda}$ se aplica mínimos cuadrados generalizados, o sea:

$$\widehat{\lambda} = \left(\boldsymbol{Z}^{*'}\boldsymbol{V}_{\lambda}^{*-}\boldsymbol{Z}^{*}\right)^{-1}\boldsymbol{Z}^{*'}\boldsymbol{V}_{\lambda}^{*-}\boldsymbol{F}\left(\boldsymbol{p}\right)^{**}$$

   donde $\boldsymbol{V}_{\lambda}^{*} = \widehat{\boldsymbol{V}}_{\lambda} \otimes \widehat{\boldsymbol{V}}_{\lambda}$.

8. Las estimaciones de $\boldsymbol{\Omega}_u$ y $\boldsymbol{\Omega}_e$ obtenidas en el paso anterior se sustituyen en:

$$\boldsymbol{V}_{\lambda} = \boldsymbol{X}\boldsymbol{\Omega}_u\boldsymbol{X}' + \boldsymbol{\Omega}_e$$

9. Se calculan nuevas estimaciones de los efectos fijos mediante el estimador mínimo cuadrado generalizado, utilizando la estimación en el paso 8 para la matriz de varianza y covarianza, obteniéndose así:

$$\widehat{\widehat{\boldsymbol{\Gamma}}} = \left(\boldsymbol{A}'\widehat{\boldsymbol{V}}_{\lambda}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}'\widehat{\boldsymbol{V}}_{\lambda}^{-1}\boldsymbol{F}\left(\boldsymbol{p}\right)$$

10. Retornar al paso 2, pero utilizando $\widehat{\widehat{\boldsymbol{\Gamma}}}$ en lugar de $\widehat{\boldsymbol{\Gamma}}$.

El algoritmo alterna entre las estimaciones de los parámetros fijos y aleatorios hasta que el procedimiento converja.

Revista Colombiana de Estadística
Diciembre 2010, volumen 33, no. 2, pp. 251 a 271

Propuesta de una prueba de rachas recortada para hipótesis de simetría

A Proposed Runs Trimming Test for the Hypothesis of Symmetry

Giovany Babativa[a], Jimmy A. Corzo[b]

Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Resumen

Combinando la teoría de rachas desarrollada por Corzo (1989) y la idea de Modarres & Gastwirth (1996), que utilizan el número de rachas que quedan después de recortar la sucesión dicotomizada, se proponen tres pruebas de rachas para la hipótesis de simetría. Utilizando la técnica de linealización de Taylor se aproxima el valor esperado y la varianza, y se realiza un estudio de aproximación de la distribución del estadístico por la distribución normal. Las pruebas propuestas son comparadas en términos de su potencia con algunas de las pruebas no paramétricas más recientes y comunes para dicho problema en tamaños de muestra $n = 10(1)25$, $n = 30$, $n = 50(50)250$ y $n = 500$. Para la comparación se utilizaron métodos de Monte Carlo, y las muestras fueron generadas de nueve distribuciones pertenecientes a la familia lambda generalizada (DLG). Las simulaciones indican que para una gran variedad de alternativas asimétricas las pruebas propuestas son más potentes que las pruebas existentes en la literatura.

Palabras clave: distribución lambda generalizada, potencia, pruebas de rachas, pruebas para simetría.

Abstract

Combining the runs theory developed by Corzo (1989) and the idea of Modarres & Gastwirth (1996), which uses the number of runs left after cutting the dichotomized succession, three families of statistics based on runs and three tests for the hypothesis of symmetry are proposed. Using the linearization Taylor's technique, the expected value and variance of two from the three proposed families is approximated. A study to aproximate the distribution of the statistics through the normal distribution for the studied sample sizes is realized. The proposed tests are compared in terms of their

[a]Egresado de maestría en estadística. E-mail: jgbabativam@unal.edu.co
[b]Profesor asociado. E-mail: jacorzos@unal.edu.co

power with some other recent and common nonparametric tests for Symmetry, for the sample sizes $n = 10(1)25$, $n = 30$, $n = 50(50)250$ and $n = 500$. For this comparison, Monte Carlo methods were used and the samples were generated from nine distributions obtained from the generalized lambda distribution. The simulations indicate that, for a wide variety of asymmetric alternatives in the generalized lambda distribution, the tests proposed are more powerful than the existing tests in literature.

Key words: Lambda distribution, Power, Runs test, Symmetry test.

1. Introducción

Entre los métodos no paramétricos existen pruebas que exigen simetría de la distribución, de la cual provienen las observaciones; este es el caso de las pruebas para la alternativa de localización en una muestra, basadas en estadísticos lineales de rangos, uno de cuyos casos particulares es la prueba del rango designado de Wilcoxon, ampliamente utilizada en diferentes áreas del conocimiento. Si este supuesto se cumple, dicha prueba es localmente más potente para la alternativa de localización. Otro caso es el análisis de regresión basada en rangos donde la simetría desempeña un papel importante en la estimación de la matriz de covarianzas de la distribución asintótica normal multivariada, a la cual converge la distribución del estimador por rangos del vector de parámetros (ver Hettmansperger 1984, pp. 241-243). En ambos casos, si el supuesto de simetría de la distribución muestreada no se cumple, las pruebas tienden a no conservar su tamaño.

Son varias las pruebas que se han propuesto para este problema bajo el supuesto de que se conoce alguna medida de localización. Lehmann (1986), Randles & Wolfe (1979) y Gibbons & Chakraborti (1992) son algunos de los autores que mencionan las pruebas de rangos más conocidas para juzgar la hipótesis de simetría de una distribución. En los últimos 20 años, varios investigadores han propuesto pruebas para determinar si la distribución muestreada es simétrica alrededor de un centro conocido. Algunos de ellos son Cohen & Menjoge (1988), McWilliams (1990), Castillo (1993), Modarres & Gastwirth (1996), Corzo & Rojas (1999) y Baklizi (2003, 2007), quienes han desarrollado pruebas de rachas; por otra parte, Tajuddin (1994) y Baklizi (2008) utilizan pruebas de rangos; Cheng & Balakrishnan (2004) emplean la información de los signos; Mira (1999) usa la medida de sesgo de Bonferroni, mientras que Modarres & Gastwirth (1998) y Thas, Rayner & Best (2005) usan información mixta combinando los signos y los rangos.

La hipótesis de simetría puede formularse como sigue: sea $X_1, \ldots, X_n$ una muestra aleatoria de una función de distribución continua F con función de densidad f con mediana conocida, la cual, sin pérdida de generalidad, se puede asumir igual a cero. Se considera el problema de testear la siguiente hipótesis:

$$H_0 : F_X(x) = 1 - F_X(-x)$$

frente a la alternativa:

$$K_1 : F_X(x) \neq 1 - F_X(-x)$$

La hipótesis nula indica que la función de distribución F es simétrica alrededor de cero; en otras palabras, el interés se centra en probar si $f(x) = f(-x)$ para todo x o si f se aparta de la hipótesis de simetría.

En la siguiente sección, presentamos tres pruebas de rachas. El procedimiento combina la teoría desarrollada por Corzo (1989) y la idea de Modarres & Gastwirth (1996) de utilizar como criterio de información el número de rachas que quedan después de hacer un recorte en la sucesión dicotomizada. Bajo la hipótesis nula de simetría se dan aproximaciones del valor esperado y de la varianza de los estadísticos de prueba utilizando la técnica de linealización de Taylor. En la sección 3, se presenta un estudio de Monte Carlo que muestra que las pruebas propuestas son más potentes que las pruebas de la literatura con las que se realizó la comparación, y por ende más potentes que las pruebas con las que los demás autores habían hecho sus comparaciones.

2. Pruebas propuestas

Sea $|X|_{(1)}, \ldots, |X|_{(n)}$ la sucesión de los valores absolutos ordenados. Definimos $|X_{D_j}| = |X|_{(j)}$ $(j = 1, \ldots, n)$, donde D_j es el antirrango de $|X|_{(j)}$; esto es, D_j es el subíndice que tenía originalmente $|X|_{(j)}$ en la sucesión de valores absolutos $|X_1|, \ldots, |X_n|$. La sucesión $\eta_1, \ldots, \eta_n$ se denomina la sucesión dicotomizada, y en ella se representan las observaciones positivas por unos y las negativas por ceros, de la siguiente manera:

$$\eta_j = \begin{cases} 1 & \text{si } X_{D_j} > 0 \\ 0 & \text{en otro caso} \end{cases} \qquad j = 1, \ldots, n \tag{1}$$

Para contar el número de cambios que hay en la sucesión dicotomizada se definen las siguientes indicadoras:

$$I_1 = 1$$

$$I_j = \begin{cases} 1 & \text{si } \eta_{j-1} \neq \eta_j \\ 0 & \text{si } \eta_{j-1} = \eta_j \end{cases} \qquad j = 2, \ldots, n$$

A partir de estas, el número de rachas hasta la j-ésima observación en la sucesión dicotomizada se calcula como:

$$r_j = \sum_{k=1}^{j} I_k, \text{ para } \; j = 1, 2, \ldots, n$$

A la sucesión $r_1, \ldots, r_n$ se le denominará la sucesión de rachas. El estadístico que cuenta el número de rachas en la sucesión dicotomizada es:

$$R^+ = r_n = \sum_{k=1}^{n} I_k$$

bajo la hipótesis alternativa, muchas observaciones negativas o muchas observaciones positivas tienden a generar agrupaciones y resultarán pocas rachas; esto significa que la hipótesis nula será rechazada para valores pequeños de $R^+ - 1$. Bajo H_0, la distribución exacta de $R^+ - 1$, dada en McWilliams (1990), es binomial con parámetros $n-1$ y $1/2$.

Como lo mencionan Modarres & Gastwirth (1996), "bajo la hipótesis nula se cumple $P(I_k = 1) = P(I_k = 0) = 1/2$, mientras que bajo la alternativa $P(I_k = 1) \neq P(I_k = 0)$ y depende de k, para $k = 2, \ldots, n$. Esto sugiere que una prueba basada en la posición relativa, k, de las rachas podría ser más potente que R^+. Teniendo en cuenta que para alternativas sesgadas, las rachas deberían aparecer en las colas, se propone modificar la prueba $R^+ - 1$ dando un mayor peso a estas rachas".

Entonces, siguiendo la metodología propuesta por Corzo (1989) y tomando la idea de recortar observaciones de la muestra presentada por Modarres & Gastwirth (1996), y Modarres & Gastwirth (1998) se proponen los siguientes tres estadísticos de prueba:

$$R_p = \frac{1}{r_n} \sum_{i=[np]+1}^{n} \phi(r_i, i)\delta_i \tag{2}$$

$$R_p^* = \frac{1}{r_n^*} \sum_{i=[np]+1}^{n} \phi(r_i, i)\delta_i \tag{3}$$

$$C_p^* = \frac{1}{r_n} \sum_{i=[np]+1}^{n} \phi(r_i, i)\delta_i^* \tag{4}$$

donde

$$\phi(r_i, i) = \begin{cases} r_i - pr_n & \text{si } \ i > np \\ 0 & \text{en otro caso} \end{cases}$$

$$\delta_i = \begin{cases} 1 & \text{si} \quad X_{D_i} > 0 \\ -1 & \text{si} \quad X_{D_i} < 0 \end{cases} \qquad i = 1, \ldots, n$$

$$\delta_i^* = \begin{cases} 1/n_1^* & \text{si} \quad X_{D_i} > 0 \\ -1/n_0^* & \text{si} \quad X_{D_i} < 0 \end{cases} \qquad i = 1, \ldots, n$$

p es una proporción de recorte; $[np]$ es la parte entera de np; n_0^* y n_1^* son el número de ceros y el número de unos en la sucesión dicotomizada después de hacer el recorte; r_n es el número total de rachas en la sucesión dicotomizada, mientras que r_n^* es el número total de rachas después de recortar la sucesión dicotomizada. En el caso de R_p y R_p^*, si $p = 0$ se obtiene el estadístico estudiado por Castillo (1993); y en el caso de C_p^*, si $p = 0$ se obtiene el estadístico estudiado en Corzo & Rojas (1999).

Los argumentos para la construcción de la región crítica son los mismos que los utilizados por McWilliams (1990). Si la hipótesis nula de simetría es cierta, entonces para $b > a \geq 0$ se cumple que $P(a < X < b) = P(-b < X < -a)$, luego en la sucesión dicotomizada $P(\eta = 0) = P(\eta = 1)$, lo cual equivale a que $P(X_{D_j} >$

$0) = P(X_{D_j} < 0) = 1/2$, y por tanto se espera que se alternen los valores positivos y negativos de los sumandos de cualquiera de los estadísticos propuestos, haciendo que tomen valores cercanos a cero. Bajo la hipótesis alternativa de asimetría $P(a < X < b) \neq P(-b < X < -a)$; por consiguiente tanto, se espera que en las colas aparezcan agrupamientos de unos o de ceros y como consecuencia de esto los valores de los estadísticos R_p, R_p^* y C_p^* estarán lejos de cero, apoyando la hipótesis alternativa.

Del análisis anterior y teniendo en cuenta que, por construcción, bajo H_0 la distribución de R_p es simétrica alrededor de cero, se concluye que la prueba rechaza la hipótesis nula de simetría a favor de la alternativa de asimetría cuando $|R_p| \geq r_{1-\alpha/2}$, donde $r_{1-\alpha/2}$ corresponde al $100(1-\alpha/2)$-ésimo percentil de la distribución de R_p; de igual manera ocurre para las pruebas R_p^* y C_p^*.

Para calcular la distribución exacta de R_p, R_p^* o C_p^* en un tamaño de muestra n, se deben considerar los 2^n arreglos distinguibles de unos y ceros para incluir todas las posibilidades de la sucesión dicotomizada. A partir de estos se calculan los valores de los mismos y se construye la distribución de frecuencias. El programa que hace los cálculos de la distribución exacta para cualquiera de los estadísticos propuestos se encuentra en *www.docentes.unal.edu.co/jacorzos/docs/PruebaSimetria/*

A continuación se enuncian dos teoremas que fueron demostrados utilizando la técnica de linealización de Taylor. Todas las demostraciones se pueden encontrar en *www.docentes.unal.edu.co/jacorzos/docs/PruebaSimetria/*

Teorema 1. *Sean $X_1, \ldots, X_n$ variables aleatorias independientes con función de distribución continua y simétrica F, función de densidad f y mediana cero, entonces:*

$$\mathbb{E}\left(R_p\right) \doteq 0$$

y

$$\mathbb{E}\left(R_p^*\right) \doteq 0$$

Teorema 2. *Sean $X_1, \ldots, X_n$ variables aleatorias independientes con función de distribución continua y simétrica F, función de densidad f y mediana cero; p un número real en el intervalo $(0,1)$; r_n el total de rachas en la sucesión dicotomizada; $\phi(r_i, i)$ y δ_i como se definieron en (2), entonces la varianza aproximada (VA) es:*

$$\begin{aligned}\mathbb{VA}\left(R_p\right) = \frac{1}{3\left(n+1\right)^2}\big\{&n(n^2+3n+2) - [np]\left([np]^2+3[np]-4\right)\\ &+3p^2(n-1)(n^2-n[np]+4)\\ &-3p\left(n^3+n^2+2n-n[np]^2-n[np]+4[np]\right)+6\big\}\end{aligned}$$

y

$$\begin{aligned}\mathbb{VA}\left(R_p^*\right) = \frac{1}{3\left(n-[np]+1\right)^2}\big\{&n(n^2+3n+2) - [np]\left([np]^2+3[np]-4\right)\\ &+3p^2(n-1)(n^2-n[np]+4)\\ &-3p\left(n^3+n^2+2n-n[np]^2-n[np]+4[np]\right)+6\big\}\end{aligned}$$

3. Estudio de Monte Carlo

En esta sección se presentan los resultados de una simulación por métodos de Monte Carlo donde se compara la potencia de las pruebas propuestas frente a otras ocho pruebas de la literatura. Las pruebas con las que se realizó la comparación son:

1. La prueba del rango signado de Wilcoxon referenciada para hipótesis de simetría por Gibbons & Chakraborti (1992):

$$W = \sum_{k=1}^{n} k\eta_k$$

2. La prueba propuesta por McWilliams (1990), basada en:

$$R^+ = \sum_{k=1}^{n} I_k$$

3. La prueba condicional de Tajuddin (1994), basada en la prueba de localización de Wilcoxon para dos muestras:

$$W_n = \sum_{k=1}^{n} k\eta_k = \sum_{k=1}^{n_1} R_k$$

donde n_1 es el número de observaciones positivas y R_k corresponde al rango de X_k en la sucesión de valores absolutos ordenados.

4. La prueba M_p propuesta por Modarres & Gastwirth (1996) que utiliza:

$$M_p = \sum_{k=[np]+2}^{n} \varphi(k) I_k$$

donde

$$\varphi(k) = \begin{cases} k - [np] & \text{si} \quad k > np \\ 0 & \text{en otro caso.} \end{cases}$$

donde $[np]$ corresponde a la parte entera de np[1].

5. La prueba híbrida en dos etapas propuesta por Modarres & Gastwirth (1998), la cual usa en la primera etapa la prueba del signo y en la segunda etapa una modificación de la prueba de Tajuddin (1994):

[1] Modarres & Gastwirth (1996) tratan a np como un entero para evitar la notación de parte entera.

Etapa I: Utilizar la prueba del signo para decidir si hay evidencias de simetría[2]:

$$Z_s = \frac{S - n/2}{\sqrt{n/4}}$$

Etapa II: Si en la etapa I se concluye que hay evidencias de simetría, utilizar la siguiente modificación de la prueba de Tajuddin (1994):

$$W_p = \sum_{k=[np]+1}^{n} \varphi(k)\eta_k$$

donde η_k es definido como en (1).

6. La prueba de Mira (1999) que detecta la asimetría de una función de distribución con media μ_F y mediana $\widetilde{\mu}_F$ desconocidas, mediante la medida de asimetría de Bonferroni. Sean $\overline{X}_n$ y $\widehat{X}_n$ la media y la mediana de una muestra de tamaño n, respectivamente. La prueba de Mira utiliza el siguiente estadístico:

$$\gamma_1(F_n) = 2\left(\overline{X}_n - \widehat{X}_n\right)$$

7. La prueba de Baklizi (2003) basada en la distribución condicional de R^+ dado el número de unos y de ceros en la sucesión dicotomizada, n_1 y n_0, respectivamente. Dicha distribución está dada en Gibbons & Chakraborti (1992) y corresponde a:

$$f_{R^+}(r_n/n_0, n_1) = \begin{cases} \dfrac{2\binom{n_1-1}{r_n/2-1}\binom{n_0-1}{r_n/2-1}}{\binom{n_1+n_0}{n_0}} & \text{si } r_n > 1 \text{ y par} \\ \dfrac{\binom{n_1-1}{(r_n-1)/2}\binom{n_0-1}{(r_n-3)/2}+\binom{n_1-1}{(r_n-3)/2}\binom{n_0-1}{(r_n-1)/2}}{\binom{n_1+n_0}{n_0}} & \text{si } r_n > 1 \text{ e impar} \end{cases}$$

si $n_1 = 0$ o $n_0 = 0$ entonces $P(r_n = 1) = 1$.

8. La prueba propuesta por Cheng & Balakrishnan (2004) que usa la información de los signos, donde el estadístico de prueba es:

$$C_6 = \eta_{n-5} + \cdots + \eta_n \tag{5}$$

Se seleccionaron nueve casos de la DLG,[3] que son los más utilizados en la literatura, ver por ejemplo: McWilliams (1990), Tajuddin (1994), Modarres & Gastwirth (1996); Modarres & Gastwirth (1998), Baklizi (2003), Cheng & Balakrishnan (2004) y Thas et al. (2005). El caso 1 representa la aproximación de la distribución normal (caso simétrico) para el cual la hipótesis nula es verdadera, mientras

[2]El artículo original de Modarres & Gastwirth (1998) se utiliza $n^{1/2}/4$ en el denominador de Z_s. Sin embargo, en el párrafo anterior es claro que S tiene una distribución binomial (n, p), $E(S) = n/2$ y $V(S) = n/4$.

[3]En el ordenamiento de los casos difiere de los utilizados por otros autores, debido a que en este trabajo se ordenaron por grupos según el coeficiente de asimetría.

que los otros ocho casos varían en su grado de asimetría y permiten comparar la potencia de las pruebas.

Las nueve funciones de densidad de la DLG seleccionadas se pueden apreciar en la figura 1. En la tabla 1 se muestran los valores de los parámetros de los nueve casos de la DLG utilizados, los coeficientes de asimetría y curtosis. Nótese que los casos 1, 2 y 3 conforman un grupo de tres densidades muy cercanas a la hipótesis nula de simetría, los casos 4 y 5 son dos densidades en las que ya se nota cierto grado de asimetría y los casos 6 al 9 son densidades que tienen un grado tal de asimetría que toda la probabilidad está acumulada en la cola del lado derecho. En los tres grupos aumentan simultaneamente los coeficientes de asimetría y curtosis.

TABLA 1: Valores de los parámetros de la DLG de los nueve casos seleccionados para el estudio de potencia.

Caso	λ_1	λ_2	λ_3	λ_4	α_3	α_4
1 (Nula)	0,000000	0,197454	0,134915	0,134915	0,0000	3,0000
2	−0,116734	−0,351663	−0,130000	−0,160000	0,8000	11,4000
3	0,000000	−1,000000	−0,100000	−0,180000	2,0000	21,2000
4	3,586508	0,043060	0,025213	0,094029	0,9000	4,2000
5	0,000000	−1,000000	−0,007500	−0,030000	1,5000	7,5000
6	0,000000	1,000000	1,400000	0,250000	0,5000	2,2000
7	0,000000	1,000000	0,000070	0,100000	1,5000	5,8000
8	0,000000	−1,000000	−0,001000	−0,130000	3,1600	23,8000
9	0,000000	−1,000000	−0,000100	−0,170000	3,8800	40,7000

Para estimar la potencia de las diferentes pruebas se realizó un programa en SAS IML. El algoritmo utilizado es el siguiente:

1. Seleccionar una muestra aleatoria $u_1, \ldots, u_n$ de tamaño n de la distribución $U(0,1)$.

2. Transformar la muestra $u_1, \ldots, u_n$ en la sucesión $x_1^*, \ldots, x_n^*$; utilizando la función percentil de la DLG, que se define por:

$$x_i^* = \lambda_1 + \frac{u_i^{\lambda_3} - (1-u_i)^{\lambda_4}}{\lambda_2}, \quad i = 1, \ldots, n \tag{6}$$

   con lo que se consigue que la sucesión $x_1^*, \ldots, x_n^*$ sea una muestra aleatoria de una DLG con parámetros $\lambda_1, \lambda_2, \lambda_3$ y λ_4.

3. Transformar $x_i = x_i^* - \theta$ para que la distribución de $x_1^*, \ldots, x_n^*$ tenga mediana cero, donde

$$\theta = \lambda_1 + \frac{0,5^{\lambda_3} - 0,5^{\lambda_4}}{\lambda_2} \tag{7}$$

4. Calcular los valores de los estadísticos que se van a comparar usando las observaciones de la muestra $x_1, \ldots, x_n$.

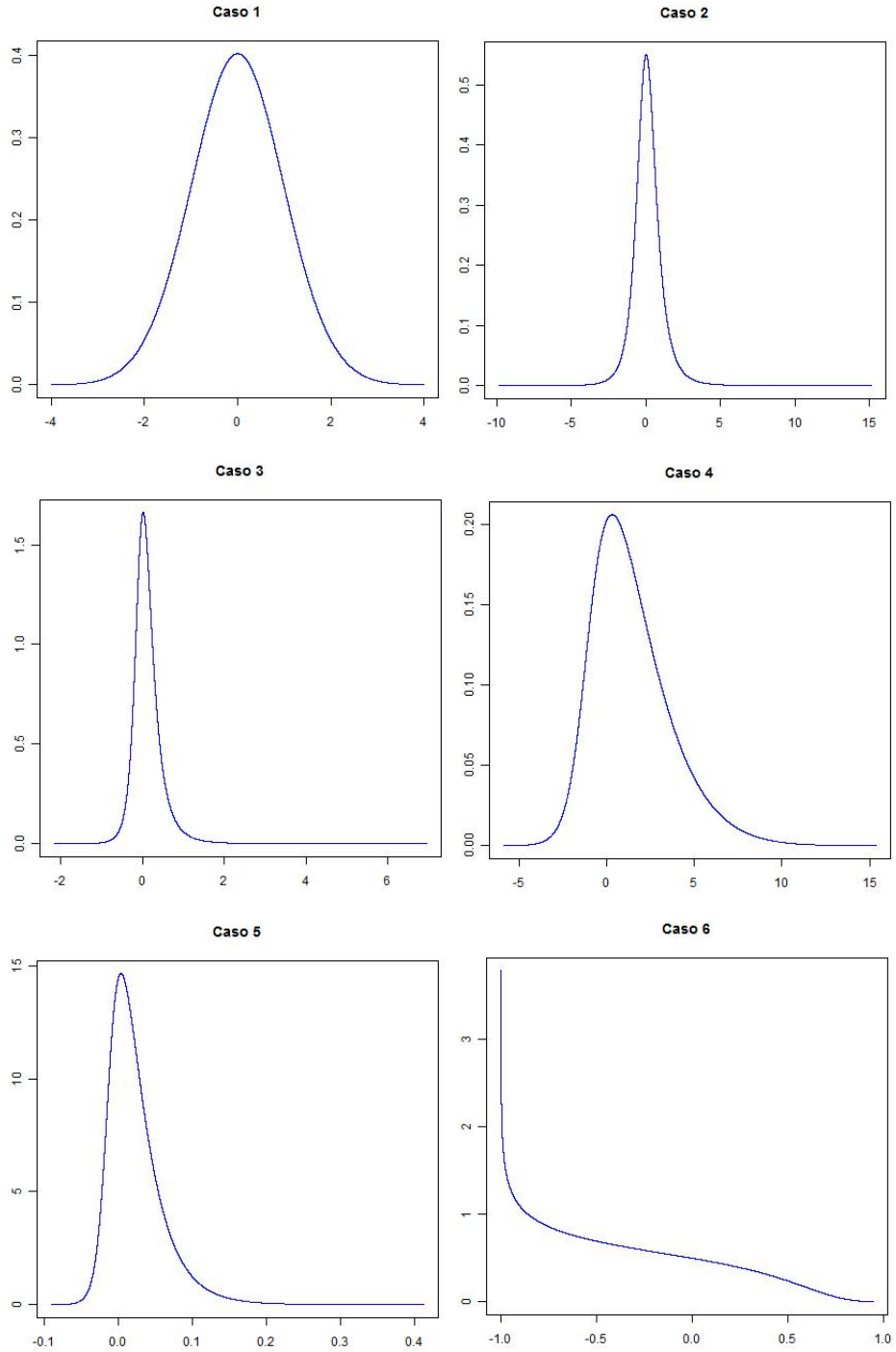


FIGURA 1: Funciones de densidad de los casos seleccionados de la DLG. Continuación

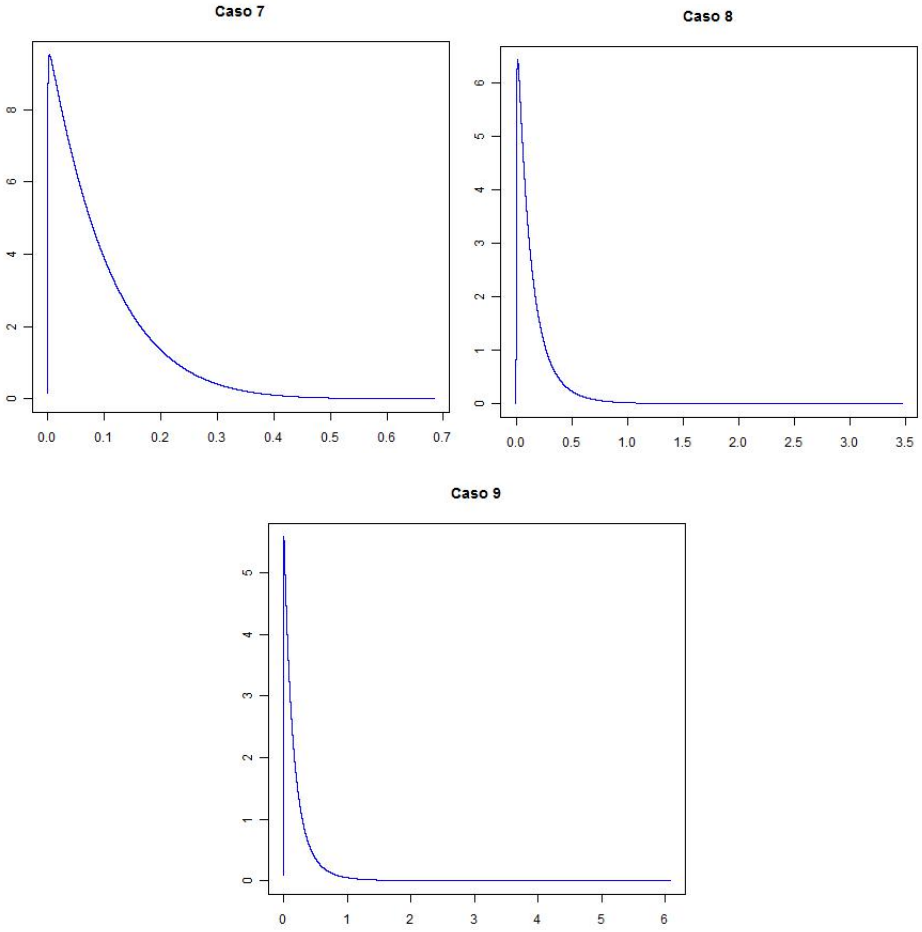


FIGURA 1: Funciones de densidad de los casos seleccionados de la DLG.

5. Realizar las respectivas pruebas de hipótesis, y para cada una determinar si se rechaza la hipótesis nula, aleatorizando la prueba.

6. Aplicar el anterior proceso m veces, y estimar la potencia de cada una de las pruebas, así:

$$\widehat{\pi} = \frac{\text{Número de rechazos en las } m \text{ réplicas}}{m}$$

Para este trabajo se estimó la potencia de todas las pruebas usando $m = 25.000$ réplicas. El máximo número de réplicas usado en los artículos consultados para este trabajo fue de 10.000 (Thas et al. 2005, Cheng & Balakrishnan 2004).

3.1. Estudio de potencia para $n \leq 25$

Para $n < 30$ las pruebas T y W_p no fueron incluidas porque están basadas en estadísticos condicionados al número de unos y ceros en la sucesión dicotomizada; esto hace que se requieran unos cálculos distintos a los hechos en este trabajo para llegar a la distribución exacta de las pruebas, lo cual implicaría programas con otros algoritmos. Por ejemplo, para calcular la distribución exacta de los estadísticos propuestos para un tamaño de muestra $n = 15$ es necesario generar 2^{15} arreglos de unos y ceros, mientras que para T y W_p es necesario generar $\binom{n}{k}$ arreglos para $k = 0, 1, \ldots, 7$, lo que implica muchos más cálculos que no estaban planeados desde el comienzo. Para $n \geq 30$ se utilizó la distribución asintótica de estas.

Por otra parte, las pruebas de Butler (1969), Rothman & Woodroofe (1972) y Hill & Rao (1977) fueron superadas ampliamente por la prueba propuesta por McWilliams (1990), razón por la que no fueron incluidas en este trabajo.

En las tablas 2 a 10 se presentan los resultados del estudio de Monte Carlo. A partir de las tablas se extraen las siguientes conclusiones:

Bajo la hipótesis nula, caso 1 (tabla 2), el tamaño de todas las pruebas está alrededor del nivel de significación $\alpha = 5\,\%$ a excepción de la prueba $\gamma_1(F_n)$ de Mira (1999) que resultó ser una prueba conservativa, para la cual se observó que su error de tipo I aún no alcanzaba el $5\,\%$ para $n = 500$. Por lo anterior, dicha prueba no fue incluida en las comparaciones de las potencias.

Para el caso 2 (tabla 3), ninguna prueba supera ampliamente a las demás. Sin embargo, la prueba $C^*_{.20}$ en general tiene las mayores potencias. Por ejemplo, para $n = 19$, las tres pruebas con las potencias más altas son $C^*_{.20}$, $C^*_{.10}$ y R^+ con el $5,6\,\%, 5,5\,\%$ y $5,46\,\%$, respectivamente.

En el caso 3 (tabla 4), la prueba $C^*_{.20}$ es la que tiene las mayores potencias para $n \leq 20$, y para los demás tamaños de muestra estudiados las pruebas $R_{.80}, R^*_{.80}$ y $C^*_{.80}$ son las que tienen ventajas, es decir, en este caso las pruebas propuestas resultan más potentes que todas las pruebas con las que se comparó.

Con muestras provenientes del caso 4 (tabla 5), nuevamente la prueba $C^*_{.20}$ es la que tiene las mayores potencias cuando $n < 20$, para $n = 20$; la prueba $R^*_{.80}$ es la que obtiene el mejor resultado y desde $n = 21$ hasta $n = 25$, las mayores potencias las tienen las pruebas $R_{.80}, R^*_{.80}$ y $C^*_{.80}$. Por ejemplo, para $n = 21$ la

potencia de las pruebas $R_{.80}, R^*_{.80}$ y $C^*_{.80}$ está alrededor del $14,5\,\%$ seguidas por $C^*_{.20}$ con el $13,3\,\%$ y posteriormente por $M_{.25}$ con el $12,6\,\%$; es decir que para este caso las pruebas propuestas resultan más potentes que las pruebas consultadas en la literatura.

Para el caso 5 (tabla 6), la prueba $C^*_{.20}$ es la que presenta las mayores potencias cuando $n < 17$; para $n = 17$ la prueba $M_{.25}$ logra ligeras ventajas sobre las demás pruebas, mientras que para $18 \leq n \leq 20$ la prueba con las mayores potencias es $R^*_{.80}$, y para los demás tamaños de muestra (hasta $n = 25$) las pruebas $R_{.80}, R^*_{.80}$ y $C^*_{.80}$ son las más potentes, lo cual reafirma que las pruebas propuestas son las que tienen las mayores potencias en los tamaños de muestra estudiados, logrando para $n = 25$ grandes diferencias con respecto a las competidoras de la literatura; $R^*_{.80}$ es la mejor prueba de las propuestas en el tamaño de muestra mencionado con una potencia del $26,8\,\%$, mientras que la mejor prueba de las competidoras es C_6 Cheng & Balakrishnan (2004) con una potencia del $23\,\%$.

En el caso 6 (tabla 7), las pruebas con el mejor desempeño para $n \leq 21$ son M_p, $p = 0, 10\,\%, 20\,\%$ y $25\,\%$, aunque con $n = 22$ las pruebas $R_{.80}$ y $C^*_{.80}$ tienen potencias del $38,6\,\%$ y $38,3\,\%$, respectivamente, alcanzando a estar en segundo y tercer lugar después de $M_{.25}$, que tiene una potencia del $39\,\%$. Si $n = 24$ o $n = 25$ las pruebas $R_{.80}, R^*_{.80}$ y $C^*_{.80}$ son las que logran las mayores potencias.

Tabla 2: Estimación de la potencia para las pruebas comparadas usando el caso 1.

$n =$	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$R_{.60}$	4,97	5,08	4,71	4,92	5,13	4,90	4,94	4,90	5,09	4,89	4,95	5,04	5,18	5,00	5,06	4,85
$R_{.80}$	4,99	5,02	4,95	4,93	5,08	5,01	5,06	4,95	4,95	4,87	4,94	5,10	5,02	4,96	4,97	5,06
$C^*_{.10}$	5,14	5,12	5,17	4,78	5,05	5,30	5,10	5,18	4,97	4,88	5,20	4,97	5,23	5,02	5,06	4,98
$C^*_{.20}$	5,24	5,15	5,31	4,94	4,84	4,94	5,18	4,90	5,00	4,86	4,98	5,07	5,21	5,00	5,20	5,06
$C^*_{.60}$	5,05	5,00	4,70	4,94	5,07	5,11	5,24	4,93	4,91	4,99	4,96	5,04	4,98	5,04	5,30	5,09
$C^*_{.80}$	5,03	4,96	5,04	4,91	5,06	4,94	5,01	4,94	4,94	4,92	4,92	5,14	5,04	4,96	4,96	5,07
$R^*_{.60}$	4,90	5,12	4,72	4,94	4,84	5,18	5,08	5,33	5,11	4,72	4,86	5,09	5,05	5,08	4,91	5,06
$R^*_{.80}$	5,03	5,12	4,97	4,99	4,91	5,10	5,00	4,84	4,97	4,84	4,81	5,10	5,07	5,03	5,04	5,04
W	5,05	5,25	4,83	4,84	5,28	5,23	4,86	4,90	5,01	4,81	4,95	4,94	5,05	5,08	5,21	4,66
R^+	5,09	4,97	4,90	4,79	5,01	4,95	4,96	4,88	5,11	4,93	5,07	5,12	5,02	4,99	5,08	5,08
R	5,25	4,99	4,91	4,88	4,86	4,98	5,01	5,06	5,17	4,95	5,11	5,09	5,01	5,09	5,00	5,26
C_6	4,93	5,22	4,85	4,97	4,99	5,21	5,02	4,81	5,06	4,79	4,88	5,13	4,98	5,03	5,08	5,07
M_0	5,15	5,01	4,77	4,68	4,93	5,12	4,96	4,85	5,06	5,03	5,05	4,93	5,08	4,95	4,81	5,00
$M_{.10}$	5,13	5,04	4,76	4,78	4,94	5,31	5,03	4,88	5,05	5,07	5,08	5,00	5,08	4,87	5,00	5,00
$M_{.20}$	5,00	5,10	4,81	4,76	5,02	5,23	5,05	4,72	5,12	5,04	5,17	5,10	4,92	4,80	4,78	5,28
$M_{.25}$	4,99	5,10	4,82	4,88	5,04	5,23	5,14	4,76	5,10	5,06	5,09	5,24	5,06	4,91	4,97	5,20
$\gamma_1(F_n)$	0,25	4,70	1,21	2,65	0,54	4,54	1,87	2,67	0,90	4,10	2,09	2,48	3,22	3,67	2,04	2,14

Tabla 3: Estimación de la potencia para las pruebas comparadas usando el caso 2.

$n =$	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$R_{.60}$	5,12	5,30	5,29	5,10	5,23	5,05	5,36	5,22	5,17	5,33	5,15	5,65	5,33	5,87	5,26	5,57
$R_{.80}$	5,15	5,16	5,08	5,23	5,26	5,27	5,25	5,30	5,32	5,26	5,41	5,65	5,47	5,89	5,54	5,91
$C^*_{.10}$	5,34	5,33	6,01	4,98	4,99	5,34	5,39	5,25	5,34	5,50	5,52	5,46	5,60	5,88	5,35	5,81
$C^*_{.20}$	5,37	5,78	5,78	5,17	5,37	5,52	5,42	5,36	5,69	5,61	5,57	5,57	5,60	5,93	5,55	5,75
$C^*_{.60}$	5,22	5,30	5,28	5,14	5,13	5,21	5,47	5,17	5,39	5,24	5,27	5,41	5,19	5,48	5,73	5,55
$C^*_{.80}$	5,08	5,27	5,19	5,08	5,17	5,17	5,23	5,24	5,29	5,24	5,40	5,64	5,43	5,87	5,53	5,93
$R^*_{.60}$	5,06	5,35	5,32	5,08	5,10	5,23	5,23	5,78	5,34	5,14	5,18	5,66	5,18	5,66	5,32	5,47
$R^*_{.80}$	5,12	5,41	5,13	5,00	5,00	5,21	5,25	5,30	5,27	5,17	5,59	5,72	5,32	5,96	5,56	5,88
W	5,00	5,05	5,03	5,09	4,97	5,05	5,29	4,78	4,96	5,26	5,12	5,33	5,07	5,13	5,11	5,01
R^+	5,24	5,44	5,29	5,03	5,18	5,22	5,18	5,31	5,28	5,46	5,21	5,13	5,20	5,17	5,11	5,28
R	5,37	5,50	5,28	5,12	5,23	5,17	5,15	5,34	5,35	5,36	5,25	5,17	5,16	5,18	5,24	5,30
C_6	5,07	5,18	5,18	5,07	5,24	5,19	5,26	5,26	5,24	5,34	5,29	5,56	5,23	5,68	5,43	5,61
M_0	5,30	5,46	5,33	5,16	5,29	5,19	5,31	5,19	5,25	5,36	5,29	5,37	5,47	5,51	5,22	5,29
$M_{.10}$	5,21	5,38	5,33	5,13	5,37	5,17	5,42	5,24	5,23	5,41	5,26	5,40	5,44	5,39	5,31	5,26
$M_{.20}$	5,16	5,35	5,38	5,02	5,47	5,15	5,53	5,25	5,25	5,33	5,35	5,45	5,32	5,38	5,07	5,40
$M_{.25}$	5,18	5,35	5,36	5,18	5,50	5,14	5,55	5,27	5,24	5,25	5,40	5,53	5,34	5,58	5,29	5,32

TABLA 4: Estimación de la potencia para las pruebas comparadas usando el caso 3.

$n=$	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$R_{.60}$	5,69	5,89	6,09	5,96	6,28	6,46	6,74	7,77	7,52	8,09	8,08	8,19	9,00	8,62	9,54	9,81
$R_{.80}$	5,55	5,96	6,18	6,32	6,43	6,60	7,25	7,62	7,76	8,14	8,13	9,14	9,93	9,98	10,72	11,22
$C^*_{.10}$	6,29	6,84	7,94	6,96	7,22	7,45	7,65	7,59	7,42	7,86	8,32	8,40	9,00	8,84	9,23	9,40
$C^*_{.20}$	6,27	6,82	7,58	6,96	7,39	7,86	7,82	8,05	8,07	8,69	8,87	8,78	9,53	9,81	10,07	10,49
$C^*_{.60}$	5,90	6,04	6,21	5,84	6,63	6,65	6,88	7,21	7,54	7,36	7,24	7,92	7,24	8,23	8,68	8,68
$C^*_{.80}$	5,70	5,31	5,23	5,58	5,86	6,26	6,84	7,41	7,58	7,98	8,05	9,01	9,82	9,95	10,67	11,20
$R^*_{.60}$	5,48	5,76	6,02	6,16	6,24	6,82	6,64	8,24	7,16	7,57	7,25	7,90	8,45	8,43	8,91	8,81
$R^*_{.80}$	5,57	5,85	6,52	6,38	6,45	6,78	7,13	7,71	7,90	8,36	8,28	9,04	9,95	9,93	10,51	11,09
W	5,21	5,41	5,45	5,20	5,03	5,31	5,43	5,46	5,34	5,58	5,32	5,42	5,71	5,50	5,74	5,47
R^+	5,75	6,08	6,10	6,18	6,35	6,27	6,05	6,76	6,52	6,98	6,79	6,86	6,95	7,21	7,15	7,28
R	6,06	6,26	6,32	6,52	6,59	6,57	6,36	6,81	6,78	7,17	7,05	7,17	7,20	7,46	7,37	7,57
C_6	5,28	5,51	5,61	5,84	6,14	6,14	6,38	7,17	7,00	7,60	7,69	8,19	8,83	8,97	9,61	9,78
M_0	5,95	6,46	6,66	6,79	6,84	6,93	6,86	7,55	7,28	7,68	7,54	8,01	8,15	8,39	8,15	8,48
$M_{.10}$	5,86	6,39	6,54	6,71	6,88	7,01	6,89	7,68	7,44	7,68	7,67	8,14	8,22	8,39	8,45	8,62
$M_{.20}$	5,76	6,24	6,72	6,72	6,88	7,00	6,95	7,76	7,57	7,65	7,67	8,38	8,16	8,36	8,23	9,04
$M_{.25}$	5,70	6,24	6,72	6,90	6,93	7,04	6,95	7,78	7,61	7,65	7,66	8,11	8,36	8,70	8,60	8,95

TABLA 5: Estimación de la potencia para las pruebas comparadas usando el caso 4.

$n=$	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$R_{.60}$	6,51	6,67	7,00	7,48	8,36	9,22	9,48	10,82	11,12	11,82	13,30	12,54	13,31	14,24	14,56	15,60
$R_{.80}$	6,19	7,08	7,53	7,92	8,24	8,58	9,66	10,68	10,97	12,11	12,72	14,71	15,89	16,90	18,02	19,08
$C^*_{.10}$	7,49	7,98	9,94	8,94	9,17	9,64	9,51	10,64	10,49	10,89	12,08	12,39	12,89	13,13	13,53	14,38
$C^*_{.20}$	7,46	9,13	9,42	9,50	9,57	10,55	10,64	11,60	11,76	12,58	13,31	13,35	13,71	14,12	14,60	16,25
$C^*_{.60}$	6,92	7,05	7,29	7,28	8,33	9,41	9,05	9,41	10,53	10,46	9,97	11,08	10,41	11,88	13,30	13,16
$C^*_{.80}$	6,48	5,54	5,64	6,48	7,17	7,78	9,01	10,19	10,52	11,79	12,54	14,46	15,68	16,76	17,96	19,05
$R^*_{.60}$	6,17	6,74	6,67	7,65	8,23	9,34	9,30	11,08	9,99	10,68	11,50	11,51	12,24	13,02	13,34	14,00
$R^*_{.80}$	6,43	6,90	7,53	7,93	8,43	8,88	9,79	11,14	11,38	12,47	13,69	14,41	15,57	16,81	18,03	19,15
W	5,58	5,47	5,13	5,69	5,48	5,87	5,69	6,19	6,01	5,78	6,24	6,48	6,10	6,47	6,40	6,44
R^+	6,73	6,88	7,08	7,74	8,27	8,30	8,17	8,70	8,79	9,31	9,17	9,36	9,50	9,71	10,12	10,27
R	7,53	7,49	7,79	8,16	8,53	8,84	8,69	9,29	9,30	9,55	9,73	9,85	10,00	10,16	10,19	10,83
C_6	5,40	5,75	6,00	6,61	7,12	8,00	8,30	9,33	9,84	10,66	11,81	12,44	13,25	14,42	15,42	16,38
M_0	7,08	7,88	8,17	8,69	9,13	9,57	9,74	10,59	10,58	11,04	11,52	11,37	12,39	12,65	12,60	13,16
$M_{.10}$	6,99	7,89	8,07	8,81	9,20	9,93	9,83	10,77	10,83	11,41	11,73	11,80	12,59	12,83	13,28	13,44
$M_{.20}$	6,57	7,77	8,27	8,87	9,35	10,14	10,34	11,16	11,10	11,64	12,00	12,24	12,72	12,94	13,29	14,22
$M_{.25}$	6,47	7,77	8,44	9,20	9,58	10,29	10,35	11,32	11,30	11,78	12,20	12,67	13,06	13,61	13,84	14,34

TABLA 6: Estimación de la potencia para las pruebas comparadas usando el caso 5.

$n=$	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$R_{.60}$	7,37	7,38	8,57	9,34	10,34	11,36	12,50	14,04	14,74	15,82	17,39	16,96	18,59	18,91	19,96	21,88
$R_{.80}$	6,51	7,93	8,46	9,06	9,30	9,85	12,41	13,34	14,12	15,07	15,87	20,14	22,01	22,68	24,52	26,35
$C^*_{.10}$	8,90	9,96	12,02	10,96	11,56	11,84	12,37	12,65	13,43	13,78	15,62	16,13	16,93	17,05	17,65	18,70
$C^*_{.20}$	8,67	10,96	11,78	11,65	12,21	13,05	13,72	14,20	14,68	15,47	17,00	17,69	18,49	18,62	19,26	21,38
$C^*_{.60}$	7,84	8,04	8,93	9,00	10,20	11,24	11,58	11,56	12,38	12,76	13,14	13,98	13,41	14,44	15,79	16,38
$C^*_{.80}$	7,21	5,90	6,14	7,05	7,89	8,68	11,15	12,39	13,45	14,58	15,63	19,64	21,70	22,50	24,40	26,30
$R^*_{.60}$	7,07	7,29	8,03	9,40	9,82	11,27	12,00	14,19	13,25	13,76	14,72	15,33	16,70	16,92	17,68	19,41
$R^*_{.80}$	6,68	7,87	8,28	9,25	9,43	10,24	12,60	14,00	15,08	16,35	17,86	19,68	21,83	22,63	24,73	26,79
W	5,63	5,61	5,86	6,01	5,81	6,21	6,16	6,48	6,72	6,61	6,73	6,86	7,36	7,18	7,15	7,26
R^+	7,54	7,95	8,22	9,11	9,80	9,72	9,86	10,34	10,42	10,97	11,43	11,54	11,80	12,19	12,52	13,23
R	8,63	8,76	9,25	9,98	10,12	10,74	10,80	11,09	11,13	11,48	12,28	12,24	12,65	12,84	12,97	14,09
C_6	5,91	6,22	6,89	7,69	8,47	9,22	10,36	11,92	12,91	14,08	15,60	16,77	18,32	19,32	21,06	22,95
M_0	8,53	9,13	9,84	10,87	11,20	11,93	12,56	13,37	13,55	14,02	14,86	15,16	16,28	16,56	16,66	17,67
$M_{.10}$	8,60	9,24	9,85	11,15	11,49	12,32	12,72	13,72	13,87	14,39	15,19	15,62	16,76	16,66	17,45	18,15
$M_{.20}$	8,09	9,12	10,29	11,16	11,72	12,74	13,20	14,22	14,39	14,73	15,60	16,46	16,94	16,98	17,66	19,14
$M_{.25}$	7,90	9,12	10,42	11,57	11,98	13,02	13,19	14,52	14,87	15,05	15,84	16,53	17,37	17,77	18,49	19,28

Para los casos 7 a 9 (tablas 8 a 10), donde la asimetría es evidente, la prueba $M_{.25}$ siempre tiene las potencias más altas. Sin embargo, para n cerca de 25 las diferencias con las pruebas $R_{.80}, R^*_{.80}$ y $C^*_{.80}$ no son grandes. Por ejemplo, en el caso 9 para $n = 25$ la mayor potencia la tiene $M_{.25}$ con el $76,4\,\%$, mientras que las pruebas $R^*_{.80}, R_{.80}$ y $C^*_{.80}$ tienen potencias del $71\,\%, 70\,\%$ y $69\,\%$, respectivamente, estando por encima de pruebas como R (Baklizi 2003), R^+ (McWilliams 1990) y W (Wilcoxon) que obtuvieron potencias del $52,6\,\%, 49,5\,\%$ y $13\,\%$, respectivamente.

Para examinar si la potencia de las pruebas R_p, R^*_p y C^*_p sigue mejorando cuando n crece, se realizaron también simulaciones para $n = 30, 50, 100, 150, 200, 250$ y 500 con $p = 60\,\%$ y $80\,\%$. Además se agregaron las pruebas T y W_p propuestas

Tabla 7: Estimación de la potencia para las pruebas comparadas usando el caso 6.

$n=$	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$R_{.60}$	8,54	8,96	10,60	12,18	13,24	17,64	17,56	20,18	22,11	24,75	26,91	24,36	26,42	28,06	30,56	32,80
$R_{.80}$	7,55	10,26	11,01	11,59	12,40	13,24	18,40	20,24	21,71	23,37	24,82	35,50	38,63	41,83	45,08	47,86
$C^*_{.10}$	10,22	11,25	13,52	12,66	13,05	13,13	13,70	14,76	14,94	15,70	17,10	17,91	17,86	18,82	19,43	19,59
$C^*_{.20}$	9,84	13,39	14,10	13,88	14,16	14,52	15,04	16,05	16,52	17,54	18,93	19,50	19,82	20,52	21,55	22,90
$C^*_{.60}$	9,54	10,27	11,44	11,90	14,12	16,65	15,37	15,36	17,43	17,38	16,97	16,89	16,03	17,21	19,90	19,40
$C^*_{.80}$	7,36	6,88	7,34	8,87	10,28	11,47	16,81	18,91	20,78	22,78	24,40	34,98	38,29	41,62	44,93	47,73
$R^*_{.60}$	7,62	8,72	9,73	12,04	14,10	17,05	17,59	21,49	20,80	22,79	25,15	24,96	27,83	28,47	30,26	34,01
$R^*_{.80}$	7,43	9,38	10,41	11,32	12,43	13,04	18,35	20,88	23,22	26,28	28,18	33,88	37,65	40,99	44,68	48,33
W	5,55	5,83	5,99	6,15	6,53	6,32	6,55	6,74	6,94	7,70	7,50	7,54	7,73	7,83	7,90	8,09
R^+	9,57	10,83	11,66	13,09	14,73	14,94	16,18	17,28	18,11	19,71	20,26	21,79	22,35	23,72	25,23	26,04
R	11,72	12,48	13,31	14,65	15,66	17,02	18,05	18,75	19,71	20,62	22,30	23,47	23,87	25,40	26,36	28,12
C_6	5,87	6,78	7,58	8,76	10,37	12,34	14,59	17,02	19,64	23,02	25,82	28,90	32,32	35,99	39,69	43,48
M_0	10,43	12,83	14,66	16,64	18,76	20,48	23,03	24,95	26,73	28,70	31,09	33,14	35,08	37,23	39,08	41,18
$M_{.10}$	10,23	12,82	14,64	17,04	19,39	21,21	23,74	25,66	27,62	29,84	32,42	34,46	36,42	38,50	41,13	43,01
$M_{.20}$	9,25	12,54	15,09	17,26	19,64	21,96	24,93	27,03	28,74	31,08	34,13	36,63	37,66	39,80	42,34	45,72
$M_{.25}$	8,94	12,54	15,34	17,75	20,18	22,60	25,12	27,53	29,43	31,83	34,77	36,64	39,01	41,56	44,19	46,41

Tabla 8: Estimación de la potencia para las pruebas comparadas usando el caso 7.

$n=$	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$R_{.60}$	11,88	12,97	16,30	20,02	22,22	29,00	28,84	32,34	35,09	39,28	42,24	38,90	41,85	43,84	47,20	51,37
$R_{.80}$	8,40	12,59	13,35	14,26	14,89	15,58	24,48	26,43	27,84	29,68	30,74	50,59	53,90	56,30	59,48	62,27
$C^*_{.10}$	14,65	15,77	20,19	18,62	19,00	19,83	20,26	21,55	22,73	23,95	25,70	27,35	28,04	28,49	28,95	30,82
$C^*_{.20}$	13,92	20,67	21,33	20,46	20,75	21,61	21,77	23,68	24,85	26,99	27,86	29,47	30,21	31,26	32,00	34,43
$C^*_{.60}$	13,78	15,37	17,79	18,77	22,29	27,29	23,98	24,39	25,73	26,45	26,24	24,25	23,04	23,14	28,00	26,78
$C^*_{.80}$	7,70	7,20	7,47	9,40	11,45	13,27	21,55	23,99	26,40	28,73	30,18	49,40	53,32	55,90	59,28	62,11
$R^*_{.60}$	10,01	11,62	14,15	17,93	21,70	26,19	27,23	32,68	32,17	34,84	37,97	39,40	43,17	44,00	45,85	51,54
$R^*_{.80}$	7,94	10,93	11,98	13,41	13,68	14,17	25,36	28,34	30,25	33,51	35,43	48,92	53,81	56,24	59,33	64,70
W	5,80	6,67	7,00	7,49	7,47	8,05	8,36	8,26	8,71	9,41	9,51	10,01	10,32	10,41	11,16	11,06
R^+	13,86	15,53	17,18	19,08	21,87	22,43	24,27	25,77	27,15	30,04	30,02	32,82	33,75	35,09	37,97	38,51
R	17,44	18,12	19,90	21,98	23,28	25,62	26,91	28,15	29,94	31,55	32,81	35,16	35,87	37,73	39,48	41,35
C_6	7,11	9,00	11,21	13,56	16,58	20,38	24,00	28,15	32,40	37,38	41,16	46,34	50,85	54,36	58,89	63,70
M_0	16,60	19,82	22,89	25,57	28,65	31,81	34,98	37,56	40,28	43,54	45,83	48,32	51,94	53,27	55,79	58,48
$M_{.10}$	16,52	20,09	23,23	26,29	29,64	32,61	36,04	38,95	41,71	45,19	47,70	50,06	53,64	54,74	58,18	60,50
$M_{.20}$	14,52	20,10	24,35	27,01	30,44	33,67	37,71	40,72	43,19	46,59	49,61	52,51	55,18	56,26	59,67	63,30
$M_{.25}$	13,89	20,10	24,84	27,97	31,19	34,51	37,82	41,30	44,21	47,50	50,53	53,85	56,64	58,19	61,42	63,97

Tabla 9: Estimación de la potencia para las pruebas comparadas usando el caso 8.

$n=$	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$R_{.60}$	13,42	15,78	19,85	25,38	28,23	35,05	35,44	39,52	43,09	46,60	50,97	46,32	49,69	52,53	56,06	59,59
$R_{.80}$	8,60	13,58	14,48	15,38	15,91	16,62	27,50	29,14	30,70	31,98	33,23	56,51	59,94	62,43	64,90	67,24
$C^*_{.10}$	17,12	17,90	23,20	22,19	22,41	24,08	23,45	25,36	26,87	27,41	30,27	31,55	32,68	33,82	35,07	35,51
$C^*_{.20}$	16,08	24,73	24,72	24,48	24,04	25,56	25,50	27,18	28,77	30,69	33,60	33,68	35,49	36,65	37,70	39,78
$C^*_{.60}$	15,64	18,73	21,77	23,26	27,69	33,36	28,96	29,49	30,83	31,98	32,49	28,21	28,25	27,73	32,71	31,88
$C^*_{.80}$	8,40	7,30	8,34	9,85	11,97	14,01	23,55	26,35	28,89	30,82	32,61	55,20	59,12	62,00	64,60	67,07
$R^*_{.60}$	11,55	13,50	16,32	21,73	25,62	30,86	32,77	39,30	38,96	41,33	45,39	46,17	50,70	52,38	54,01	59,49
$R^*_{.80}$	8,14	11,73	12,55	13,94	13,85	14,20	28,17	31,70	33,02	35,03	36,47	54,73	59,83	62,88	65,47	69,55
W	6,70	7,31	7,77	8,65	9,06	8,71	9,21	9,26	10,48	10,63	10,89	11,28	12,23	12,60	12,70	12,88
R^+	16,21	18,92	20,38	23,17	26,20	27,57	28,75	30,77	33,18	36,23	37,34	38,88	40,39	42,04	45,16	46,06
R	20,24	21,78	23,75	26,48	27,67	31,19	31,68	33,71	36,65	37,52	40,62	41,47	42,92	45,03	46,61	49,14
C_6	8,40	10,62	13,39	17,13	20,87	24,82	29,79	35,04	40,39	44,65	49,72	54,33	59,27	63,66	67,74	71,75
M_0	20,07	24,44	27,93	31,65	34,82	38,66	41,93	45,11	48,82	51,22	54,90	57,09	60,16	61,88	64,23	67,07
$M_{.10}$	20,34	24,99	28,50	32,36	36,11	39,73	43,13	46,33	50,28	52,87	56,72	58,84	61,94	63,68	66,56	69,16
$M_{.20}$	18,24	24,86	29,87	33,30	37,02	40,60	45,04	48,29	51,91	54,53	58,53	61,06	63,59	65,28	68,02	71,83
$M_{.25}$	17,47	24,86	30,51	34,78	37,84	41,61	45,25	49,09	52,89	55,32	59,31	61,94	65,07	66,92	69,74	72,66

por Tajuddin (1994) y Modarres & Gastwirth (1998), respectivamente, para la comparación.

3.2. Estudio de potencia para $n = 30, 50, 100, 150, 200, 250$ y 500

Con base en las tablas 11 a 17, se concluye que:

Para tamaños de muestra $n \geq 150$, la prueba C^*_p no fue incluida porque su desempeño desmejoró frente a las pruebas $R_{.80}$ y $R^*_{.80}$, sin que necesariamente la potencia de la prueba $C^*_{.80}$ estuviera por debajo de algunas de las competidoras de la literatura.

Tabla 10: Estimación de la potencia para las pruebas comparadas usando el caso 9.

$n=$	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$R_{.60}$	14,17	16,32	21,13	27,37	30,50	38,19	37,74	41,54	44,73	49,18	53,39	48,45	51,35	54,73	59,11	61,72
$R_{.80}$	8,70	13,99	14,85	15,68	16,34	16,86	28,29	30,00	31,44	32,91	33,99	59,23	61,69	64,46	67,00	68,97
$C^*_{.10}$	18,09	18,39	24,20	23,11	23,97	25,23	24,64	25,89	27,14	28,09	31,00	31,91	33,22	34,68	35,71	36,69
$C^*_{.20}$	16,95	25,35	25,98	25,10	25,39	26,37	26,76	27,56	29,21	31,54	34,26	34,33	35,74	36,74	38,39	40,04
$C^*_{.60}$	16,47	19,53	23,03	24,75	29,84	36,08	31,59	31,92	33,08	33,85	34,03	29,92	29,86	29,96	34,90	32,86
$C^*_{.80}$	8,52	7,26	8,45	9,77	12,11	14,15	24,43	27,20	29,50	31,65	33,28	57,79	60,92	63,98	66,75	68,84
$R^*_{.60}$	11,90	14,21	17,74	22,84	27,68	33,55	34,33	40,97	40,36	43,56	48,27	49,56	53,11	54,88	57,40	62,00
$R^*_{.80}$	8,34	12,38	13,28	13,97	13,64	13,74	28,31	32,16	34,84	36,62	37,48	56,80	61,31	64,09	66,75	70,99
W	6,69	7,47	7,83	8,78	9,02	9,06	9,71	9,82	10,78	10,93	11,61	11,92	12,02	12,77	13,29	12,98
R^+	17,60	19,47	21,54	23,95	28,21	29,51	31,56	33,42	35,49	39,15	39,57	42,24	43,67	45,94	49,23	49,55
R	21,99	22,59	25,10	27,71	29,88	33,19	34,62	36,30	39,07	40,62	42,90	44,86	46,24	49,15	50,81	52,62
C_6	8,67	11,12	14,33	18,07	22,63	27,31	31,94	36,87	42,20	47,74	52,58	57,93	62,06	66,67	70,99	74,49
M_0	21,65	25,43	29,54	32,94	37,58	41,58	45,30	48,85	51,53	55,25	57,86	60,80	64,18	66,74	69,11	71,41
$M_{.10}$	21,98	26,13	30,25	33,91	38,83	42,94	46,61	50,35	53,23	56,98	59,78	62,76	65,84	68,38	71,42	73,36
$M_{.20}$	19,68	26,27	31,75	35,06	39,88	44,30	48,26	52,30	54,88	58,57	61,92	65,04	67,56	69,88	73,05	75,94
$M_{.25}$	18,85	26,27	32,50	36,57	40,72	45,28	48,72	53,01	56,21	59,46	62,79	66,70	68,83	71,63	74,47	76,46

La prueba W, referenciada para la hipótesis de simetría por Gibbons & Chakraborti (1992), es la que tiene el desempeño más bajo de todas las pruebas que se compararon.

Existen grandes diferencias entre las potencias de las pruebas $R_{.80}$ y $R^*_{.80}$ con respecto a algunas de sus competidoras. Por ejemplo, para $n = 200$ con muestras provenientes del caso 4, la potencia de las pruebas $R_{.80}$ y $R^*_{.80}$ es del $98,3\,\%$ y $97,9\,\%$, respectivamente, mientras que las pruebas R^+, R, T y W tienen potencias del $33,2\,\%$, $34,1\,\%$, $64,9\,\%$ y $21,4\,\%$, respectivamente, las pruebas M_0, $M_{.10}$, $M_{.20}$ y $M_{.25}$ tienen potencias del $56\,\%$, $58\,\%$, $61\,\%$ y $62\,\%$, respectivamente, y la potencia de la prueba C_6 es del $93\,\%$, notándose claramente el dominio de las pruebas propuestas sobre las competidoras mencionadas.

Se podría decir que las pruebas W_{70} y W_{80} son las competidoras más directas que se tienen, pues siguiendo con el ejemplo de $n = 200$, estas tuvieron potencias del $98\,\%$ y $99\,\%$, respectivamente. Sin embargo, no existe una diferencia clara entre las pruebas propuestas y estas; es más, en la tabla 11 (potencia para $n = 30$) se puede observar que las diferencias son mínimas para los casos 1 al 5 donde las potencias de W_{70} son muy similares a las de $R_{.80}$ y $R^*_{.80}$, y que para los demás casos (6 al 9) las pruebas propuestas tienen las mayores potencias. Además, recordemos que la prueba W_p utiliza dos estadísticos por ser una prueba híbrida, lo que hace un poco más complicado usarla.

4. Conclusiones y discusión

En general, para los tamaños de muestra menores o iguales a 25 se concluye:

En los casos 2, 3, 4 y 5 de la DLG y para $n < 20$, la potencia de la prueba $C^*_{.20}$ es mayor que la potencia de las pruebas M_p, R, R^+, C_6 y W.

En los casos 2 al 6 y para $20 < n \leq 25$, se observa que la potencia de las pruebas $R_{.80}, R^*_{.80}$ y $C^*_{.80}$ es mayor que la potencia de las pruebas M_p, R, R^+, C_6 y W consultadas en la literatura.

Para los casos 7 al 9, las potencias de las pruebas propuestas $R_{.80}, R^*_{.80}$ y $C^*_{.80}$ son mayores que las potencias de las pruebas R (Baklizi 2003), R^+ (McWilliams

Tabla 11: Potencias para los nueve casos de la DLG y $n = 30$.

	1	2	3	4	5	6	7	8	9
$R_{.60}$	5,13	5,61	10,99	18,47	25,48	36,86	54,93	64,70	66,67
$R_{.80}$	5,19	6,02	13,23	25,65	36,17	71,07	86,86	91,43	93,53
$R^*_{.60}$	4,97	5,46	10,00	16,89	22,86	41,51	59,94	68,57	71,62
$R^*_{.80}$	5,22	5,92	12,87	24,77	34,34	69,57	84,01	88,43	90,20
$C^*_{.60}$	5,21	5,74	9,89	15,91	20,23	21,46	28,25	33,28	34,28
$C^*_{.80}$	5,32	6,49	12,05	21,52	30,08	62,24	80,71	87,08	89,90
R^+	5,16	5,28	7,68	11,49	14,40	30,17	45,30	54,33	57,78
R	5,31	5,33	8,08	12,19	15,48	33,03	48,45	57,37	61,00
W	4,93	5,06	5,98	6,93	7,63	9,04	12,72	15,03	15,69
T	4,68	5,51	9,55	13,91	18,72	23,30	39,10	48,56	50,91
C_6	5,17	5,81	11,64	21,65	30,42	60,96	79,37	85,74	88,43
M_0	5,15	5,54	9,36	15,53	20,89	50,62	68,66	77,15	80,84
$M_{.10}$	5,11	5,53	9,61	15,79	21,55	53,28	71,11	79,01	82,81
$M_{.20}$	4,89	5,21	9,39	15,79	21,62	54,63	72,53	80,18	84,19
$M_{.25}$	4,95	5,27	9,41	16,05	22,17	55,97	73,63	81,06	84,89
$W_{.70}$	4,18	5,31	12,93	25,45	36,45	60,58	82,10	88,18	90,39
$W_{.80}$	3,70	4,81	11,94	24,07	34,03	60,94	74,38	77,77	78,90

Tabla 12: Potencias para los nueve casos de la DLG y $n = 50$.

CASO =	1	2	3	4	5	6	7	8	9
$R_{.60}$	4,85	6,54	16,91	31,66	43,87	54,19	75,11	82,54	83,19
$R_{.80}$	4,87	7,07	22,10	45,91	62,68	88,90	97,34	98,77	99,18
$R^*_{.60}$	4,93	6,32	15,54	29,81	41,55	63,69	83,84	89,66	90,72
$R^*_{.80}$	4,87	6,98	20,39	42,24	57,95	91,46	97,84	98,87	99,29
$C^*_{.80}$	4,97	5,96	9,17	16,43	22,39	57,50	81,71	89,38	91,48
R^+	4,86	5,34	8,86	14,48	20,22	49,04	67,22	76,62	81,24
R	5,23	5,64	9,47	15,61	21,66	51,54	69,73	78,76	82,88
W	4,88	5,08	6,81	8,92	10,65	12,22	18,61	23,31	24,23
T	4,76	5,92	12,76	21,72	29,85	37,15	60,23	70,65	73,29
C_6	4,81	6,81	19,89	43,10	59,30	95,13	98,99	99,58	99,73
M_0	4,79	5,54	11,20	21,96	31,13	76,14	90,32	94,75	96,55
$M_{.10}$	4,76	5,67	11,55	22,79	32,42	78,70	91,88	95,82	97,34
$M_{.20}$	4,75	5,84	12,04	23,92	33,99	81,35	93,34	96,77	97,81
$M_{.25}$	4,64	5,68	12,14	24,19	34,49	82,33	93,88	97,09	98,01
$W_{.70}$	4,45	6,66	22,15	46,12	63,67	87,16	97,47	99,06	99,37
$W_{.80}$	4,09	6,33	21,67	48,77	67,02	94,52	99,21	99,77	99,86

1990) y W (Wilcoxon), y se mantienen cerca de las potencias obtenidas por la prueba $M_{.25}$ que fue la que alcanzó los mejores resultados.

Para los casos 7 al 9 de la DLG, donde las pruebas $C^*_{.20}$, $R_{.80}$, $R^*_{.80}$ y $C^*_{.80}$ no tienen las mayores potencias (aunque están por encima de pruebas reconocidas) es más fácil, por su forma distribucional, hacer un análisis descriptivo y detectar la asimetría; es decir que la asimetría se puede detectar de forma sencilla para los casos donde las pruebas propuestas no tienen su mejor desempeño, mientras que para los casos donde la asimetría no es tan severa y se requiere una prueba, la mayor potencia se obtiene usando las pruebas propuestas.

En general, para los tamaños de muestra mayores que 25 se concluye:

Tabla 13: Potencias para los nueve casos de la DLG y $n = 100$.

CASO =	1	2	3	4	5	6	7	8	9
$R_{.60}$	5,16	8,40	32,15	58,21	73,70	77,87	93,67	96,49	97,01
$R_{.80}$	5,12	10,53	45,62	80,79	92,40	98,97	99,95	99,99	99,99
$R^*_{.60}$	5,07	8,13	30,60	57,44	74,12	87,81	97,77	99,01	99,29
$R^*_{.80}$	5,06	9,66	41,50	76,86	90,34	99,70	99,99	100,00	100,00
$C^*_{.80}$	4,94	6,58	13,57	22,00	23,60	54,44	85,82	93,50	96,08
R^+	4,73	5,87	11,27	21,28	31,72	78,60	92,96	96,63	98,00
R	4,50	5,58	10,91	20,64	31,36	78,37	92,97	96,54	97,94
W	4,82	5,34	9,19	13,00	16,80	20,46	35,34	42,29	43,99
T	4,84	7,10	21,75	38,32	52,38	63,24	88,32	94,30	95,24
C_6	4,89	8,82	37,15	74,99	88,72	99,98	100,00	100,00	100,00
M_0	4,84	6,47	15,55	35,27	51,37	97,34	99,67	99,94	99,99
$M_{.10}$	4,86	6,51	16,24	36,81	53,56	98,00	99,81	99,97	99,99
$M_{.20}$	4,83	6,66	16,92	38,54	55,62	98,62	99,87	99,98	100,00
$M_{.25}$	4,91	6,63	17,33	39,36	56,72	98,84	99,89	99,99	100,00
$W_{.70}$	5,08	10,31	46,12	79,97	92,77	99,28	99,99	100,00	100,00
$W_{.80}$	4,99	10,88	49,53	85,63	95,70	99,92	100,00	100,00	100,00

Tabla 14: Potencias para los nueve casos de la DLG y $n = 150$.

CASO =	1	2	3	4	5	6	7	8	9
$R_{.60}$	4,82	9,94	45,42	75,58	88,96	88,82	98,24	99,49	99,58
$R_{.80}$	4,77	12,76	62,72	93,84	98,74	99,89	100,00	100,00	100,00
$R^*_{.60}$	4,90	9,89	44,62	76,64	90,21	95,16	99,55	99,88	99,94
$R^*_{.80}$	4,46	11,62	58,39	92,13	98,34	99,99	100,00	100,00	100,00
R^+	4,96	5,82	12,99	28,28	41,34	92,01	98,80	99,53	99,81
R	4,73	5,42	12,43	27,47	40,56	91,80	98,77	99,53	99,81
W	4,86	5,69	11,19	17,18	23,82	28,93	48,31	57,73	60,57
T	5,10	8,01	30,21	53,49	70,29	80,31	97,05	99,09	99,48
C_6	4,64	10,38	47,63	87,75	96,31	100,00	100,00	100,00	100,00
M_0	5,24	6,61	19,53	46,97	67,04	99,76	99,99	100,00	100,00
$M_{.10}$	5,34	6,73	20,40	49,07	69,53	99,85	100,00	100,00	100,00
$M_{.20}$	5,33	6,72	21,26	51,52	71,83	99,93	100,00	100,00	100,00
$M_{.25}$	5,28	6,66	21,74	52,77	73,16	99,95	100,00	100,00	100,00
$W_{.70}$	5,04	12,57	64,02	93,70	98,78	99,97	100,00	100,00	100,00
$W_{.80}$	4,91	13,27	68,15	96,47	99,54	100,00	100,00	100,00	100,00

La prueba W, referenciada para la hipótesis de simetría por Gibbons & Chakraborti (1992), es la que tiene el desempeño más bajo de todas las pruebas que se compararon.

Las pruebas $R_{.80}$ y $R^*_{.80}$ tienen las mayores potencias en vecindades de la hipótesis nula, lo que permite conjeturar que las pruebas propuestas son localmente más potentes.

Para $n \geq 30$, como se sospechaba, la potencia de las pruebas $R_{.80}$ y $R^*_{.80}$ mejoró para los casos 7 al 9, tanto así que para $n = 30$ resultan ser las pruebas con el mejor desempeño en cualquiera de los casos de la DLG seleccionados, es decir, que la potencia de las pruebas $R_{.80}$ y $R^*_{.80}$ es siempre mayor que la potencia de todas sus competidoras de la literatura en el tamaño de muestra mencionado.

Tabla 15: Potencias para los nueve casos de la DLG y $n = 200$.

CASO =	1	2	3	4	5	6	7	8	9
$R_{.60}$	5,30	12,46	58,51	87,10	95,86	94,28	99,54	99,91	99,92
$R_{.80}$	5,17	16,05	76,59	98,27	99,83	99,99	100,00	100,00	100,00
$R^*_{.60}$	5,16	11,89	57,42	87,72	96,49	98,07	99,91	99,98	99,99
$R^*_{.80}$	5,15	15,09	73,47	97,94	99,80	100,00	100,00	100,00	100,00
R^+	5,31	6,17	14,79	33,25	50,46	97,08	99,77	99,96	99,99
R	5,52	6,41	15,22	34,12	51,27	97,27	99,80	99,97	99,99
W	5,10	6,09	13,09	21,40	29,85	36,67	60,37	71,11	72,85
T	5,01	9,47	39,36	64,87	82,23	90,16	99,35	99,93	99,94
C_6	5,03	11,85	54,85	93,03	98,34	100,00	100,00	100,00	100,00
M_0	5,20	6,91	22,66	56,14	77,68	99,98	100,00	100,00	100,00
$M_{.10}$	5,12	7,06	23,54	58,38	79,85	99,99	100,00	100,00	100,00
$M_{.20}$	5,05	7,10	24,64	60,96	82,11	100,00	100,00	100,00	100,00
$M_{.25}$	5,08	7,09	25,49	62,50	83,29	100,00	100,00	100,00	100,00
$W_{.70}$	4,86	15,78	77,18	98,03	99,83	100,00	100,00	100,00	100,00
$W_{.80}$	4,89	16,53	80,80	99,12	99,98	100,00	100,00	100,00	100,00

Tabla 16: Potencias para los nueve casos de la DLG y $n = 250$.

CASO =	1	2	3	4	5	6	7	8	9
$R_{.60}$	5,04	14,08	67,21	92,36	98,23	98,25	99,90	100,00	100,00
$R_{.80}$	4,66	18,85	84,36	99,55	99,96	100,00	100,00	100,00	100,00
$R^*_{.60}$	5,03	13,66	66,93	93,16	98,67	99,20	99,99	100,00	100,00
$R^*_{.80}$	4,71	17,78	82,17	99,39	99,96	100,00	100,00	100,00	100,00
R^+	4,85	6,40	16,59	38,83	57,82	99,07	99,97	100,00	100,00
R	4,72	6,29	16,37	38,48	57,58	99,06	99,97	100,00	100,00
W	5,11	6,43	15,69	25,76	36,41	44,10	70,09	80,04	81,97
T	4,90	10,48	47,40	74,31	89,63	95,42	99,88	99,97	99,99
C_6	4,85	12,78	60,36	95,56	99,14	100,00	100,00	100,00	100,00
M_0	4,96	7,05	26,71	64,76	84,84	100,00	100,00	100,00	100,00
$M_{.10}$	4,92	7,10	28,23	67,27	86,68	100,00	100,00	100,00	100,00
$M_{.20}$	4,97	7,23	29,54	70,07	88,57	100,00	100,00	100,00	100,00
$M_{.25}$	4,97	7,25	30,23	71,46	89,60	100,00	100,00	100,00	100,00
$W_{.70}$	4,75	18,90	85,47	99,47	99,98	100,00	100,00	100,00	100,00
$W_{.80}$	4,66	19,98	88,95	99,88	100,00	100,00	100,00	100,00	100,00

Recientemente, Baklizi (2007) sugiere el uso de la longitud de la racha más larga en la cola derecha de la sucesión dicotomizada, y en su trabajo de 2008 propone eliminar la primera etapa de la prueba de Modarres & Gastwirth (1998), con lo cual se logra incrementar su potencia. En la tabla 18 (Apéndice A), se incluyen los valores de las potencias de las pruebas propuestas en los dos artículos mencionados anteriormente y las potencias de las pruebas propuestas. Comparando con las pruebas L^*, $L^*_{n,0,8}$ y L propuestas en Baklizi (2007), se observa que:

- Para $n = 20$, las pruebas propuestas tienen un mejor desempeño en algunos casos.

- Para $n = 30$, las pruebas propuestas tienen las mayores potencias en todos los casos.

Tabla 17: Potencias para los nueve casos de la DLG y $n = 500$.

CASO =	1	2	3	4	5	6	7	8	9
$R_{.60}$	4,74	23,81	92,40	99,74	99,99	99,92	100,00	100,00	100,00
$R_{.80}$	4,86	34,09	98,96	100,00	100,00	100,00	100,00	100,00	100,00
$R^*_{.60}$	4,78	23,60	92,60	99,82	100,00	99,98	100,00	100,00	100,00
$R^*_{.80}$	5,02	32,94	98,76	100,00	100,00	100,00	100,00	100,00	100,00
R^+	5,12	6,72	24,41	61,40	83,08	100,00	100,00	100,00	100,00
R	5,10	6,81	24,48	61,70	83,24	100,00	100,00	100,00	100,00
W	4,95	8,00	26,54	46,09	62,11	72,76	93,85	97,56	98,16
T	4,91	16,66	76,19	95,97	99,53	99,90	100,00	100,00	100,00
C_6	4,91	16,39	74,95	99,09	99,93	100,00	100,00	100,00	100,00
M_0	5,25	7,68	41,58	89,33	98,40	100,00	100,00	100,00	100,00
$M_{.10}$	5,21	7,73	43,63	91,16	98,85	100,00	100,00	100,00	100,00
$M_{.20}$	5,16	7,95	45,78	92,77	99,18	100,00	100,00	100,00	100,00
$M_{.25}$	5,16	8,15	46,91	93,45	99,35	100,00	100,00	100,00	100,00
$W_{.70}$	5,06	33,94	99,22	100,00	100,00	100,00	100,00	100,00	100,00
$W_{.80}$	4,98	36,41	99,58	100,00	100,00	100,00	100,00	100,00	100,00

- Para $n = 50$ y $n = 100$, las pruebas propuestas superan las pruebas de Baklizi en los casos 2 al 5, y en los casos 6 al 9 coincide la potencia de las pruebas comparadas.

[Recibido: marzo de 2009 — Aceptado: octubre de 2010]

Referencias

Baklizi, A. (2003), 'A Conditional Distribution Runs Test for Symmetry', *Journal of Nonparametric Statistics* **15**(6), 713–718.

Baklizi, A. (2007), 'Testing Symmetry Using A Trimmed Longest Run Statistic', *Australian & New Zealand Journal of Statistics* **49**(4), 339–347.

Baklizi, A. (2008), 'Improving the Power of the Hybrid Test of Symmetry', *International Journal of Contemporary Mathematical Sciences* **3**(10), 497–499.

Butler, C. (1969), 'A test for symmetry using the sample distribution function', *The Annals of Mathematical Statistics* **40**, 2209–2210.

Castillo, O. (1993), Una prueba de rachas para simetría, Tesis de maestría, Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia.

Cheng, W. H. & Balakrishnan, N. (2004), 'A Modified Sign Test for Symmetry', *Communications in Statistics Simulation and Computation* **33**, 703–709.

Cohen, J. & Menjoge, S. (1988), 'One-sample runs test of symmetry', *Journal of Statistical Planning and Inference* **18**(1), 93–100.

Corzo, J. (1989), Verallgemeinerte Runtests für Lage- und Skalenalternativen, Tesis de doctorado, Fachbereich Statistik, Universität Dortmund, Alemani.

Corzo, J. & Rojas, A. (1999), 'Una prueba basada en rachas para simetría alrededor de una mediana específica', *Revista Colombiana de Estadística* **22**(2), 39–53.

Gibbons, J. D. & Chakraborti, S. (1992), *Nonparametric Statistical Inference*, CRC Press, New York.

Hettmansperger, T. (1984), *Statistical Inference Based on Ranks*, John Wiley & Sons, New York.

Hill, D. & Rao, P. (1977), 'Test of Symmetry Based on Cramér-Von Mises', *Biometrika* **64**, 489–494.

Lehmann, E. L. (1986), *Testing Statistical Hyphoteses*, second edn, John Wiley.

McWilliams, P. (1990), 'A Distribution-Free Test for Symmetry Based on a Runs Statistic', *Journal of the American Statistical Association* **85**(412), 1130–1133.

Mira, A. (1999), 'Distribution-Free Test for Symmetry Based on Bonferroni's Measure', *Journal of Applied Statistics* **26**(8), 959–972.

Modarres, R. & Gastwirth (1996), 'A modified runs tes for symmetry', *Statistics and probability* **25**(5), 575–585.

Modarres, R. & Gastwirth, J. L. (1998), 'Hybrid Test for the Hypothesis of Symmetry', *Journal of Applied Statistics* **25**(6), 777–783.

Randles, R. & Wolfe, D. (1979), *Introduction to the Theory of Nonparametric Statistics*, John Wiley & Sons, New York.

Rothman, E. & Woodroofe, M. (1972), 'A Cramér-Von Mises Type Statistic for Testing Symmetry', *The Annals of Mathematical Statistics* **43**, 2035–2038.

Tajuddin, I. (1994), 'Distribution-Free Test for Symmetry Based on the Wilcoxon Two-Sample Test', *Journal of Applied Statistics* **21**(5), 409–414.

Thas, O., Rayner, J. C. W. & Best, D. J. (2005), 'Tests for Symmetry Based on the One-Sample Wilcoxon Signed Rank Statistic', *Communications in Statistics - Simulation and Computation* **34**, 957–973.

Apéndice A. Desempeño de las pruebas propuestas con las pruebas de Baklizi (2007, 2008)

TABLA 18: Potencia de las pruebas recientemente sugeridas por Baklizi y de las pruebas propuestas.

Caso	n	L^*	$L^*_{n,0.8}$	L	R_{2008}	R^*_{80}	R_{80}
1	20	5	5,5	5,5	4,5	4,8	4,9
1	30	5,1	5,4	4,8	4,5	5,2	5,2
1	50	4,9	5	5,8	5	4,9	4,9
1	100	5,5	5,4	4,9	5,1	5,1	5,1
2	20	5,7	6,2	4,6	5,8	5,6	5,4
2	30	5	6,3	5,8	6,2	5,9	6,0
2	50	5,8	6,2	5,3	7,5	7,0	7,1
2	100	5,3	6,6	6,5	10,6	9,7	10,5
3	20	5,7	8,8	7	11,4	8,3	8,1
3	30	6,9	12,8	8,8	15,6	12,9	13,2
3	50	7,5	14,5	14,1	25,3	20,4	22,1
3	100	12,2	22,8	23,3	48,6	41,5	45,6
4	20	7,2	13,8	8,7	18,7	13,7	12,7
4	30	10	22,9	14,2	28,7	24,8	25,7
4	50	14,4	30,3	31,4	49,9	42,2	45,9
4	100	33,3	57,2	59,5	81,8	76,9	80,8
5	20	9,5	16,5	12	25,4	17,9	15,9
5	30	13,1	33,3	21,9	40,4	34,3	36,2
5	50	23	45,7	45,5	66,8	58,0	62,7
5	100	55,6	76,9	78,4	93,9	90,3	92,4
6	20	13,6	26,7	19,1	42	28,2	24,8
6	30	25,6	64,9	46,9	65,3	69,6	71,1
6	50	64,2	90,8	91,4	89,4	91,5	88,9
6	100	99,5	100	100	99,4	99,7	99,0
7	20	22,6	33	31,4	59,3	35,4	30,7
7	30	44,4	82,3	68,9	85,4	84,0	86,9
7	50	86	97,8	98	98	97,8	97,3
7	100	100	100	100	100	100,0	100,0
8	20	29,5	36,6	38,9	66,6	32,6	33,2
8	30	55,5	87,6	79	91,3	88,4	91,4
8	50	93	99,1	99,1	99,3	98,9	98,8
8	100	100	100	100	100	100,0	100,0
9	20	32,2	35,9	42,8	69,2	37,5	34,0
9	30	60,8	89,8	82,1	92,4	90,2	93,5
9	50	95,3	99,3	99,5	99,5	99,3	99,2
9	100	100	100	100	100	100,0	100,0

Revista Colombiana de Estadística
Diciembre 2010, volumen 33, no. 2, pp. 273 a 293

Análisis de correspondencias a partir de una muestra probabilística

Analysis of Correspondence from a Probabilistic Sample

Javier Ramírez[a], Guillermo Martínez[b]

Departamento de Matemáticas y Estadística, Facultad de Ciencias Básicas e Ingenierías, Universidad de Córdoba, Montería, Colombia

Resumen

A partir del análisis de correspondencias clásico aplicado a las tablas denominadas de correspondencias, se desarrolla la teoría para dicho análisis a partir de una muestra probabilística. El enfoque de esta teoría se encamina a la estimación de los valores y vectores propios asociados a las matrices por diagonalizar, ya sea en el análisis simple o en el múltiple, para luego establecer las estimaciones de los vectores propios que conducen a los ejes factoriales, permitiéndose una representación gráfica para mejorar la interpretación en el análisis. Se realizan además estimaciones de las medidas de calidad asociadas a la representación, como son: inercia, contribuciones y cosenos cuadrados.

Palabras clave: análisis de correspondencias, muestreo probabilístico, *Bootstrap*, *Jackknife*.

Abstract

From the classic analysis of correspondences applied to the denominated tables of correspondences, the theory for this analysis from a probabilistic sample is developed. The approach of this theory directs to the estimation of eigenvalues and eigenvectors associated to the matrices to be diagonalized, either in a simple analysis or in the multiple one, to establish estimations of the eigenvectors that lead to the factorial axes, allowing a graphical representation to improve performance in the analysis. Estimates of quality measures associated to the representation are made, such as inertia, contributions and squares cosines.

Key words: Correspondence analysis, Probability sampling, *Bootstrap*, *Jackknife*.

[a]Profesor asistente. E-mail: javierramirez@sinu.unicordoba.edu.co
[b]Profesor asociado. E-mail: gmartinez@sinu.unicordoba.edu.co

1. Introducción

Una de las técnicas de los métodos factoriales que analiza la asociación entre dos o más variables categóricas es el denominado análisis de correspondencias. A través del análisis de correspondencias simples (ACS) aplicado a las tablas de contingencia, se construyen las representaciones de las asociaciones entre filas y columnas de estas tablas, basados en la distancia χ^2. Se trata de tablas de efectivos obtenidos cruzando las modalidades de dos variables cualitativas definidas sobre una misma población de n individuos Escofier & Pagés (1992). Por otra parte, con el análisis de correspondencias múltiples el cual es una extensión del dominio de aplicación del ACS, se describen grandes tablas de variables categóricas, representando las categorías de las variables como puntos en un espacio de pocas dimensiones Clausen (1998).

Ahora bien, un requisito fundamental para este tipo de análisis es la obtención de los valores y vectores propios, y por ende las coordenadas sobre los ejes factoriales que permiten la interpretación de las asociaciones entre las variables categóricas. En este trabajo se presenta una metodología de estimación de los valores y vectores propios de las matrices por diagonalizar en los análisis de correspondencias simples y múltiples, a partir de una muestra probabilística. Con ellos se obtienen los ejes, las coordenadas factoriales y las relaciones de transición entre los espacios, la estimación de la inercia, las contribuciones y los cosenos cuadrados. Lo que se tiene entonces es una complementación entre los diseños de muestreo probabilístico y el análisis de correspondencias, lo que permite describir no solo el comportamiento o la asociación entre variables categóricas obtenidas a través de una muestra probabilística tomada de alguna población bajo estudio, sino también inferir acerca de dicho comportamiento y el grado de asociación entre las variables de estudio, siguiendo la metodología dada por Martínez (1998).

En la sección 2 se presenta la propuesta de estimación de los elementos de base en el análisis de correspondencias simples y múltiples, al igual que las demás medidas que intervienen en el análisis; por otra parte, se propone el cálculo de la varianza de los valores propios estimados mediante las técnica *Jackknife* y *Bootstrap*, donde en la sección 3 se muestra un ejemplo de aplicación, en el que se comparan estas dos técnicas y se llega a discusiones importantes. Por último en la sección 4, se presentan los métodos computacionales utilizados, y en la sección 5 se dan a conocer las conclusiones del trabajo.

2. Resultados y discusión

2.1. Análisis general

El procedimiento para efectuar un análisis factorial para métricas y matrices de peso cualesquiera es diagonalizar la matriz $A = X'LXM$, donde M corresponde a la métrica y L a la matriz, de masa o peso, para encontrar los q valores propios más grandes de dicha matriz y a partir de estos obtener las coordenadas factoriales necesarias para llevar a cabo el análisis.

En general, si se desea efectuar un procedimiento de análisis factorial como el de correspondencias a partir de una muestra probabilística el interés se centra en la diagonalización de la matriz estimada, $\widehat{A} = X'\widehat{L}X\widehat{M}$ a partir del diseño muestral empleado, para así obtener los valores propios estimados de esta matriz (Lebart, Morineau & Piron 2000).

El polinomio característico dado por la ecuación (1)

$$|A - \lambda I| = 0 \tag{1}$$

resulta ser de la forma

$$p(\lambda) = (-1)^r \left(\lambda^r + b_{r-1}\lambda^{r-1} + \cdots + b_1\lambda + b_0\right) = 0 \tag{2}$$

el cual es posible estimarlo con la expresión

$$p\left(\widehat{\lambda}\right) = \left|\widehat{A} - \widehat{\lambda}I\right| = (-1)^r \left(\widehat{\lambda}^r + b_{r-1}\widehat{\lambda}^{r-1} + \cdots + b_1\widehat{\lambda} + b_0\right) = 0 \tag{3}$$

donde r es el número de modalidades de estudio y $b_{r-1}, \ldots, b_0$ son valores numéricos que se pueden escribir como funciones de totales poblacionales estimados. De esta manera se puede establecer que los valores estimados de λ son de la forma

$$\widehat{\lambda} = f\left(\widehat{t}_1, \widehat{t}_2, \ldots, \widehat{t}_r\right) \tag{4}$$

donde $\widehat{t}_i$ son los estimadores de los totales poblacionales que conforman la matriz A.

Así, obtenidos estos valores, es posible efectuar por completo el análisis factorial a partir de la información de la muestra, dado que con los valores y vectores propios estimados se pueden construir las demás componentes del análisis (Lebart et al. 2000).

2.2. Análisis de correspondencias simples a partir de una muestra probabilística

Dada la población $U = \{1, \ldots, N\}$, suponga que a los elementos de U se les miden dos variables, digamos Z_1 y Z_2 con p_1 y p_2 modalidades, respectivamente. La matriz de datos, resultado de la medición de las variables sobre los N individuos, es como sigue

$$Z = \left[Z_1 \vdots Z_2\right]$$

donde

$$Z_m = \begin{bmatrix} z_{11} & \cdots & z_{1i} & \cdots & z_{1p_m} \\ \vdots & & \vdots & & \vdots \\ z_{l1} & \cdots & z_{li} & \cdots & z_{lp_m} \\ \vdots & & \vdots & & \vdots \\ z_{N1} & \cdots & z_{Ni} & \cdots & z_{Np_m} \end{bmatrix}$$

con

$$z_{li} = \begin{cases} 1, & \text{si el sujeto } l \text{ seleccionó la modalidad } i \text{ de la pregunta } Z_m \\ 0, & \text{si el sujeto } l \text{ no seleccionó la modalidad } i \text{ de la pregunta } Z_m \end{cases}$$

así, para $m = 1, 2$, $z_{li} = 1$ ó $z_{li} = 0$ para $l = 1, 2, \ldots, N$ e $i = 1, 2, \ldots, p_m$.

2.2.1. Tabla de contingencia

La tabla de contingencia a partir de las matrices Z_1 y Z_2 es

$$\mathbf{C} = \mathbf{Z}_1^T \mathbf{Z}_2$$

es decir

$$\mathbf{C} = \begin{bmatrix} k_{11} & \ldots & k_{1j} & \ldots & k_{1p_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ k_{i1} & \ldots & k_{ij} & \ldots & k_{ip_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ k_{p_1 1} & \ldots & k_{p_1 j} & \ldots & k_{p_1 p_2} \end{bmatrix}$$

donde

$$k_{ij} = \sum_{l=1}^{N} z_{ijl} \tag{5}$$

corresponde a un total de un dominio, en este caso el total de individuos que respondieron la modalidad i de la pregunta Z_1 y la modalidad j de la pregunta Z_2 simultáneamente con:

$$z_{ijl} = z_{il} \times z_{jl} = \begin{cases} 1, & \text{si} \quad z_{il} = 1 \quad \text{y} \quad z_{jl} = 1 \\ 0, & \text{si} \quad z_{il} = 0 \quad \text{ó} \quad z_{jl} = 0 \end{cases}$$

2.2.2. Tabla de contingencia estimada

Basados en una muestra probabilística S obtenida a través del diseño $p(\cdot)$ (ver Särndal, Swensson & Wretman 1992), con probabilidades de inclusión π_l para los elementos de U, podemos estimar cada total en la ecuación (5), a través de un π estimador de la siguiente forma:

$$\widehat{k}_{ij\pi} = \sum_{l \in s} \frac{z_{ijl}}{\pi_l} \tag{6}$$

donde los π_l son las probabilidades de inclusión de cada individuo; así, la matriz de correspondencias estimadas es:

$$\widehat{C} = Z_n^T \Pi^{-1} Z_n \tag{7}$$

donde n corresponde a los individuos de la muestra, ahora la matriz de código binario asociado a las dos variables es

$$Z_n \begin{bmatrix} z_{11} & \cdots & z_{1i} & \cdots & z_{1p_m} \\ \vdots & & \vdots & & \vdots \\ z_{l1} & \cdots & z_{li} & \cdots & z_{lp_m} \\ \vdots & & \vdots & & \vdots \\ z_{n1} & \cdots & z_{ni} & \cdots & z_{np_m} \end{bmatrix}$$

para $m = 1, 2$ con matriz de probabilidades de inclusión

$$\Pi = \begin{bmatrix} \pi_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & \pi_l & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \pi_{n_s} \end{bmatrix}$$

entonces la matriz de correspondencias estimada (7), tendrá la siguiente forma

$$\widehat{C} = \begin{bmatrix} \widehat{k}_{11\pi} & \ldots & \widehat{k}_{1j\pi} & \ldots & \widehat{k}_{1p_2\pi} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{k}_{i1\pi} & \ldots & \widehat{k}_{ij\pi} & \ldots & \widehat{k}_{ip_2\pi} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{k}_{p_1 1\pi} & \ldots & \widehat{k}_{p_1 j\pi} & \ldots & \widehat{k}_{p_1 p_2\pi} \end{bmatrix}$$

con $\widehat{k}_{ij\pi}$ representando la estimación de cada total de la matriz de correspondencias, definida en la ecuación (6), donde

$$\widehat{k}_{\pi} = \sum_{l \in s} \frac{1}{\pi_l} = \widehat{N} \tag{8}$$

2.2.3. Criterio por maximizar y matriz por diagonalizar

En el espacio de las columnas $\mathbb{R}^{p_2}$, el interés es maximizar la suma ponderada de los cuadrados de las proyecciones sobre el eje, es decir, dada la muestra s, maximizar la ecuación

$$\text{Máx}_{\widehat{u}} \left\{ \sum_i \widehat{f}_i \widehat{d}^2(i, O) \right\} \tag{9}$$

lo que es equivalente a maximizar la expresión

$$\widehat{u}' \widehat{D}_{p2}^{-1} \widehat{F}' \widehat{D}_{p1}^{-1} \widehat{F} \widehat{D}_{p2}^{-1} \widehat{u} \tag{10}$$

con la restricción

$$\widehat{u}' \widehat{D}_{p2}^{-1} \widehat{u} = 1 \tag{11}$$

donde $\widehat{u}$ es el vector propio de la matriz estimada

$$\widehat{S} = \widehat{F}'\widehat{D}_{p1}^{-1}\widehat{F}\widehat{D}_{p2}^{-1} \tag{12}$$

asociado al valor propio estimado $\widehat{\lambda}$ más grande diferente de 1. La matriz $\widehat{F}$ corresponde a la matriz de frecuencias relativas estimadas de término general $\widehat{f}_{ij}$, es decir

$$\widehat{F} = \{\widehat{f}_{ij}\} \tag{13}$$

donde

$$\widehat{f}_{ij} = \frac{\widehat{k}_{ij\pi}}{\widehat{k}} \tag{14}$$

y las matrices de márgenes filas $\widehat{f}_{i\cdot}$ y columna $\widehat{f}_{\cdot j}$ están dadas por:

$$\widehat{D}_{p1} = \text{diag}\{\widehat{f}_{i\cdot}\} \tag{15}$$

para $i = 1, 2, \ldots, p_1$ y $j = 1, 2, \ldots, p_2$

$$\widehat{D}_{p2} = \text{diag}\{\widehat{f}_{\cdot j}\} \tag{16}$$

respectivamente, donde:

$$\widehat{f}_{i\cdot} = \sum_{j=1}^{p_2} \frac{\widehat{k}_{i\cdot\pi}}{\widehat{k}} \qquad \widehat{f}_{\cdot j} = \sum_{i=1}^{p_1} \frac{\widehat{k}_{\cdot j\pi}}{\widehat{k}}$$

con

$$\widehat{k}_{i\cdot\pi} = \sum_{j=1}^{p_2} \widehat{k}_{ij\pi} \qquad \widehat{k}_{\cdot j\pi} = \sum_{i=1}^{p_1} \widehat{k}_{ij\pi}$$

De esta forma se tiene en $\mathbb{R}^{p_2}$ que la métrica $\widehat{M}$ es $\widehat{D}_{p2}^{-1}$ y la matriz de pesos $\widehat{N}$ es $\widehat{D}_{p1}^{-1}$; así, la matriz por diagonalizar es:

$$\widehat{S} = \{\widehat{S}_{jj'}\} \tag{17}$$

de término general

$$\widehat{S}_{jj'} = \sum_{i=1}^{p1} \frac{\widehat{f}_{ij}}{\widehat{f}_{i\cdot}} \frac{\widehat{f}_{ij'}}{\widehat{f}_{\cdot j'}} = \sum_{i=1}^{p1} \frac{\widehat{k}_{ij}}{\widehat{k}_{i\cdot}} \frac{\widehat{k}_{ij'}}{\widehat{k}_{\cdot j'}} \quad \text{para} \quad j, j' = 1, 2, \ldots, p_2 \tag{18}$$

De la misma forma, en el espacio de las filas estimadas $\mathbb{R}^{p1}$, se busca maximizar la cantidad

$$\widehat{v}'\widehat{D}_{p1}^{-1}\widehat{F}\widehat{D}_{p2}^{-1}\widehat{F}'\widehat{D}_{p1}^{-1}\widehat{v} \tag{19}$$

con la restricción

$$\widehat{v}'\widehat{D}_{p1}^{-1}\widehat{v} = 1 \tag{20}$$

donde $\widehat{v}$ es el vector propio asociado a un valor propio de la siguiente matriz:

$$\widehat{T} = \widehat{F}\widehat{D}_{p2}^{-1}\widehat{F}'\widehat{D}_{p1}^{-1} \tag{21}$$

Así, de acuerdo con Lebart et al. (2000), la métrica $\widehat{M}$ en $\mathbb{R}^{p1}$ es $\widehat{D}_{p1}^{-1}$ y la matriz de pesos $\widehat{N}$ es $\widehat{D}_{p2}^{-1}$. Para la diagonalización de la matriz $\widehat{S}$ a partir de una muestra probabilística, definida en (12), podemos obtener un estimador de λ según la metodología definida en Särndal et al. (1992), sección 5. Entonces, siguiendo a Särndal et al. (1992), λ se puede escribir como una función de totales estimados de la forma

$$\widehat{\lambda} = f\left(\widehat{k}_{11\pi}, \widehat{k}_{12\pi}, \ldots, \widehat{k}_{1p2\pi}, \widehat{k}_{21\pi}, \widehat{k}_{22\pi}, \ldots, \widehat{k}_{2p2\pi}, \ldots, \widehat{k}_{p_1 1\pi}, \widehat{k}_{p_1 2\pi}, \ldots, \widehat{k}_{p_1 p_2 \pi}\right)$$

El estimador anterior se encuentra resolviendo el polinomio característico

$$\left|\widehat{S} - \widehat{\lambda} I_{p2}\right| = (-1)^{p2}\left(\widehat{\lambda}^{p2} + \widehat{b}_{p-1}\widehat{\lambda}^{p2-1} + \cdots + \widehat{b}_1\widehat{\lambda} + \widehat{b}_0\right) = 0 \tag{22}$$

donde $\widehat{k}_{ij\pi}$ es definido en la ecuación (6) y

$$\widehat{b}_r = f\left(\widehat{k}_{11}, \widehat{k}_{12}, \ldots, \widehat{k}_{1p2}, \widehat{k}_{21}, \widehat{k}_{22}, \ldots, \widehat{k}_{2p2}, \ldots, \widehat{k}_{p_1 1}, \widehat{k}_{p_1 2}, \ldots, \widehat{k}_{p_1 p_2}\right) \tag{23}$$

2.3. Estimación de la varianza (ACS)

De acuerdo con Särndal et al. (1992, cap. 5) el π estimador propuesto

$$\widehat{\lambda} = f\left(\widehat{k}_{11\pi}, \widehat{k}_{12\pi}, \ldots, \widehat{k}_{1p2\pi}, \widehat{k}_{21\pi}, \widehat{k}_{22\pi}, \ldots, \widehat{k}_{2p2\pi}, \ldots, \widehat{k}_{p_1 1\pi}, \widehat{k}_{p_1 2\pi}, \ldots, \widehat{k}_{p_1 p_2 \pi}\right)$$

es un estimador aproximadamente insesgado para λ.

Esto se puede demostrar a través de la técnica de linealización de primer orden de la serie de Taylor alrededor del punto k_{ij}, para $i = 1, 2, \ldots, p_1$ y $j = 1, 2, \ldots, p_2$; procediendo como en Martínez (1998), se tiene por aproximación de primer orden de Taylor una aproximación al estimador $\widehat{\lambda}$ por un pseudo estimador $\widehat{\lambda}_0$ de la forma

$$\widehat{\lambda} \doteq \widehat{\lambda}_0 = \lambda + \sum_{i,j} a_{ij}\left(\widehat{k}_{ij\pi} - k_{ij}\right) \tag{24}$$

donde

$$a_{ij} = \left.\frac{\partial f}{\partial \widehat{k}_{ij\pi}}\right|_{\left\{\widehat{k}_{ij\pi}\right\} = \{k_{ij}\}_{i=1,\ldots,p_1;\, i=1,\ldots,p_2}}$$

Como $\widehat{k}_{ij\pi}$ es el total de un dominio de estudio, entonces

$$E\left(\widehat{k}_{ij\pi}\right) = E\left(\sum_{l\in s}\frac{z_{ijl}}{\pi_l}\right) = \sum_{l\in U}\frac{z_{ijl}}{\pi_l}\pi_l = \sum_{l\in U} z_{ijl} = k_{ij} \tag{25}$$

luego, en vez de (24), se tiene que

$$\begin{aligned} E\left(\widehat{\lambda}\right) &\doteq E\left(\widehat{\lambda}_0\right) \\ &\doteq \lambda + \sum_{i,j} a_{ij}(E(\widehat{k}_{ij\pi}) - k_{ij}) \\ &\doteq \lambda + \sum_{i,j} a_{ij}(k_{ij} - k_{ij}) \\ &\doteq \lambda \end{aligned}$$

Por tanto, $E\left(\widehat{\lambda}\right) \approx E\left(\widehat{\lambda}\right)$.

Nuestro objetivo ahora es obtener una medida de la calidad de la estimación de λ. Entonces, siguiendo el método de funciones de totales dado en Särndal et al. (1992), definimos $u_l = \sum_{i,j} a_{ij} z_{ijl}$ y $\check{u}_l = u_l/\pi_l$ donde a_{ij} está dado en (24); así, la varianza y un estimador de la varianza Horvith-Thompson para funciones de totales es

$$AV\left(\widehat{\lambda}\right) = \sum_{l\in U}\sum_{l'\in U} \Delta_{ll'}\check{u}_l\check{u}_{l'} \tag{26}$$

Dado que las cantidades u_l dependen de valores desconocidos, el estimador de la aproximación de la varianza de Horvitz-Thompson $AV\left(\widehat{\lambda}\right)$ bajo la técnica es

$$\widehat{V}\left(\widehat{\lambda}\right) = \sum_{l\in s}\sum_{l'\in s} \check{\Delta}_{ll'}\frac{\widehat{u}_l}{\pi_l}\frac{\widehat{u}_{l'}}{\pi_{l'}} \tag{27}$$

donde

$$\widehat{u}_l = \sum_{i,j} \widehat{a}_{ij} z_{ijl}$$

y los coeficientes $\widehat{a}_{ij}$ se obtienen como

$$\widehat{a}_{ij} = \frac{\partial f}{\partial \widehat{k}_{ij\pi}}. \tag{28}$$

2.3.1. El estimador *Jackknife*

Dada la cantidad de parámetros por calcular para estimar la varianza de Horvitz-Thompson, se estudian los métodos de *Jackknife* y *Bootstrap* para la estimación de la varianza, los cuales son usados para este tipo de situaciones dada su simplicidad de cálculo y los supuestos para su aplicación; por tanto, la estimación para la varianza de $\widehat{\lambda}$, según el método *Jackknife* presentado en Wolter (1985), se define para este estimador como

$$v_{jk} = \frac{n-1}{n}\sum_{l=1}^{n}\left(\widehat{\lambda}_{n-1,l} - \frac{1}{n}\sum_{l'=1}^{n}\widehat{\lambda}_{n-1,l'}\right)^2 \tag{29}$$

Luego este estimador se conoce como el estimador *Jackknife* (delete-1) de $V\left(\widehat{\lambda}_n\right)$, donde

$$\widehat{\lambda}_{n-1,l} = f\left(\widehat{k}_{11\pi_{(n-1)}}, \widehat{k}_{12\pi_{(n-1)}}, \ldots, \widehat{k}_{p_1 1\pi_{(n-1)}}, \widehat{k}_{p_1 2\pi_{(n-1)}}, \ldots, \widehat{k}_{p_1 p_2 \pi_{(n-1)}}\right) \quad (30)$$

y

$$\widehat{k}_{ij\pi(n-1,l)} = \sum_{l' \in S-\{l\}} \frac{z_{ijl'}}{\pi_{l'}} \quad (31)$$

Es decir, $\widehat{\lambda}_{n-1,l}$ es el estimador del valor propio correspondiente, basado en la muestra de tamaño $n-1$ que resulta luego de eliminar el individuo l-ésimo de la muestra, y $\widehat{k}_{ij\pi(n-1,l)}$ es la estimación de un dominio eliminando la misma observación.

2.3.2. El estimador *Bootstrap*

Teniendo en cuenta la importancia de utilizar el método de remuestreo *Bootstrap* para estimar valores propios, Milan & Whittaker (1995) realizan una aplicación del *Bootstrap* paramétrico a modelos que incorporan valores singulares, donde se desarrollan discusiones importantes sobre el efecto de la variación de muestreo en las estimaciones.

Para calcular los valores propios mediante el método el remuestreo *Bootstrap*, se realizan los siguientes pasos:

1. Dada la muestra de tamaño n, calcular $\widehat{\lambda}$. La distribución de esta muestra se considera equivalente a la distribución de la población y $\widehat{\lambda}$ es el estimador muestral del parámetro poblacional λ.

2. Generar B muestras *Bootstrap* de tamaño n mediante muestreo con remplazo de la muestra original, y calcular los correspondientes valores $\widehat{\lambda}^{*1}, \widehat{\lambda}^{*2}, \ldots, \widehat{\lambda}^{*B}$ para cada una de las B muestras *Bootstrap*.

3. Estimar el error estándar del parámetro estimado $\widehat{\lambda}$ calculando la desviación estándar de las B réplicas *Bootstrap*.

Así, obtenemos que el error estándar es

$$\sigma^*_{\lambda} = \sqrt{\frac{\sum_{b=1}^{B}\left(\lambda^{b^*} - \bar{\widehat{\lambda}}^*\right)^2}{(B-1)}} = \sigma_{BOOT} \quad (32)$$

donde

$$\bar{\widehat{\lambda}}^* = \frac{1}{B}\sum_{b=1}^{B}\lambda^{b^*} \quad (33)$$

corresponde al promedio de los valores propios calculados en cada remuestra.

2.4. Estimación de las coordenadas factoriales

Siguiendo la metodología definida en Lebart et al. (2000), para análisis factorial ponderado, es decir el caso general, podemos obtener los cosenos cuadrados, inercia, para el análisis de correspondencias a partir de una muestra probabilística. Las coordenadas factoriales estimadas, analizando los perfiles fila $\frac{\widehat{f}_{ij}}{\widehat{f}_{i.}}$ en el espacio de las columnas $\mathbb{R}^{p2}$ y los perfiles columna $\frac{\widehat{f}_{ij}}{\widehat{f}_{.j}}$ en el espacio de las filas $\mathbb{R}^{p1}$ vendrán dados respectivamente por:

$$\widehat{\psi}_{\alpha} = \widehat{D}_{p1}^{-1}\widehat{F}\widehat{D}_{p2}^{-1}\widehat{u}_{\alpha} \tag{34}$$

$$\widehat{\varphi}_{\alpha} = \widehat{D}_{p2}^{-1}\widehat{F}'\widehat{D}_{p2}^{-1}\widehat{v}_{\alpha} \tag{35}$$

con términos generales

$$\widehat{\psi}_{\alpha i} = \sum_{j=1}^{p_2} \frac{\widehat{f}_{ij}}{\widehat{f}_{i.}\widehat{f}_{.j}}\widehat{\mathbf{u}}_{\alpha j} \tag{36}$$

$$\widehat{\varphi}_{\alpha j} = \sum_{i=1}^{p_1} \frac{\widehat{f}_{ij}}{\widehat{f}_{i.}\widehat{f}_{.j}}\widehat{\mathbf{v}}_{\alpha i} \tag{37}$$

respectivamente. Ahora se presentan las relaciones entre los espacios, fruto de la estimación de los vectores propios asociados a los valores propios estimados.

$$\widehat{\mathbf{v}}_{\alpha} = \frac{1}{\sqrt{\lambda_{\alpha}}}\widehat{\mathbf{F}}\widehat{\mathbf{D}}_{p_2}^{-1}\widehat{\mathbf{u}}_{\alpha}$$

$$\widehat{\mathbf{u}}_{\alpha} = \frac{1}{\sqrt{\lambda_{\alpha}}}\widehat{\mathbf{F}}'\widehat{\mathbf{D}}_{p_1}^{-1}\widehat{\mathbf{v}}_{\alpha}$$

2.4.1. Relaciones de transición

Las relaciones fundamentales existentes entre los puntos fila y puntos columna sobre el eje α son llamadas relaciones de transición, y se calculan así:

$$\widehat{\psi}_{\alpha i} = \frac{1}{\sqrt{\widehat{\lambda}_{\alpha}}}\sum_{j=1}^{p_2} \frac{\widehat{f}_{ij}}{\widehat{f}_{i.}}\widehat{\varphi}_{\alpha j} \tag{38}$$

$$\widehat{\varphi}_{\alpha j} = \frac{1}{\sqrt{\widehat{\lambda}_{\alpha}}}\sum_{i=1}^{p_1} \frac{\widehat{f}_{ij}}{\widehat{f}_{.j}}\widehat{\psi}_{\alpha i} \tag{39}$$

donde $\frac{1}{\sqrt{\widehat{\lambda}_{\alpha}}}$ es el coeficiente de dilatación estimado.

2.4.2. Estimación de la inercia, contribuciones y cosenos cuadrados

Siguiendo a Lebart et al. (2000), se presentan a continuación los estimadores muestrales de la inercia, las contribuciones a los ejes factoriales y los cosenos cuadrados estimados en el análisis de correspondencias.

Inercia. La inercia total estimada para el caso de análisis de correspondencias simple es

$$\widehat{I} = \sum_{\alpha=1}^{p-1} \widehat{\lambda}_\alpha \tag{40}$$

donde $\widehat{\lambda}_\alpha$ es valor propio estimado definido en la sección 2.3.

Contribuciones. La estimación de las contribuciones mediante los perfiles fila, es decir en el espacio $\mathbb{R}^{p_2}$, es:

$$\widehat{Cr}_\alpha(i) = \frac{\widehat{f}_{i\cdot}\widehat{\psi}_{\alpha i}^2}{\widehat{\lambda}_\alpha} = \frac{\widehat{k}\left(\sum_{j=1}^{p_2} \frac{\widehat{k}_{ij}}{\widehat{k}_{\cdot j}}\widehat{\mathbf{u}}_{\alpha j}\right)^2}{\widehat{k}_{i\cdot}\widehat{\lambda}_\alpha} \tag{41}$$

y la estimación para los perfiles columna en el espacio $\mathbb{R}^{p_1}$

$$\widehat{Cr}_\alpha(j) = \frac{\widehat{f}_{\cdot j}\widehat{\varphi}_{\alpha j}^2}{\widehat{\lambda}_\alpha} = \frac{\widehat{k}\left(\sum_{j=1}^{p_1} \frac{\widehat{k}_{ij}}{\widehat{k}_{i\cdot}}\widehat{\mathbf{v}}_{\alpha i}\right)^2}{\widehat{k}_{\cdot j}\widehat{\lambda}_\alpha} \tag{42}$$

Cosenos cuadrados. Los cosenos cuadrados estimados para los perfiles fila son:

$$\widehat{Cos}_\alpha^2(i) = \frac{\widehat{\psi}_{\alpha i}^2}{\widehat{d}^2(i,G)} = \frac{\widehat{k}^2\left(\sum_{j=1}^{p_2} \frac{\widehat{k}_{ij}}{\widehat{k}_{\cdot j}}\widehat{\mathbf{u}}_{\alpha j}\right)^2}{\widehat{k}_{i\cdot}^2\widehat{d}^2(i,G)} \tag{43}$$

y para los perfiles columna

$$\widehat{Cos}_\alpha^2(j) = \frac{\widehat{\varphi}_{\alpha j}^2}{\widehat{d}^2(j,G)} = \frac{\widehat{k}^2\left(\sum_{j=1}^{p_1} \frac{\widehat{k}_{ij}}{\widehat{k}_{i\cdot}}\widehat{\mathbf{v}}_{\alpha i}\right)}{\widehat{k}_{\cdot j}^2\widehat{d}^2(j,G)} \tag{44}$$

donde en los perfiles fila la distancia de un punto i al centro de gravedad tendrá la siguiente estimación

$$\widehat{d}^2(i,G) = \sum_{j=1}^{p_2} \frac{1}{\widehat{f}_{\cdot j}}\left(\frac{\widehat{f}_{ij}}{\widehat{f}_{i\cdot}} - \widehat{f}_{\cdot j}\right)^2 = \sum_{j=1}^{p_2} \frac{\widehat{k}}{\widehat{k}_{\cdot j}}\left(\frac{\widehat{k}_{ij}}{\widehat{k}_{i\cdot}} - \frac{\widehat{k}_{\cdot j}}{\widehat{k}}\right)^2 \tag{45}$$

y para los perfiles columna

$$\widehat{d}^2(j,G) = \sum_{i=1}^{p_1} \frac{1}{\widehat{f}_{i\cdot}}\left(\frac{\widehat{f}_{ij}}{\widehat{f}_{\cdot j}} - \widehat{f}_{i\cdot}\right)^2 = \sum_{i=1}^{p_1} \frac{\widehat{k}}{\widehat{k}_{i\cdot}}\left(\frac{\widehat{k}_{ij}}{\widehat{k}_{\cdot j}} - \frac{\widehat{k}_{i\cdot}}{\widehat{k}}\right)^2 \tag{46}$$

2.5. Análisis de correspondencia múltiple a partir de una muestra probabilística

Extendiendo el análisis de correspondencias a partir de una muestra probabilística del espacio $U = \{1, \ldots, N\}$ al caso de m variables con $p_1, \ldots, p_m$ modalidades respectivamente, se obtienen las expresiones a este caso más general a través de un diseño $p(\cdot)$; luego la matriz S definida en caso simple puede ser extendida al caso múltiple como:

$$S = F' D_N^{-1} F D_p^{-1} = \frac{1}{m} Z' Z D^{-1} = \frac{1}{m} B D^{-1} \tag{47}$$

donde

$$\mathbf{Z} = \begin{bmatrix} z_{11} & \ldots & z_{1j} & \ldots & z_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{l1} & \ldots & z_{lj} & \ldots & z_{lp} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{N1} & \ldots & z_{Nj} & \ldots & z_{Np} \end{bmatrix}$$

con

$$z_{lj} = \begin{cases} 1, & \text{si el sujeto } l \text{ seleccionó la modalidad } j \\ 0, & \text{si el sujeto } l \text{ no seleccionó la modalidad } j \end{cases}$$

Entonces $z_{lj} = 1$ ó $z_{lj} = 0$ con $l = 1, 2, \ldots, N$ y $j = 1, 2, \ldots, p$. La matriz $\mathbf{D}$ es diagonal de orden (p, p) obtenida a partir de la matriz de Burt, $\mathbf{B} = \mathbf{Z}'\mathbf{Z}$, sin pérdida de generalidad, definimos $p = \sum_{j=1}^{m} p_j$. Así, la matriz de Burt está dada por:

$$B = Z'Z = \begin{bmatrix} \sum_{i=1}^N z_{i1} z_{i1} & \sum_{i=1}^N z_{i1} z_{i2} & \cdots & \sum_{i=1}^N z_{i1} z_{ip} \\ \sum_{i=1}^N z_{i2} z_{i1} & \sum_{i=1}^N z_{i2} z_{i2} & \cdots & \sum_{i=1}^N z_{i2} z_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^N z_{ip} z_{i1} & \sum_{i=1}^N z_{ip} z_{i2} & \cdots & \sum_{i=1}^N z_{ip} z_{ip} \end{bmatrix}$$

Entonces los elementos de S son de la forma

$$s_{jj'} = \frac{1}{m z_{.j'}} \sum_{i=1}^{N} z_{ij} z_{ij'} \tag{48}$$

donde

$$z_{.j'} = \sum_{i=1}^{N} z_{ij'}$$

Así, $s_{jj'}$ es una función de totales poblacionales de las variables $z_{j'}$ y $z_{jj'} = z_j z_{j'}$. Entonces el polinomio característico de S de acuerdo con la ecuación (1) viene dado por

$$p(\lambda) = \mid \mathbf{S} - \lambda \mathbf{I} \mid = (-1)^p \left(\lambda^p + b_{p-1} \lambda^{p-1} + \cdots + b_1 \lambda + b_0 \right) \tag{49}$$

donde cada b_j en el polinomio característico es función de los valores $s_{jj'}$ que a su vez son funciones de totales poblacionales, así

$$b_r = f\left(t_{z_1}, \ldots, t_{z_{j'}}, \ldots, t_{z_p}, \ldots, t_{z_{11}}, \ldots, t_{z_{jj'}}, \ldots, t_{z_{pp}}\right) \tag{50}$$

donde

$$t_{z_{j'}} = \sum_{i=1}^{N} z_{ij'}$$

y

$$t_{z_{jj'}} = \sum_{i=1}^{N} z_{ij} z_{ij'}$$

Por tanto se puede asumir que λ es también función de totales poblacionales

$$\lambda = f\left(t_{z_1}, \ldots, t_{z_{j'}}, \ldots, t_{z_p}, \ldots, t_{z_{11}}, \ldots, t_{z_{jj'}}, \ldots, t_{z_{pp}}\right) \tag{51}$$

Así, t_{zi} puede ser estimado a través de un π estimador basado en una muestra probabilística s de tamaño n, como sigue

$$\widehat{t}_{zj'} = \sum_{i \in s} \frac{z_{ij'}}{\pi_i}$$

y

$$\widehat{t}_{zjj'} = \sum_{i \in s} \frac{z_{ij} z_{ij'}}{\pi_i}$$

De esta manera, se puede establecer que un estimador aproximado para λ es de la forma

$$\widehat{\lambda} = f\left(\widehat{t}_{z_1}, \ldots, \widehat{t}_{z_{j'}}, \ldots, \widehat{t}_{z_p}, \ldots, \widehat{t}_{z_{11}}, \ldots, \widehat{t}_{z_{jj'}}, \ldots, \widehat{t}_{z_{pp}}\right) \tag{52}$$

resultado de resolver el polinomio característico estimado a partir de la muestra s de la matriz

$$\widehat{S} = Z_n' \Pi^{-1} Z_n \widehat{D}^{-1} = \widehat{B} \widehat{D}^{-1} \tag{53}$$

dado por

$$p(\lambda) = \left|\widehat{S} - \widehat{\lambda} I\right| = (-1)^p \left(\widehat{\lambda}^p + \widehat{b}_{p-1} \widehat{\lambda}^{p-1} + \cdots \widehat{b}_1 \widehat{\lambda} + \widehat{b}_0\right) \tag{54}$$

donde los b_r están definidos en la ecuación (50), la matriz de Burt estimada en la ecuación (53) es

$$\widehat{B} = Z_n' \Pi^{-1} Z_n \tag{55}$$

la matriz diagonal estimada obtenida a partir de la matriz de Burt, corresponde a una matriz de orden (p, p), dada por

$$\widehat{D} = \text{diag}\{Z_n' \Pi^{-1} Z_n\} \tag{56}$$

y Π es la matriz diagonal de probabilidades de inclusión

$$\Pi = \text{diag}\{\pi_1, \ldots, \pi_n\} \tag{57}$$

entonces

$$\mathbf{Z}_n = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}$$

donde

$$\widehat{\mathbf{B}} = \mathbf{Z}_n' \Pi^{-1} \mathbf{Z}_n = \begin{bmatrix} \sum_{i=1}^n \frac{z_{i1} z_{i1}}{\pi_i} & \sum_{i=1}^n \frac{z_{i1} z_{i2}}{\pi_i} & \cdots & \sum_{i=1}^n \frac{z_{i1} z_{ip}}{\pi_i} \\ \sum_{i=1}^n \frac{z_{i2} z_{i1}}{\pi_i} & \sum_{i=1}^n \frac{z_{i2} z_{i2}}{\pi_i} & \cdots & \sum_{i=1}^n \frac{z_{i2} z_{ip}}{\pi_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n \frac{z_{ip} z_{i1}}{\pi_i} & \sum_{i=1}^n \frac{z_{ip} z_{i2}}{\pi_i} & \cdots & \sum_{i=1}^n \frac{z_{ip} z_{ip}}{\pi_i} \end{bmatrix}$$

y la matriz diagonal estimada

$$\widehat{\mathbf{D}} = diag\left\{\mathbf{Z}_n' \Pi^{-1} \mathbf{Z}_n\right\} = \begin{bmatrix} \sum_{i=1}^n \frac{z_{i1}^2}{\pi_i} & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^n \frac{z_{i2}^2}{\pi_i} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{i=1}^n \frac{z_{ip}^2}{\pi_i} \end{bmatrix}$$

2.6. Estimación de la varianza (ACM)

El π estimador propuesto

$$\widehat{\lambda} = f\left(\widehat{t}_{z_1}, \ldots, \widehat{t}_{z_{j'}}, \ldots, \widehat{t}_{z_p}, \ldots, \widehat{t}_{z_{11}}, \ldots, \widehat{t}_{z_{jj'}}, \ldots, \widehat{t}_{z_{pp}}\right)$$

es un estimador aproximadamente insesgado para λ.

Esta prueba es análoga como en (27); entonces para el caso múltiple

$$\widehat{u}_i = \sum_r \widehat{a}_r z_{ri}$$

donde

$$\widehat{a}_r = \frac{\partial f}{\partial \widehat{t}_r}$$

Nuevamente las formas explícitas para $\widehat{u}_i$ son muy complicadas de calcular en la práctica; ende tanto se hace necesario establecer un estimador para la varianza del π estimador $\widehat{\lambda}$, a través del método *Jackknife* o *Bootstrap*, como sigue.

2.6.1. El estimador *Jackknife*

La expresión del estimador para la varianza es como en (29), donde

$$\widehat{\lambda}_{n-1,i} = f\left(\widehat{t}_{z_1(n-1,i)}, \ldots, \widehat{t}_{z_p(n-1,i)}, \ldots \widehat{t}_{z_{11}(n-1,i)}, \ldots, \widehat{t}_{z_{jj'}(n-1,i)} \ldots, \widehat{t}_{z_{pp}(n-1,i)}\right)$$

con

$$\widehat{t}_{zp\pi(n-1,i)} = \sum_{l \in S-\{i\}} \frac{z_{lj}}{\pi_l} \tag{58}$$

También es posible utilizar aproximaciones "apropiadas" de v_{JK} que requieran menos cálculos.

Shao & Tu (1995) introducen dos métodos computacionales para desarrollar el *Jackknife* delete-1 en la estimación de varianza, el método de agrupamiento y el método de submuestreo aleatorio; sin embargo en este trabajo se compara *Jackknife* con *Bootstrap*, mostrando que el remuestreo *Bootstrap* resulta ser más eficiente.

2.6.2. El estimador *Bootstrap*

El estimador de la varianza es similar como en la ecuación (32), donde se define la desviación *Bootstrap*; el procedimiento para la estimación es análogo.

2.7. Estimación de los elementos de base en el análisis de correspondencias múltiples

Construcción de las nubes. Suponga el conjunto de las coordenadas fila una nube de N puntos en el espacio de las p columnas. Cada punto i tiene por coordenadas en $\mathbb{R}^p$ $\left\{\frac{z_{ij}}{\widehat{z}_{i\cdot}}; j = 1, 2, \ldots, p\right\}$ y cada punto j tiene por coordenadas en $\mathbb{R}^N$ $\left\{\frac{z_{ij}}{\widehat{z}_{\cdot j}}; i = 1, 2, \ldots, N\right\}$

Selección de distancias. En $\mathbb{R}^N$ la distancia estimada χ^2 entre modalidades es

$$\widehat{d}^2(j, j') = \sum_{i=1}^{N} \widehat{N}\left(\frac{z_{ij}}{\widehat{z}_{\cdot j}} - \frac{z_{ij'}}{\widehat{z}_{\cdot j'}}\right)^2 \tag{59}$$

y en $\mathbb{R}^p$ la distancia estimada entre dos individuos es

$$\widehat{d}^2(i, i') = \frac{1}{m} \sum_{i=1}^{N} \frac{\widehat{N}}{\widehat{z}_{\cdot j}} \left(z_{ij} - z_{i'j}\right)^2 \tag{60}$$

con $\widehat{z}_{\cdot j} = \widehat{t}_{z_j} = \sum_{i \in s} \frac{z_{ij}}{\pi_i}$ y $\widehat{N} = \sum_{j=1}^{p} \sum_{i \in s} \frac{z_{ij}}{\pi_i}$

2.7.1. Coordenadas factoriales estimadas

Retomando los resultados del análisis de correspondencias se tiene

$$\begin{aligned}
\widehat{\mathbf{F}} &= \frac{1}{\widehat{N}m}\mathbf{Z} \text{ de término general } \widehat{f}_{ij} = \frac{z_{ij}}{\widehat{N}p} \\
\widehat{\mathbf{D}}_p &= \frac{1}{\widehat{N}m}\widehat{\mathbf{D}} \text{ de término general } \widehat{f}_{.j} = \delta_{ij}\frac{\widehat{z}_{.j}}{\widehat{N}m} \\
\widehat{\mathbf{D}}_N &= \frac{1}{\widehat{N}}\mathbf{I}_N \text{ de término general } \widehat{f}_{i.} = \frac{\delta_{ij}}{\widehat{N}}
\end{aligned}$$

donde $\mathbf{I}_{\widehat{N}}$ es la matriz identidad de orden $(\widehat{N}, \widehat{N})$ y $\delta_{ij} = 1$ si $i = j$ y cero si no.

Para encontrar los ejes factoriales $\widehat{u}_\alpha$ se diagonaliza la matriz

$$\widehat{\mathbf{S}}^+ = \frac{1}{m^2}\widehat{\mathbf{D}}^{-\mathbf{1}}\widehat{\mathbf{B}}\widehat{\mathbf{D}}^{-\mathbf{1}}\widehat{\mathbf{B}}$$

donde $\widehat{\mathbf{D}}$ es la matriz diagonal de orden (p, p) de la matriz $\widehat{\mathbf{B}} = \mathbf{Z}'\mathbf{Z}$.

En $\mathbb{R}^p$, la ecuación del α-ésimo eje factorial estimado es

$$\frac{1}{m}\mathbf{Z}'_n\Pi^{-1}\mathbf{Z}_n\widehat{\mathbf{D}}^{-1}\widehat{\mathbf{u}}_\alpha = \widehat{\lambda}_\alpha\widehat{\mathbf{u}}_\alpha \tag{61}$$

la ecuación del α-ésimo factor estimado es

$$\frac{1}{m}\widehat{\mathbf{D}}^{-1}\mathbf{Z}'_n\Pi^{-1}\mathbf{Z}_n\widehat{\varphi}_\alpha = \widehat{\lambda}_\alpha\widehat{\varphi}_\alpha \tag{62}$$

del mismo modo se escribe el α-ésimo factor estimado en $\mathbb{R}^N$ como

$$\frac{1}{m}\left[\mathbf{Z}_n\Pi^{-1}\widehat{\mathbf{D}}^{-1}\mathbf{Z}'_n\right]\widehat{\psi}_\alpha = \widehat{\lambda}_\alpha\widehat{\psi}_\alpha \tag{63}$$

Las coordenadas factoriales estimadas de un individuo i sobre el eje α están dadas por

$$\widehat{\psi}_{\alpha i} = \frac{1}{m\sqrt{\widehat{\lambda}_\alpha}} = \sum_{j\in p(i)}\widehat{\varphi}_{\alpha i} \tag{64}$$

$$\widehat{\varphi}_{\alpha i} = \frac{1}{\sqrt{\widehat{\lambda}_\alpha}} = \sum_{i\in I(j)}\widehat{\psi}_\alpha \tag{65}$$

donde $p(i)$ designa al conjunto de modalidades seleccionadas por el individuo i, y por otra parte $I(j)$ designa al conjunto de los individuos que seleccionaron la modalidad j en la muestra.

2.7.2. Relaciones de transición

Los factores estimados $\widehat{\varphi}_\alpha$ y $\widehat{\psi}_\alpha$ de norma $\widehat{\lambda}_\alpha$ representan las coordenadas estimadas de los puntos fila y los puntos columna sobre el eje factorial α; luego las

relaciones de transición vienen dadas por:

$$\widehat{\varphi}_\alpha = \frac{1}{\sqrt{\widehat{\lambda}_\alpha}} \widehat{\mathbf{D}}^{-1} \mathbf{Z}_n' \Pi^{-1} \widehat{\psi}_\alpha \tag{66}$$

y

$$\widehat{\psi}_\alpha = \frac{1}{m\sqrt{\widehat{\lambda}_\alpha}} \left[\mathbf{Z}_n \Pi^{-1}\right]' \widehat{\varphi}_\alpha \tag{67}$$

respectivamente.

2.7.3. Estimación de la inercia, contribuciones y cosenos cuadrados

Por otra parte, la estimación de los elementos de base en el análisis, como la inercia, las contribuciones y los cosenos cuadrados, se presenta a continuación:

Inercia. Para el caso del análisis de correspondencias multiple, la inercia $I(j)$ de una modalidad j se puede estimar a través de la expresión:

$$\widehat{I}(j) = \frac{1}{m}\left(1 - \frac{\sum_{i\in s} \frac{z_{ij}}{\pi_i}}{\widehat{N}}\right) \tag{68}$$

donde

$$\widehat{N} = \sum_{j_q=1}^{p_q} \sum_{i\in S} \frac{z_{ij_q}}{\pi_i}$$

Con z_{ij_q} variable dicótoma, así, es igual a uno, si el individuo i respondió la modalidad j de la pregunta q, y cero en otra modalidad de la misma pregunta. La inercia $I(q)$ de una pregunta q esta dada por:

$$\widehat{I}(q) = \sum_{j=1}^{p_q} \widehat{I}(j)$$

Entonces la inercia total estimada es:

$$I = \sum_q \widehat{I}(q)$$

Contribuciones. Las contribuciones estimadas en el análisis de correspondencias para los puntos fila ($\mathbb{R}^p$) y columna ($\mathbb{R}^n$), respectivamente, son:

$$\widehat{C}r_\alpha(i) = \frac{\widehat{\psi}_{\alpha i}^2}{\widehat{N}\widehat{\lambda}_\alpha} \quad y \quad \widehat{C}r_\alpha(j) = \frac{\widehat{z}_{\cdot j}\widehat{\varphi}_{\alpha j}^2}{\widehat{N}m\widehat{\lambda}_\alpha} \tag{69}$$

Cosenos cuadrados. Los cosenos cuadrados para los puntos fila ($\mathbb{R}^p$) y columna ($\mathbb{R}^n$) son:

$$\widehat{Cos}^2_\alpha(i) = \frac{\widehat{\psi}^2_{\alpha i}}{\widehat{d}^2(i,G)} \quad y \quad \widehat{Cos}^2_\alpha(j) = \frac{\widehat{\varphi}^2_{\alpha j}}{\widehat{d}^2(j,G)} \tag{70}$$

respectivamente, donde en los puntos fila la distancia de un punto i al centro de gravedad tendrá la siguiente estimación:

$$\widehat{d}^2(i,G) = \sum_{j=1}^{p_2} \frac{1}{\widehat{f}_{\cdot j}} \left(\frac{\widehat{f}_{ij}}{\widehat{f}_{i\cdot}} - \widehat{f}_{\cdot j} \right)^2 = \frac{\widehat{N}}{\widehat{z}_{\cdot j}} - 1 \tag{71}$$

3. Ejemplo de aplicación

Esta aplicación se refiere a las condiciones de vida de una población que se encuentra descrita por 150 UPM (unidades primarias de muestreo), las cuales se emplean para realizar un estudio por muestreo probabilístico a través de una muestra correspondiente al 30 % de las UPM, seleccionada bajo un diseño de muestreo bietápico con MAS en la primera etapa y MAS en la segunda etapa. En la siguiente tabla figuran las etiquetas de las modalidades de las cuatro preguntas:

Tabla 1: Descripción de variables.

Id	Preguntas	Modalidades
1	Se siente bien en el hogar	FA01. "Sí" FA02. "No"
2	Los gastos de vivienda son	DL01. "Despreciable" DL02. "Sin problema" DL03. "Gran carga" DL04. "Carga muy pesada"
3	Ha sufrido de dolor de espalda	MA01. "Sí" MA02. "No"
4	Se impone restricciones	RE01. 'Sí" RE02. "No"

El objetivo de este ejemplo es comparar los resultados de aplicar un análisis de correspondencias múltiple a la población descrita anteriormente y a una muestra aleatoria seleccionada de dicha población a través de un diseño de MAS-MAS. Se verifica luego que hay seis valores propios poblacionales y estimados no nulos ($6 = 10 - 4 = p - m$) que se muestran a continuación:

Tabla 2: Comparación de valores propios poblacionales y estimados.

Valores poblacionales				Valores estimados *Jackknife*				
No	V.P.	% Ine.	% Acum.	V.P.	% Ine.	% Acum.	ECM_{JK}	cve
1	0,11531	29,04	29,04	0,11927	30,14	30,14	0,0000417	0,0442
2	0,09634	24,26	53,31	0,08895	22,47	52,61	0,0001527	0,1028
3	0,06356	16,01	69,31	0,06331	16,00	68,61	0,0000071	0,0409
4	0,05298	13,34	82,66	0,05327	13,46	82,07	0,0000351	0,1114
5	0,04353	10,96	93,62	0,04402	11,12	93,19	0,0000142	0,0850
6	0,02532	6,38	100	0,02694	6,81	100	0,0000076	0,0869

En la tabla 2 notamos que el primer plano factorial poblacional, conformado por los dos primeros valores propios diferentes de cero, proporcionan un porcentaje

de inercia del 53,31 %, mientras que el primer plano factorial estimado presenta un porcentaje de variación del 52,61 % y coeficiente de variación del 10 %, lo cual es un buen indicio de estimación; además se puede observar la precisión en las estimaciones de manera puntual a cada uno de los valores propios, y finalmente es de resaltar la calidad de las estimaciones, a través de los coeficientes de variación, ya que resultan bastante efectivos.

Tabla 3: Comparación de valores propios poblacionales y estimados.

Valores poblacionales				Valores estimados *Bootstrap*				
No	V.P.	% Ine.	% Acum.	V.P.	% Ine.	% Acum.	ECM_{Boot}	cve
1	0,11531	29,04	29,04	0,11666	29,25	29,25	0,0000158	0,0442
2	0,09634	24,26	53,31	0,10082	25,27	54,52	0,0000421	0,1028
3	0,06356	16,01	69,31	0,06351	15,92	70,44	0,0000090	0,0409
4	0,05298	13,34	82,66	0,04982	12,49	82,93	0,0000480	0,1114
5	0,04353	10,96	93,62	0,04307	10,80	93,73	0,0000132	0,0850
6	0,02532	6,38	100	0,02501	6,27	100	0,0000061	0,0869

Se aprecia en la tabla 3 que utilizando el método *Bootstrap*, las estimaciones resultan ser un poco más eficientes que mediante el método *Jackknife* teniendo en cuenta la simulación realizada, en particular para los valores propios más grandes basándose en el error cuadrático medio y demás estimaciones. Cabe destacar que se realizaron 2000 simulaciones con $B = 500$ remuestras *Bootstrap*; estos valores se consideraron teniendo en cuenta que no se presentaron cambios significativos para un número mayor de simulaciones en las estimaciones.

Las estimaciones de los elementos de base en el análisis de correspondencias se presentan en el anexo.

4. Métodos computacionales

Una vez seleccionada la muestra probabilistica de la población de interés, y determinados los factores de expansión a partir del diseño de muestreo seleccionado, es posible aplicar algunas rutinas de programas estadísticos como SAS, SPAD o XLSTAT para estimar los elementos de base del análisis de correspondencias. Esto se logra incluyendo la variable "Peso" en el análisis correspondiente (en nuestro contexto, a los factores de expansión del diseño). Más explícitamente: mediante SAS, consiste en cargar el paquete PROC CORRESP con la opción Weigth p, donde p es la columna que contiene los factores de expansión del análisis por muestreo probabilístico. Mediante SPAD, consiste en establecer una columna en la base de datos de la muestra como los factores de expansión, la cual se colocará como "ponderación" en los parámetros del software. Mediante XLSTAT, consiste en insertar la matriz de Burt, Disyuntiva Completa, o Individuos Variable, en la opción que especifica la matriz con la cual se realizará el análisis, en nuestro caso la matriz de Burt, y después en la opción Peso se insertan los factores de expansión.

5. Conclusiones

- Dado que los valores propios por estimar se pueden escribir como funciones de las variables observadas en la muestra, como totales, razones y dominios poblacionales, es posible determinar medidas de calidad para las estimaciones. La varianza de estos estimadores se puede obtener a través de técnicas de aproximación de varianza, como es el caso de la técnica de linealización de primer orden de Taylor, y su estimación se puede realizar mediante métodos robustos como el *Jackknife* y *Bootstrap*, siendo este último más eficiente teniendo en cuenta la simulación realizada.

- Los diseños de muestreo probabilístico pueden ser utilizados para estimar los elementos de base en el análisis de correspondencias dando resultados confiables.

- Es posible utilizar diseños de muestreo más complejos, a través del muestreo en dos fases, utilizando como estimador de la varianza el estimador *Jackknife*, presentado por Pacheco & Martínez (2007).

Agradecimientos

Los autores del presente trabajo agradecen de manera muy especial a todas aquellas personas que contribuyeron en la elaboración y corrección del mismo. En particular agradece al Departamento de Matemáticas y Estadística de la Universidad de Córdoba.

[Recibido: abril de 2009 — Aceptado: octubre de 2010]

Referencias

Clausen, S. E., ed. (1998), *Applied Correspondence Analylis: An Introducction*, number 121 *in* 'Series on Quantitative Applications in the Social Sciences', Sage University Papers, Thousand Oaks, California.

Escofier, B. & Pagés, J. (1992), *Análisis factoriales simples y múltiples: objetivos, métodos e interpretación*, Universidad del País Vasco, Bilbao.

Lebart, L., Morineau, A. & Piron, M. (2000), *Statistique Exploratoire Multidimensionnelle*, Dunod, Francia.

Martínez, G. (1998), Estimación de los coeficientes de un análisis en componentes principales a partir de una muestra probabilística, Trabajo final, Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia.

Milan, L. & Whittaker, R. J. (1995), 'Application of the Parametric Bootstrap to Models that Incorporate a Singular Value Decomposition', *Applied Statistics* **44**(1), 31–49.

Pacheco, M. & Martínez, G. (2007), 'Un estimador jackknife de varianza en muestreo en dos fases con probabilidades desiguales', *Revista Colombiana de Estadística* **30**(2), 203–212.

Shao, J. & Tu, D. (1995), *The Jackknife and Bootstrap*, Springer-Verlag, New York.

Särndal, C. E., Swensson, B. & Wretman, J. H. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Wolter, K. M. (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York.

Apéndice A. Tablas

Tabla 4: Comparación de las coordenadas poblacionales y estimadas.

Modalidad	Coordenadas poblacionales					Coordenadas estimadas				
	1	2	3	4	5	1	2	3	4	5
FA01	−0,05	0,25	0,06	−0,15	−0,16	0,12	0,25	0,05	−0,11	−0,21
FA02	0,11	−0,54	−0,12	0,33	0,36	−0,22	−0,48	−0,09	0,21	0,40
DL01	−0,60	0,81	0,25	1,15	−0,07	−0,03	1,04	0,03	1,18	0,09
DL02	−0,30	−0,23	−0,06	−0,21	0,09	−0,37	−0,03	−0,04	−0,21	−0,02
DL03	0,62	0,19	−0,36	0,06	−0,20	0,65	−0,24	−0,42	0,05	−0,07
DL04	0,66	−0,11	1,71	−0,14	0,23	0,50	−0,21	1,53	−0,05	0,25
MA01	0,05	0,40	−0,07	−0,10	0,32	0,26	0,33	−0,08	−0,20	0,32
MA02	−0,04	−0,37	0,07	0,09	−0,29	−0,21	−0,26	0,07	0,16	−0,26
RE01	0,40	−0,02	0,01	0,02	0,01	0,33	−0,13	0,01	0,03	−0,04
RE02	−0,60	0,03	−0,01	−0,04	−0,02	−0,59	0,24	−0,02	−0,06	0,07

Tabla 5: Comparación de los cosenos cuadrados poblacionales y estimados.

Modalidad	Coordenadas poblacionales					Coordenadas estimadas				
	1	2	3	4	5	1	2	3	4	5
FA01	0,02	0,52	0,03	0,19	0,23	0,10	0,47	0,02	0,09	0,33
FA02	0,02	0,52	0,03	0,19	0,23	0,10	0,47	0,02	0,09	0,33
DL01	0,14	0,27	0,02	0,54	0,01	0,01	0,43	0,01	0,56	0,01
DL02	0,41	0,23	0,02	0,20	0,04	0,64	0,01	0,01	0,21	0,01
DL03	0,57	0,05	0,19	0,05	0,06	0,54	0,07	0,23	0,01	0,01
DL04	0,12	0,01	0,83	0,05	0,02	0,09	0,02	0,84	0,01	0,02
MA01	0,01	0,57	0,02	0,04	0,40	0,20	0,33	0,02	0,13	0,31
MA02	0,01	0,57	0,02	0,04	0,37	0,20	0,33	0,02	0,13	0,31
RE01	0,82	0,01	0,01	0,01	0,01	0,68	0,11	0,01	0,01	0,01
RE02	0,82	0,01	0,01	0,01	0,01	0,68	0,11	0,01	0,01	0,01

Revista Colombiana de Estadística
Diciembre 2010, volumen 33, no. 2, pp. 295 a 305

Funciones de varianza y correlación bicuadrática para distribuciones normales

Biweight Variance and Correlation Functions for Normal Distributions

Carlos Eduardo Alonso[a], Jorge Martínez[b]

Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Resumen

En este trabajo se analiza el comportamiento del funcional ϱ asociado al estimador de correlación bicuadrático $-\widehat{\varrho}-$, asumiendo que se observan vectores aleatorios con distribución normal bivariada. Esto, con el objetivo de verificar si este estimador robusto es un estimador insesgado del coeficiente de correlación $-\rho-$.

El trabajo se desarrolló a partir de las propiedades de la función generadora de momentos de una distribución.

De acuerdo con los resultados, $\varrho > \rho$ cuando $\rho < 0$, $\varrho < \rho$ cuando $\rho > 0$, y $\varrho = 0$ cuando $\rho = 0$, e indican que el estimador propuesto $\widehat{\varrho}$ no es un estimador insesgado del coeficiente de correlación.

Lo anterior plantea como reto modificar el estimador $\widehat{\varrho}$ con el objetivo de obtener un estimador robusto insesgado o asintóticamente insesgado del coeficiente de correlación.

***Palabras clave*:** coeficiente de correlación, distribución truncada, estimación robusta, estimador M.

Abstract

In this paper, we have analized the behavior of the functional ϱ, associated to the biweight correlation estimator $-\widehat{\varrho}-$, assuming the sampled population has a bivariate normal distribution. The purpose is to verify if the estimator $\widehat{\varrho}$ is an unbiased estimator of the correlation coefficient ρ.

The results show $\varrho > \rho$ when $\rho < 0$, $\varrho < \rho$ when $\rho > 0$, y $\varrho = 0$ when $\rho = 0$. This results indicate $\widehat{\varrho}$ is not an unbiased estimator of the correlation coefficient.

***Key words*:** Correlation coefficient, M-estimate, Truncated distribution, Robust estimation.

[a]Profesor asistente. E-mail: cealonsom@unal.edu.co
[b]Profesor especial. E-mail: jmartinezc@unal.edu.co

1. Coeficiente de correlación bicuadrático

En la práctica es importante estudiar el desempeño de los procedimientos estadísticos ante incumplimientos de los supuestos, variaciones en los supuestos que son comunes en la cotidianidad de un usuario. Esto con el objetivo de hallar situaciones en las cuales no se recomienda usar la herramienta, y al tiempo plantear herramientas no tan sensibles ante el incumplimiento de los supuestos.

En este sentido, en Wei (2006) y Valcárcel (2007) se ha mostrado que el estimador clásico de la función de autocorrelación (FAC) es altamente sensible ante la presencia de valores extremos, sensibilidad que contrasta con la relevancia de este estimador en el análisis de series de tiempo, porque a partir de los valores estimados de la FAC se puede identificar el modelo, se construyen los estimadores de los parámetros y se analizan los residuales, entre otros. De lo anterior se ha planteado un estimador robusto del coeficiente de correlación, y posteriormente un estimador robusto de la FAC, estimador que se presenta a continuación.

Dada una muestra aleatoria $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, el estimador del coeficiente de correlación propuesto está dado por,

$$\widehat{\varrho}_{xy} = \frac{\delta_{xy}^2}{\delta_{xx}\delta_{yy}} \tag{1}$$

con

$$\delta_{xy}^2 = nk^2[MAD_x][MAD_y]\frac{\sum_{i=1}^{n}\psi(z_{x_i})\psi(z_{y_i})}{\left(\sum_{i=1}^{n}\psi'(z_{x_i})\right)\left(\sum_{i=1}^{n}\psi'(z_{y_i})\right)} \tag{2}$$

donde k es una constante positiva de estandarización[1], $\psi(\cdot)$ y $\psi'(\cdot)$ son la función de bicuadrática planteada por Beaton & Tukey (1974) y su derivada, respectivamente. La función bicuadrática está dada por

$$\psi(z) = \begin{cases} z(1-z^2)^2 & \text{para } |z| < 1 \\ \\ 0 & \text{para } |z| \geqq 1 \end{cases} \tag{3}$$

con $z_t = \frac{x_t - Med_x}{kMAD_x}$, $Med_x = mediana\{x_1, x_2, \ldots, x_n\}$ y $MAD_x = mediana|x_t - Med_x|$. Para utilizar el estimador $\widehat{\varrho}_{xy}$ en la estimación de la FAC $-\rho_h-$, para una serie de tiempo estacionaria $y_1, y_2, \ldots, y_T$, la ecuación (2) se transforma en

$$\varphi_h = Tk^2MAD_y^2\frac{\sum_{t=1}^{T-|h|}\psi(z_{y_t})\psi(z_{y_{t+|h|}})}{\left(\sum_{t=1}^{T}\psi'(z_{y_t})\right)\left(\sum_{j=1}^{T-|h|}\psi'(z_{y_{t+|h|}})\right)} \tag{4}$$

[1]La propuesta de Lax (1975) es $k = 9$, valor planteado desde las propiedades de una distribución $N(\mu, \sigma^2)$, donde $MAD \approx \frac{2}{3}\sigma$; así, si $k = 3d$, se tiene que $kMAD \approx d \times \sigma$, es decir al construir el estimador no se tienen en cuenta las observaciones a más de d desviaciones estándar de la media.

A partir de (4) el estimador de ρ_h está dado por

$$\widehat{\rho}_h = \frac{\varphi_h}{\varphi_0} \tag{5}$$

2. Resultados

Uno de los resultados intermedios del trabajo es la generalización del concepto de covarianza, resultado que se muestra porque a partir de este se desarrolla el funcional asociado a $\widehat{\varrho}_{xy}$.

2.1. Generalización de la covarianza del coeficiente de correlación

Definición 1 (ψ-*Covarianza*)**.** Dadas dos variables aleatorias X y Y, con función distribución conjunta y segundos ψ-momentos finitos ($E\left[\psi^2(X)\right] < \infty$ y $E\left[\psi^2(Y)\right] < \infty$), la ψ-covarianza entre X y Y se define como:

$$\gamma_{\psi_{XY}} = \frac{E\left[\psi(Z_x)\psi(Z_y)\right]}{E\left[\psi'(Z_x)\right]E\left[\psi'(Z_y)\right]} \tag{6}$$

Un caso particular de esta definición se obtiene haciendo $\psi(x) = x$, $\psi'(x) = \frac{\partial\psi(x)}{\partial x} = 1$, $Z_x = X - EX$ y $Z_y = Y - EY$, de donde $\gamma_{\psi_{XY}} = E\left[(X - EX)(Y - EY)\right]$, que es la definición de covarianza clásica[2].

Definición 2 (ψ-*Correlación*)**.** Dadas dos variables aleatorias X y Y, con $\gamma_{\psi_{XX}} < \infty$ y $\gamma_{\psi_{YY}} < \infty$, la ψ-correlación entre X y Y se define como:

$$\varrho_{\psi_{XY}} = \frac{\gamma_{\psi_{XY}}}{\left(\gamma_{\psi_{XX}}\gamma_{\psi_{YY}}\right)^{\frac{1}{2}}} \tag{7}$$

De lo desarrollado, es claro que el coeficiente de correlación de Pearson es un caso particular de esta definición.

2.2. Funcional asociado a $\widehat{\varrho}_{xy}$

Se define

$$\begin{aligned}\widehat{\tau}^2_{xy} &= n\frac{\sum_{i=1}^{n}\psi(z_{x_i})\psi(z_{y_i})}{\left(\sum_{i=1}^{n}\psi'(z_{x_i})\right)\left(\sum_{i=1}^{n}\psi'(z_{y_i})\right)} \\ &= \frac{\int\psi(z_{x_i})\psi(z_{y_i})dF_n}{\left(\int\psi'(z_{x_i})dF_n\right)\left(\int\psi'(z_{y_i})dF_n\right)} = T(F_n)\end{aligned} \tag{8}$$

[2]Esto se plantea inicialmente para polinomios.

donde F_n es la función de distribución muestral. Asociado a $\widehat{\tau}^2_{xy}$ se tiene el funcional

$$\tau^2_{xy} = \frac{\int \psi(z_{x_i})\psi(z_{y_i})dF}{\left(\int \psi'(z_{x_i})dF\right)\left(\int \psi'(z_{y_i})dF\right)} = T(F) \tag{9}$$

donde F es la función de distribución poblacional. Unido a lo anterior se puede mostrar que

$$\widehat{\varrho}_{xy} = \frac{\widehat{\tau}^2_{xy}}{\widehat{\tau}_{xx}\widehat{\tau}_{yy}} \tag{10}$$

A partir de las ecuaciones (9) y (10), se tiene que el funcional asociado al estimador planteado en (1) es el coeficiente de ψ-*Correlación* presentado en la definición 2, ecuación (7), con $\psi(\cdot)$ la función bicuadrática definida en (3).

El objetivo de este trabajo es analizar el comportamiento de este funcional, es decir

$$\varrho_{\psi_{XY}} = \frac{\gamma_{\psi_{XY}}}{\gamma_{\psi_{XX}}\gamma_{\psi_{YY}}} \tag{11}$$

asumiendo que el vector (X, Y) tiene distribución $N(0, 0, 1, 1, \rho)$. En lo que sigue se mencionarán $\gamma_{\psi_{XX}}$ y $\gamma_{\psi_{XY}}$ como funciones de varianza y covarianza bicuadrática respectivamente. En este mismo sentido, $\varrho_{\psi_{XY}}$ se llama en adelante coeficiente de correlación bicuadrático[3].

2.3. Varianza bicuadrática

Si se asume que el vector (X, Y) tiene distribución $N(0, 0, 1, 1, \rho)$, las variables aleatorias X y Y tienen distribución univariada $N(0, 1)$, y

$$Med_X = Med_Y = 0 \qquad \text{y} \qquad MAD_X = MAD_Y = 0,67448975$$

resultados que conllevan a que las variables estandarizadas estén dadas por $Z_x = \frac{X}{l}$ y $Z_y = \frac{Y}{l}$, $l = k \times 0.67448975$, k constante de estandarización definida en (2)[4].

Si se define la variable aleatoria $M = I_{\{X \leq l\}}X$ la varianza bicuadrática de X, se puede escribir en términos de M como

$$\gamma_{\psi_{XX}} = \frac{\frac{E_2}{l^2} - 4\frac{E_4}{l^4} + 6\frac{E_6}{l^6} - 4\frac{E_8}{l^8} + \frac{E_{10}}{l^{10}}}{\left(1 - 6\frac{E_2}{l^2} + 5\frac{E_4}{l^4}\right)^2} \tag{12}$$

donde $E_r = E(M^r)$. La función de densidad de M está dada por $f_M(m) = \frac{\phi(m)}{c_1} I_{\{|m|<l\}}$, $f_M(\cdot)$ es la función de densidad resultante de truncar a dos colas la función de densidad asociada a una distribución normal estándar $\phi(\cdot)$, y $c_1 = P(-l < X < l)$.

[3] Esto indica una distribución normal bivariada con parámetros $E(X) = E(Y) = 0$, $V(X) = V(Y) = 1$ y $Corr(X, Y) = \rho$.

[4] Los valores de k con los cuales se realizó este trabajo son $k = 3, 6, 9$.

Si se nota la derivada de orden r de la función generatriz de momentos de la variable M, por $m^{(r)}(t) = \frac{\partial E(e^{tM})}{\partial t^r}$, esta derivada cumple con la siguiente recurrencia

$$\begin{aligned} m^{(r)}(t) =& (r-1)m^{(r-2)}(t) + tm^{(r-1)}(t) \\ &+ (-l)^{r-1}\frac{e^{-\frac{1}{2}l^2}\left[e^{-lt} + (-1)^k e^{lt}\right]}{c_1(2\pi)^{\frac{1}{2}}} \quad \text{para} \quad r \geq 2 \end{aligned} \tag{13}$$

Esta ecuación permite obtener el valor de los momentos de la variable aleatoria $\mathbb{M}$; no es difícil observar que los momentos de orden impar son cero.

2.3.1. Varianza bicuadrática

A partir de la ecuación (13) se obtienen los momentos de la variable M, y a partir de estos, usando la ecuación (12), se hallan valores de la varianza bicuadrática para $k = 9, 6, 3$, resultados que se presentan en la tabla 1.

TABLA 1: Valores de la varianza bicuadrática para una distribución $N(0,1)$.

	Valor de k		
	3	6	9
$\gamma_{\psi_{XX}}$	0,491377330220	0,066819506434	0,027637604055

2.4. Coeficiente de correlación bicuadrático

Si se define el vector aleatorio $\mathbb{M} = (M_1, M_2)^T$, con $\mathbb{M} = I_{\{|X|<l,|Y|<l\}}\mathbb{X}$, $\mathbb{X} = (X, Y)$ vector aleatorio con distribución $N(0, 0, 1, 1, \rho)$, la función de covarianza bicuadrática de $\mathbb{X}$ en función del vector $\mathbb{M}$ está dada por,

$$\gamma_{\psi_{XY}} = \frac{\frac{E_{1,1}}{l^2} - 4\frac{E_{3,1}}{l^4} + 2\frac{E_{5,1}}{l^6} + 4\frac{E_{3,3}}{l^6} - 4\frac{E_{5,3}}{l^8} + \frac{E_{5,5}}{l^{10}}}{\left[1 - 6\frac{E_2}{l^2} + 5\frac{E_4}{l^4}\right]^2} \tag{14}$$

con $E_{r,h} = E(M_1^r M_2^h)$, $E_r = E(M_1^r) = E(M_2^r)$, para r y h enteros[5]. Análogo a lo realizado para el caso de la varianza bicuadrática, la distribución del vector aleatorio $\mathbb{M}$ es resultado de truncar una distribución normal bivariada; de lo anterior la función de densidad conjunta de $\mathbb{M}$ está dada por $f_{\mathbb{M}}(m_1, m_2) = \frac{\phi(m_1,m_2)}{c_2} I_{\{|X|<l,|Y|<l\}}$, donde $\phi(\cdot,\cdot)$ es la función de densidad conjunta de una distribución $N(0, 0, 1, 1, \rho)$, y $c_2 = P(|X| < l, |Y| < l)$.

[5] $E(M_1^r) = E(M_2^r)$, dado que X y Y tienen la misma distribución.

2.4.1. Covarianza bicuadrática - caso $\rho = 0$

Si se supone $\rho = 0$, de la definición del vector $\mathbb{M} = (M_1, M_2)$, se sigue que las variables aleatorias M_1 y M_2 son independientes. Este resultado conlleva a que la ecuación (14) se transforme en

$$\gamma_{\psi_{XY}/\rho=0} = \frac{\frac{E^2(M_1)}{l^2} - 4\frac{E(M_1^3)E(M_1)}{l^4} + 2\frac{E(M_1^5)E(M_1)}{l^6}}{\left[1 - 6\frac{E_2}{l^2} + 5\frac{E_4}{l^4}\right]^2} + \frac{4\frac{E^2(M_1^3)}{l^6} - 4\frac{E(M_1^5)E(M_1^3)}{l^8} + \frac{E^2(M_1^5)}{l^{10}}}{\left[1 - 6\frac{E_2}{l^2} + 5\frac{E_4}{l^4}\right]^2} = 0 \tag{15}$$

El denominador es distinto de cero, y en el numerador sólo se tienen momentos de orden impar. Dado que la función de densidad de M_1 es simétrica alrededor de cero, los valores esperados en el numerador son cero, de donde se tiene $\gamma_{\psi_{XY}/\rho=0} = 0$.

2.4.2. Covarianza bicuadrática - caso $\rho \neq 0$

A partir de la ecuación (14), el camino por seguir para el caso $\rho \neq 0$, es calcular los momentos conjuntos del vector $\mathbb{M}$, tarea que se realiza usando la función generatriz. El valor de los momentos univariados de M_1 y M_2, ya se desarrollaron en la sección 2.3.

El trabajo con distribuciones normales truncadas no es nuevo. Pearson inicialmente trabajó en los años de 1930 sobre estas distribuciones, con el propósito de generar algunas tablas (tomado de Rosenbaum 1961, p. 405); posteriormente trabajaron sobre este tipo de distribuciones Cohen (1955), Singh (1960), Rosenbaum (1961), Tallis (1961), Finney (1962) y Khatri & Jaiswal (1963). Si se nota $\mathbf{m} = (m_1, m_2)^T$ y $\mathbf{t} = (t_1, t_2)^T$, la función generatriz del vector $\mathbf{M}$ está dada por

$$G_{\mathbb{M}}(\mathbf{t}) = E\left(e^{\mathbf{t}^T\mathbb{M}}\right) = \frac{e^{\frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}}}{c_2}\int_{-l}^{l}\int_{-l}^{l}\frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)}e^{-\frac{1}{2}\left[(\mathbf{m}-\Sigma\mathbf{t})^T\Sigma^{-1}(\mathbf{m}-\Sigma\mathbf{t})\right]}dm_1dm_2$$

donde $|\cdot|$ indica determinante[6]. Para hacer más corta la escritura, las derivadas de la función generatriz se notan como $D^{(h,r)} = \frac{\partial^{r+h}G_{\mathbb{M}}(\mathbf{t})}{\partial t_2^r \partial t_1^h}$. Los momentos conjuntos de orden r y h de $\mathbb{M}$ (r y h enteros nonegativos), alrededor al origen, están dados por

$$E(M_1^h M_2^r) = D^{(h,r)}\Big|_{t_1=0,t_2=0} \tag{16}$$

[6]El valor de acotamiento $l = k \times 0,67448975$ usado en el cálculo de varianza bicuadrática, sección 2.2.

A partir de que la distribución normal cumple con las condiciones de regularidad (ver Bickel & Docksum 1977, p. 378), se tiene que

$$D^{(1,0)} = G_{\mathbb{M}}(\mathbf{t})(t_1 + \rho t_2) - \frac{\beta_1(l,t_1,t_2) - \beta_2(-l,t_1,t_2) - \rho\left[\beta_3(l,t_2,t_1) - \beta_4(-l,t_2,t_1)\right]}{c_2(2\pi)^{\frac{1}{2}}} \tag{17}$$

con

$$\beta_j(v,u,w) = e^{-\frac{v^2}{2(1-\rho^2)}+vu+\frac{\left[\rho v+(1-\rho^2)w\right]^2}{2(1-\rho^2)}} P(-|v| < \xi_j < |v|) \quad j = 1,2,3,4$$

donde ξ_1, ξ_2, ξ_3 y ξ_4 son variables aleatorias cuyas distribuciones se presentan a continuación:

$$\xi_1 \sim N\left(\rho l + (1-\rho^2)t_2; 1-\rho^2\right), \qquad \xi_2 \sim N\left(-\rho l + (1-\rho^2)t_2; 1-\rho^2\right)$$
$$\xi_3 \sim N\left(\rho l + (1-\rho^2)t_1; 1-\rho^2\right) \quad \text{y} \quad \xi_4 \sim N\left(-\rho l + (1-\rho^2)t_1; 1-\rho^2\right)$$

y para $h \geq 2$ y $r \geq 1$ se tiene

$$D^{(h,r)} = r\rho D^{(h-1,r-1)} + (h-1)D^{(h-2,r)} + D^{(h-1,r)}(t_1 + \rho t_2) - \frac{\left[l^{h-1}(\beta_1 + (-1)^h\beta_2)^{(0,r)} + \rho l^r(\beta_3 - (-1)^r\beta_4)^{(h-1,0)}\right]}{c_2(2\pi)^{\frac{1}{2}}} \tag{18}$$

donde

$$\frac{\partial^{r+s}(\beta_1 \pm \beta_2)}{\partial t_2^r \partial t_1^s} = (\beta_1 \pm \beta_2)^{(s,r)} \qquad \text{y} \qquad \frac{\partial^{m+n}(\beta_3 \pm \beta_4)}{\partial t_2^m \partial t_1^n} = (\beta_3 \pm \beta_4)^{(n,m)}$$

2.4.3. Valores de $\varrho_{\psi_{XY}}$

A partir de los desarrollos mostrados en la sección 2.4.2, se obtienen los valores de los momentos conjuntos del vector $\mathbb{M}$. Una vez calculados estos, se consiguieron los valores de la covarianza bicuadrática utilizando la ecuación (14), valores que hacen posible calcular el coeficiente de correlación bicuadrático mediante la ecuación (11). Los valores de la varianza bicuadrática ya habían sido obtenidos (ver sección 2.3.1).

Los resultados muestran que el valor de $\varrho_{\varphi_{XY}}(\rho) \to \rho$ cuando k crece[7] (ver figuras 1, 2 y tabla 2). Para $k = 9$ las diferencias entre $\varrho_{\varphi_{XY}}$ y ρ son de tal magnitud que las líneas se superponen, razón por la cual se muestra una ampliación de la misma gráfica en el cuadrante (0,4; 0,7) (ver figura 2).

[7]La línea delgada indica la identidad, es decir $\rho = \rho$, los puntos el valor de ϱ; lo ideal es que $\varrho \approx \rho$, es decir que las dos líneas coincidan.

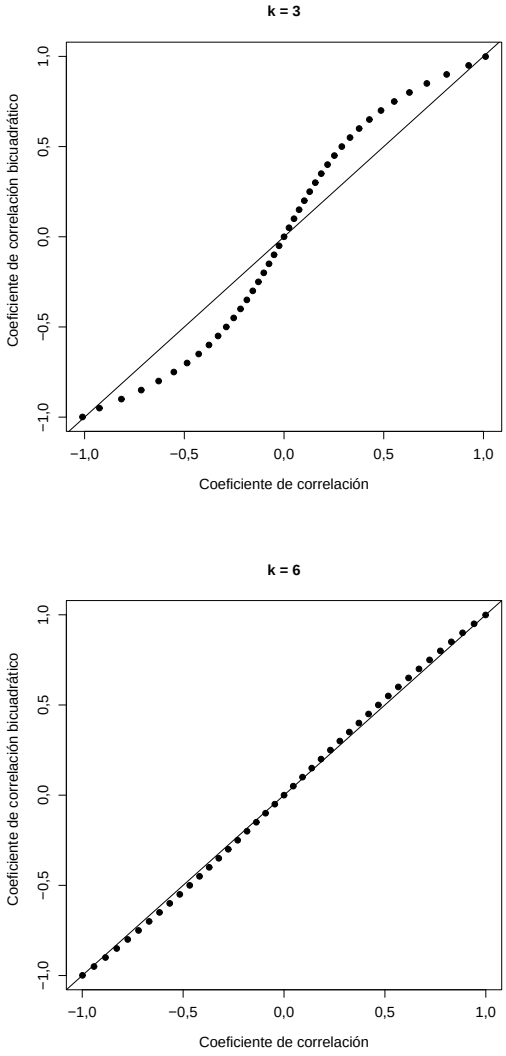


FIGURA 1: Valores de ϱ para $k = 3$ y $k = 6$.

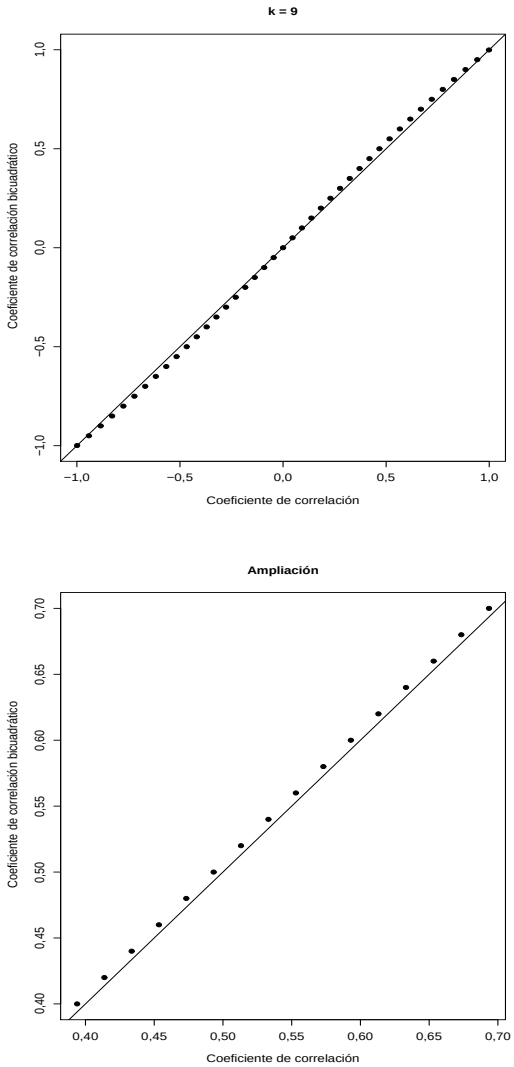


FIGURA 2: Valores de ϱ para $k = 9$.

Tabla 2: Valores de ϱ de acuerdo con los valores de ρ y k.

	Valor de k		
Valor de ρ	3	6	9
0,001	0,000497	0,000914	0,000982
0,100	0,049966	0,091461	0,098207
0,200	0,101565	0,183422	0,196523
0,300	0,156766	0,276384	0,295054
0,400	0,218223	0,370858	0,393911
0,500	0,289691	0,467366	0,493201
0,600	0,376515	0,566446	0,593034
0,700	0,486181	0,668655	0,693520
0,800	0,628749	0,774574	0,794769
0,900	0,815515	0,884808	0,896891
0,999	1,010647	0,998821	0,998964

3. Conclusiones

Asumiendo que el estimador del correlación bicuadrático presenta un comportamiento análogo al comportamiento del funcional aquí estudiado, los resultados sugieren que el estimador bicuadrático subestima el valor ρ cuando $\rho > 0$, y sobreestima su valor cuando $\rho < 0$.

[Recibido: marzo de 2010 — Aceptado: octubre de 2010]

Referencias

Beaton, A. & Tukey, J. (1974), 'The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data', *Technometrics* **16**(2), 147–185.

Bickel, P. & Docksum, K. (1977), *Mathematical Statistics, Basic Ideas and Selected Topics*, Holden-day Inc., San Francisco.

Cohen, C. (1955), 'Restriction and Selection in Samples from Bivariate Normal Distributions', *Journal of the American Statistical Association* **50**(271), 884–893.

Finney, D. (1962), 'Cumulants of Truncated Multi-Normal Distributions', *Journal of the Royal Statistical Society, serie B* **24**(2), 535–536.

Khatri, C. & Jaiswal, M. (1963), 'Estimation of Parameters of a Truncated Bivariate Normal Distribution', *Journal of the American Statistical Association* **58**(302), 519–526.

Lax, D. (1975), An Interim Report of a Monte Carlo Study of Robust Estimators of Widthers, Technical report, Department of Statistics, Princeton University.

Rosenbaum, S. (1961), 'Moments of a Truncated Bivariate Normal Distribution', *Journal of the Royal Statistical Society* **23**(2), 405–408.

Singh, N. (1960), 'Estimation of Parameters of a Multivariate Normal Population from Truncated and Censured Samples', *Journal of the Royal Statistical Society, serie B* **22**(2), 307–311.

Tallis, G. (1961), 'The Moment Generating Function of the Truncated Multinormal Distribution', *Journal of the Royal Statistical Society, serie B* **23**(1), 223–229.

Valcárcel, H. (2007), Propuesta de una función de autocorrelación con base en la función bicuadrática, Trabajo de grado, Departamento de Estadística, Universidad Nacional de Colombia, Bogotá.

Wei, W. (2006), *Time Series Analysis: Univariate and Multivariate Methods*, second edn, Addison Wesley, Boston.

Revista Colombiana de Estadística
Diciembre 2010, volumen 33, no. 2, pp. 307 a 319

The Size Problem of Bootstrap Tests when the Null is Non- or Semiparametric

El problema del tamaño de los contrastes bootstrap cuando la hipótesis nula es No- o semiparamétrica

Jorge Barrientos-Marín[1,a], Stefan Sperlich[2,b]

[1]Departamento de Economía, Facultad de Ciencias Económicas, Universidad de Antioquia, Medellín, Colombia

[2]Institut für Statistik und Ökonometrie, Georg August Universität Göttingen, Göttingen, Germany

Abstract

In non- and semiparametric testing, the wild bootstrap is a standard method for determining the critical values of tests. If the null hypothesis is also semi- or nonparametric, then we know that at least asymptotically oversmoothing is necessary in the pre-estimation of the null model for generating the bootstrap samples. See Härdle & Marron (1990, 1991). However, in practice this knowledge is of little help. In this note we highlight that this bandwidth choice problem can become quite serious. As an alternative, we briegly discuss the possibility of subsampling.[1]

***Key words*:** Bandwidth choice, Bootstrap tests, Nonparametric specification tests.

Resumen

En contrastes no- y semiparamétricos el *wild-bootstrap* es un método estándar para la determinación de los valores críticos de los estadísticos de contrastes. Si la hipótesis nula es no o semiparamétrica, sabemos que al menos asintóticamente es necesaria una sobre-suavización en la pre-estimación del modelo bajo la nula para generar las muestras bootstrap, ver por ejemplo Härdle & Marron (1990, 1991).

No obstante, en la práctica este conocimiento es de poca o ninguna ayuda. En este artículo, ponemos de manifiesto que el problema de la selección de la banda de suavidad para procedimientos de contraste puede ser muy serio. Como alternativa, discutimos brevemente la posibilidad de usar submuestras.

***Palabras clave*:** ancho de banda, contrastes de especificación no-paramétricos, contrastes bootstrap.

[a]Profesor. E-mail: jbarr@economicas.udea.edu.co

[b]Professor. E-mail: stefan.sperlich@wiwi.uni-goettingen.de

[1]The authors gratefully acknowledge very helpful comments of two anonymous referees as well as financial support from the Spanish MTM2008-03010 and the Deutsche Forschungsgemeinschaft FOR916.

1. Introduction

In both applied and mathematical statistics, non- and semiparametric specification testing is still quite a popular research field. Unfortunately, only a few papers address the problem of choosing an appropriate smoothing parameter. This a problem is fundamental for the reasonable use of these methods. There has been a growing amount of literature on adaptive testing where the adaptiveness refers to the smoothness of the alternative and deals with the smoothing of the test or the alternative.

However, these papers typically concentrate on testing problems where the null hypothesis is fully parametric. Here we are interested in testing qualitative restrictions, i.e. where the null hypothesis is semi- or nonparametric; think e.g. of additivity tests. When bootstrap is used to determine the critical value, these tests entail at least one more parameter choice problem: pre-estimating the model under the null hypothesis to later generate the bootstrap samples. This is necessary as in most cases the bandwidths for the estimation and the bootstrap should have different rates. See Härdle & Marron (1990, 1991). As in practical applications this problem has hardly been addressed, in most published procedures for testing or constructing confidence bands with a semi- or nonparametric null hypothesis, there is no guarantee that the bands meet the nominal coverage probability. This has been confirmed in the work of Dette, von Lieres, Wilkau & Sperlich (2005). In the latter paper, the problem is avoided by using subsampling.

To study the problem outlined in more detail, we concentrate on the problem of testing additivity. We limit ourselves to two test statistics proposed in Dette et al. (2005) and Roca & Sperlich (2007) but we extended this to different modifications including subsampling. The aim is not to find the most efficient additivity test or to propose new ones. Our focus is only directed at highlighting the size problem when the null hypothesis and the resampling method are non- or semiparametric. After a review of the additivity tests considered here, we study some of the typically proposed procedures for bandwidth choice. Unfortunately, we have not found a generally valid method. Our conclusion is that further research is necessary to find a proper bootstrap bandwidth.

2. Estimators and test statistics for additive models

Assume we face (not necessarily) independent and identically distributed (i.i.d.) data $\{(X_i, Y_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$, where

$$Y_i = m(X_i) + u_i \qquad i = 1, 2, \ldots, n, \tag{1}$$

with $m : \mathbb{R}^d \to \mathbb{R}$ an unknown function of interest, $m(x) = E(Y \mid X = x)$, and u_i i.i.d. random errors with $E[u_i] = 0$ and finite variance $\sigma^2(x_i)$. The internalized

Nadaraya-Watson estimator is defined as

$$\widehat{m}_k(x) = \sum_{i=1}^{n} v_k(x, X_i) Y_i, \text{ with } v_k(x, X_i) = \left(\widehat{f}_k(X_i)\right)^{-1} \mathbf{K}_k(x - X_i) \qquad (2)$$

where $\widehat{f}_k(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_k(X_j - X_i)$ is a kernel density estimator with a multiplicative kernel, i.e. for $w = (w_1, \ldots, w_d) \in I\!R^d$ we think of $\mathbf{K}_k(w) = \prod_{\alpha=1}^{d} K_k(w_\alpha)$, $K_k(w_\alpha) = k^{-1} K\left(w_\alpha k^{-1}\right)$. Commonly, the kernel is assumed to be Lipschitz continuous with compact support and $\int |K(x)|\, dx < \infty, \int K(x)dx = 1$. Furthermore, k is the bandwidth, assumed to go to zero for sample size n going to infinity, but nk_n^d going to infinity. Let V_k be the $n \times n$ matrix whose (j, i) element is $v_k(X_j, X_i)$, then $\widehat{m}(x) = V_k(x)\, Y$.

We are interested in the additive model, which we write in terms of

$$E(Y \mid X = x) = m_S(x) = \psi + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha) \qquad (3)$$

where we set $E_{X_\alpha}\{m_\alpha(X_\alpha)\} = \int m_\alpha(x) f_\alpha(x) dx = 0\ \forall \alpha$ for identification. Here, m_α, $\alpha = 1, \ldots, d$ are the marginal impact functions for each regressor. Therefore, ψ is a constant equal to the unconditional expectation of Y. Writing $m(X) = m_\alpha(X_\alpha) + m_{-\alpha}(X_{-\alpha})$ where $X_{-\alpha}$ is the vector X of all explanatory variables without X_α, i.e. $X_{-\alpha} = \left(X_{i1}, \ldots, X_{i(\alpha-1)}, X_{i(\alpha+1)}, \ldots, X_{id}\right)$, we can use the identification condition directly to estimate m_α. The so called marginal integration idea is based on that for x_α fix we have

$$E_{X_{-\alpha}}[m(x_\alpha, X_{-\alpha})] = \int m(x_\alpha, x_{-\alpha}) f_{-\alpha}(x_{-\alpha}) \prod_{\beta \neq \alpha} dx_\beta = \psi + m_\alpha(x_\alpha)$$

Substituting for $m(\cdot)$ a nonparametric pre-estimator such as the one given in (2), a sample average for the expectation, and for ψ simply $\widehat{\psi} = \frac{1}{n}\sum_{i=1}^{n} y_i$ gives

$$\widehat{m}_\alpha(x_\alpha) = \sum_{i=1}^{n} w_{\alpha h}(x_\alpha, X_{i\alpha})\, Y_i$$

where for a bandwidth h (the one fixing the smoothness of our H_0 model)

$$w_h(x_\alpha, X_{\alpha i}) = K_h(x_\alpha - X_{i\alpha}) \frac{\widehat{f}_{-\alpha}(X_{i,-\alpha})}{\widehat{f}(X_{i\alpha}, X_{i,-\alpha})} \qquad (4)$$

Finally, we set $\widehat{m}_S(X_j) = \widehat{\psi} + \sum_{\alpha=1}^{d} \widehat{m}_\alpha(X_{j\alpha})$ for each $j = 1, 2, \ldots, n$. Note that defining $W_h = \sum_{\alpha=1}^{d} W_{\alpha h}(x_\alpha)$ with $W_{\alpha h}(x_\alpha)$ being the $n \times n$ matrices with $w_{\alpha h}(X_j, X_i)$ as elements, one has $\widehat{m}_S(x) = \psi + W_h(x)\, Y$.

As mentioned before, we do not introduce new testing procedures but rather study two modified statistics which have already been studied in the above mentioned papers, and which performed excellently in the study by Roca & Sperlich

(2007) though in a different context. The null hypothesis of interest is $H_0 : m(\cdot) = m_S(\cdot)$ versus $H_1 : m(\cdot) \neq m_S(\cdot)$. We consider the following two test statistics:

$$\tau_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}\left(X_i\right) - \widehat{m}_S\left(X_i\right)\right)^2 w(X_i)$$

$$\tau_2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{nk^d}\sum_{j=1}^{n}\mathbf{K}_k\left(X_i - X_j\right)\left(Y_j - \widehat{m}_S\left(X_j\right)\right)\right]^2 w(X_i)$$

where $\widehat{e}_i = Y_i - \widehat{m}_S\left(X_i\right)$, i.e. the residuals under the null hypothesis, and $\widehat{u}_i = Y_i - \widehat{m}\left(X_i\right)$, the residuals without restrictions. We included also a weight function $w(\cdot)$ which typically is just used for trimming at the boundaries or regions where data are sparse. Note that in our simulation study we will make use of the trimming at the boundaries. Obviously, τ_1 calculates directly the integrated squared difference between the null and alternative models. Alternatively, τ_2 seeks to mitigate the bias problem inherited from the estimate $\widehat{m}$, which suffers from the "curse of dimensionality". In Dette et al. (2005) it is proved that, for both tests τ_j, the $nk^{\frac{d}{2}}\left(\tau_j - \mu_j\right)$ converge under the null to a normal variable with mean zero and variances v_j^2 for $j = 1, 2$ with

$$\mu_1 = E_{H_0}\left\{\tau_1\right\} = \frac{1}{nk^d}\int \sigma^2(x)w(x)dx \int \mathbf{K}^2(x)dx + o\left(\frac{1}{nk^d}\right)$$

$$\mu_2 = E_{H_0}\left\{\tau_2\right\} = \int \left(\mathbf{K} * \mathbf{K}\right)^2(x)\,dx \int \sigma^2(x)f^2(x)w(x)\,dx$$

and

$$v_1^2 = Var_{H_0}\left\{\tau_1\right\} = 2\int \sigma^4(x)w^2(x)dx \int \left(\mathbf{K} * \mathbf{K}\right)^2(x)\,dx$$

$$v_2^2 = Var_{H_0}\left\{\tau_2\right\} = \int \sigma^4(x)f^4(x)w^2(x)\,dx$$

All tests have been proven to be consistent in the sense that under the alternative they converge with n to infinity. Let us also mention that we have studied many more test statistics, e.g. those given in Dette et al. (2005) or Roca & Sperlich (2007) but not presented here. These, however, showed even less satisfactory performance, so we have skipped them in our presentation.

3. The resampling

Asymptotic expressions are of little help in practice for several reasons: Bias and variance contain unknown expressions which have to be estimated nonparametrically, and the convergence rate is quite slow for large d. For this reason, it is common to use resampling-mostly bootstrap-methods to approximate the critical value for the particular sample statistic. These can be bootstrap methods or subsampling procedures. Unfortunately, for the bootstrap it is not known how to

choose the smoothing parameter in practice for the pre-estimation of the model that is used to generate the bootstrap samples. From theory, it is known that one should somewhat oversmooth.

We give the general bootstrap procedure first and then discuss the details:

1. With bandwidth h, calculate the estimate $\widehat{m}_S$ under the null hypothesis of additivity and its resulting residuals $\widehat{e}_i$, $i = 1, \ldots, n$.

2. With bandwidth k, calculate the estimator $\widehat{m}$ for the conditional expectation without the additivity restriction, and the corresponding residuals $\widehat{u}_i$, $i = 1, \ldots, n$.

3. With the results from step 1 and 2, we can calculate our test statistics τ_1 and τ_2.

4. Repeat step 1 with a bandwidth h_b. We call the outcome $\widehat{m}_S^b$, respectively $\epsilon_i = Y_i - \widehat{m}_S^b(X_i)$, $i = 1, \ldots, n$.

5. Draw random variables e_i^* with $E[(e_i^*)^j] = u_i^j$ (respectively $\widehat{e}_i^j$ or ϵ_i^j, see discussion below) for $j = 1, 2, 3$ (respectively $j = 1, 2$, see below again). Set $Y_i^* = \widehat{m}_S^b(X_i) + e_i^*$, $i = 1, \ldots, n$, i.e. generate wild bootstrap samples. Repeat this B times. This defines B different bootstrap samples $\{(X_i, Y_i^{*,b})\}_{i=1}^n$, $b = 1, \ldots, B$.

6. For each bootstrap sample from steps 4 and 5, calculate the test statistics $\tau_j^{*,b}$, $j = 1, 2$, $b = 1, \ldots, B$. Then, for each test statistic τ_j, $j = 1, 2$, the critical value is approximated by the corresponding quantiles of the distribution of the B bootstrap analogues: $F^*(u) = \frac{1}{B} \sum_{b=1}^{B} I\{\tau_j^{*,b} \leq u\}$. Recall that they are generated under the null hypothesis.

In step 2, the bandwidth k has simply to obey the different assumptions required for each specific test. It can be chosen in such a way that it maximizes the power of the test for a given size. Therefore, different from Dette et al. (2005) we apply the adaptive testing approach introduced in Spokoiny (1998, 1996). He considers simultaneously a family of tests $\{\tau^k, k \in \mathfrak{K}\}$, where $\mathfrak{K} = \{k_1, k_2, \ldots, k_P\}$ is a finite set of reasonable bandwidths. The theoretical maximal number P depends on n, but is of no practical relevance. For details, see Horowitz & Spokoiny (2001). They define

$$\tau^{\max} = \max_{k \in \mathfrak{K}} \frac{\tau^k - E_0[\tau^k]}{Var^{1/2}[\tau^k]}$$

where $E_0[\cdot]$ indicates the expectation under H_0. A particularity of the resampling analogues of $\tau^{\max}$ is that one first needs to calculate the resampling statistics $(\tau^k)^{*,b}$ for all $k \in \mathfrak{K}$ to afterwards get $(\tau^{\max})^{*,b}$. Note that for each k, the empirical moments of the resampling statistics $(\tau^k)^{*,b}$ can be used as a substitute for $E_0[\tau^k]$, respectively $Var^{1/2}[\tau^k]$, in practice.

In step 5, the wild bootstrap (see Härdle & Mammen 1993) it is let open which residuals should be taken $\widehat{u}_i$, $\widehat{e}_i$ or ϵ_i. While theory says clearly that the best

power can be reached when taking the residuals of the alternative, i.e. $\widehat{u}_i$, our simulations (not shown) confirm the findings of Dette et al. (2005) that in practice ϵ_i should be taken. Next, it is often sufficient if we allow for heteroscedasticity of an unknown form using $e_i^* = \varepsilon_i \epsilon_i$, where the ε_i are i.i.d., drawn either from the golden-cut distribution, i.e.

$$\varepsilon_i = \begin{cases} -(\sqrt{5}+1)/2 & \text{with probability } \ p = (\sqrt{5}+1)/(2\sqrt{5}) \\ (\sqrt{5}+1)/2 & \text{with probability } \ 1-p \end{cases}$$

or from the Gaussian normal $N(0,1)$. This answers the question up to order the moment of the bootstrap errors have to coincide with the residual moments. In the simulation section, we will compare golden-cut with Gaussian bootstrap.

In step 4, bandwidth h_b has to be chosen along the arguments of Härdle & Marron (1990, 1991): For the mean of $\widehat{m}_h(x) - m(x)$ under the conditional distribution of $Y_1, \ldots, Y_n \mid X_1, \ldots, X_n$, respectively of $\widehat{m}_h^*(x) - \widehat{m}_{h_b}(x)$ under the conditional distribution of $Y_1^*, \ldots, Y_n^* \mid X_1, \ldots, X_n$, it is well known that

$$E^{Y|X}(\widehat{m}_h(x) - m(x)) \approx h^2 \frac{\mu(K)}{2} m''(x) \tag{5}$$

$$E^{*}(\widehat{m}_h^*(x) - \widehat{m}_h(x)) \approx h^2 \frac{\mu(K)}{2} \widehat{m}''_{h_b}(x) \tag{6}$$

where $\mu(K) = \int u^2 K(u) du$. Obviously, we need that $\widehat{m}''_{h_b}(x) - m''(x) \to 0$. The optimal bandwidth h_b for estimating the second derivative must to be larger (in rates) than bandwidth h for estimating the function itself. We can even give the optimal rate. For example, the optimal rate to estimate m''_S is of the order $n^{-1/9}$ (instead of $n^{-1/5}$), an observation we make use of in our simulation studies. There it will be seen that the typical comment *h_b has to be oversmoothing*, is unhelpful in practice. Intuitively, one may think that a proper choice for h_b depends strongly on h. This might be true numerically, looking at equations (5) and (6) in the asymptotics the "h-effect" seems to cancel out as long as h/h_b goes to zero (a necessary condition for the consistency of bootstrap inference here) for n going to infinity. As one wants check whether the best possible additive model is an adequate fit, one therefore can concentrate on those bandwidth selectors for h which aim to optimize $\widehat{m}_S$ like cross validation or some plug-in methods do.

After all, it might be interesting to also have a look at subsampling as an alternative to bootstrapping (see Politis, Romano & Wolf 1999). Neumeyer & Sperlich (2006) introduce subsampling in a slightly context, other than that we discuss here, because there the bootstrap failed. There exists an automatic choice of the adequate subsample size m. As we remodeled this method to serve as a procedure for finding h_b, we introduce subsampling and the automatic choice of the subsample size m in more detail:

Let $\mathcal{Y} = \{(X_i, Y_i) \mid i = 1, \ldots, n\}$ be the original sample, and denoted by $\tau(\mathcal{Y})$ the original statistic calculated from this sample, leaving aside index $j = 1, 2, 3$ for a moment. To determine the critical values we need to approximate

$$Q(z) = P\left(n\sqrt{k^d}\tau(\mathcal{Y}) \leq z\right) \tag{7}$$

Recall that under H_0 this distribution converges to an $N(\mu_j, v_j^2)$, for μ_j and v_j, $j = 1, 2$, see above. For finite sample size n, drawing B subsamples $\mathcal{Y}_b$ -each of size m- we can approximate Q under H_0 by

$$\widehat{Q}(z) = \frac{1}{B}\sum_{b=1}^{B} I\left(m\sqrt{k_m^d}\tau^{k_m}(\mathcal{Y}_m) \leq z\right) \tag{8}$$

Note that the awkward notation comes from we have to adjust all bandwidths for the new sample size m. For example, imagine $k = k_0 \cdot n^{-\delta}$ for k_0 being constant. Then, τ^{k_m} is calculated like τ but with bandwidth $k_m = k_0 n^{\delta} m^{-\delta}$.

Certainly, under the alternative H_1, both $n\sqrt{k^d}\tau(\mathcal{Y})$ and $m\sqrt{k_m^d}\tau^{k_m}(\mathcal{Y}_m)$ converge to infinity. When demanding $m/n \to 0$ guarantees that $n\sqrt{k^d}\tau(\mathcal{Y})$ converges (much) faster to infinity than the subsample analogues. Then, $\widehat{Q}$ underestimates the quantiles of Q, which yields the rejection of H_0.

The optimal m is actually a function of the level α. Again, one applies resampling methods: Draw some pseudo sequences $\mathcal{Y}^{*,l}$, $l = 1, \ldots, L$ of $\mathcal{Y}$ of size n with the same distribution as $\mathcal{Y}$. For the desired level α, test $H_0^* : m(x) - m_S(x) = \widehat{m}(x) - \widehat{m}_S(x)$ the same way as you want to test $H_0 : m(x) = m_S(x)$, i.e. applying your particular test statistic to H_0^* and using subsampling. From the L repetitions you can determine the empirical rejection level (estimated size) for your given α. Now, find an m such that this empirical rejection level is $\approx \alpha$. In practice, you choose from a grid of possible m the one whose estimated rejection level for H_0^* is closest to α from below. Note that H_0^* is always true up to an estimation error that should be almost the same as in your original test. The only drawback of this procedure is the enormous computational effort. For further details and examples, see Politis et al. (1999) or Delgado, Rodríguez & Wolf (2001).

4. Simulation results

We give here only a summary of our large simulation study. The model considered is as follows: As in Dette et al. (2005), we draw $n = 100$ i.i.d. $X \in \mathbb{R}^3$ with

$$X_i \sim N(0, \Sigma_X) \quad \text{with} \quad \Sigma_X = \begin{pmatrix} 1 & 0.2 & 0.4 \\ 0.2 & 1 & 0.6 \\ 0.4 & 0.6 & 1 \end{pmatrix}$$

to generate

$$Y_i = X_{1,i} + X_{2,i}^2 + 2\,\sin(\pi X_{3,i}) + vX_{2,i}X_{3,i} + e_i, \quad i = 1, \ldots, n$$

with i.i.d. standard normal errors e_i, $v = 0$ being an additive separable model, or $v = 2$ for an alternative.

In both test, statistics we use the weighting function $w(\cdot)$ for a possible trimming: We cut the outer 5% or nothing (0%) of the sample, where "outer" refers to the tails of the explanatory variables. This is done to get rid of the boundary

effects in the statistics. To speed up our simulation studies, the presented results are calculated from 250 replications using only 200 bootstrap samples (or subsamples respectively). We used the multiplicative quartic kernel throughout but note that we know from our simulations in Dette et al. (2005) as well as from three years simulation experiences for the studies in Barrientos (2007), that the results change hardly for larger bootstrap samples.

We first looked for an average cross validation bandwidth h, which turned out to be $h_{opt} = 0.78$ for the direction of interest, and $6h_{opt}$ for the nuisance directions, cf. Dette et al. (2005). This was done not only for computational reasons but also because otherwise the size of the tests would also depend on the randomness induced by the estimation of h. For the k-adaptive test procedure, k ran over an equispaced grid of 10 bandwidths from $k_{\min} = 0.1 \cdot range(X_1)$ to $k_{\max} = range(X_1)$.

We will study now the results for several choices of h_b with different bootstrap generating methods, i.e. golden-cut vs. Gaussian bootstrap errors. To have h_b as a function of h, to take also into account $h/h_b \to 0$, and validate the rate $n^{-1/9}$ (motivated above) we set $h_b = hn^{1/5-1/\kappa}$ and try different $\kappa \leq 9$.

Table 1 shows the results for the k-adaptive bootstrap tests. We compare the size and power for different h_b, golden-cut vs Gaussian bootstrap, trimming boundary effects vs no trimming, and finally also a bit τ_1 vs τ_2 (though the latter is not the aim of this paper).

First, the results basically show that the size problem is not solved simply by different smoothing in the pre-estimation. Oversmoothing, in contrast to the theoretical findings, seems to go in the wrong direction, at least for τ_1. In particular, the hope that the ideas of Härdle & Marron (1990, 1991) (see equations (5) and (6)) might give us a hint or even provide a rule of thumb for the choice of h_b is not confirmed here.

Second, following to some extent the findings of (Delgado et al. 2001), we find a clear improvement for the Gaussian compared to the golden-cut bootstrap. Actually, when using the golden-cut method, then τ_1 does not hold the size for several h_b (κ respectively). Even worse, it rejects more often under H_0 than it does under H_1. This phenomenon is not observed for the simpler Gaussian wild bootstrap.

Third, boundary effects seem not to be the reason of our size and power problems. Surely, we get different numerical results for different weighting (i.e. trimming) functions, but cutting at the boundaries does not substantially change our general findings.

Finally, it is obvious that τ_2 outperforms τ_1 throughout. When recalling the motivation of the construction of τ_2, cf. Neumeyer & Sperlich (2006) and Roca & Sperlich (2007), it is obvious that the size problem comes from the bias rather from the variance. Or, in other words, bootstrap can capture pretty well the variance of a statistic but not its bias. There are two possible reasons for the surprising fact that τ_1 sometimes rejects more under H_0 than under H_1. First, while it is clear that the bias distorts the rejection level, it is not clear in what direction; moreover, the distortion effect certainly changes with the true underlying data generation

TABLE 1: Rejection levels of the two k-adaptive test statistics with and without trimming. Critical values are determined with golden-cut respectively Gaussian wild bootstrap, using $h_b = hn^{1/5-1/\kappa}$ for the pre-estimation.

			Golden Cut				Gaussian Residuals			
			H_0 $(v=0)$		H_1 $(v=2)$		H_0 $(v=0)$		H_1 $(v=2)$	
Trim	α%	κ	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2	τ_1	τ_2
0%	5	4	.000	.024	.016	.364	.000	.024	.004	.440
		5	.012	.020	.016	.332	.008	.024	.020	.380
		6	.056	.020	.028	.344	.056	.020	.100	.376
		7	.136	.024	.044	.332	.124	.028	.168	.368
		8	.196	.016	.072	.360	.172	.020	.244	.384
		9	.244	.016	.088	.388	.216	.016	.320	.396
	10	4	.000	.068	.064	.572	.012	.088	.068	.672
		5	.024	.052	.068	.464	.024	.060	.076	.508
		6	.100	.048	.084	.440	.032	.036	.076	.492
		7	.188	.040	.092	.452	.036	.036	.096	.464
		8	.252	.040	.104	.468	.056	.036	.108	.488
		9	.308	.040	.124	.476	.068	.036	.132	.508
5%	5	4	.004	.024	.060	.352	.004	.020	.008	.420
		5	.024	.020	.048	.324	.028	.020	.040	.348
		6	.112	.016	.068	.336	.096	.020	.144	.360
		7	.180	.016	.100	.316	.164	.020	.236	.348
		8	.276	.012	.132	.348	.216	.016	.328	.360
		9	.360	.012	.152	.372	.292	.016	.436	.388
	10	4	.016	.072	.108	.568	.064	.088	.120	.664
		5	.036	.048	.100	.460	.052	.052	.108	.500
		6	.164	.044	.116	.428	.068	.036	.104	.460
		7	.256	.036	.144	.448	.092	.036	.144	.460
		8	.356	.036	.184	.460	.120	.032	.176	.484
		9	.432	.036	.228	.476	.128	.040	.224	.504

process. Second, making the tests $k-$adaptive entails a normalization by the estimated variance. In the unfortunate situation where the variance estimation is getting larger, the power of the test decreases. Both effects together lead here to the counter-intuitive performance of τ_1.

In the last section we introduced subsampling as an alternative resampling method to bootstrap. Therefore, we also provide a simulation study where the critical values are approximated by subsampling, trying several subsample sizes m. Recall that the different subsample sizes have a similar effect here like it has the choice of h_b for bootstrap tests. The results are given in Table 2 for k-adaptive tests. For τ_1 we see here basically the same bad behavior we observed when using golden-cut bootstrap to determine the critical values. In contrast, τ_2 seems to

TABLE 2: Rejection levels of the two k-adaptive test statistics with and without trimming. Critical values are determined with subsampling, using subsamples of sizes m.

			H_0 $(v=0)$		H_1 $(v=2)$	
Trim	α%	m	τ_1	τ_2	τ_1	τ_2
0%	5	80	.000	.000	.000	.000
		70	.000	.000	.000	.000
		60	.056	.000	.020	.036
		50	.276	.020	.076	.292
		40	.516	.212	.168	.732
	10	80	.000	.000	.000	.000
		70	.020	.000	.016	.000
		60	.272	.008	.072	.144
		50	.584	.104	.256	.644
		40	.816	.476	.480	.912
5%	5	80	.000	.000	.000	.000
		70	.000	.000	.000	.000
		60	.016	.000	.000	.032
		50	.060	.020	.012	.276
		40	.152	.216	.024	.712
	10	80	.000	.000	.000	.000
		70	.004	.000	.000	.000
		60	.060	.008	.016	.164
		50	.200	.092	.024	.636
		40	.380	.460	.120	.908

work-though with less power than we observed when using Gaussian bootstrap, cf. Table 1.

Recall that our main focus is the size distortion of resampling tests. Therefore our last two studies are about the automatic choice of m in subsampling and h_b in (Gaussian) bootstrap, respectively.

A quite time consuming simulation study evaluating the automatic choice of m indicates that this procedure does unfortunately not work at all. Nevertheless, our last study is to apply this idea for getting an automatic choice of h_b. In order to do so, we first have to adjust the procedure for an automatic choice of the subsample size m to now find an adequate bootstrap bandwidth h_b.

This can be done as follows, described here in detail for τ_2. To make notation and calculation easier, we consider the non-k-adaptive version but fix $k = range(X_1)/2$. Let now $\{Y_i^*, x_i^*\}_{i=1}^n := \mathcal{Y}^*$ be a member of the pseudo sequence introduced above. Then, for testing $H_0^* : m(x) - m_S(x) = \widehat{m}(x) - \widehat{m}_S(x)$ with

sample $\mathcal{Y}^*$, an analogue to τ_2 would be

$$\tau_2^\sharp = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{nk^d}\sum_{j=1}^{n}\mathbf{K}_h(X_i^* - X_j^*)\left\{Y_j^* - \widehat{m}_S(X_j^*)\right\}\right.$$
$$\left. - \mathbf{K}_h(X_i^* - X_j)\left\{Y_j - \widehat{m}_S(X_j)\right\}\right]^2 w(X_i^*) \quad (9)$$

Other statistics are thinkable certainly, e.g.

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{nk^d}\sum_{j=1}^{n}\mathbf{K}_h(X_i - X_j^*)\{Y_j^* - \widehat{m}_S(X_j^*)\}\right.$$
$$\left. - \mathbf{K}_h(X_i - X_j)\{Y_j - \widehat{m}_S(X_j)\}\right]^2 w(X_i)$$

but they should all be asymptotically equivalent to (9). The procedure was performed with only $L = 100$ pseudo samples $\mathcal{Y}^*$. As the results varied widely we were forced either to enlarge L considerably or to reduce σ_e considerably. For computational reasons we decided on the second option and repeated the study with $\sigma_e = 0.1$.

Some results are summarized in Table 3. As et can be seen, this time we emphasize the possibility of undersmoothing much more. You first have to look at $\tau_2^\sharp$ to find the κ giving the rejection level closest to $\alpha = 5\%$ from below. Here, this is always $\kappa = 3$. Note that this might also change depending on the trimming, α, sample size, etc. It is important to understand that the lines of τ_2^* can always be calculated, i.e. without knowing the true data generating process. Therefore we call this method fully automatic. Now look at the lines for τ_2, the test of interest. Obviously, $\kappa = 3$ is indeed the best possible choice; it has the strongest power among all κ respecting the nominal level. This could be taken as indicating that our suggestion for selecting h_b works. Unfortunately, this method does not work that well for all possible α; specifically, it becomes quite incorrect for $\alpha \geq 10\%$. Even worse, it did not work for τ_1 (not shown).

5. Conclusions

Our main focus is the bootstrap and its size distortion in practice when the sample size is small or moderate. These points are illustrated along the popular problem of additivity testing. Naturally, one looks for an optimal trade-off between controlling for size under the null hypothesis H_0 and maximizing power. Even though these problems have already been discussed and studied in theory, as yet, it is unclear how to set the smoothing parameter for the bootstrap prior estimates in practice. We show that theory is not just unhelpful here; at present, a reasonable application of bootstrap tests of these kinds is questionable.

TABLE 3: Rejection levels of τ_2 and $\tau_2^\sharp$ for $\alpha = 5\%$, with and without trimming, using Gaussian bootstrap with $h_b = hn^{1/5-1/\kappa}$ for the pre-estimation, and $k = range(X_1)/2$.

			κ						
Trim			1	2	3	4	5	6	7
H_0	0%	$\tau_2^\sharp$	.012	.063	.028	.030	.032	.031	.029
$(v = 0)$		τ_2	.680	.392	.032	.012	.012	.012	.016
	5%	$\tau_2^\sharp$	.012	.062	.028	.030	.032	.031	.029
		τ_2	.676	.380	.024	.012	.012	.012	.020
H_1	0%	$\tau_2^\sharp$	.001	.019	.042	.022	.015	.011	.009
$(v = 2)$		τ_2	.972	.932	.632	.380	.272	.260	.264
	5%	$\tau_2^\sharp$	.001	.019	.042	.023	.015	.011	.010
		τ_2	.968	.936	.620	.368	.260	.252	.264

Further, we have shown that subsampling is an interesting alternative to bootstrap which in addition provides a procedure for the analogue problem of subsample size choices.

Finally we introduced the idea of extending the procedure of subsample size selection to smoothing parameter (h_b) selection in bootstrap testing problems. However, further research is necessary to provide reliable procedures for the nonparametric testing problems considered here.

[Recibido: marzo de 2010 — Aceptado: octubre de 2010]

References

Barrientos, J. (2007), Some Practical Problems of Recent Nonparametric Procedures: Testing, Estimation and Application, Tesis doctoral, Departamento de Fundamentos del Análisis Económico, Universidad de Alicante, España.

Delgado, M. A., Rodríguez, J. M. & Wolf, M. (2001), 'Subsampling Cube Root Asymptotics with an Application to Manski's MSE', *Economics Letters* **73**, 241–250.

Dette, H., von Lieres, C., Wilkau, C. & Sperlich, S. (2005), 'A Comparison of Different Nonparametric Method for Inference on Additive Models', *Journal of Nonparametric Statistics* **17**, 57–81.

Härdle, W. & Mammen, E. (1993), 'Comparing Nonparametric Versus Parametric Regression Fits', *Annals of Statistics* **21**(1926-1947).

Härdle, W. & Marron, J. S. (1990), 'Semiparametric Comparison of Regression Curves', *Annals of Statistics* **18**, 63–89.

Härdle, W. & Marron, J. S. (1991), 'Bootstrap Simultaneous Bars For Nonparametric Regression', *Annals of Statistics* **19**, 778–796.

Horowitz, J. L. & Spokoiny, V. (2001), 'An adaptive, rate-optimal test of parametric mean-regression model against a nonparametric alternative', *Econometrica* **69**, 599–631.

Neumeyer, N. & Sperlich, S. (2006), 'Comparison of separable components in different samples', *Scandinavian Journal of Statistics* **33**, 477–501.

Politis, D. N., Romano, J. P. & Wolf, M. (1999), *Subsampling*, Springer Series in Statistics, Springer-verlag, New York.

Roca, J. & Sperlich, S. (2007), 'Testing the Link when the Index is Semiparamtric - A Comparison Study', *Computational Statistics and Data Analysis* **12**, 6565–6581.

Spokoiny, V. (1996), 'Adaptive Hypothesis Testing using Wavelets', *Annals of Statistics* **24**, 2477–2498.

Spokoiny, V. (1998), 'Adaptive and spatially adaptive testing of a nonparametric hypothesis', *Mathematical Methods of Statistics* **7**(245-273).

Revista Colombiana de Estadística
Diciembre 2010, volumen 33, no. 2, pp. 321 a 339

Una revisión de los algoritmos de partición más comunes en el análisis de conglomerados: un estudio comparativo

A Review of the Most Common Partition Algorithms in Cluster Analysis: A Comparative Study

Susana A. Leiva-Valdebenito[a], Francisco J. Torres-Avilés[b]

Departamento de Matemática y Ciencia de la Computación, Facultad de Ciencia, Universidad de Santiago de Chile, Santiago, Chile

Resumen

Este estudio está enfocado en comparar diversos métodos de partición del análisis de conglomerados, usualmente conocidos como métodos no jerárquicos. En este trabajo, se realizan estudios de simulación para comparar los resultados obtenidos al implementar los algoritmos k-medias, k-medianas, PAM y Clara cuando los datos son multivariados y de tipo continuo. Adicionalmente, se efectúa un estudio de simulación con el fin de comparar algoritmos de partición para datos cualitativos, confrontando la eficiencia de los algoritmos PAM y k-modas. La eficiencia de los algoritmos se compara usando el índice de Rand ajustado y la tasa de correcta clasificación. Finalmente, se aplican los algoritmos a bases de datos reales, las cuales poseen clases predefinidas.

Palabras clave: algoritmos de conglomerados, medida de similaridad, simulación.

Abstract

This study is oriented to compare several partition methods in the context of cluster analysis, which are also called non hierarchical methods. In this work, a simulation study is performed to compare the results obtained from the implementation of the algorithms k-means, k-medians, PAM and CLARA when continuous multivariate information is available. Additionally, a study of simulation is presented to compare partition algorithms qualitative information, comparing the efficiency of the PAM and k-modes algorithms. The efficiency of the algorithms is compared using the Adjusted Rand Index and the correct classification rate. Finally, the algorithms are applied to real databases with predefined classes.

Key words: Clustering algorithm, Similarity measure, Simulation.

[a]Estudiante de Ingeniería Estadística. E-mail: susanaleivav@gmail.com
[b]Profesor asistente. E-mail: francisco.torres@usach.cl

1. Introducción

El análisis de conglomerados es una técnica de análisis exploratorio, definido dentro de los métodos multivariantes de clasificación, que permite separar en diferentes clases o grupos a un conjunto de objetos o individuos, de modo que todos los que pertenecen a una misma clase son homogéneos entre sí y diferentes de aquellos objetos que pertenecen a una clase distinta. Este método de agrupación es muy utilizado hoy en día en diferentes áreas, tales como, estudios de segmentación de clientes en el área financiera (Abonyi & Feil 2007), biología (Quinn & Keough 2002, Der & Everitt 2006), ecología (McGarigal, Cushman & Stafford 2000), entre otros, puesto que la mayoría de las veces no utiliza ningún supuesto estadístico para llevar a cabo el proceso de agrupación. Los algoritmos de agrupamiento más conocidos son los métodos jerárquicos y los métodos de partición o no jerárquicos, aunque existen otros métodos basados en densidades o modelos probabilísticos (Kamber & Han 2006).

En los métodos de partición, el conjunto de datos es inicialmente distribuido en un número pre-especificado de conglomerados k, el que puede ser aleatorio, y luego iterativamente se asignan las observaciones a los conglomerados hasta que se satisface algún criterio de parada. Entre estos métodos, el más utilizado es el algoritmo k-medias (MacQueen 1967, Anderberg 1973), sin embargo, existen otros algoritmos denominados k-medianas, PAM y Clara (Kaufman & Rousseeuw 1990), los cuales no han sido ampliamente difundidos, razón por la cual no se utilizan con frecuencia.

El presente trabajo tiene como propósito entregar un panorama acerca de los métodos de partición más comunes, abordando su estructura algorítmica e implementación. La contribución está relacionada con comparar la efectividad de los algoritmos para determinar y separar los grupos, usando distintos esquemas de simulación que incluyen datos generados a partir de distribuciones normales y Skew-Normales para el caso cuantitativo y distribuciones multinomiales para el caso cualitativo (ver Leiva 2008).

En una primera etapa, el presente trabajo entrega resultados relacionados con la comparación de los métodos de clasificación k-medias, k-medianas, PAM y Clara cuando se dispone de datos numéricos continuos, en diversos escenarios de agrupación. Luego, se comparan los algoritmos k-modas y PAM cuando se dispone de datos cualitativos. A partir de estudios de simulación se extraen las características y diferencias más relevantes, para posteriormente aplicarlos a bases de datos reales expuestos en la literatura. La implementación y aplicación se realizan usando el software libre R (R Development Core Team 2010).

El trabajo estará organizado de la siguiente manera. En la sección 2 se definirán los algoritmos de partición que intervendrán en el presente estudio. La sección 3 presentará los esquemas de simulación y resultados bajo los cuales se implementaron considerando la presencia de datos continuos y categóricos separadamente. La sección 4 mostrará los resultados de aplicar los algoritmos a bases de datos reales, expuestos en la literatura, para finalmente, en la sección 5, presentar las conclusiones más relevantes.

2. Algoritmos de partición

Se definen como algoritmos que permiten construir k particiones de las observaciones, donde cada partición representará a un conglomerado, segmento o grupo, y son útiles cuando existe ignorancia del investigador frente a la clasificación de observaciones, y éste dispone simultáneamente de un significativo número de variables o características. El algoritmo inicia considerando una división inicial, luego, busca encontrar el mejor agrupamiento reubicando los objetos de un grupo a otro, hasta que se optimice una función objetivo específica, la cual actúa como criterio de parada (Kamber & Han 2006). Intuitivamente, una clasificación adecuada debería considerar que la dispersión dentro de los grupos sea la menor posible.

Estos algoritmos son en general de implementación rápida, pero sufren inconvenientes en la especificación de semillas o particiones iniciales (Hartigan 1975). Este problema puede ser solucionado aplicando algún método jerárquico previo, tal como se hará más adelante. En las subsecciones siguientes se presenta la mecánica y las principales características de los algoritmos involucrados en el presente estudio.

2.1. k-medias

El algoritmo k-medias es uno de los métodos de partición más difundidos y populares. La génesis del algoritmo menciona a MacQueen (1967) como su precursor, a partir del cual se presentan diversas variaciones tales como aquellas propuestas por Anderberg (1973) y Hartigan (1975).

El algoritmo funciona como sigue. Dado un número inicial de conglomerados k, el objetivo del algoritmo es minimizar la distancia euclídea de los elementos dentro de cada conglomerado, respecto a su centro. La similitud de los objetos dentro de cada conglomerado es medida respecto a su vector de promedios, llamado generalmente centroide. El criterio de parada usado es el error cuadrático medio definido por

$$E = \sum_{l=1}^{k} \sum_{\mathbf{X} \in c_l} (\mathbf{X} - M_l)'(\mathbf{X} - M_l) \tag{1}$$

donde $\mathbf{X}$ es el punto en el espacio que representa a los objetos dados y M_l es el vector de promedios del conglomerado c_l, ambos vectores de $\mathbb{R}^p$. Existen varias formas de implementar el algoritmo, pero básicamente sigue los pasos expuestos en el algoritmo 2.1.1.

2.1.1. Algoritmo

1. Seleccionar arbitrariamente los k objetos que serán los centros o centroides iniciales de los conglomerados.

2. Se asigna cada objeto al conglomerado con el centroide más cercano, con base en el valor medio de los objetos en el conglomerado.

3. Se recalculan los centros de los conglomerados es decir se actualiza la media.

4. Se iteran los pasos 2 y 3 hasta que se alcance la convergencia del criterio de parada, o hasta que los centroides se modifiquen levemente.

Características relevantes de este algoritmo es que a menudo termina en un óptimo local, es escalable y eficiente en procesos que involucran grandes conjuntos de datos, y la complejidad computacional del algoritmo es del orden de nkt, donde t es el número de iteraciones y n el número de objetos o individuos (Han, Kamber & Tung 2001). Es de cálculo rápido y trabaja bien con valores faltantes o "missing"; sin embargo, es sensible a valores extremos, ya que distorsiona la media y el criterio de parada.

Entre sus mayores debilidades podemos señalar su sensibilidad a la selección de los centroides iniciales. Más aún, si las elecciones de diferentes centroides producen finales diferentes o la convergencia es muy lenta, quizás no existan agrupaciones naturales en los datos. La mayoría de las variaciones que existen del algoritmo difieren en la elección de los centroides iniciales, no obstante, es usual que esta selección se realice de forma aleatoria dentro del conjunto total de datos. Otra opción está relacionada con que el propio investigador especifique los centroides.

Una alternativa más objetiva para seleccionar el número de centroides iniciales se basa en determinarlos a partir de la aplicación previa de algún método de conglomerados jerárquico aglomerativo (Peña 2002), induciendo al algoritmo 2.1.2.

2.1.2. Algoritmo

1. Aplicar un método de conglomerado jerárquico y guardar los k centros que resulten del análisis.

2. Se asigna cada objeto al conglomerado con el centroide más cercano, con base en el valor medio de los objetos en el conglomerado.

3. Se recalculan los centros de los conglomerados, es decir, se actualiza la media.

4. Se iteran los pasos 2 y 3 hasta que se alcance la convergencia del criterio de parada, o hasta que los centroides se modifiquen levemente.

El algoritmo 2.1.2 será denominado más adelante k-medias jerárquico.

2.2. k-medianas

Descriptivamente, la mediana es una medida más robusta que la media, puesto que no se ve influida por valores extremos; el algoritmo k-medianas funciona de forma similar al algoritmo k-medias, sustituyendo el vector de promedios por el correspondiente vector de medianas como centro del conglomerado. En este caso se utiliza la distancia de Manhattan en vez de la distancia euclídea al cuadrado como medida de disimilitud (Anderson, Gross, Musicant, Ritz, Smith & Steinberg 2006).

Los vectores de medianas de los objetos en cada conglomerado serán denotados por $Me = \{Me_1, \ldots, Me_k\}$.

Usando la medida de distancia de Manhattan, la función por minimizar está dada por la siguiente expresión:

$$P(W, Me) = \sum_{l=1}^{k} \sum_{\mathbf{X} \in c_l} W_l' |\mathbf{X} - Me_l| \tag{2}$$

donde Me_l es el vector de medianas del l-ésimo conglomerado y $W = [W_1, \ldots, W_k]$ es una matriz de pesos con dimensión $n \times k$, cuyos vectores columnas son $W_l = (w_{1l}, \ldots, w_{kl})'$, para $l = 1 \ldots, k$, con $\sum_{l=1}^{k} w_{il} = 1$, donde $w_{il} \in (0, 1)$, para todo $i = 1, \ldots, n$, $j = 1, \ldots, k$.

2.2.1. Algoritmo

1. Seleccionar arbitrariamente los k objetos que serán los centros o centroides iniciales de los conglomerados. El tipo de selección inicial de centroides es análogo a los presentados en el algoritmo k-medias.

2. Asignar cada punto al conglomerado con el centroide más cercano, a través de la distancia de Manhattan.

3. Calcular el nuevo conjunto de centroides de los conglomerados, calculando la mediana de los nuevos grupos formados.

4. Se iteran los pasos 2 y 3 hasta que se minimice la función objetivo (2), o hasta que los centroides apenas se modifiquen.

El algoritmo k-medianas tiene la misma complejidad computacional que el algoritmo k-medias. Al ser bastante similar al algoritmo k-medias, es también sensible a la selección de los centroides iniciales; la ventaja que presenta es que la mediana no está influida por los valores extremos, por lo que se logra un método, en teoría, más robusto. Nótese que este algoritmo no está implementado en los softwares tradicionales.

2.3. k-medoids

Estos métodos fueron introducidos por Kaufman & Rousseeuw (1987). Están basados en el uso de objetos actuales del conjunto de datos para ser los representantes de los conglomerados, denominados medoids. Estos medoids se definen como los puntos localizados lo más al centro posible de cada conglomerado y son representativos de la estructura de los datos. Cada objeto restante es agrupado con el medoid más cercano, e iterativamente estos algoritmos realizan todos los intercambios posibles entre los objetos representativos y los que no lo son, hasta que se minimice una medida de disimilitud entre los k-medoids y los vectores de observaciones que forman los conglomerados (Kamber & Han 2006).

En este grupo encontramos los algoritmos "Partition Around Medoids" (PAM) y "Clustering Large Applications" (Clara). Tanto el algoritmo PAM como el Clara se encuentran implementados en el software R-gui y permiten ingresar una matriz de disimilitudes definida por el analista. Esto presenta una ventaja, pues es posible incorporarlo al proceso de partición cuando se dispone de datos cualitativos.

2.3.1. Algoritmo PAM

El algoritmo "Partitioning Around Medoids" (PAM) es un método tipo k-medoid que intenta determinar k particiones de n objetos determinando los objetos representativos de cada conglomerado (Ng & Han 1994). Para encontrar los k medoids, PAM empieza con una selección arbitraria de k objetos representativos. En cada iteración hace un intercambio entre un objeto seleccionado, O_i, y uno no seleccionado, O_h, si y solo si el intercambio mejora la calidad del agrupamiento.

El efecto de tal intercambio entre O_i y O_h se mide a través de una función de costo, es decir, el algoritmo calcula los costos C_{jih} para todos los objetos no seleccionados O_j. Según el caso en el cual O_j se encuentre, C_{jih} puede ser definido por una de las siguientes expresiones:

Caso 1: Suponga que actualmente O_j pertenece al conglomerado representado por O_i. Además, O_j es más similar a O_{j2} que a O_h, donde O_{j2} es el segundo medoid más similar a O_j. Entonces, si O_i es remplazado por O_h como un medoid, O_j pertenecería al conglomerado representado por O_{j2}. De esta forma el costo del intercambio es dado por:

$$C_{jih} = d(O_j, O_{j2}) - d(O_j, O_i)$$

Esta ecuación siempre da un valor no negativo.

Caso 2: Suponga que O_j pertenece actualmente al conglomerado representado por O_i. En este caso O_j es más similar a O_h que a O_{j2}. Entonces, si O_i es remplazado por O_h, O_j pertenecería al conglomerado representado por O_h. El costo es dado por:

$$C_{jih} = d(O_j, O_h) - d(O_j, O_i)$$

A diferencia del caso anterior, C_{jih} puede ser positivo o negativo.

Caso 3: Suponga que O_j pertenece actualmente al conglomerado representado por O_{j2} y O_j es más similar a O_{j2} que a O_h. Entonces, aun si O_i es remplazado por O_h, O_j permanecería en el conglomerado representado por O_{j2}. De esta manera el costo sería:

$$C_{jih} = 0$$

Caso 4: Suponga que O_j pertenece actualmente al conglomerado representado por O_{j2}, pero O_j es menos similar a O_{j2} que a O_h. Entonces, remplazando O_i por O_h causaría que O_j fuese representado por O_h desde el conglomerado O_{j2}. De esta manera el costo es dado por:

$$C_{jih} = d(O_j, O_h) - d(O_j, O_{j2})$$

Este costo es siempre negativo.

Combinando los cuatro casos, el costo total de remplazar O_i por O_h está dado por $TC_{ih} = \sum_j C_{jih}$. La función $d(\cdot)$ es una la medida de distancia o disimilitud que se utiliza entre los objetos. Por defecto se usa la distancia euclídea. La estructura del algoritmo PAM se puede visualizar en el algoritmo 2.3.2.

2.3.2. Algoritmo

1. Seleccionar arbitrariamente k objetos representativos, los cuales serán los k-medoids iniciales.

2. Calcular TC_{ih} para todos los pares de objetos O_i, O_h donde O_i es actualmente un medoid, y O_h no lo es.

3. Seleccionar el par O_i, O_h el cual corresponda al $\text{mín}_{O_iO_h}\{TC_{ih}\}$. Si el mínimo TC_{ih} es negativo, se intercambia O_i con O_h; regresar al paso 2.

4. Repetir el paso 2 y 3 hasta que no haya cambio.

5. Asignar cada objeto a su medoid más cercano.

La principal ventaja del algoritmo PAM es la robustez del método en presencia de ruido u "outliers", pues el cálculo del medoid está menos influido por ellos u otros valores extremos. PAM comienza a ser muy costoso a medida que el tamaño muestral n y el número de iteraciones k aumentan, siendo una de las principales desventajas de este algoritmo, razón por la cual es eficiente sólo para bases de datos pequeñas (Han et al. 2001).

2.3.3. Algoritmo Clara

El algoritmo Clara, separa múltiples muestras de la base completa y aplica el algoritmo PAM sobre cada una de ellas; luego, encuentra los conjuntos de k-medoids de las muestras. El principal motivo de Kaufman & Rousseeuw (1990) para proponer este algoritmo fue debido a la deficiencia del algoritmo PAM para trabajar con bases de datos con grandes volúmenes de información. Si estas muestras son realmente representativas de toda la base de datos, los medoids de las muestras deberían acercarse a aquellos que se hubiesen escogidos de la base de datos completa. Según estos autores, los resultados experimentales indican que 5 muestras con $(40 + 2k)$ objetos cada una, producen resultados satisfactorios. La calidad del agrupamiento es medida con la disimilitud media de todos los datos, y no sólo aquellos objetos considerados en las muestras (Ng & Han 1994).

2.3.4. Algoritmo

1. Para $i = 1$ a 5 repetir los siguientes pasos.

2. Seleccionar una muestra aleatoria de $s = (40 + 2k)$ objetos de la base completa.

3. Ejecutar el algoritmo PAM sobre la muestra, para encontrar los k medoids de esta muestra.

4. Para cada objeto O_j de la base completa, determinar su medoid más cercano y agruparlos.

5. Calcular la disimilaridad media del agrupamiento obtenido. Si este valor es menor al mínimo actual, usar este valor como el mínimo actual y conservar los k medoids obtenidos en el paso 3 como el mejor conjunto de medoids obtenidos.

6. Retornar al paso 1 y comenzar con la próxima iteración.

La efectividad de Clara depende tanto del tamaño de la muestra como de su calidad. Note que PAM busca los mejores k medoids entre un conjunto total de datos, mientras que Clara busca los mejores k medoids entre las muestras seleccionadas del conjunto total de datos. Clara no podría encontrar la mejor agrupación si los mejores k medoids no son seleccionados dentro de las muestras.

El algoritmo Clara presenta una complejidad mayor en cada iteración, puesto que depende adicionalmente del tamaño de la muestra seleccionada (Han et al. 2001).

2.4. k-modas

La mayoría de las aplicaciones realizadas bajo los algoritmos de partición se ha centrado en datos numéricos, cuyas propiedades pueden ser explotadas para definir naturalmente las funciones de distancia entre los objetos. Sin embargo, las aplicaciones en segmentación contienen conjuntos de datos que incluyen datos categóricos cuyas funciones de distancia o disimilitud no están definidas naturalmente.

Una debilidad importante del algoritmo k-medias es su incapacidad de trabajar con datos que no sean numéricos. Huang (1998) presentó un algoritmo para trabajar con un entorno categórico. La idea de Huang fue extender el algoritmo k-medias al ámbito categórico denominándolo k-modas, teniendo en cuenta algunas modificaciones, tales como: usar una medida de disimilitud de correspondencia simple para datos categóricos, remplazar las medias de los grupos por sus respectivas modas y utilizar un método basado en frecuencias para actualizarlas.

El algoritmo k-modas utiliza el mismo proceso que el k-medias, lo que preserva su eficiencia y es altamente deseable en los análisis de agrupación de datos.

Es así como $X = (X_1, \ldots, X_n)$ representa un conjunto de n objetos descritos por un conjunto de m atributos y frecuencias denotadas por A_j y p_j, respectivamente. Entonces, los objetos X_i son representados por un vector del tipo $[x_{i1}, x_{i2}, \ldots, x_{im}]$. Además $X_i = X_r$, si y solo si $x_{ij} = x_{rj}$ para $1 \leq j \leq m$, que significa que los dos objetos tienen iguales categorías para los distintos atributos.

Como medida de disimilitud entre dos objetos categóricos, se usará el total de discordancias de los correspondientes valores de los atributos entre dos objetos (Kaufman & Rousseeuw 1990), definido por:

$$d(X_i, X_j) = \sum_{k=1}^{m} \delta(x_{ik}, x_{jk}) \tag{3}$$

donde

$$\delta(x_{ik}, x_{jk}) = \begin{cases} 0, & (x_{ik} = x_{jk}) \\ 1, & (x_{ik} \neq x_{jk}) \end{cases}$$

Al usar la medida de disimilitud para objetos categóricos (3), la función objetivo por minimizar es:

$$P(W, Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{il} \delta(x_{ij}, q_{lj}) \tag{4}$$

sujeto a

$$\sum_{l=1}^{k} w_{il} = 1 \text{ y } w_{il} \in (0, 1) \text{ con } 1 \leq i \leq n, 1 \leq l \leq k$$

donde $w_{il} \in W$, W es una matriz de $n \times k$, y los elementos $w_{il} = 1$ indica que el objeto X_i es asignado al clúster C_l. Además, $Q = [Q_1, \ldots, Q_k]$ es un conjunto de vectores de modas representantes de los k grupos, es decir, la matriz Q actúa como centroide, donde $Q_l = (q_{l1}, \ldots, q_{lm})'$, $l = 1, \ldots, k$. De acuerdo con lo propuesto por Andreopoulos, An & Wang (2006), en la práctica es utilizado el algoritmo 2.4.1.

2.4.1. Algoritmo

1. Seleccionar las k modas iniciales para cada clúster.
2. Para cada objeto X
   - Calcular la disimilitud entre el objeto X y las modas de todos los clústers.
   - Insertar el objeto X dentro del clúster cuya moda sea la más similar al objeto X
   - Actualizar las modas de los clústers.
3. Calcular de nuevo la disimilitud de los objetos con las modas actuales. Si un objeto es más similar a la moda de otro clúster que su actual moda, reasignar el objeto a ese clúster y actualizar ambas modas.
4. Repetir 3 hasta que los objetos no hayan cambiado después de probar el ciclo completo del conjunto de datos.

Tal como el algoritmo k-medias, los resultados arrojados por el k-modas produce óptimos locales, éstos dependen de las modas iniciales y del orden de los objetos en el conjunto de datos. Su complejidad computacional es igual a la del algoritmo k-medias.

El algoritmo k-modas original (Huang 1998) presenta ciertas debilidades tanto en la asignación de los objetos como en la selección de los centros de los conglomerados, ya que la medida de disimilitud usada no considera la relación implícita integrada en los valores categóricos y esto produce grupos con una similitud interna más débil (He, Xu & Deng 2007).

Por las razones antes mencionadas se han realizado diversos estudios para mejorar el algoritmo k-modas original, como He, Deng & Xu (2005) que modifica el algoritmo considerando las frecuencias de los valores del atributo en la medida de disimilitud. Ng, Li, Huang & Zengyou (2007) abordan de manera más pertinente el algoritmo presentado en He et al. (2005). Recientemente He et al. (2007) presentaron diversos esquemas para ponderar el valor del atributo en el agrupamiento k-modas.

3. Simulación

3.1. Simulación para datos continuos

Los algoritmos por comparar en esta sección son k-medias, k-medianas, PAM y Clara, y adicionalmente un quinto algoritmo, el cual es una variación del algoritmo k-medias, pero con la elección de los centroides iniciales resultantes de un agrupamiento por método jerárquico (método de Ward), que para efectos de notación será denominando "k-medias jer.". La literatura usual aborda la aplicación de estos métodos considerando sólo datos reales. Recientemente, un estudio desarrollado por Velmurugan & Santhanam (2010), compara la eficiencia de los algoritmos k-medias y k-medoids a través de datos simulados en términos del tiempo, concluyendo que el algoritmo k-medias demora más que k-medoids.

Lo anterior motivó a realizar el estudio usando siete tipos de esquemas de conglomerados y aplicándolos a 5 y 8 grupos, respectivamente, tal como se ilustra en la figura 1, siguiendo aquellos expuestos en algunos textos especializados y las referencias que éstos contienen (SAS Institute Inc. 2008). El trabajo de Velmurugan & Santhanam (2010) no presenta los distintos esquemas analizados en esta investigación y que efectivamente se pueden dar en la práctica. Esto hace la diferencia de otros artículos, donde la aplicación de éstos considera sólo tres conglomerados en datos reales y simulados (Hae & Chi 2009), siendo diez el máximo de grupos analizados bajo sólo dos esquemas de conglomerados, encontrado en Velmurugan & Santhanam (2010). Para esta investigación, la eficiencia de los algoritmos se mide en cada escenario, a través de los índices de Rand ajustado o "ARI" (Hubert & Arabie 1985) y las respectivas tasas de correcta clasificación (TCC) considerando un determinado algoritmo. El total de simulaciones por cada esquema fue de 50, considerando que se obtuvieron los mismos resultados al aplicar una cantidad mayor de éstas.

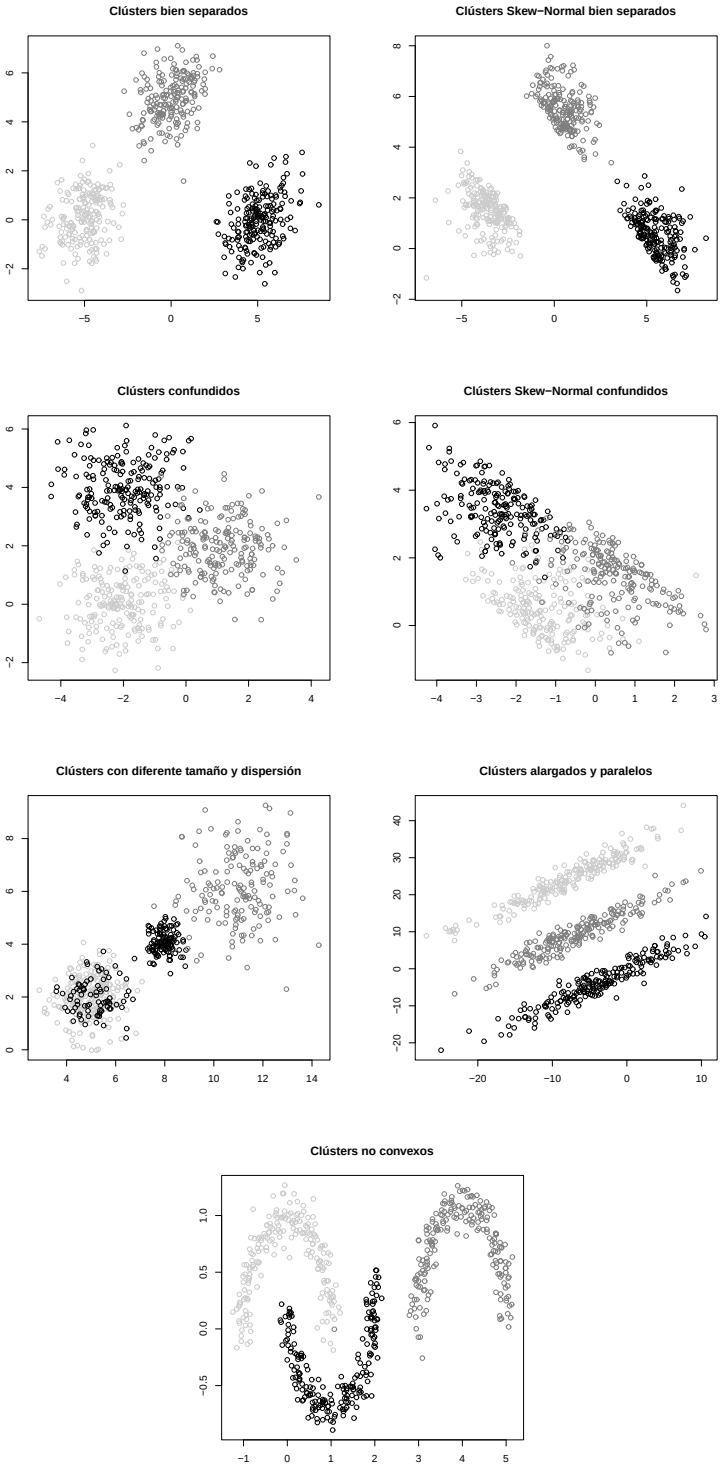


FIGURA 1: Esquemas de simulación para datos continuos.

3.1.1. Resultados de las simulaciones

Analizando las tablas 1 y 2 en forma global, es posible observar que el algoritmo que tiene, en promedio, ambos índices menores tanto para cinco como para ocho grupos, considerando todos los esquemas, es el algoritmo k-medias. A pesar de no existir grandes diferencias entre los otros cuatro algoritmos, el que presenta índices medio mayores es el algoritmo PAM y k-medias jerárquico.

Tabla 1: Medias del índice de Rand ajustado y de la tasa de correcta clasificación en 5 grupos. Datos simulados - caso cuantitativo.

		5 clústers				
Esquema		k-medias	k-medias jer.	k-medianas	PAM	Clara
Bien separados	ARI	0,9141	*0,9991*	*0,9991*	*0,9991*	0,9985
	TCC	0,9279	*0,9996*	*0,9996*	*0,9996*	0,9995
Skew-Normal	ARI	0,8706	0,9995	0,9995	*0,9996*	0,9995
Bien Sep.	TCC	0,8726	*0,9998*	*0,9998*	*0,9998*	0,9998
Confundidos	ARI	0,7843	*0,7884*	0,7820	0,7873	0,7611
	TCC	0,9035	*0,9084*	0,9053	*0,9078*	0,8945
Skew-Normal	ARI	0,8156	*0,8286*	0,8067	0,8281	0,8049
confundidos	TCC	0,9173	*0,9257*	0,9155	0,9219	0,9139
Diferente Tam.	ARI	0,8490	0,9537	0,9529	*0,9547*	0,9504
y dispersión	TCC	0,8604	0,9729	0,9721	*0,9736*	0,9712
Alargados	ARI	0,6181	0,6389	0,6130	*0,7251*	0,7122
	TCC	0,7449	0,7716	0,7568	*0,8762*	0,8673
No convexos	ARI	0,6927	0,7181	0,7535	*0,7553*	0,7356
	TCC	0,8193	0,8641	0,8859	*0,8848*	0,8721
Media	ARI	0,7921	0,8466	0,8438	**0,8642**	0,8517
	TCC	0,8637	0,9203	0,9193	**0,9377**	0,9312
SD	ARI	0,1041	0,1423	0,1454	0,1177	0,1268
	TCC	0,0642	0,0825	0,0851	0,0528	0,0580

Al realizar la comparación entre el ARI y el TCC por cada esquema, se nota que las conclusiones sobre los algoritmos son semejantes, teniendo en cuenta que los valores de las tasas de correcta clasificación son valores más altos que los del índice de Rand ajustado. Considerando lo anterior, a continuación se analizan los resultados de forma independiente para cada esquema sólo por medio del índice de Rand ajustado.

Se analizan los índices medio obtenidos por cada algoritmo en cinco y ocho clústers, resaltando el algoritmo k-medias jerárquico para los cuatro primeros esquemas, seguido muy de cerca por el algoritmo basado en medoids, PAM. Este último presenta los mejores índices para los esquemas de grupos con "Diferente tamaño y dispersión" y "Alargados". Finalmente, el esquema de grupos "No convexo" muestra que el índice asociado al algoritmo k-medianas es el que presenta una mejor agrupación de los datos simulados.

TABLA 2: Medias del índice de Rand ajustado y de la tasa de correcta clasificación en 8 grupos. Datos simulados - caso cuantitativo.

		8 clusters				
Esquema		k-medias	k-medias jer.	k-medianas	PAM	Clara
Bien separados	ARI	0,8740	*0,9997*	*0,9997*	*0,9997*	0,9994
	TCC	0,8972	*0,9999*	*0,9999*	*0,9999*	0,9997
Skew-Normal	ARI	0,8171	*0,9994*	0,9990	*0,9994*	0,9989
Bien Sep.	TCC	0,8148	*0,9998*	0,9996	0,9997	0,9995
confundidos	ARI	0,7272	*0,7396*	0,7264	0,7371	0,6468
	TCC	0,8550	*0,8714*	0,8600	0,8695	0.7877
Skew-Normal	ARI	0,7691	*0,8333*	0,8156	0,8316	0,7899
Confundidos	TCC	0,8548	*0,9213*	0,9127	0,9202	0,8925
diferente Tam.	ARI	0,8901	0,9655	0,9653	*0,9665*	0,9600
y Dispersión	TCC	0,9083	0,9799	0,9826	*0,9831*	0,9797
Alargados	ARI	0,5986	0,5925	0,5901	*0,6878*	0,6743
	TCC	0,6956	0,7521	0,7431	*0,8354*	0,8049
No convexos	ARI	0,7368	0,8364	0,8514	*0,8498*	0,7429
	TCC	0,7659	0,9167	0,9274	*0,9235*	0,8175
Media	ARI	0,7733	0,8523	0,8497	**0,8674**	0,8303
	TCC	0,8274	0,9201	0,9179	**0,9330**	0,8974
SD	ARI	0,0997	0,1511	0,1538	0,1262	0,1534
	TCC	0,0756	0,0884	0,0927	0,0649	0,0955

3.2. Simulación para datos cualitativos

En presencia de datos cualitativos, se propone comparar los algoritmos k-modas y PAM, puesto que este último permite la incorporación de la matriz de distancias o disimilitudes, en lugar de los datos originales. Respecto a la simulación, estas variables se generaron asumiendo distribuciones multinomiales con 2, 3 y 4 categorías. Además, se consideraron probabilidades para las categorías, de tal forma que tres grupos estén bien definidos, variando el número de objetos en cada grupo ($k = 3$).

En la figura 2 se presenta un resumen de los esquemas de simulación utilizados. Para comparar ambos algoritmos se emplea una matriz de disimilitudes construida a través del cálculo del coeficiente general propuesto por Gower (1971), medida adecuada en presencia de este tipo de datos.

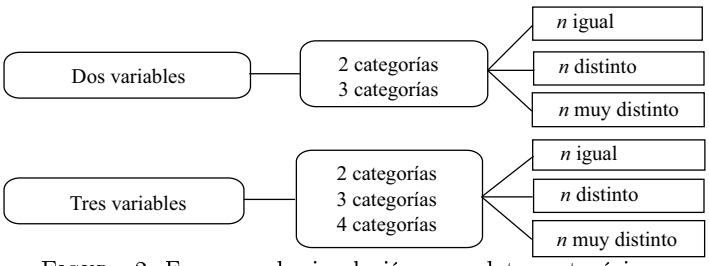


FIGURA 2: Esquema de simulación para datos categóricos.

3.2.1. Resultados de las simulaciones

El índice de Rand ajustado y la tasa de correcta clasificación media obtenida en los diferentes esquemas se presentan en la tabla 3.

Tabla 3: Medias del índice de Rand ajustado y tasa de correcta clasificación. Datos simulados - caso categórico.

		Dos variables		Tres variables	
n		k-moda	PAM	k-moda	PAM
Igual	ARI	0,5234	0,3977	*0,5598*	0,5539
	TCC	*0,8140*	0,7469	*0,8286*	0,8215
Distinto	ARI	0,4486	0,4109	0,4028	*0,4592*
	TCC	*0,7737*	0,7582	0,7613	*0,7811*
Muy distinto	ARI	0,3984	0,3419	0,2970	*0,3006*
	TCC	*0,7547*	0,7332	*0,6718*	0,6685
Media	ARI	**0,4568**	0,3835	0.4199	**0,4379**
	TCC	**0,7808**	0,7461	0,7539	**0,7570**
SD	ARI	0,0629	**0,0366**	0,1322	**0,1280**
	TCC	0,0302	**0,0126**	**0,0787**	0,0793

El valor medio del índice es relativamente bajo en todos los esquemas, aproximadamente menor a 0,56; sin embargo, se puede decir que para el caso de dos variables, el algoritmo que obtuvo mejores resultados fue el k-modas, y al aumentar la cantidad de variables a tres, se presenta una mejora en los índices medios obtenidos por PAM, considerando los esquemas en donde n es "Distinto" o "Muy distinto". Los índices medios más altos fueron obtenidos cuando n era igual para todos los grupos, y apunta a que el algorimto k-modas presenta la mejor agrupación de la información. Esto último sucede de la misma manera con la tasa de correcta clasificación, con la diferencia de que toma valores más altos, alcanzando un 83 % aproximado como máximo, en el caso de tres variables con n igual.

4. Aplicaciones

Las bases de datos reales se obtuvieron del UCI Machine Learning Repository (Asuncion & Newman 2007). Para el caso continuo se utilizan seis conjuntos de datos del UCI Repository, de los cuales todos contienen atributos continuos y atributos de clase. En la implementación categórica se utilizan cinco conjuntos de datos, utilizados por He et al. (2007), de los cuales todos contienen solo atributos categóricos y atributos de clase. Para ambos casos, los resultados son presentados considerando el respectivo índice de Rand ajustado y la tasa de clasificación correcta.

4.1. Resultados obtenidos de bases de datos continuos

La distribución de las frecuencias de las clases se puede observar en la tabla 4. Una vez que se aplican las metodologías, tal como se observa en la tabla 5, los

cinco algoritmos no obtuvieron grandes diferencias en las tres primeras bases de datos; sin embargo, en "Wine" y "Ecoli" PAM presenta una disminución importante en ARI y en su TCC. Estas dos bases de datos tienen las frecuencias con las clases menos homogéneas, lo que puede provocar que PAM no entregue óptimos resultados. En la última base de datos, "Image Segment", PAM presenta mejores resultados en comparación con el resto de los algoritmos.

Tabla 4: Frecuencias de las clases en las bases de datos continuas.

Base de datos	Clases	Atributos	Casos	Distribución de las clases
Iris	3	4	150	50 c/u
Breast Cancer	2	30	569	357/212
Bank	2	6	200	100 c/u
Wine	3	13	178	71/59/48
Ecoli	4	7	336	143/116/52/25
Imag. Segment.	7	19	210	30 c/u

Tabla 5: Índice de Rand ajustado y tasa de correcta clasificación para datos continuos.

Base de datos	k-medias	k-medias jer.	k-medianas	PAM	Clara
Iris	0,893	0,893	0,887	0,893	*0,900*
Breast Cancer	0,854	0,854	0,866	*0,868*	*0,868*
Bank	*1*	*1*	0,990	0,990	0,980
Wine	*0,966*	*0,966*	0,961	0,910	0,938
Ecoli	0,738	0,866	*0,880*	0,676	0,735
Imag. Segment	0,519	0,580	0,595	*0,633*	0,491
Media	0,828	0,860	**0,863**	0,828	0,819

4.2. Resultados obtenidos de bases de datos categóricos

Tal como lo muestra la tabla 6, la distribución de frecuencias de las primeras bases de datos contiene frecuencias homogéneas para cada uno de los atributos. A diferencia de los primeros ejemplos, las dos últimas bases de datos contienen frecuencias desbalanceadas, con valores extremos en algunos casos.

Tabla 6: Frecuencias de las bases de datos categóricas.

Base de datos	Clases	Atributos	Casos	Distribución de las clases
Voting	2	16	435	168/267
Breast Cancer	2	9	699	241/458
Soybean	4	35	47	10/10/10/17
Lymphography	4	18	148	2/4/61/81
Zoo	7	17	101	4/5/8/10/13/20/41

Como se observa en la tabla 7, para "Breast Cancer", "Soybean" y la base de datos de los votos del congreso ("Voting"), el algoritmo PAM muestra mejores resultados que los entregados por el algoritmo k-modas. Sin embargo, en los dos últimos conjuntos de datos, el algoritmo k-modas funciona mejor que el algoritmo de los medoids.

Tabla 7: Tasa de clasificación correcta de datos categóricos.

Base de datos	PAM	k-modas
Voting	*0,864*	0,859
Breast Cancer	*0,937*	0,85
Soybean	*0,936*	0,819
Lymphography	0,412	*0,661*
Zoo	0,743	*0,829*
Media	0,778	**0,804**

Una razón por la cual el uso del algoritmo PAM con datos categóricos no da buenos resultados en los dos últimos conjuntos de datos podría deberse a que las frecuencias de las clases que componen los datos difieren mucho entre sí, es decir, las frecuencias de las clases de "Lymphography" y "Zoo" son menos homogéneas que las del resto.

5. Conclusiones

Dada la importancia que ha tomado el análisis de conglomerados en los diversos estudios de aplicación, se ha querido abordar la presentación, implementación y comparación de los algoritmos de partición de mayor uso en la literatura. En este trabajo se presenta un estudio de simulación y aplicación con el fin de exponer y comparar los algoritmos de partición, según distintos esquemas de agrupación.

En el estudio de simulación realizado, en el caso continuo, los algoritmos que dieron mejores resultados fueron el k-medias jerárquico y PAM, en donde este último sobresale en los esquemas con grupos con "Diferente tamaño y dispersión" y "Alargados". El algoritmo k-medianas da resultados sobresalientes en los esquemas "No convexos", esquema que en general tiene problemas con los valores iniciales y con la convergencia a una única solución. El algoritmo Clara presentó resultados similares a PAM en el esquema de grupos con "Diferentes tamaños y dispersión"; sin embargo, es más inestable ya que trabaja con base a muestras. Para el caso categórico, el algoritmo k-modas obtuvo mejores resultados que PAM; no obstante los resultados de PAM mejoraron al aumentar la cantidad de variables. Cabe acotar que k-modas, a diferencia de PAM, presenta resultados coherentes al considerar esquemas donde las frecuencias de las clases son aproximadamente iguales. Es evidente que si el número de observaciones o el tamaño muestral aumenta, la complejidad de los algoritmos también lo hará, lo que llevará a proponer la aplicación de otros métodos para optimizar el tiempo de ejecución y reducir el costo computacional.

En las aplicaciones realizadas a datos continuos en promedio el algoritmo más débil fue k-medias y en cuatro conjuntos de datos los demás algoritmos dan resultados muy semejantes. Por otro lado, PAM fues el que obtuvo un índice menor que el resto en los datos asociados a "Wine" y "Ecoli". Sin embargo, en "Imag. segment." el algoritmo PAM es el que da una mejor tasa de buena clasificación. Al comparar PAM con k-modas sucede que en las bases de datos en donde la distribución de las frecuencias de las clases son relativamente homogéneas, PAM entrega mejores

resultados que k-modas, pero PAM no es eficiente al tener las frecuencias de las clases muy heterogéneas.

Agradecimientos

Los autores agradecen a los árbitros por tan importantes comentarios que ayudaron a mejorar el contenido de este trabajo.

[Recibido: mayo de 2010 — Aceptado: octubre de 2010]

Referencias

Abonyi, J. & Feil, B. (2007), *Clustering Analysis for Data Mining and System Identification*, Birkhauser Verlag AG, Berlin, Germany.

Anderberg, M. (1973), *Cluster Analysis for Applications*, Academic Press, New York, United States.

Anderson, B., Gross, D., Musicant, D., Ritz, A., Smith, T. & Steinberg, L. (2006), Adapting K-Medians to Generate Normalized Cluster Centers, *in* 'Proceedings of the 2006 SIAM International Conference on Data Mining', Bethesda, pp. 165–175.

Andreopoulos, B., An, A. & Wang, X. (2006), Clustering Mixed Numerical and Low Quality Categorical Data: Significance Metrics on a Yeast Example, *in* 'ACM SIGMOD Workshop on Information Quality in Information Systems, IQIS 2005 Statistics Clustering Session', Baltimore, pp. 87–98.

Asuncion, A. & Newman, D. J. (2007), 'UCI machine learning repository'. *http://www.ics.uci.edu/∼mlearn/MLRepository.html

Der, G. & Everitt, B. S. (2006), *Statistical Analysis of Medical Data using SAS*, CRC Press, Boca Raton, United States.

Gower, J. C. (1971), 'A General Coefficient of Similarity and Some of its Properties', *Biometrics* **27**, 623–637.

Hae, P. & Chi, J. (2009), 'A simple and Fast Algorithm for K-medoids Clustering', *Expert Systems with Applications* **36**, 3336–3341.

Han, J., Kamber, M. & Tung, A. K. H. (2001), Spatial Clustering Methods in Data Mining: A Survey, *in* H. J. Miller & J. Han, eds, 'Geographic Data Mining and Knowledge Discovery', Taylor & Francis.

Hartigan, J. (1975), *Clustering Algorithms*, John Wiley & Sons, Nueva York, United States.

He, Z., Deng, S. & Xu, X. (2005), *Improving K-modes Algorithm Considering Frequencies of Attribute Values in Mode*, Vol. 3801 of *Lecture Notes in Computer Science*, Springer Berlin - Heidelberg, pp. 157–162.

He, Z., Xu, X. & Deng, S. (2007), 'Attribute Value Weighting in K-Modes Clustering', *Computer Science e-Prints* **1**, 1–15.

Huang, Z. (1998), 'Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values', *Data Mining and Knowledge Discovery* **2**, 283–304.

Hubert, L. & Arabie, P. (1985), 'Comparing Partitions', *Journal of Classification* **2**, 193–218.

Kamber, M. & Han, J. (2006), *Data Mining Concepts and Techniques*, Morgan Kaufman Publishers, San Francisco, United States.

Kaufman, L. & Rousseeuw, P. (1987), Clustering by Means of Medoids, *in* D. Y., ed., 'Statistical Data Analysis Based on the L_1 Norm and Related Methods', North-Holland, pp. 405–416.

Kaufman, L. & Rousseeuw, P. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley and Sons, New York, United States.

Leiva, S. (2008), Algoritmos de partición en el análisis de conglomerados: un estudio comparativo, Trabajo de grado, Universidad de Santiago de Chile, Chile.

MacQueen, J. (1967), Some Methods for classification and Analysis of Multivariate Observations, *in* 'Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, Symposium on mathematics, pp. 281–297.

McGarigal, K., Cushman, S. & Stafford, S. (2000), *Multivariate Statistics for Wildlife and Ecology Research*, Springer Verlag, New York, United States.

Ng, M. K., Li, M. J., Huang, J. Z. & Zengyou, H. (2007), 'On the Impact of Dissimilarity Measure in k-Modes Clustering Algorithm', *, IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**(3), 503–507.

Ng, R. & Han, J. (1994), Efficient and Effective Clustering Methods for Spatial Data Mining, *in* 'Proceeding of the 20^{th} International Conference on Very Large Databases', pp. 144–155.

Peña, D. (2002), *Análisis de datos multivariantes*, McGraw-Hill, Madrid, España.

Quinn, G. & Keough, M. (2002), *Experimental Design and Data Analysis for Biologists*, Cambridge University Press, Cambridge, UK.

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

SAS Institute Inc. (2008), *SAS/STAT 9.2 User's Guide*, SAS Publishing, Cary, Carolina del Norte.

Velmurugan, T. & Santhanam, T. (2010), 'Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points', *Journal of Computer Science* **6**(3), 363–368.

Revista Colombiana de Estadística

Índice de autores del volumen 33, 2010

Información para los autores

La **Revista Colombiana de Estadística** publica artículos originales de carácter teórico o aplicado en cualquiera de las ramas de la estadística. Se consideran también artículos divulgativos de gran calidad de exposición sobre metodologías o técnicas estadísticas aplicadas en diferentes campos del saber. Únicamente se publican artículos en español e inglés, si el autor escribe en una lengua diferente a la nativa debe enviar un certificado de un traductor oficial o de un corrector de estilo que haya revisado el texto.

El Comité Editor únicamente acepta trabajos para evaluación que no han sido publicados previamente y que no están siendo propuestos simultáneamente para publicación en otros medios, ni lo serán sin previo consentimiento del Comité, a menos que, como resultado de la evaluación, se decida no publicarlos en la Revista. Se supone además que cuando los autores hacen entrega de un documento con fines de publicación en la **Revista Colombiana de Estadística**, conocen las condiciones anteriores y que están de acuerdo con ellas.

Material

Los artículos remitidos a la **Revista Colombiana de Estadística** deben ser presentados en archivo PDF o PS, con textos, gráficas y tablas en color negro y, además, los autores deben agregar una versión del artículo sin nombres ni información de los autores, que se utilizará para el arbitraje. Se debe enviar una carta firmada por cada uno de los autores, donde manifiesten estar de acuerdo con someter el artículo y con las condiciones de la Revista. Si un artículo es aceptado, los autores deben poner a disposición del Comité Editorial los archivos: fuente en LaTeX y de gráficas en formato EPS en blanco y negro.

Para facilitar la preparación del material publicado se recomienda utilizar MiKTeX[1], usando los archivos de la plantilla y del estilo *revcoles* disponibles en la página Web de la Revista[2] y siguiendo las instrucciones allí incorporadas.

Todo artículo debe incluir:

- Título en español y su traducción al inglés.
- Los nombres completos y el primer apellido, la dirección postal o electrónica y la afiliación institucional de cada autor.
- Un resumen con su versión en inglés (*abstract*). El resumen en español no debe pasar de 200 palabras y su contenido debe destacar el aporte del trabajo en el tema tratado.
- Palabras clave (*Key words*) en número entre 3 y 6, con su respectiva traducción al inglés, siguiendo las recomendaciones del *Current Index to Statistics* (CIS)[3].

[1] http://www.ctan.org/tex-archive/systems/win32/miktex/
[2] http://www.estadistica.unal.edu.co/revista
[3] http://www.statindex.org/CIS/homepage/keywords.html

- Cuando el artículo se deriva de una tesis o trabajo de grado debe indicarse e incluirse como una referencia.
- Si se deriva de un proyecto de investigación, se debe indicar el título del proyecto y la entidad que lo patrocina.
- Referencias bibliográficas, incluyendo solamente las que se hayan citado en el texto.

Referencias y notas al pie de página

Para las referencias bibliográficas dentro del texto se debe utilizar el formato autor-año, dando el nombre del autor seguido por el año de la publicación dentro de un paréntesis. La plantilla LaTeX suministrada utiliza, para las referencias, los paquetes BibTeX y Harvard[4]. Se recomienda reducir el número de notas de pie de página, especialmente las que hacen referencia a otras notas dentro del mismo documento y no utilizarlas para hacer referencias bibliográficas.

Tablas y gráficas

Las tablas y las gráficas, con numeración arábiga, deben aparecer referenciadas dentro del texto mediante el número correspondiente. Las tablas deben ser diseñadas en forma que se facilite su presentación dentro del área de impresión de la Revista. En este sentido, los autores deben considerar en particular la extensión de las tablas, los dígitos representativos, los títulos y los encabezados. Las gráficas deben ser visualmente claras y debe ser posible modificar su tamaño. Cuando el artículo sea aceptado para su publicación, los autores deben poner la versión definitiva a disposición del Comité Editorial. Todos los elementos como barras, segmentos, palabras, símbolos y números deben estar impresos en color negro.

Responsabilidad legal

Los autores se hacen responsables por el uso de material con propiedad intelectual registrada como figuras, tablas, fotografías, etc.

Arbitraje

Los artículos recibidos serán revisados por el Comité Editorial y sometidos a arbitraje por pares especializados en el tema respectivo. El arbitraje es "doble ciego" (árbitros anónimos para los autores y viceversa). El Comité Editorial decide aceptar, rechazar o solicitar modificaciones a los artículos con base en las recomendaciones de los árbitros.

[4] http://tug.ctan.org/tex-archive/macros/latex/contrib/harvard