Revista Colombiana de Estadística

Volumen 33. Número 1 - junio - 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: Universidad Nacional de Colombia, Editorial, Tel. 57-1-3165000, Ext. 19645, Bogotá.

Revista Colombiana de Estadística	Bogotá	Vol. 33	Nº 1
ISSN 0120 - 1751	COLOMBIA	junio-2010	Págs. 1-166

Contenido

Editorial[1]

Beatriz Piedad Urdinola[a]

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

Las encuestas de opinión, tan de moda en este año electoral, deberían tener un anuncio parecido al de las cajas de cigarrillos que diga "léase con precaución". De esta manera se evitarían muchas de las innecesarias discusiones que se dan a diario en los medios de comunicación, en casi todos los casos por desconocimiento estadístico, en particular del área de muestreo. Esta ignorancia estadística, del público en general, ha llevado incluso a cambios en el diseño de las encuestas de opinión electoral en Colombia, en la contienda actual, pero con efectos desconocidos.

Es normal, hasta cierto punto, que los medios y el público en general desconozcan el significado académico detrás del argot del muestreo. Lo paradójico es que precisamente una entidad que no es técnica en estos temas, la Comisión Nacional Electoral (CNE), sea quien legisla sobre los parámetros de las firmas encuestadoras y las condiciones estadísticas de las encuestas electorales, tal como se puede verificar en la normatividad de la CNE (resoluciones 23 de 1996 y 50 de 1997). Este hecho sería intrascendente si detrás de dicha norma existiera una entidad técnica asesora, pero desafortunadamente la legislación más reciente refleja la ausencia de dicha asesoría.

Esto lo aseguro, pues un análisis de las encuestas de opinión electoral sobre las elecciones presidenciales estadounidenses de 2000, realizado por Erikson et al. (2004), muestra que quienes tienen mayor probabilidad de abstención en su voto en la elección real, tienden a votar por el partido demócrata. El seguimiento de dichas encuestas por tres meses de 2000 muestra cambios importantes en el tiempo de la composición de votantes clasificados como: 1. registrados, 2. probables y 3. poco probables. También muestra preferencias muy marcadas entre los dos principales contendores, para cada uno de estos tres tipos de votantes, y que el cambio en el tiempo del potencial ganador depende en gran medida de esta clasificación de los votantes. En Colombia, por el contrario, hasta el momento no se conoce ningún estudio similar que haya servido como palanca técnica en la última medida de la CNE de exigencia de filtros adicionales a las firmas encuestadoras de las encuestas de opinión.

Todo el revuelo que causa la situación en las conciencias éticas de los estadísticos no es exclusivo de Colombia. La mayoría de países donde se permiten las encuestas electorales sufren similares condiciones de interpretación por la prensa,

[1] Escrito en mayo de 2010.

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

legisladores y público en general. En respuesta se han creado organizaciones de corte académico que buscan acabar con dicha ignorancia estadística en la materia. Por ejemplo, la American Association for Public Opinion Research (AAPOR), de Estados Unidos (http://www.aapor.org), presenta continuamente documentos técnicos, y otros no tanto, sobre conceptos, técnicas de muestreo y diseño de encuestas.

Espero que este corto editorial motive a nuestros estudiantes y egresados en pro de una alternativa como la que tienen los estadounidenses, y que no dejen solo en manos de la formación que se imparte en las aulas y de la ética profesional de las empresas de muestreo este espacio que tanto necesita de romper con el círculo vicioso de ignorancia estadística en temas relevantes de la vida nacional.

Referencias

Erikson, R., Panagopoulus, C. & Wlezien, C. (2004), 'Likely (and unlikely) Voters and the Assessments of Campaign Dynamics', *Public Opinion Quarterly* **68**(4), 588–601.

Revista Colombiana de Estadística
Junio 2010, volumen 33, no. 1, pp. 1 a 11

A Probability Model for the Child Mortality in a Family

Un modelo probabilístico para la mortalidad en la infancia en una familia

Himanshu Pandey[a], Jai Kishun[b]

Department of Mathematics and Statistics, Gorakhpur University, Gorakhpur, India

Abstract

This paper proposed, under assumptions of inflated type fixed displaced geometric model, the distribution pattern of families according to number of child deaths within the first five years of life. The proposed model involves several parameters related to child mortality in a family, which is estimated with Method of Moments and Maximum Likelihood Estimation techniques. The proposed models fitted the observed data showing a better approximation at the survey area and draw some vital conclusions.

***Key words*:** F distribution, G-estimation, M-estimation, t-distribution, Mortality, Death, Contraception, Probability model.

Resumen

Este documento presenta, bajo el supuesto de un modelo geométrico de desplazamiento fijo, patrones de distribución de las familias, de acuerdo con el número de defunciones de sus hijos menores de cinco años. El modelo emplea diferentes parámetros relacionados con la mortalidad en la infancia en una familia, estimada con el método de momentos y de máxima verosimilitud. Los modelos propuestos ajustan los datos observados, mostrando mejor aproximación a la encuesta de área y describe algunas conclusiones vitales.

***Palabras clave*:** distribución F, estimación G, estimación M, Distribución t, mortalidad, anticoncepción, modelos de probabilidad.

[a]Professor. E-mail: himanshupandey@is.iiita.ac.in
[b]Professor. E-mail: jaikishan.stat@gmail.com

1. Introduction

Considerable interest has been shown in the past by several researchers to measure the levels of child mortality. Currently in the Developing Nations, the force of child mortality is still high at the younger ages particularly during the infancy. Infant and child mortality remain disturbingly high in developing countries despite the significant decline in most parts of the developed world. The state of the world's children indicated that about 12.9 million children die every year in the developing world (UNI 1987). Mortality for infants and child under the age of 5 years are expressed as the number of deaths in a given period. Infant mortality is defined as death during the first year of life and child mortality as that between the first and fifth birthdays. The deaths during childhood suffer from substantial degree of errors. Usually errors occurs due to recall laps which result in omission of events, misplacement of deaths and the distortion of reports on the duration of vital events.

The most important factors for child mortality are food shortages, contaminated water crowded and sub-standard housing, unchecked infectious diseases, the absence of day care facilities for the children of working mothers and the lack of minimally adequate and free medical care. The influences of most of these factors are absent in the present areas of these countries under study. However, some unmeasured genetic, environmental and behavioural components still remain nonnegligible.

In demography Child Mortality are useful as a Sensitive Index of a Nation's Health Conditions and as guided for the structuring of Public Health Programmes. Child Mortality is interrelated to social, cultural, economic, physiological and other factor. The high rate of infant and child mortality shows a low-level development of the health programme and also for the Nation's. Infant and Child mortality has been of interest of researchers and demographers because of its apparent relationship with fertility and indirect relationship with the acceptance of modern contraceptive methods (Kabir & Amir 1993).

Some attempts have been made to estimate the levels of child mortality by using data available from the different survey and other specific sources. Hill & Devid (1989) have suggested an approach for estimating child mortality from all births which have taken place in last five years before the survey. However, the estimate obtained through this method also suffers from the problem of under reporting (Pathak et al. 1991). In these circumstances, a number of attempts have been made to study the age pattern of mortality by using models (Goldblatt 1989, Heligman & Pollard 1980, Krishnan & Jin 1993, Ronald & Lawrence 1992, Thiele 1972). Initially, Keyfit (1977) used a hyperbolic function to study the infant and child mortality. Later, Arnold (1993) used pareto distribution; Krishnan & Jin (1993) and Chauhan (1997) applied finite range model for the same.

Brass (1995*a*, 1995*b*) is one of the proponents of indirect method of mortality estimation. He based his mortality estimate on retrospective data given by women of reproductive age on the number of children ever born and their status (either death or living). Other contributor in this line includes Preston & Palloni

(1977). However, indirect infant and child mortality estimates result from poor, inadequate and incomplete data, especially in developing countries. Most deaths outside hospital premises were not recorded and that many people do not record infant deaths because they only keep track of such occurrence as misfortunes, and when recorded, the age at death were either under or overstated.

The direct measures of mortality being not reliable, the problem may be overcome by the recently developed model building approaches which make it possible to obtain estimates from information other than vital statistics. In this connection, Srivastava (2001) has been proposed a probability model for the distribution of family according to number of child deaths (Deaths within the first five years of life). The main objective of this paper is to modify the Srivastava (2001) model by taking the finite range.

2. Probability Model

Let x denote the number of child deaths in a family at the survey point. Then the distribution of x is derived under the following assumption.

1. Only those families are considered in which at least one birth prior to the survey has occurred.
2. At the survey point, a family either has experienced a child loss or not. Let α and $(1-\alpha)$ be the respective proportions.
3. Out of α proportion of families, let β be the proportion of families in which only one child death has occurred.
4. Remaining $(1-\beta)\alpha$ proportion of families, experiencing multiple child deaths, follows a displaced geometric distribution with parameter p according to the number of child deaths.

Under these assumptions, the probability distribution of X is given by

$$\left.\begin{aligned} P[X=0] &= (1-\alpha),\ k=0 \\ P[X=1] &= \alpha\beta,\ k=1 \\ P[X=k] &= \frac{(1-\beta)\alpha p q^{k-2}}{1-q^N},\ k=2,3,\ldots,N \end{aligned}\right\} \tag{1}$$

Where p denotes the success of child deaths in a family and $q=1-p$. If we put $q^N=0$, then the proposed model (1) reduced to Srivastava (2001). The proposed model (1) is an improvement over the Srivastava (2001) model by taking a concept of finite range model i.e. $k=0,1,2,3,\ldots,N$.

2.1. Estimation Method of moment

The method of moments is discussed to estimate the four parameters α, β, p and N of the probability function (1). Now it is difficult to estimate all these

parameters, so it is assumed that N is the maximum numbers of child deaths occurred. Let (x, f) be a frequency distribution whose parameters to be estimated from the observed distributions of families according to the number of child deaths. The following estimation technique is employed for estimating rest of the parameter.

Equating proportions of zeroth cell, first cell frequency and simple mean to their corresponding observed values respectively, which is converted into the following equations,

$$\frac{f_0}{f} = 1 - \alpha \tag{2}$$

$$\frac{f_1}{f} = \alpha\beta \tag{3}$$

$$\overline{X} = \alpha\beta + (1-\beta)\alpha \left[\frac{\left(1 - q^{N-1}\right)}{p(1-q^N)} - \frac{Nq^{N-1}}{(1-q^N)} + \frac{1}{(1-q^N)} \right] \tag{4}$$

Where,

$f_0 =$ Observed value of zeroth cell frequency

$f_1 =$ Observed value of first cell frequency

$f =$ Total number of observations $= \sum_i f_i$

$\overline{X} =$ Sample mean of the observed values.

2.2. Method of Maximum Likelihood

The proposed model involves four parameters α, β, p and N to be estimated from the observed distribution of families according to the number of child deaths, but it cannot be possible to estimate all these simultaneously by this method, so the value of N has been taken as method of moments. Let $x_1, x_2, x_3, \ldots, x_n$ be a random sample of size N from the population (1). The likelihood function L for the given sample can be expressed as

$$L = (1-\alpha) f_0 (\alpha\beta) f_1 \times \left[\frac{(1-\beta)\alpha p}{1-q^N} \right]^{f_2} \left[\frac{\alpha\{1-\beta-(1-\beta)p\}}{1-q^N} \right]^{f-f_0-f_1-f_2} \tag{5}$$

Taking log both sides, we get

$$\log L = f_0 \log(1-\alpha) + f_1 \log(\alpha\beta) + f_2 \log \left[\frac{(1-\beta)\alpha p}{1-q^N} \right] + (f - f_0 - f_1 - f_2) \log \left[\alpha(1-\beta) \left\{ 1 - \frac{p}{1-q^N} \right\} \right]$$

Now, partially differentiating w.r.t. α, β and p respectively and equating to zero.

$$\frac{\partial \log L}{\partial \alpha} = -\frac{f_0}{(1-\alpha)} + \frac{f - f_0}{\alpha} = 0 \tag{6}$$

$$\frac{\partial \log L}{\partial \beta} = \frac{f_1}{\beta} + \frac{f - f_0 - f_1}{1-\beta} = 0 \tag{7}$$

$$\frac{\partial \log L}{\partial p} = \frac{f_2\left[\left(1-q^N\right) - pN(1-p)^{N-1}\right]}{p\left(1-q^N\right)} -$$
$$(f - f_0 - f_1 - f_2)\left[\frac{\left(1-q^N\right) - pN(1-P)^{N-1}}{\left[\left(1-p^N\right) - p\right]\left(1-p^N\right)}\right] = 0 \tag{8}$$

Solving equations (6), (7) and (8), the estimate of α, β and p can easily be obtained as

$$\alpha = \frac{f - f_0}{f}$$

$$\beta = \frac{f_1}{f - f_0}$$

$$\frac{p}{\left(1-q^N\right)} = \frac{f_2}{f - f_0 - f_1}$$

The second partial derivatives of $\log L$ obtained is

$$\frac{\partial^2 \log L}{\partial \alpha^2} = -\frac{f_0}{(1-\alpha)^2} - \frac{f - f_0}{\alpha^2} \tag{9}$$

$$\frac{\partial^2 \log L}{\partial \beta^2} = -\frac{f_1}{\beta^2} - \frac{f - f_0 - f_1}{(1-\beta)^2} \tag{10}$$

$$\frac{(\partial^{\uparrow} 2 \log L)}{(\partial p^{\uparrow} 2)} = \frac{(f_{\downarrow} 2[N(1-N)(1-p)^{\uparrow}(N-2)])}{(1-q^{\uparrow}N)}$$
$$- \frac{(f - f_{\downarrow}0 - f_{\downarrow}1 - f_{\downarrow}2)\left[\left\{\left\{(1-q^{\uparrow})^{\uparrow}2p(1-q^{\uparrow})\right\}\left\{N(N-1)p(1-p)^{\uparrow}(N-2)\right\}\right\}\right]}{\left[(1-q^{\uparrow}N)^{\uparrow}2\right]} \tag{11}$$

Now, partial derivative of $\frac{\partial \log L}{\partial \alpha}$, $\frac{\partial \log L}{\partial \beta}$ and $\frac{\partial \log L}{\partial p}$ w.r. to, β, p and α respectively we get as following,

$$\frac{\partial^2 \log L}{\partial \alpha \partial \beta} = \frac{\partial^2 \log L}{\partial \beta \partial p} = \frac{\partial^2 \log L}{\partial p \partial \alpha} = 0 \tag{12}$$

Here,

$$E(f_0) = f(1-\alpha)$$

$$E(f_1) = f\alpha\beta$$

$$E(f_2) = \frac{f(1-\beta)\alpha p}{1-q^N}$$

$$E(f - f_0 - f_1 - f_2) = f\alpha(1-\beta)\left\{1 - \frac{p}{1-q^N}\right\}$$

Using the above facts, the expected value of the second partial derivatives obtained as

$$\phi_{11} = E\frac{\left[\frac{-\partial^2 \log L}{\partial \alpha^2}\right]}{f} = \left[\frac{1}{1-\alpha} + \frac{1}{\alpha}\right] \tag{13}$$

$$\phi_{22} = E\frac{\left[\frac{-\partial^2 \log L}{\partial \beta^2}\right]}{f} = \alpha\left[\frac{1}{1-\beta} + \frac{1}{\beta}\right] \tag{14}$$

$$\phi_{22} = E\frac{\left[\frac{-\partial^2 \log L}{\partial \beta^2}\right]}{f} = \alpha\left[\frac{1}{1-\beta} + \frac{1}{\beta}\right] \tag{15}$$

the covariance between the estimators becomes zero since

$$E\left(\frac{\partial^2 \log L}{\partial \alpha \partial \beta}\right) = E\left(\frac{\partial^2 \log L}{\partial \beta \delta p}\right) = E\left(\frac{\partial^2 \log L}{\partial \alpha \partial p}\right) = 0 \tag{16}$$

Thus, the asymptotic variances of the estimator can be obtained as

$$V(\widehat{\alpha}) = \frac{1}{\phi_{11}}, V\left(\widehat{\beta}\right) = \frac{1}{\phi_{22}}, V(\widehat{p}) = \frac{1}{\phi_{33}} \tag{17}$$

3. Application

The suitability of the proposed model is examined to the study that has been conducted in North-Eastern Libya stretching from Benghazi to Emsaad. From the study area, 7 localities out of 27 have been selected by probability proportional to numbers of families in the localities. The data on fertility and mortality under age 5 along with some other demographic characteristics have been collected from 1,252 couples of childbearing ages of selected localities. About one-third (35.7 percent) of the investigated mothers have lost at least one child. The percentage of multiple child loss mothers is 11.3 and these mothers have given, one an average, 10 or more births. The differential in child loss by fertility level is highly significant. However, child mortality to mothers having lower and differential in child mortality by fertility in north-eastern Libya 325 medium (6 ever born children) fertility is similar. This study indicates that high parity and high mortality move in the same direction (Bhuyan & Deogratias 1999) and one set of data has been taken from a Household Sample Survey in Brazil in 1987. Details are given in Sastry (1997). Other two set of sample data were collected under a Survey entitled "Effect of breastfeeding on fertility in North Rural India" in 1995 and "A Demographic survey on fertility and mortality in rural Nepal: A Study of Palpa and Rupandehi Districts" in 2000. The details of these two set of data are given in Srivastava (2001).

The parameters of the proposed model have been estimated by the method of moment and method of maximum likelihood. The estimated values of different parameters are given in tables 1 to 4 for the child deaths.

The estimated value of α are 0.3753, 0.2139, 0.2683 and 0.3570 for India, Nepal, North East Brazil and North East Libya, respectively. It represents that the proportion of families experiencing a child loss was found slightly higher in India (0.3753) than North East Libya (0.3570), North East Brazil (0.2683) and Nepal (.2139). The estimate of β are 0.5857, 0.7528, 0.6560 and 0.6846, respectively, for India, Nepal, North East Brazil and North East Libya. It means that the proportion of families having only one child death was found greater for Nepal (0.7527) as compared to other countries. The estimated values for the probability of success of death p are 0.5889, 0.6257, 0.6490 and 0.6630 by the method of moment and 0.6022, 0.7110, 0.6178 and 0.6592 by the maximum likelihood, respectively, for the above mentioned countries. The average number of child death per family $\alpha\beta + (1-\beta)\alpha\left[\frac{(1-q^{N-1})}{p(1-q^N)} - \frac{Nq^{(N-1)}}{(1-q^N)} + \frac{1}{(1-q^N)}\right]$ for Eastern Uttar Pradesh (India), Nepal, North East Brazil and North East Libya were found to be 0.63, 0.30, 0.41 and 0.53 respectively. This show that, on an average, the child mortality is high in Eastern Uttar Pradesh (India). The exact variances of the estimators obtained by maximum likelihood method are also given. for Eastern Uttar Pradesh (India), Nepal, North East Brazil and North East Libya were found to be 0.63, 0.30, 0.41 and 0.53 respectively. This show that on an average the child mortality is high in Eastern Uttar Pradesh (India). The exact variances of the estimators obtained by maximum likelihood method are also given.

Changes in levels of mortality may be attributed to socioeconomic factors such as improvements in primary health care services, control of epidemics, availability of health care facilities, and with the improvement in economic condition among lower parity women, there is a downward shift in child mortality. However economic condition and mortality move in the same direction among high parity women. Differential impacts of age of female spouse at marriage are observed among mothers of different parity level.

A inflated geometric distribution for finite range provides a suitable description of child mortality at micro level, i.e. at the family level (Tables 1 to 4). The value of χ^2 are insignificant at 5 percent level of significance for all set of data. The proposed model fitted satisfactorily and described the pattern of child mortality to several sets of sample data in Indian Subcontinents.

TABLE 1: Distribution of Observed and Expected Number of Families, according to the Number of Child Deaths in Eastern Uttar Pradesh (India).

Number of child dead	Observed number families	Method of Moment (Expected no. of families)	Method of Maximumum Likelihood (Expected no. of families)
0	506	506.0070	506.0070
1	178	178.0490	178.0490
2	76	74.3124	75.9659
3	32	30.5520	30.2177
4	8	12.5609	12.0200
5	6	5.1633	4.7792
6	3	2.1226	1.9007
7	1	1.1628	1.0529
Total	810	810.0000	810.0000
$\widehat{\alpha}$		0.3753	0.3753
$\widehat{\beta}$		0.5857	0.5857
$\widehat{p}$		0.5889	0.6023
$V(\widehat{\alpha})$			0.000289
$V(\widehat{\beta})$			0.000798
$V(\widehat{p})$			0.003006
χ^2		2.2841	2.3983
$d.f.$		4	4

Source: Srivastava (2001)

TABLE 2: Distribution of Observed and Expected number of Families, According to the Number of Child Deaths in Nepal.

Number of child dead	Observed number families	Method of Moment (Expected no. of families)	Method of Maximumum Likelihood (Expected no. of families)
0	669	668.9711	668.9711
1	137	137.0314	137.0314
2	32	28.1500	31.6996
3	6	10.5365	9.2479
4	3	3.9438	2.6726
5	2	1.4762	0.7724
6	2	0.5525	0.5232
7	0	0.3385	0.0768
Total	851	851.0000	851.0000
$\widehat{\alpha}$		0.2139	0.2139
$\widehat{\beta}$		0.7528	0.7528
$\widehat{p}$		0.6257	0.7110
$V(\widehat{\alpha})$			0.000196
$V(\widehat{\beta})$			0.0001022
$V(\widehat{p})$			0.006352
χ^2		7.0227	7.3799
$d.f.$		4	4

Source: Srivastava (2001)

TABLE 3: Distribution of Observed and Expected Number of Families, according to the Number of Child Deaths in North East Brazil.

Number of child dead	Observed number families	Method of Moment (Expected no. of families)	Method of Maximumum Likelihood (Expected no. of families)
0	769	769.0167	769.0167
1	185	184.9810	184.9810
2	60	62.9986	59.9999
3	26	22.1125	22.9320
4	9	7.7615	8.7646
5	1	2.7243	3.3498
6	1	0.9562	1.2803
7	0	0.4492	0.6757
Total	1051	1051.0000	1051.0000
$\widehat{\alpha}$		0.2683	0.2683
$\widehat{\beta}$		0.6560	0.6560
$\widehat{p}$		0.6490	0.6178
$V(\widehat{\alpha})$			0.000215
$V(\widehat{\beta})$			0.000800
$V(\widehat{p})$			0.004051
χ^2		2.5663	2.8021
$d.f.$		4	4

Source: Sastry (1997)

TABLE 4: Distribution of Observed and Expected Number of Families, according to the Number of Child Deaths in North East Libya.

Number of child dead	Observed number families	Method of Moment (Expected no. of families)	Method of Maximumum Likelihood (Expected no. of families)
0	805	805.0360	805.0360
1	306	305.9916	305.9916
2	93	93.5115	92.9755
3	36	31.5134	31.6953
4	7	10.6200	10.8049
5	2	3.5789	3.6834
6	1	1.2061	1.2557
7	2	0.5425	0.5576
Total	1252	1252.0000	1252.0000
$\widehat{\alpha}$		0.3570	0.3570
$\widehat{\beta}$		0.6846	0.6846
$\widehat{p}$		0.6630	0.6592
$V(\widehat{\alpha})$			0.000172
$V(\widehat{\beta})$			0.000483
$V(\widehat{p})$			0.002457
χ^2		6.5231	6.4771
$d.f.$		4	4

Source: Bhuyan and Deogratias (1999)

4. Conclusion

From the above discussion of proposed model related with child mortality it is concluded that infant and child mortality are still higher in Developing countries like India, Nepal, N-E Brazil and N-E Libya as shown in Tables 1 to 4. Changes in the level of child mortality is directly or indirectly associated with socioeconomic factors prevailing in that countries.

To overcome the child mortality spread of education is first and foremost essential factor. By imparting education to the parents and family members they will be conscious and aware to the health of the child. To keep their child healthy and sound they will follow the advice of the doctors from the date when the mother is conceived. Parents will adhere the time schedule to vaccinate the child to control different diseases. They will also care for hygienic atmosphere and cleanness.

Keeping all viewpoints, the proposed model is an essential tool for predicting the child mortality level of any country and it can be used as an indicator of good health condition of the society.

Acknowledgement

The authors are very thankful to referees for their valuable suggestions.

[Recibido: julio de 2009 — Aceptado: diciembre de 2009]

References

Arnold, B. C. (1993), 'Pareto Distributions', *Statistical Distributions* **5**.

Bhuyan, K. C. & Deogratias, R. (1999), 'On a Probability Model for Child Mortality Pattern in North East Libya', *Turkish Journal of Population Studies* (21), 33–38.

Brass, W. (1995*a*), A Simple approximation for the time Location of Estimates of Child Mortality from Proportions dead by age of Mother, *in* 'Advances in Methods for Estimating Fertility and Mortality from Limited and Defective Data', London School of Hygiene and Tropical Medicine, London, pp. 1–16.

Brass, W. (1995*b*), The Derivative of Life Tables from Retrospective Estimates of Child and Adult Mortality, *in* 'Advances in methods for estimating fertility and mortality from limited and defective data', London School of Hygiene and Tropical Medicine, London.

Chauhan, R. K. (1997), 'Graduation of Infant Deaths by Age', *Demography India* **2**(26), 261–274.

Goldblatt, P. O. (1989), 'Mortality by Social Class, 1971-85', *Population Trends* (56), 6–15.

Heligman, L. & Pollard, J. H. (1980), 'Age Pattern of Mortality', *Journal of Institute of Actuaries* (117), 49–80.

Hill, A. G. & Devid, H. P. (1989), Measuring Child Mortality in the Third World, *in* N. Sources & N. Approaches, eds, 'IUSSP Proceeding of International Conference', New Delhi, India.

Kabir, M. & Amir, R. (1993), 'Factors Influencing Child Mortality in Bangladesh and Their Implications for the National Health Programme', *Asia-Pacific Population Journal* **8**(3), 31–46.

Keyfit, N. (1977), *Applied Mathematical Demography*, John Wiley, New York, United States.

Krishnan, P. & Jin, Y. (1993), 'A Statistical Model of Infant Mortality', *Janasamkhya* **11**(2), 67–71.

Pathak, K. B., Pandey, A. & Mishra, U. S. (1991), 'On Estimating current Levels of Fertility and Child Mortality from the Data on Open Birth Interval and Survival Status of the last Child', *Janasamkhya* **9**(1), 15–24.

Preston, S. H. & Palloni, A. (1977), 'Fine-Tuning Brass Type Mortality Estimates with Data on Ages of Surviving Children', *Population Bulletin of the United Nations* (10).

Ronald, D. L. & Lawrence, R. C. (1992), 'Modeling and Forecasting U.S. Mortality', *Journal of American Statistical Associations* (87), 659–675.

Sastry, N. (1997), 'A Nested Frailty Model for Survival Data, with an Application to the Study of Child Survival in North East Brazil', *Journal of the American Statistical Association* **92**.

Srivastava, S. (2001), Some Mathematical Models in Demography and Their Application, Doctoral thesis, Banaras Hindu University, Varanasi, India.

Thiele, P. N. (1972), 'On Mathematical Formula to Express the Rate of Mortality Throughout the Whole of Life', *Journal of Institute of Actuaries* (16), 313.

UNI (1987), *The State of the World's Children*.

Revista Colombiana de Estadística
Junio 2010, volumen 33, no. 1, pp. 13 a 24

Distribución de probabilidad que involucra algunas funciones hipergeométricas generalizadas

Probability Distributions Involving on Generalized Hypergeometric Functions

Rafael Alfonso Meléndez[a], Jaime Antonio Castillo[b], Carlos Jesús Jiménez[c]

Centro de Investigaciones, Grupo de Investigación GIMA, Universidad de la Guajira, Riohacha, Colombia

Resumen

Se define una nueva función de probabilidad que involucra algunas funciones hipergeométricas generalizadas; se encontraron algunas propiedades y casos especiales como la gamma y la exponencial. Se establecieron algunas funciones básicas asociadas a la nueva distribución de probabilidad, como la media, momentos, función característica, y se obtienen representaciones gráficas para esta nueva función de probabilidad.

***Palabras clave*:** función de densidad de probabilidad, función hipergeométrica generalizada, función generadora de momento, función característica.

Abstract

We define a new function of probability that involves some generalized hypergeometric functions, we found some properties and special cases such as gamma and exponential. We establish some basic functions associated with the new probability distribution like mean, the moments, characteristic function and several graphic representations are obtained for this new function of probability.

***Key words*:** Probability function, Generalized hypergeometric functions, The moments, Characteristic function.

[a]Profesor asociado. E-mail: melendez24@hotmail.com
[b]Profesor titular. E-mail: jacas68@yahoo.es
[c]Profesor asociado. E-mail: carlosj114@gmail.com

1. Introducción

Muchas funciones especiales de matemáticas aplicadas pueden expresarse en términos de funciones hipergeométricas, las cuales son clases importantes de funciones especiales. La función hipergeométrica y sus generalizaciones han sido usadas en varios problemas de la estadística (Lebedev 1965, Nakhi & Kalla 2005), particularmente en el estudio de nuevas funciones de densidad de probabilidad generalizadas y sus propiedades estadísticas, las cuales tienen diversas aplicaciones no solo en la teoría de confiabilidad, sino también en algunos problemas asociados con tasas demográficas, biomedicina, datos de tráfico y fallas de equipos electrónicos (Virchenko et al. 2001). Consideremos el problema de resolver la ecuación diferencial lineal

$$z(1-z)u'' + [\gamma - (\alpha+\beta+1)z]\,u' - \alpha\beta u = 0 \tag{1}$$

donde z es una variable compleja, y γ,α,β son parámetros que pueden tomar valores reales o complejos. Reduciendo (1) a la forma estándar dividiendo por el coeficiente u', obtenemos una ecuación cuyos coeficientes son funciones analíticas de z en el dominio $0 < |z| < 1$. Esto sigue de la teoría general de ecuaciones diferenciales lineales, donde (1) tiene una solución particular (Virchenko et al. 2001).

$$u = z^s \sum_{k=0}^{\infty} c_k z^k$$

donde $c_0 \neq 0$, s es número convenientemente escogido, y así la serie de potencia converge en $|z| < 1$.

Para valores de $\gamma \neq 0, -1, -2, \ldots$, una solución particular está dada por

$$u = F(\alpha, \beta; \gamma; z) = \sum_{k=0}^{\infty} \frac{(\alpha)_k (\beta)_k}{(\gamma)_k} \frac{z^k}{k!}$$

que se conoce como la serie hipergeométrica de Gauss.

A continuación veremos algunas distribuciones de probabilidad establecidas recientemente por diferentes autores. Good (1953) introdujo la siguiente distribución gaussiana inversa

$$g(t) = \frac{1}{A(\alpha, a, b)} t^{\alpha-1} e^{-at-b/t}$$

$$a, b, t > 0;\ -\infty < \alpha < \infty$$

donde

$$A(\alpha, a, b) = \left[\int_0^{\infty} t^{\alpha-1} e^{-at-b/t} dt\right]^{-1}$$

esta distribución gaussiana inversa se plantea como la función de densidad de primer paso de tiempo con movimiento browniano con derivada positiva (Jorgensen 1982). Tales modelos han sido usados por Hoem (1976) y Jorgensen (1982) en la teoría de confiabilidad y teoría de tasas demográficas; este último estudió varias

aplicaciones de la distribución anterior, asociadas con daños de equipos de aire acondicionados y datos de tráfico. En Lebedev (1965) y Mathais (1993) se presentan otras aplicaciones de las funciones especiales a la teoría de confiabilidad. En un trabajo reciente, Agarwal & Kalla (1996) desarrollaron una nueva distribución tipo gamma generalizada, con función de densidad:

$$f(x) = \frac{\beta\alpha^{m/\beta}}{\Gamma_\lambda(m/\beta, n)} x^{m-1} \left(\alpha x^\beta + n\right)^{-\lambda} e^{-\alpha x^\beta}, \quad \alpha, m, n < 0$$

donde

$$\Gamma_\lambda(m, n) = \int_0^\infty x^{m-1} e^{-x} (x+n)^{-\lambda} \, dx, \quad m > 0$$

siendo esta la función gamma generalizada de Kobayashi (1991), la cual es esencialmente una función hipergeométrica confluente de segunda clase (Agarwal & Kalla 1996). Motivados por sus resultados, Agarwal & Kalla (1996) y Ghitany (1998) obtienen algunas propiedades adicionales para esta distribución. Recientemente, Al-Musallam & Kalla (1998), Al-Saqabi et al. (2002),Virchenko et al. (2001) y Virchenko (1999) definieron y desarrollaron algunas funciones hipergeométricas-τ y confluente-τ que son generalizaciones de las funciones hipergeométricas de Gauss y funciones hipergeométricas confluentes Kummer.

El presente trabajo tiene como objeto definir una nueva función de densidad generalizada a partir de algunas funciones hipergeométricas generalizadas; para esto se hará uso de las representaciones integrales y en serie doble; a partir de esta función de densidad $f(x)$ se encuentran algunas propiedades que permiten caracterizarla, como la función generadora de momento, los momentos, la función característica, la función tasa de riesgo y algunos casos especiales. Se muestran algunas figuras las cuales corresponden a la simulación de esta nueva función de densidad para diferentes valores de los parámetros.

Galué et al. (2005) definen algunas generalizaciones que involucran a cuatro series de Appell definidas por Humbert (1920), introducen dos nuevos parámetros τ y τ' y definen sus representaciones en serie e integrales. A partir de estas se obtienen nuevas distribuciones de probabilidad que involucran estas generalizaciones. Con esta nueva función de densidad generalizada se encuentran casos especiales como una generalización con menos parámetros, la gamma y la exponencial, los momentos y sus casos particulares como el valor esperado y la varianza, la función generadora de momento, función característica, la función tasa de riesgo.

2. Generalización de algunas funciones hipergeométricas de dos variables

La generalización de las funciones hipergeométricas de dos variables está asociada con la generalización de la función hipergeométrica de Gauss propuesta por Virchenko (1999), quien la introdujo de la siguiente forma:

$${}_2R_1^\tau(z) \equiv {}_2R_1(a, b; c; \tau; z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{k=0}^{\infty} \frac{\Gamma(a+k)\Gamma(b+\tau k)}{\Gamma(c+\tau k)} \frac{z^k}{k!} \tag{2}$$

$$\tau > 0, |z| < 1, c \neq 0, -1, -2, \ldots$$

Esta función tiene la siguiente representación integral:

$$ {}_2R_1(a, b; c; \tau; z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-b)} \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-zt^{\tau})^{-a} dt \tag{3}$$

$$\tau > 0, \Re(c) > \Re(b) > 0$$

Para

$$\tau = 1,\ {}_2R_1(a, b; c; \tau; z) = {}_2F_1(a, b; c; z)$$

donde ${}_2F_1$ es la función hipergeométrica de Gauss. Similarmente la función hipergeométrica confluente se define como

$$\Phi^{\tau} = \Phi^{\tau}(a; b; c) = \frac{\Gamma(c)}{\Gamma(a)} \sum_{k=0}^{\infty} \frac{\Gamma(a+k)}{\Gamma(c+\tau k)} \frac{z^k}{k!} \tag{4}$$

$$\tau > 0, |z| < 1, c \neq 0, -1, -2, \ldots$$

Anteriormente se presentaron algunas generalizaciones de funciones hipergeométricas de dos variables donde se introduce el parámetro τ y a continuación se relacionan unas generalizaciones que involucran algunas funciones de Humbert.

2.1. Generalización de algunas funciones de Humbert

Siete formas confluentes de las cuatro series de Appell fueron definidas por Humbert (1920), denotadas por: $\Phi_1, \Phi_2, \Phi_3, \Psi_1, \Psi_2, \Xi_1, \Xi_2$.

Recientemente Galué et al. (2005) consideraron una extensión de las funciones de Humbert Ψ_1, Ψ_2, Ξ_1 y Ξ_2 introduciendo parámetros adicionales τ, τ', y establecieron sus representaciones en serie e integral.

Las generalizaciones τ de las funciones confluentes de dos variables Ξ_1 y Ξ_2 pueden expresarse en términos de la función ${}_2R_1(a, b; c; \tau; w)$ en la forma siguiente:

$$\Xi_1^{\tau,\tau'}(a, a', b; c; w, z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(a')} \sum_{k,l=0}^{\infty} \frac{\Gamma(a+\tau k)\Gamma(a'+\tau' l)}{\Gamma(c+\tau k+\tau' l)} \frac{w^k}{k!} \frac{z^l}{l!} \tag{5}$$

$$\Xi_1^{\tau,\tau'}(a, a', b; c; w, z) = \frac{\Gamma(c)}{\Gamma(a')} \sum_{l=0}^{\infty} \frac{\Gamma(a'+\tau' l)}{\Gamma(c+\tau' l)}\ {}_2R_1(b, a; c+\tau' l; \tau; w) \frac{z^l}{l!} \tag{6}$$

$$\tau, \tau' > 0, |w| < 1, c + \tau' l \neq 0, -1, -2, \ldots$$

$$\Xi_2^{\tau}(a, b; c; w, z) = \frac{\Gamma(c)}{\Gamma(a)} \sum_{k,l=0}^{\infty} \frac{\Gamma(a+\tau k)(b)_k}{\Gamma(c+\tau k+l)} \frac{w^k}{k!} \frac{z^l}{l!} \tag{7}$$

$$\Xi_2^{\tau}(a,b;c;w,z) = \sum_{l=0}^{\infty} \frac{1}{(c)_l}\ {}_2R_1(b,a;c+l;\tau,w)\frac{z^l}{l!} \tag{8}$$

$$\tau > 0, |w| < 1, c \neq 0, -1, -2, \ldots$$

donde $(c)_n$ denota el símbolo de Pochhammer $(c)_n = \Gamma(c+n)/\Gamma(c)$.

2.2. Funciones de Bessel

A continuación se define la función de Bessel modificada de primera clase

$$I_{\upsilon}(z) = \sum_{k=0}^{\infty} \frac{(z/2)^{\upsilon+2k}}{\Gamma(k+1)\Gamma(k+\upsilon+1)}, \qquad |z| < \infty, \quad |\arg\ z| < \pi \tag{9}$$

A continuación se muestra una representación integral definida por Galué para la función Ξ_2 y algunas propiedades, las cuales permitirán encontrar la nueva distribución de probabilidad generalizada y sus propiedades.

2.3. Representación integral

Galué et al. (2005) presentó además la representación integral para la generalización τ de la función de Humbert Ξ_2, de la siguiente forma:

$$\int_0^{\infty} x^{\alpha-1}\ e^{-px}\ \Xi_2^{\tau}(a,b;c;w,xz)dx = \frac{\Gamma(\alpha)}{p^{\alpha}}\ \Xi_1^{\tau,1}\left(a,\alpha,b;c;w,\frac{z}{p}\right) \tag{10}$$

$$\tau, \operatorname{Re} p, \operatorname{Re}\alpha, \operatorname{Re}(p-z) > 0, |w| < 1$$

Algunas propiedades. Galué et al. (2005) establecieron también algunas propiedades para las extensiones de Humbert de la siguiente manera:

$$\Xi_1^{\tau,\tau'}(a,a',b;c;w,0) = {}_2R_1(b,a;c;\tau;w);\ |w| < 1 \tag{11}$$

$$\Xi_2^{\tau}(a,b;c;w,0) = {}_2R_1(b,a;c;\tau;w);\ |w| < 1 \tag{12}$$

$$\Xi_1^{\tau,\tau'}(a,a',b;c;0,z) = {}_1\Phi_1^{\tau'}(a';c;z) \tag{13}$$

$$\Xi_2^{\tau}(a,b;c;0,z) = \Gamma(c)\ z^{-(c-1)/2}\ I_{c-1}(2\sqrt{z}) \tag{14}$$

2.4. Función gamma incompleta y gamma generalizada

Una nueva función gamma generalizada puede considerarse utilizando Ξ_2^τ definida en (7), de la siguiente manera:

$$ {}_\tau\Gamma(\alpha, p; a, b; c; w, x) = \int_0^\infty t^{\alpha-1}\ e^{-pt}\Xi_2^\tau(a, b; c; w, tz)\, dt \tag{15}$$

$$\tau \in \mathbb{R}, \tau, \operatorname{Re} p, \operatorname{Re}\alpha, \operatorname{Re}(p - z) > 0$$

Definimos la siguiente función gamma generalizada incompleta

$$ {}_\tau^w\Gamma_0(\alpha, p; a, b; c; w, x) = \int_0^w t^{\alpha-1}\ e^{-pt}\ \Xi_2^\tau(a, b; c; w, tz)\, dt \tag{16}$$

$$\tau, \operatorname{Re} p, \operatorname{Re}\alpha, \operatorname{Re}(p - z) > 0$$

La función gamma incompleta generalizada complementaria se define como

$$ {}_\tau^\infty\Gamma_w(\alpha, p; a, b; c; w, x) = \int_w^\infty t^{\alpha-1}\ e^{-pt}\ \Xi_2^\tau(a, b; c; w, tz)\, dt \tag{17}$$

$$\tau, \operatorname{Re} p, \operatorname{Re}\alpha, \operatorname{Re}(p - z) > 0$$

3. Una función de densidad de probabilidad generalizada

En esta sección usaremos la generalización de la función hipergeométrica de dos variables Ξ_2^τ, establecida por Galué et al. (2005), para definir la siguiente función de densidad de probabilidad.

$$f(x) = \frac{p^\alpha\ x^{\alpha-1}e^{-px}\ \Xi_2^\tau(a, b; c; w, xz)}{\Gamma(\alpha)\ \Xi_1^{\tau,1}(a, \alpha, b; c; w, z/p)} \tag{18}$$

$$\tau, \operatorname{Re} p, \operatorname{Re}\alpha, \operatorname{Re}(p - z) > 0, |w| < 1, x > 0$$

Propiedades:

i)	$f(x) = 0$	para $\alpha > 1$ y $x = 0$
ii)	$f(x) = \frac{p^\alpha\ x^{\alpha-1}\ e^{\alpha-1}}{\Gamma(\alpha)}$	para $b = \alpha > 1$ y $z = 0$
iii)	$f(x) \to \infty$	cuando $x \to 0^+$ y $\alpha < 1$
iv)	$f(x) \to 0$	cuando $x \to \infty$ y $\alpha < 1$

3.1. Algunos casos especiales

1. Para $\tau = 1$, obtenemos una nueva función de densidad de probabilidad involucrando la generalización τ de la función confluente de dos variables Ξ_2.

$$f(x) = \frac{p^\alpha \; x^{\alpha-1} \; e^{-px} \;\; \Xi_2(a, b; c; w, xz)}{\Gamma(\alpha) \; \Xi_1^{\tau,1}(a, \alpha, b; c; w, z/p)} \tag{19}$$

$$\operatorname{Re} p, \operatorname{Re} \alpha, \operatorname{Re}(p - z) > 0, |w| < 1$$

2. Para $\tau' = 1, \;\; w = 0$, y utilizando las propiedades (13) y (14) en (18) se obtiene una distribución con cuatro parámetros

$$f(x) = \frac{p^\alpha \; x^{2(\alpha-c-1)} \; e^{-px} \; \Gamma(c) \; z^{-(c-1)/2} \; I_{c-1}(2\sqrt{xz})}{\Gamma(\alpha) \; {}_1\Phi_1(a'; c; z/p)} \tag{20}$$

$$\operatorname{Re} p, \operatorname{Re} \alpha, \operatorname{Re}(p - z) > 0$$

3. Para $z = 0$ y $a' = b = \alpha$ y utilizando las propiedades (10) y (11) en (18) obtenemos la distribución gamma

$$f(x) = \frac{p^\alpha \; x^{\alpha-1}}{\Gamma(\alpha)} \; e^{-px} \tag{21}$$

$$\operatorname{Re} p, \operatorname{Re} \alpha > 0, |w| < 1$$

4. Para $\alpha = 1$ en (21)

$$f(x) = p \; e^{-px} \tag{22}$$

$$\operatorname{Re} p > 0, x > 0$$

se obtiene la bien conocida distribución exponencial.

3.2. Los momentos

El n-ésimo momento μ'_n con respecto al origen de una variable aleatoria continua X con función de densidad $f(x)$ se define como

$$\mu'_n = \int_{-\infty}^{\infty} x^n \; f(x) dx$$

Para la función de densidad $f(x)$, dada por (18) se consideran distribuciones de soporte positivo, dado que estas involucran funciones tipo gamma, la cual es continua sobre los reales positivos.

$$\mu'_n = E(x^n) = \int_0^\infty x^{n+\alpha-1} e^{-px} \frac{\Gamma(c)}{\Gamma(a)} \sum_{k,l=0}^{\infty} \frac{\Gamma(a + \tau k)(b)_k}{\Gamma(c + \tau k + l)} \frac{w^k}{k!} \frac{(xz)^l}{l!} dx.$$

Usando la expresión de la serie de la función Ξ_2^τ dada por (7) se tiene

$$\Xi_2^\tau(a,b;c;w,z) = \frac{\Gamma(c)}{\Gamma(a)} \sum_{k,l=0}^{\infty} \frac{\Gamma(a+\tau k)(b)_k}{\Gamma(c+\tau k+l)} \frac{w^k}{k!} \frac{z^l}{l!} \int_0^\infty x^{n+\alpha+l-1} e^{-px} dx$$

resolviendo la integral se tiene el siguiente resultado

$$\mu'_n = \frac{(\alpha)_n}{p^n} \frac{\Xi_1^{\tau,1}(a,\alpha+n;c;w,z/p)}{\Xi_1^{\tau,1}(a,\alpha,b;c;w,z/p)} \tag{23}$$

$$\tau, \operatorname{Re} p, \operatorname{Re} \alpha, \operatorname{Re}(p-z) > 0, |w| < 1, n = 1,2,3,\ldots$$

Casos especiales. El momento μ'_n para $n = 1$, denotado por $E(x)$, llamado la media, está dado por

$$\mu'_1 = E(x) = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\ p} \frac{\Xi_1^{\tau,1}(a,\alpha+1;c;w,z/p)}{\Xi_1^{\tau,1}(a,\alpha;c;w,z/p)} \tag{24}$$

La varianza de una variable aleatoria X de la función de probabilidad $f(x)$ definida por (13) con media μ'_1, está dada por

$$Var(x) = E(x^2) - [E(x)]^2$$

donde

$$E(x^2) = \mu'_2 = \frac{(\alpha)_2}{\Gamma(\alpha)p^2} \frac{\Xi_1^{\tau,1}(a,\alpha+2;c;w,z/p)}{\Xi_1^{\tau,1}(a,\alpha;c;w,z/p)} \tag{25}$$

$$\tau, \operatorname{Re} p, \operatorname{Re} \alpha, \operatorname{Re}(p-z) > 0, |w| < 1, x > 0$$

3.3. Función generadora de momento

La función generadora de momento, de una variable aleatoria X, es definida para cada real t, se denota por $M_x(t)$ es definida por

$$M_x(t) = E(e^{xt}) = \int_{-\infty}^{\infty} e^{xt}\ f(x)\, dx \tag{26}$$

Para la función de densidad $f(x)$ definida por (18), se tiene la función generadora de momento

$$M_x(t) = \int_0^\infty p^\alpha\ x^{\alpha-1}\ e^{-(p-t)x} \frac{\Xi_2^\tau(a,b;c;w,xz)}{\Gamma(\alpha)\ \Xi_1^{\tau,1}(a,\alpha,b;c;w,z/p)}\, dx \tag{27}$$

Aquí se han tenido en cuenta las mismas consideraciones dadas en 3.2 para tener distribuciones de soporte positivo.

Resolviendo la integral en (27) usando la definición (7) de Ξ_2^{τ} se obtiene finalmente la función generadora de momento para $f(x)$ definida en (18)

$$M_x(t) = \left(\frac{p}{p-t}\right)^{\alpha} \frac{\Xi_1^{\tau,1}(a, \alpha+n; c; w, z/(p-t))}{\Xi_1^{\tau,1}(a, \alpha, b; c; w, z/p)} \tag{28}$$

$$\tau, \operatorname{Re} p, \operatorname{Re} \alpha, \operatorname{Re}(p-t) > 0, |w| < 1$$

3.4. Función característica

La función característica de X está dada por

$$E(e^{itx}) = \int_0^{\infty} e^{itx}\ f(x)\ dx \tag{29}$$

usando (29) y la función $f(x)$ definida en (18)

$$\digamma\left[f(x)\right]_{(t)} = E(e^{itx}) = \frac{p^{\alpha}}{\Gamma(\alpha)} \frac{\frac{\Gamma(c)}{\Gamma(a)} \sum_{k,l=0}^{\infty} \frac{\Gamma(a+\tau k)(b)_k}{\Gamma(c+\tau k+l)} \frac{w^k}{k!} \frac{z^l}{l!}}{\Xi_1^{\tau,1}(a, \alpha, b; c; w, z/p)} \int_0^{\infty} x^{\alpha+l-1}\ e^{-(p-it)x}\ dx$$

resolviendo la integral y utilizando la definición (5) de $\Xi_1^{\tau,1}$ se tiene

$$\digamma\left[f(x)\right]_{(t)} = \left(\frac{p}{p-it}\right)^{\alpha} \frac{\Xi_1^{\tau,1}(a, \alpha; c; w, z/(p-it))}{\Xi_1^{\tau,1}(a, \alpha, b; c; w, z/p)} \tag{30}$$

3.5. La función tasa de riesgo

La función tasa de riesgo se define como

$$h(x) = \frac{f(x)}{S(x)} \tag{31}$$

donde $S(x)$ es la función de sobrevida de x

$$S(x) = 1 - F(x), \quad \text{para} \quad x > 0 \tag{32}$$

siendo $F(x)$ la función de densidad acumulada

$$F(x) = \int_0^x f(u)\, du$$

La función $S(x)$ tiene origen en la teoría de confiabilidad. En este caso la función de densidad $f(x)$ definida por (18) donde ${}_{\tau}\overset{w}{\Gamma}_0(\alpha, p; a, b; c; w, x)$ está definida por (16)

$$F(x) = \frac{p^{\alpha}\ {}_{\tau}\overset{w}{\Gamma}_0(\alpha, p; a, b; c; w, x)}{\Gamma(\alpha)\ \Xi_1^{\tau,1}(a, \alpha; c; w, z/p)} \tag{33}$$

luego la función de sobrevida $S(x)$ está dada por

$$S(x) = \frac{\Gamma(\alpha)\ \Xi_1^{\tau,1}(a,\alpha;c;w,z/p) - p^{\alpha}\ {}_{\tau}\overset{w}{\Gamma}_0(\alpha,p;a,b;c;w,x)}{\Gamma(\alpha)\ \Xi_1^{\tau,1}(a,\alpha;c;w,z/p)}$$

la función de tasa de riesgo está expresada por

$$h(x) = \frac{p^{\alpha}\ x^{\alpha-1}\ e^{-px}\ \Xi_2^{\tau,1}(a,b;c;w,xz)}{\Gamma(\alpha)\ \Xi_1^{\tau,1}(a,\alpha;c;w,z/p) - p^{\alpha}\ {}_{\tau}\overset{w}{\Gamma}_0(\alpha,p;a,b;c;w,x)} \tag{34}$$

3.6. Representaciones gráficas

Las siguientes figuras representan la función de densidad de probabilidad (*fdp*) generalizada dada por (18) para diferentes valores tanto de α como de τ; se observa la variación en los gráficos cuando se consideran valores distintos en dichos parámetros.

Tomando los valores $p = 2.5, a = 1.5$, $b = 2$, $c = 3$, $w = 0.8, \alpha = 2.8$ y $\tau = 2.2$ y $p = 2.5$, $a = 1.5$, $b = 3$, $c = 1,5$, $w = 0.8$, $\alpha = 3.4$ y $\tau = 3.4$, se tienen respectivamente las figuras 1 y 2.

En la figura 3 se consideran los mismos parámetros de la figura 1 y se gráfica el semilog de la función de densidad de probabilidad dada en (18).

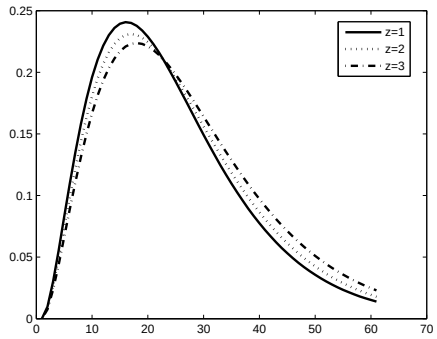

FIGURA 1: *fdp* para diferentes valores de z y $\tau = 2.2$.

4. Conclusiones

Este trabajo contiene una nueva función de densidad de probabilidad generalizada, la cual se obtuvo a partir de funciones generalizadas de tipo hipergeométrico desarrolladas recientemente; se encuentran algunas propiedades que permiten caracterizarla, como la función generadora de momento, los momentos, la función característica, la función tasa de riesgo y algunos casos especiales tales como exponencial y la gamma.

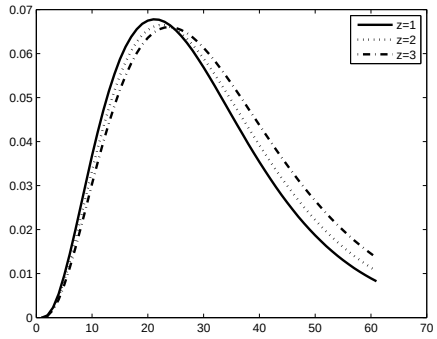

FIGURA 2: *fdp* para diferentes valores de z y $\tau = 3.4$.

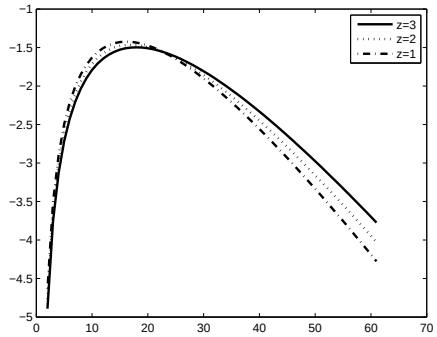

FIGURA 3: *semilog(fdp)* para diferentes valores de z y $\tau = 2.2$.

[Recibido: diciembre de 2008 — Aceptado: enero de 2010]

Referencias

Agarwal, S. K. & Kalla, S. L. (1996), *A Generalized Gamma Distribution and its Applications in Reliability Comm Statist*, Theory Method, Oxford.

Al-Musallam, F. & Kalla, S. L. (1998), 'Further Results on a Generalized Gamma Functions Ocurring in Diffraction Theory', *Statist, Theory Methods* **7**, 175–190.

Al-Saqabi, B. N., Kalla, S. L. & Shafea, A. (2002), 'On a Probability Distribution Involving a τ-confluent Hypergeometric Function of Two Variables', *Algebras Groups and Geometries* **19**(2), 254–257.

Galué, L., Al-Zamel, A. & Kalla, S. L. (2005), 'An Extension of Some Humbert's Functions', *International Journal of Applied Mathematics* **17**, 91–106.

Ghitany, M. E. (1998), 'On a Recent Generalization of Gamma Distribution', *Statist, Theory Methods* **27**, 223–233.

Good, I. J. (1953), 'The Population Frequencies of Species and the Estimation of Population Parameters', *Biometrika* **40**, 237–260.

Hoem, J. N. (1976), 'The Statistical Theory of Demographics Rates', *Scandinavian Journal of Statistics* **3**, 160–185.

Humbert, P. (1920), 'The Confluent Hypergeometric Functions of Two Variables', *Proceedings Royal Society of Edinburgh* **41**, 73–96.

Jorgensen, B. (1982), *Statistical Propertiers of Generalized Inverse Gaussians Distributions*, Lecture Notes in Statistics, New York.

Kobayashi, K. (1991), 'On a Generalized Gamma Functions Occurring in Diffraction Theory', *Journal of the Physical Society of Japan* **60**, 1501–1512.

Lebedev, N. N. (1965), *Special Functions and Their Applications*, primera edn, Dover Publications, Inc., New York.

Mathais, A. M. (1993), *A Handbook of Special Functions for Statistical and Physical Sciences*, Clarendon Press, Oxford.

Nakhi, B. & Kalla, S. L. (2005), 'On a Generalized Mixture Distribution', *Applied Mathematics and Computation* **169**, 943–952.

Virchenko, N. (1999), 'On Some Generalizations of the Functions Hypergeometric Type', *Integral Transforms and Special Functions* **2**, 233–244.

Virchenko, N., Kalla, S. L. & Al-Zamel, A. (2001), 'Some Result on a Generalized Hypergeometric Function', *Integral Transforms and Special Functions* **12**, 89–100.

Revista Colombiana de Estadística
Junio 2010, volumen 33, no. 1, pp. 25 a 41

Nonparametric Time Series Analysis of the Conditional Mean and Volatility Functions for the COP/USD Exchange Rate Returns

Análisis de series de tiempo no paramétrico de las funciones de media y varianza condicional de los retornos de la tasa de cambio COP/USD

Santiago Gallón[1,3,a], Karoll Gómez[2,3,b]

[1]Departamento de Estadística y Matemáticas - Departamento de Economía, Facultad de Ciencias Económicas, Universidad de Antioquia, Medellín, Colombia

[2]Departamento de Economía, Facultad de Ciencias Humanas y Económicas, Universidad Nacional de Colombia, Medellín, Colombia

[3]Grupo de Econometría Aplicada, Facultad de Ciencias Económicas, Universidad de Antioquia, Medellín, Colombia

Abstract

The modeling and estimation of the conditional volatility associated with a stochastic process usually have been based on parametric ARCH-type and stochastic volatility models. These time series models are very powerful in representing the dynamic stochastic properties of the data generating process only if the parametric functions are correctly specified. The nonparametric approach acquires importance as a complementary and flexible method to explore these properties without imposing particular functional forms on the conditional moments of process. This paper presents an application of nonparametric time series methods to estimate the conditional volatility function of the COP/USD exchange rate returns. Additionally, we estimate the conditional mean function under this approach.

***Key words*:** Nonparametric regression, Local polynomial regression, Nonlinear time series, Variance function estimation, Autoregressive conditional heteroscedasticity, Time series analysis.

Resumen

La modelación y estimación de la volatilidad condicional asociada a un proceso estocástico ha estado basada en los modelos paramétricos tipo ARCH y de volatilidad estocástica. Estos modelos son muy poderosos para representar las propiedades dinámicas estocásticas del proceso generador de

[a]Profesor asistente. E-mail: santiagog@udea.edu.co
[b]Profesor auxiliar. E-mail: kgomezp@unal.edu.co

datos solo si las funciones paramétricas están correctamente especificadas. En este sentido, el enfoque no paramétrico adquiere importancia como un método complementario y flexible para explorar dichas propiedades al no imponer formas funcionales particulares en los momentos condicionales del proceso. Este documento presenta una aplicación de los métodos no paramétricos de series de tiempo para estimar la función de volatilidad condicional de los retornos de la tasa de cambio COP/USD. Además, se estima la función de media condicional bajo este enfoque.

Palabras clave: regresión no paramétrica, regresión polinomial local, series de tiempo no lineales, estimación de la función de varianza, heterocedasticidad condicional autorregresiva, análisis de series de tiempo.

1. Introduction

In numerous publications researchers have written about the important role that associated volatility plays in a stochastic process, particularly in economics and finance. For example, the estimation of a conditional volatility measure that approximates its principal empirical features such as cluster volatility, asymmetries, leverage effects, and long memory, among others, is crucial for different issues in finance, like financial risk management, asset pricing, and efficient portfolio allocation. Subsequently, the development of models to adequately approximate the volatility process has concentrated the attention of researchers in the past two decades (Andersen, Bollerslev & Diebold 2009, Straumann 2005).

In this way, most volatility models have concentrated the attention on the parametric approach assuming an explicit functional form to the volatility process. This being said since Engle's (1982) Autoregressive Conditional Heteroscedasticity -ARCH- specification, where he explicitly expresses conditional volatility as a linear function of past squared innovations of the process, there has been an exponential growth of different parametric specifications. A short list of these specifications: Bollerslev's (1986) Generalized ARCH -GARCH- model, Engle & Bollerslev's (1986) Integrated GARCH -IGARCH- model, Nelson's (1991) Exponential GARCH -EGARCH- model, Ding et al. (1993) Asymmetric Power ARCH -APARCH- model, Baillie et al. (1996) Fractionally Integrated GARCH -FIGARCH- model, and Davidson's (2004) Hyperbolic GARCH -HYGARCH- model, among others. For a complete review of ARCH-type models, see Bollerslev et al. (1992), Bollerslev et al. (1994), and Andersen, Davis, Kreiß & Mikosch (2009). The estimation of ARCH-type models is commonly done by maximum likelihood under different distribution functions such as the usual Gaussian distribution, the Student-t distribution, the Generalized Error distribution (GED), and the skewed-Student distribution.

Jointly with ARCH-type models are also the Stochastic Volatility -SV- models. This class of parametric models presents, unlike ARCH models, an alternative approach to the specification of the volatility function where the standard specification contains an unobserved variance component (latent state variable)

which is modeled directly as a linear stochastic process, such as an autoregression (Harvey et al. 1994). See Ghysels et al. (1996), Shephard (2005), and Andersen, Davis, Kreiß & Mikosch (2009) for a complete overview about SV models. The estimation of SV models covers a wide range of estimation procedures, for instance quasi-maximum likelihood, applying the Kalman filter, Bayesian estimation, generalized method of moments, and efficient method of moments.

However, as it is well-known, the parametric time series models are very powerful in representing the stochastic dynamical properties of the data generating process if the parametric functions are correctly specified (Hardle & Linton 1994, Fan & Yao 2005), and searching for a parametric functional form is critical and not always is a simple task, especially when the process has nonlinear characteristics, as is the case of financial time series variables. Thus the nonparametric approach gains importance as a way of searching more flexible models without imposing particular functional forms of the conditional moments such as a mean, variance, or density function of process. The nonparametric estimates may be used as an end product or, perhaps more importantly, as a guide to identifying a parametric model to be used in a subsequent stage or to validate an existing one (Masry & Tjostheim 1995). Additionally, the estimation of nonparametric regression functions is not always complicated; on the contrary, it usually takes much less time estimation with respect to some more complicated parametric models where convergence problems are commonly found in their estimation algorithms.

Although the use of nonparametric methods in time series analysis has a long tradition, it has obtained popularity with modern nonparametric techniques, particularly in the analysis of nonlinear time series, due to the existence of large data sets and computational advances (Hardle et al. 1997). Some references about the development of nonparametric time series theory and its applications are: books by Hardle (1990), Fan & Gijbels (1996), and Fan & Yao (2005), and the articles by Robinson (1983), Hardle et al. (1997), Tjostheim (1999), and references therein. This approach is applied to a vast range of areas in economics, and has come to obtain great popularity in financial econometrics, for example, in modeling the drift and diffusion process underlying asset returns, among other issues in empirical finance. See, for example, Pagan & Ullah (1988), Diabold & Nason (1990), Mizrach (1992), Bossaerts et al. (1995), Bossaerts et al. (1996), Ait-Sahalia (1996*a*), Ait-Sahalia (1996*b*), Ait-Sahalia & Lo (1998), and Ait-Sahalia & Lo (2000). Particular studies using the nonparametric approach to estimate the conditional volatility function are Engle & Gonzalez-Rivera (1991), Bossaerts et al. (1995), Bossaerts et al. (1996), Masry & Tjostheim (1995), Hardle & Tsybakov (1997), Fan & Yao (1998), and Ziegelmann (2002).

In this paper, we apply nonparametric time series methods to estimate the conditional mean and volatility functions for the Colombian Peso/US Dollar - COP/USD- exchange rate returns. The fundamental reason for studying this variable is that the Colombian exchange rate has had significant variation episodes generating great uncertainty, and with large and severe costs on various economic sectors. Additionally, international asset pricing theories and international portfolio management depend on the expected foreign exchange rate movements (Bollerslev et al. 1992); therefore this paper can be a contribution to properly

understand the foreign exchange rate dynamics using the advantages of the nonparametric time series methods. The reason being that in Colombia, almost all analyses about foreign exchange rate has been concentrated on parametric models. The only nonparametric study for the COP/USD exchange rate is by Julio et al. (2005). For international nonlinear analysis on exchange rates using nonparametric procedures, see the studies by Meese & Rose (1990), Diabold & Nason (1990), LeBaron (1990), Bossaerts et al. (1995), Bossaerts et al. (1996), and Hardle & Tsybakov (1997).

The paper is organized as follows: Section 2 makes a short description of the nonparametric time series model and its different estimation methods. Section 3 applies the nonparametric model to estimate both the conditional mean and volatility functions for the returns of the COP/USD exchange rate process. Finally, Section 4 concludes.

2. The Nonparametric Time Series Model

The starting point of the data generating process of a strictly stationary discrete-time stochastic process $\{X_t\}$ defined on some probability space $(\Omega, \mathcal{F}, P)$ is the general univariate nonlinear stochastic regression model given by

$$X_t = m(X_{t-1}, \ldots, X_{t-p}) + \sigma(X_{t-1}, \ldots, X_{t-p})\varepsilon_t, \quad t = 1, \ldots, T \tag{1}$$

where $m(x_{t-1}, \ldots, x_{t-p}) = \mathbb{E}(X_t \mid X_{t-1} = x_1, \ldots, X_{t-p} = x_p)$ is the nonlinear autoregressive conditional mean (smooth) function, $\sigma^2(x_{t-1}, \ldots, x_{t-p}) = \mathrm{Var}(X_t \mid X_{t-1} = x_1, \ldots, X_{t-p} = x_p)$ represents the nonlinear autoregressive conditional variance (smooth) function, and $\{\varepsilon_t\}$ is an independent and identically distributed (i.i.d.) sequence of random variables with $\mathbb{E}(\varepsilon_t \mid X_{t-1}, \ldots, X_{t-p}) = 0$, $\mathrm{Var}(\varepsilon_t \mid X_{t-1}, \ldots, X_{t-p}) = 1$, and independent of $\{X_{t-1}, X_{t-2}, \ldots\}$.

The model (1) is known as the Conditional Heteroscedastic Autoregressive Nonlinear -CHARN- model; see Bossaerts et al. (1996), or the Nonparametric Autoregressive Conditional Heteroscedastic -NARCH- model; see Fan & Yao (2005).

This model is the most flexible nonparametric time series model because it does not impose any (parametric) particular form on the conditional mean and volatility functions. However, due to the well-known "curse of dimensionality" problem, the estimation of equation (1) is complicated.[1] As a consequence, it is necessary to assume a certain level of structure on the conditional functions $m(\cdot)$ and $\sigma(\cdot)$.[2]

[1] Nonparametric regression estimators are very flexible, but their statistical accuracy decreases greatly if there are several explicatory variables in the model (Hardle et al. 2004). Additionally, their estimation is difficult unless the sample size is excessively large (Fan & Yao 2005), and (Fan & Gijbels 1996).

[2] A very popular nonparametric model is the Functional-Coefficient Autoregressive -FAR- model (Chen & Tsay 1993), where the conditional mean and volatility functions are specified as

$$m(X_{t-1}, \ldots, X_{t-p}) = a_1(X_{t-d})X_1 + \cdots + a_p(X_{t-d})X_{t-p}$$
$$\sigma^2(X_{t-1}, \ldots, X_{t-p}) = b_1(X_{t-d})X_1^2 + \cdots + b_p(X_{t-d})X_{t-p}^2$$

The usual assumption is: suppose $p = 1$ such that the model (1) becomes

$$X_t = m(X_{t-1}) + \sigma(X_{t-1})\varepsilon_t \tag{2}$$

Following Hardle & Tsybakov (1997), and Fan & Yao (1998) if $\{X_t\}$ is a stationary process, the conditional variance function can be decomposed as

$$\begin{aligned} \sigma^2(x) &= \mathbb{E}(X_t^2 \mid X_{t-1} = x) - \{\mathbb{E}(X_t \mid X_{t-1} = x)\}^2 \\ &= g(x) - \{m(x)\}^2 \end{aligned} \tag{3}$$

such that the conditional variance estimate is based on the nonparametric estimation of $g(x)$ and $m(x)$ given by $\widehat{\sigma}_T^2(x) = \widehat{g}_T(x) - \{\widehat{m}_T(x)\}^2$.

2.1. Nonparametric Kernel Estimation

A way to obtain estimates of functions $m(x)$ and $g(x)$ is by applying the popular Nadaraya-Watson estimator given by:

$$\begin{aligned} \widehat{m}_T(X_{t-1}) &= \frac{\sum_{t=2}^{T} K\left((X_{t-1} - x)/h_T\right) X_t}{\sum_{t=2}^{T} K\left((X_{t-1} - x)/h_T\right)} \\ \widehat{g}_T(X_{t-1}) &= \frac{\sum_{t=2}^{T} K\left((X_{t-1} - x)/h_T\right) X_t^2}{\sum_{t=2}^{T} K\left((X_{t-1} - x)/h_T\right)} \end{aligned} \tag{4}$$

where $K(\cdot) : \mathbb{R} \to \mathbb{R}$ is the (continuous, bordered, symmetric, and integrating to one) Kernel function and $h_T > 0$ is the bandwidth parameter (also smoothing parameter), $h_T \to 0$ as $T \to \infty$. The Nadaraya–Watson estimator is a special case of the local polynomial estimation explained below. The Kernel functions most commonly used are the Gaussian, Quartic, and Epanechnikov Kernels.

These estimators are strongly consistent and asymptotically normal for α-mixing observations;[3] see Robinson (1983), and Masry & Tjostheim (1995).

---

with $a_i(\cdot)$ and $b_i(\cdot)$, $i = 1, \ldots, p$ one-dimensional unknown functions, and X_{t-d} is the model-dependent variable. Another common model is the Additive Autoregressive -*AAR*- model (Jones 1978) which assumes an additive structure for conditional mean and variance,

$$\begin{aligned} m(X_{t-1}, \ldots, X_{t-p}) &= m_1(X_{t-1}) + \cdots + m_p(X_{t-p}) \\ \sigma^2(X_{t-1}, \ldots, X_{t-p}) &= \sigma_1(X_{t-1}^2) + \cdots + \sigma_p(X_{t-p}^2) \end{aligned}$$

where $m_i(\cdot)$ and $\sigma_i(\cdot)$, $i = 1, \ldots, p$ are univariate unknown functions. For other nonparametric models such as Partially Linear models and Single-Index Models, see Hardle & Tsybakov (1997), Hardle et al. (2004), Fan & Yao (2005), and Gao (2007).

[3] A sequence is said to be α-mixing if $\alpha(n) \to 0$ when $n \to \infty$, with $\alpha(n)$ defined as

$$\alpha(n) = \sup_{A \in \mathcal{F}_{-\infty}^{k}, B \in \mathcal{F}_{k+n}^{\infty}} \mid P(A \cap B) - P(A)P(B) \mid, \quad n = 1, 2, \ldots,$$

where $\mathcal{F}_i^j$ is the σ-field generated by $X_i, \ldots, X_j$. See Robinson (1983), and Fan & Yao (2005).

2.2. Local Polynomial Regression

Another nonparametric technique used to estimate the functions $m(x)$ and $g(x)$ is proposed by Hardle & Tsybakov (1997), who applied the local polynomial regression method. The estimates for $m(x)$ and $g(x)$ functions are derived through the solution of the following weighted least-squares problems:

$$\begin{aligned} c_T(x) &= \underset{c\in\mathbb{R}^l}{\arg\min} \sum_{t=1}^{T} (X_t - c'U_{tT})^2 K\left((X_{t-1}-x)/h_T\right) \\ \overline{c}_T(x) &= \underset{\overline{c}\in\mathbb{R}^l}{\arg\min} \sum_{t=1}^{T} (X_t^2 - \overline{c}'U_{tT})^2 K\left((X_{t-1}-x)/h_T\right) \end{aligned} \tag{5}$$

where $K(\cdot)$ and $h_T > 0$ are again the Kernel function and bandwidth parameter, respectively, and $U_{tT} = F(u_{tT})$ with $F(u) = (1, u, \ldots, u^{l-1}/(l-1)!)'$ and $u_{tT} = (X_{t-1} - x)/h_T$ (the symbol $'$ denotes the transpose of a row vector). Note that when $l = 1$, the local polynomial fit is reduce to the Nadaraya-Watson estimator.

The estimators of $m(x)$ and $g(x)$ are given by $\widehat{m}_T(x) = \widehat{c}_T(x)'F(0)$ and $\widehat{g}_T(x) = \widehat{\overline{c}}_T(x)'F(0)$, respectively; such that the estimator of the conditional variance functions is defined as

$$\widehat{\sigma}_T^2(x) = \widehat{\overline{c}}_T(x)'F(0) - \{\widehat{c}_T(x)'F(0)\}^2 \tag{6}$$

Hardle & Tsybakov (1997) establish the asymptotic normality of local polynomial estimators for conditional mean and variance.

In the application of the local polynomial nonparametric regression method to estimate the volatility function to DM/USD and YEN/USD foreign exchange rates Hardle & Tsybakov (1997) use a local linear approximation ($l = 2$), such that

$$\begin{aligned} c_T(x) &= \underset{c\in\mathbb{R}^2}{\arg\min} \sum_{t=1}^{T} \left(X_t - c_1 - c_2(X_{t-1}-x)\right)^2 K\left((X_{t-1}-x)/h_T\right) \\ \overline{c}_T(x) &= \underset{\overline{c}\in\mathbb{R}^2}{\arg\min} \sum_{t=1}^{T} \left(X_t^2 - \overline{c}_1 - \overline{c}_2(X_{t-1}-x)\right)^2 K\left((X_{t-1}-x)/h_T\right) \end{aligned} \tag{7}$$

Note that in the minimization problems in (5), employed to obtain local approximation estimates of $m(x)$ and $g(x)$, the Kernel function and bandwidth parameter are commons in both equations. This strategy is used by Hardle & Tsybakov (1997) to avoid nonnegative estimators of $\sigma^2(x)$ and to reduce bias. Therefore, Fan & Yao (1998) propose a residual-based estimator to conditional variance based on local linear regression.

From (2) we have that $r_t^2 = \{X_t - m(X_{t-1})\}^2 = \sigma^2(X_{t-1})\varepsilon_t^2$, such that its conditional expectation is $\mathbb{E}(r_t^2 \mid X_{t-1} = x) = \sigma^2(x)$. This therefore shows that it is natural to estimate $\sigma^2(x)$ using the estimated residuals. Consequently

the estimates of $m(x)$ and $\sigma^2(x)$ are derived from the solutions of the following minimization problems:

$$\begin{aligned}\widehat{a}_T(x) &= \underset{a\in\mathbb{R}^2}{\arg\min} \sum_{t=1}^{T}\{X_t - a_1 - a_2(X_{t-1}-x)\}^2 K\left((X_{t-1}-x)/h_{1T}\right)\\ \widehat{b}_T(x) &= \underset{b\in\mathbb{R}^2}{\arg\min} \sum_{t=1}^{T}\{\widehat{r}_t^2 - b_1 - b_2(X_{t-1}-x)\}^2 W\left((X_{t-1}-x)/h_{2T}\right)\end{aligned} \tag{8}$$

where $K(\cdot)$ and $W(\cdot)$ are the Kernel functions, $h_{1T} > 0$ and $h_{2T} > 0$ the bandwidth parameters, and $\widehat{r}_t^2 = \{X_t - \widehat{m}_T(x)\}^2$ the estimated residuals.

The estimate of $m(x)$ is given by $\widehat{m}_T(x) = \widehat{a}_T(x)'e = \widehat{a}_1$ where $e = (1,0)'$ such that the residuals are $\widehat{r}_t^2 = \{X_{t-1} - \widehat{a}_1\}^2$ which are used in the above second minimization problem to obtain the estimator of $\sigma^2(x)$ given by

$$\widehat{\sigma}_T^2(x) = \widehat{b}_T(x)'e = \widehat{b}_1 \tag{9}$$

Fan & Yao (1998) demonstrate the asymptotic normality and efficiency of $\widehat{\sigma}_T^2(x)$, and apply their method to estimate the conditional mean and volatility functions to yields of the three-month Treasury Bill.

The reason for using the local polynomial regression, especially the local linear estimator applied in the Hardle & Tsybakov (1997) and Fan & Yao (1998), is because the local polynomial estimator has diverse statistical properties. Among these are: Agreeable nice asymptotic properties such as asymptotic minimax efficiency (Fan 1993), good finite sampling and design-adaptation properties, and it overcomes the drawbacks of the Nadaraya-Watson estimator and other nonparametric estimators such as large biases due to boundary effects. See Fan (1993), Fan & Gijbels (1996), and Fan & Yao (2005) for a complete derivation and description of statistical properties of the local polynomial estimator.

Note that the implementation of the above estimators depends on the appropriate selection of both bandwidth parameter and Kernel function. For example, for local linear estimator, Hardle & Tsybakov (1997) applied the cross-validation method to choose the bandwidth parameter using the Quartic Kernel, and Fan & Yao (1998) applied the data-driven bandwidth selection method using the Epanechnikov Kernel. See Fan & Yao (2005) for a complete description of bandwidth parameter selection methods for dependent processes.

3. Empirical Application for the COP/USD Exchange Rate Returns

In this section we apply the previously described different nonparametric estimators to estimate both the conditional mean and variance functions for the COP/USD exchange rate returns, S_t, defined as $X_t = \log(S_t/S_{t-1})$. The data consists of 3515 commercial daily observations (from 2 January 1995-20 June

2008). There are two reasons for choosing this period. Firstly, the applied work on COP/USD exchange rate usually studies this series without the exchange rate band regime, and due to the nonparametric time series methods are very effective in modeling potential nonlinearities (for example, due to interventions), we consider that is important to use the whole period. Secondly, the estimation and bandwidth selection methods are difficult unless the sample size is large (Fan & Yao 2005, Fan & Gijbels 1996). The statistical source of the database is the Central Bank of Colombia.

The graphs of the COP/USD exchange rate, its returns and respectively squared returns are presented in Figure 1. As we can see, the COP/USD exchange rate returns present the well-known cluster volatility phenomena. This is related to the excess kurtosis as is ilustrated by the Kernel density estimation for the COP/USD exchange rate returns in Figure 1.

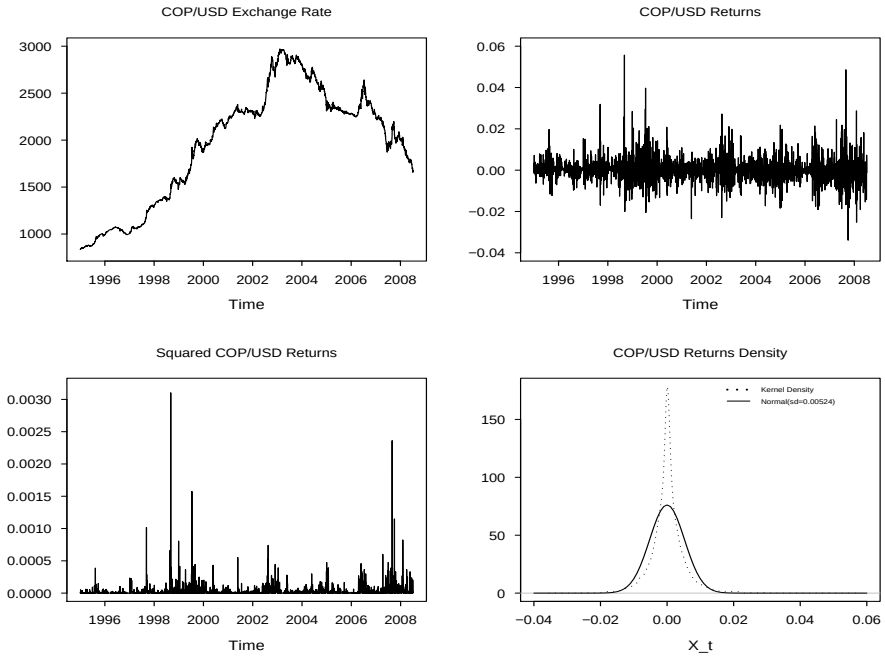

FIGURE 1: COP/USD exchange rate, COP/USD exchange rate returns, squared COP/USD exchange rate returns, and Kernel density estimation for COP/USD exchange rate returns compared with a mean zero normal density with standard deviation, $\widehat{\sigma} = 0.00524$.

To test for the existence of nonlinearity in the COP/USD returns and its squares, we applied the popular BDS test (see Brock et al. 1996), which can be considered a misspecification test in time series analysis. This test has power against a wide range of linear and nonlinear alternatives. The results displayed in Table 1 show that the null hypothesis of i.i.d. is rejected for most combinations of m and ϵ for both variables. Since there appears to be no discernible linear

structure in the returns and its squares, the results suggest that there may be a nonlinear structure.

TABLE 1: BDS test statistics for nonlinearity of X_t and X_t^2.

	X_t				X_t^2			
$m\backslash\epsilon$	0.0026	0.0052	0.0079	0.0105	0e+00	1e−04	1e−04	2e−04
2	16.508	17.078	15.279	13.424	11.163	9.718	7.513	6.103
	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
3	22.301	21.550	19.088	17.442	14.018	12.615	11.318	11.644
	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
4	28.976	25.196	21.426	19.114	15.720	13.521	12.439	12.864
	(0)	(0) (0)	(0)	(0)	(0)	(0)	(0)	
5	38.839	28.877	23.283	20.095	16.937	13.770	12.835	13.305
	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)

m: embedding dimension, ϵ: close point.
p-values in parenthesis.

The Kernel function used in all estimations for both the conditional mean and volatility functions was the Epanechnikov Kernel, given by $K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1)$. In addition, we developed all estimations by employing other kernel functions such as the Quartic Kernel, $K(u) = \frac{15}{18}(1-u^2)^2 I(|u| \leq 1)$, and the Gaussian Kernel, $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$; and by not necessarily using the same kernel function for the conditional mean and volatility, we obtained very similar results. For optimal bandwidth parameter selection, we always choose it by applying the following bandwidth selection methods: cross-validation, rule of thumb, the pre-asymptotic substitution method by Fan & Gijbels (1995), and the plug-in approach. After the empirical comparison of the estimations obtained by means of the different bandwidth selection methods, the optimal bandwidth parameters for the conditional mean were 0.0134 and 0.0140 to Nadaraya-Watson and local linear polynomial estimators, respectively, and for the conditional variance were 0.0134, 0.0127, and 0.0149 to the Nadaraya-Watson, Hardle & Tsybakov (1997), and Fan & Yao (1998) estimators, respectively. All estimations and computations were carried out using the Xplore software version 4.8 and the `locpol R` package developed by Cabrera (2008). We also applied a robust local polynomial regression proposed by Cleveland (1979) to guard against deviant points (outliers) which may have had a distorting effect on the smoothing. As the results were very similar, they have not been illustrated to save space.[4]

Figure 2 shows the scatterplots of X_t against X_{t-1}, and the conditional mean function estimated by (a) Nadaraya-Watson (local constant estimator), and (b) local linear polynomial estimators denoted by $\widehat{m}_T(X_{t-1})$. The dashed lines correspond to pointwise 95% asymptotic confidence intervals.[5] As we can see, the shape of both estimate curves is almost equal, except for on the left edge where there is a boundary effect in the local constant fit thus generating bias problems in the lineal direction on the edges.

[4]The plots are available upon request.

[5]See Fan & Yao (2005) for a complete construction of the confidence intervals for dependent data.

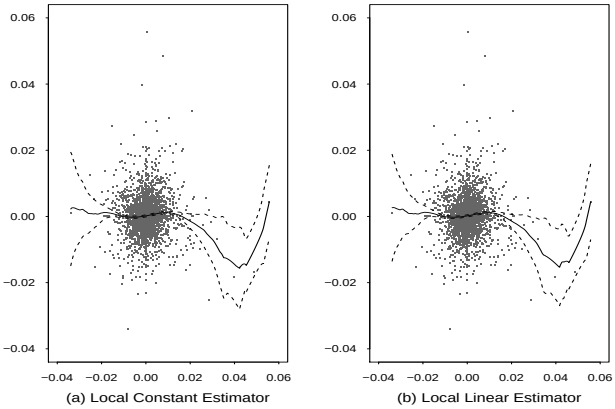

FIGURE 2: (a) X_t against X_{t-1}, and local constant polynomial fit $\widehat{m}_T(X_{t-1})$. (b) X_t against X_{t-1} and local linear polynomial fit $\widehat{m}_T(X_{t-1})$.

The most important segment in the graphs corresponds at the center: between -0.02 and 0.02, because there in that interval are most of the observations, and therefore more efficiency in the estimation of the conditional mean function. As we can see in that segment, the slope is almost zero which that possibly means that an efficient exchange rate market exists. The results found in the literature on exchange rate markets are diverse. Some show evidence of low negative slopes (mean reversion) (Hardle & Tsybakov 1997) and others low positive slopes (Julio et al. 2005).

The graphs of the estimated residuals, $\widehat{r}_t = \{X_t - \widehat{m}_T(x)\}$, where $\widehat{m}_T(x)$ is the estimated conditional mean function obtained from the local linear polynomial estimation, and its squares, $\widehat{r}_t^2$, are illustrated in Figure 3. The latter is employed in the estimation for the conditional variance function using the local linear estimator of Fan & Yao (1998). Additionally the scatterplot between $\widehat{r}_t$ against X_{t-1}, including the estimated conditional mean curve, and the Kernel density estimation for residuals compared with a mean zero normal density with standard deviation, $\widehat{\sigma} = 0.00524$ are shown in Figure 3.

Figures 4(a) and 4(b) depict the scatterplots of the squared returns, X_t^2, against X_{t-1}, and the estimated regression curve of the conditional variance, denoted by $\widehat{\sigma}_T^2(X_{t-1})$, estimated by the Nadaraya-Watson and the local linear estimator by Hardle & Tsybakov (1997). Figure 4(c) shows the scatterplot of the squared residuals, $\widehat{r}_t$, against X_{t-1}, and the local linear estimator of Fan & Yao (1998).

Figure 5 shows the volatility functions, $\widehat{\sigma}_T(X_{t-1})$, estimated by (a) the Nadaraya-Watson, (b) the Hardle & Tsybakov (1997), and (c) the Fan & Yao (1998) estimators. The results show that the conditional volatility function estimated by means of the Fan and Yao residual-based local linear estimator is smoother than the local constant and linear estimators using the squared returns.

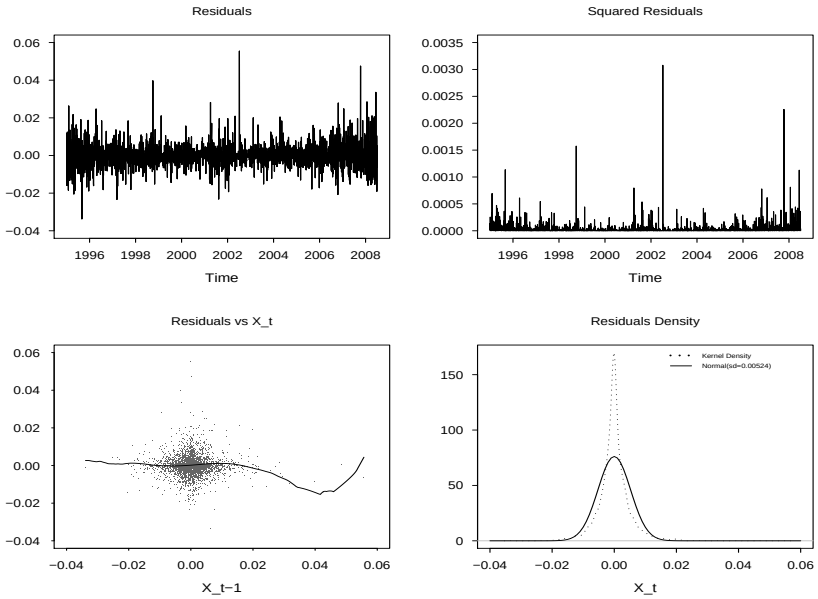

FIGURE 3: Residuals, squared residuals, residuals against X_{t-1}, and Kernel density estimation for residuals compared to a mean zero normal density with standard deviation, $\widehat{\sigma} = 0.00524$.

Furthermore, the results do not to show evidence of volatility asymmetries. On the contrary, the well-known U-shape in the conditional volatility function (except to the Fan and Yao's estimator) is present. This result concurs with findings in other studies that employ the parametric approach (Castaño et al. 2008, Maya & Gómez 2008). This indicates, for example, that it is proper to carry out an option evaluation on the COP/USD exchange rate with the symmetric volatility "smile". However, this symmetric U-shape is particularly clear in the central area of the graphs where the majority of observations are, this is, in the segment between -0.02 and 0.02 (see Figure 3). As expected, this symmetry in volatility is broken due to boundary effects on the right and left edges where there are few observations for correct smoothing (Hardle 1990, Hardle & Tsybakov 1997, Fan & Gijbels 1996, Hardle et al. 2004).

Finally, the results are along the lines of the findings by Julio et al. (2005), who applied the local lineal polynomial regression.[6] The goal of their study was to determine the features of the "volatility U-shape" and mean response functions, and the market effect of central bank interventions on those functions. They found that "discretional interventions" tend to change the concavity of the "volatility U-shape". However, that change was moderated and it never produced "volatility skew".

[6] In their study they used a different sample, from September 27th 1999-March 31st 2005. Additionally, they fit a local linear approximation to conditional mean and local quadratic approximation to conditional variance.

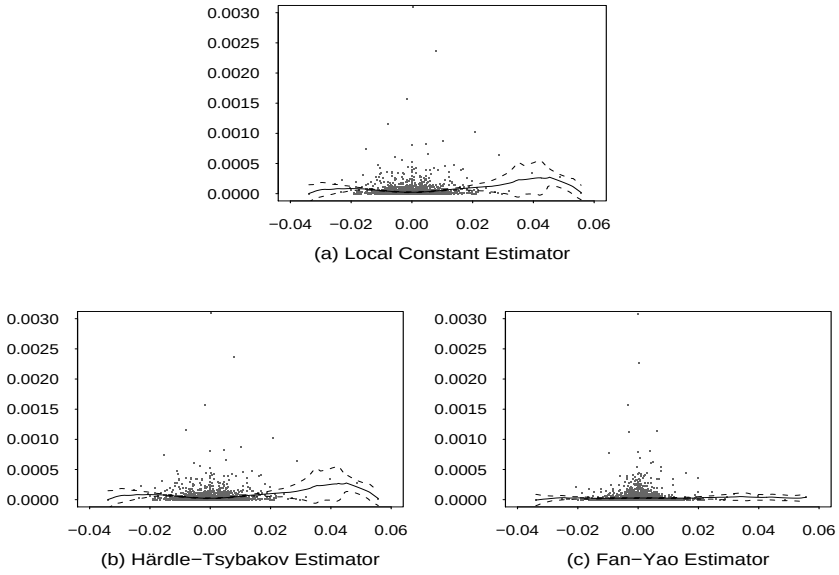

FIGURE 4: (a) X_t^2 against X_{t-1}, and local constant polynomial fit $\widehat{\sigma}_T^2(X_{t-1})$. (b) X_t^2 against X_{t-1}, and local linear polynomial fit $\widehat{\sigma}_T^2(X_{t-1})$. (c) $\widehat{r}_t^2$ against X_{t-1}, and local linear polynomial fit $\widehat{\sigma}_T^2(X_{t-1})$.

4. Conclusions

The role of volatility associated with a stochastic process is well-known, particularly in economics and finance. However, most of the volatility models have focused their attention on the parametric approach to represent the stochastic dynamic properties of the data generating process, assuming explicit functional forms for the mean and variance processes. In this paper a nonparametric time series analysis to the conditional mean and variance functions was carried out on the Colombian Peso/US Dollar -COP/USD- exchange rate returns.

Two nonparametric estimators were applied to estimate the conditional mean function: the local constant polynomial (Nadaraya-Watson) and local linear polynomial estimators; whereas three were used for the conditional variance function: the local constant and linear estimators based on the squared returns and the residual-based local linear estimator. The results show no evidence of asymmetries in the volatility of COP/USD exchange rate. On the contrary, we found the "volatility U-shape". Additionally, the results indicate that there is mean reversion according to the existence of a lineal function relationship to conditional mean.

Finally, as Bossaerts et al. (1995) point out, the nonparametric analysis can be extended considering a less restricted data generating process on the conditional mean function as well as on the conditional variance function including more lags

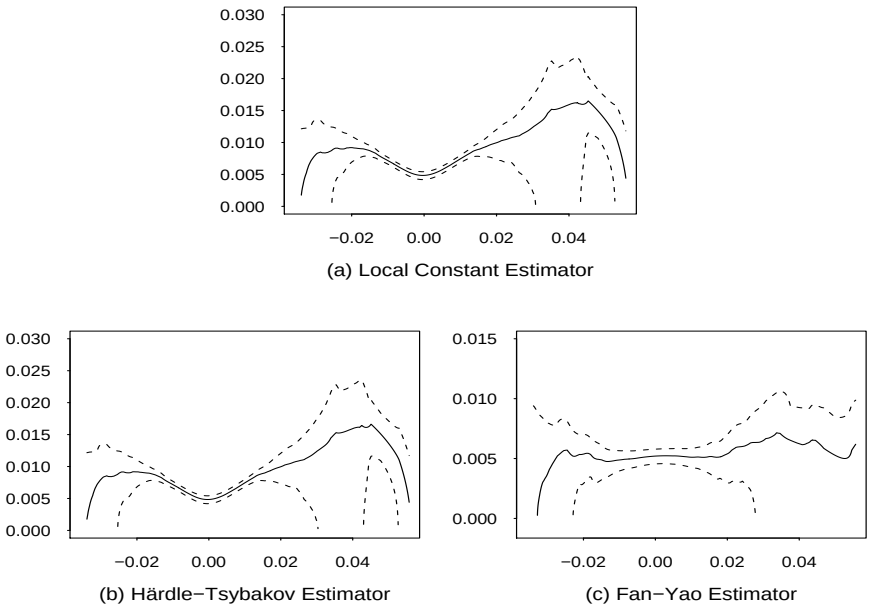

FIGURE 5: (a) $\widehat{\sigma}_T(X_{t-1})$ against X_{t-1} (Nadaraya-Watson). (b) $\widehat{\sigma}_T(X_{t-1})$ against X_{t-1} (Hardle-Tsybakov). (c) $\widehat{\sigma}_T(X_{t-1})$ against X_{t-1} (Fan-Yao).

in both functions. However this implies the well-known "curse of dimensionality" problem. Moreover, other nonparametric models can be attempted; for instance the Functional-Coefficient Autoregressive model, the Additive Autoregressive model, and among others. At this moment this extension is been performed jointly with a multivariate analysis to modeling portfolios of exchange rates and forecast future returns over short horizons.

Acknowledgements

We are grateful to the professor S. Sperlich for their very helpful comments and constructive suggestions. We also wish to thank to participants in the "III Encuentro del Grupo de Series de Tiempo," at the Universidad Nacional de Colombia, and to two anonymous referees for their valuable comments.

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

References

Ait-Sahalia, Y. (1996*a*), 'Nonparametric Pricing of Interest Rate Derivative Securities', *Econometrica* **64**, 527–560.

Ait-Sahalia, Y. (1996*b*), 'Testing Continuous-Time Models of the Spot Interest Rate', *The Review of Financial Studies* **9**, 385–426.

Ait-Sahalia, Y. & Lo, A. (1998), 'Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices', *Journal of Finance* **53**, 499–548.

Ait-Sahalia, Y. & Lo, A. (2000), 'Nonparametric Risk Management and Implied Risk Aversion', *Journal of Econometrics* **94**, 9–51.

Andersen, T., Bollerslev, T. & Diebold, F. (2009), Parametric and Nonparametric Volatility Measurement, *in* L. P. Hansen & Y. Ait-Sahalia, eds, 'Handbook of Financial Econometrics', Vol. 1, North Holland, Amsterdam.

Andersen, T., Davis, R., Kreiß, J. & Mikosch, T. (2009), *Handbook of Financial Time Series*, Springer, New York.

Baillie, R., Bollerslev, T. & Mikkelsen, H. (1996), 'Fractionally Integrated Generalized Autoregressive Conditional Heteroscedasticity', *Journal of Econometrics* **74**, 3–30.

Bollerslev, T. (1986), 'Generalized Autoregressive Conditional Heteroskedasticity', *Journal of Econometrics* **31**, 307–327.

Bollerslev, T., Chou, R. & Kroner, K. (1992), 'ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence', *Journal of Econometrics* **52**, 5–59.

Bollerslev, T., Engle, R. & Nelson, D. (1994), ARCH Models, *in* R. Engle & D. McFadden, eds, 'Handbook of Econometrics', Vol. 4, North-Holland: Amsterdam.

Bossaerts, P., Hardle, W. & Hafner, C. (1995), 'A New Method for Volatility Estimation with Applications in Foreign Exchange Rate Series', *Proceedings of the 5th. Karlsruher Okonometrie-Workshop* . Universit at Karlsruhe.

Bossaerts, P., Hardle, W. & Hafner, C. (1996), Foreign Exchange-Rates Have Surprising Volatility, *in* P. M. Robinson & M. Rosenblatt, eds, 'Athens Conference on Applied Probability and Time Series Analysis', Vol. 2, Springer, pp. 55–72.

Brock, W., Dechert, W., Scheinkman, J. & LeBaron, B. (1996), 'A Test for Independence Based on the Correlation Dimension', *Econometric Reviews* **15**, 197–235.

Cabrera, J. (2008), *The locpol Package: Kernel Local Polynomial Regression*, R Foundation for Statistical Computing.
*http://www.R-project.org

Castaño, E., Gómez, K. & Gallón, S. (2008), 'Pronóstico y estructuras de volatilidad multiperiodo de la tasa de cambio del peso colombiano', *Cuadernos de Economía* **48**, 241–266.

Chen, R. & Tsay, R. (1993), 'Functional-Coefficient Autoregressive Models', *Journal of the American Statistical Association* **88**, 298–308.

Cleveland, W. (1979), 'Robust Locally Weighted Regression and Smoothing Scatterplots', *Journal of the American Statistical Association* **74**, 829–836.

Davidson, J. (2004), 'Moment and Memory Properties of Linear Conditional Heteroscedasticity Models, and a New Model', *Journal of Business and Economics Statistics* **22**, 16–29.

Diabold, F. & Nason, J. (1990), 'Nonparametric Exchange Rate Prediction?', *Journal of International Economics* **28**, 315–332.

Ding, Z., Granger, C. & Engle, R. (1993), 'A Long Memory Property of Stock Market Returns and A New Model', *Journal of Empirical Finance* **1**, 83–106.

Engle, R. (1982), 'Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation', *Econometrica* **50**, 987–1008.

Engle, R. & Bollerslev, T. (1986), 'Modelling the Persistence of Conditional Variances', *Econometric Reviews* **5**, 1–50.

Engle, R. & Gonzalez-Rivera, G. (1991), 'Semi-Parametric ARCH Models', *Journal of Business and Economic Statistics* **9**, 345–359.

Fan, J. (1993), 'Local Linear Regression Smoothers and their Minimax Efficiency', *Annals of Statistics* **21**, 196–216.

Fan, J. & Gijbels, I. (1995), 'Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation', *Journal of the Royal Statistical Society* **B57**, 371–394.

Fan, J. & Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall/CRC, London.

Fan, J. & Yao, Q. (1998), 'Efficient Estimation of Conditional Variance Functions in Stochastic Regression', *Biometrika* **85**, 645–660.

Fan, J. & Yao, Q. (2005), *Nonlinear Time Series, Nonparametric and Parametric Methods*, Springer, New York.

Gao, J. (2007), *Nonlinear Time Series: Semiparametric and Nonparametric Methods*, Chapman and Hall/CRC, London.

Ghysels, E., Harvey, A. & Renault, E. (1996), Stochastic Volatility, *in* C. R. Rao & G. S. Maddala, eds, 'Handbook of Statistics', Vol. 14, Butterworth-Heinemann, Amsterdam.

Hardle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Hardle, W. & Linton, O. (1994), Applied Nonparametric Methods, *in* R. Engle & D. McFadden, eds, 'Handbook of Econometrics', Vol. 4, Elsevier Science & Technology Books, Amsterdam.

Hardle, W., Lutkepohl, H. & Chen, R. (1997), 'A Review of Nonparametric Time Series Analysis', *International Statistical Review* **65**, 49–72.

Hardle, W., Muller, M., Sperlich, S. & Werwatz, A. (2004), *Nonparametric and Semiparametric Models*, Springer, New York.

Hardle, W. & Tsybakov, A. (1997), 'Local Polynomial Estimators of the Volatility Function in Nonparametric Autoregression', *Journal of Econometrics* **81**, 223–242.

Harvey, A., Ruiz, E. & Shephard, N. (1994), 'Multivariate Stochastic Variance Models', *The Review of Economic Studies* **61**, 47–264.

Jones, D. A. (1978), 'Non-linear Autoregressive Processes', *Journal of the Royal Statistical Society* **A360**, 71–95.

Julio, J., Rodríguez, N. & Zárate, H. (2005), 'Estimating the COP Exchange Rate Volatility Smile and the Market Effect of Central Bank Interventions: A CHARN Approach', *Borradores de Economía* (347). Banco de la República.

LeBaron, B. (1990), Forecast Improvements using a Volatility Index, Working paper, University of Wisconsin-Maddison.

Masry, E. & Tjostheim, D. (1995), 'Nonparametric Estimation and Identification of Nonlinear ARCH Time Series: Strong Convergence and Asymptotic Normality', *Econometric Theory* **11**, 258–289.

Maya, C. & Gómez, K. (2008), 'What Exactly is Bad News in Foreign Exchange Markets? Evidence from Latin American Markets', *Cuadernos de Economía* **45**, 161–183.

Meese, R. A. & Rose, A. K. (1990), 'Nonlinear, Nonparametric, Nonessential Exchange Rate Estimation', *American Economic Review Paper and Proceedings* **80**, 192–196.

Mizrach, B. (1992), 'Multivariate Nearest-neighbour Forecasts of EMS Exchange Rates', *Journal of Applied Econometrics* **7**, S151–63.

Nelson, D. (1991), 'Conditional Heteroskedasticity in Asset Returns: A New Approach', *Econometrica* **59**, 347–370.

Pagan, A. & Ullah, A. (1988), 'The Econometric Analysis of Models with Risk Terms', *Journal of Time Series Analysis* **3**, 87–105.

Robinson, P. M. (1983), 'Nonparametric Estimators for Time Series', *Journal of Time Series Analysis* **4**, 185–207.

Shephard, N. (2005), *Stochastic Volatility: Selected Readings*, Oxford University Press, New York.

Straumann, D. (2005), *Estimation in Conditionally Heteroscedastic Time Series Models*, Vol. 181 of *Lecture Notes in Statistics*, Springer, Berlin.

Tjostheim, D. (1999), Nonparametric Specification Procedures for Time Series, *in* S. Ghosh, ed., 'Asymptotics, Nonparametrics, and Time Series', Vol. 158, Dekker, Marcel Inc., New York, pp. 149–199.

Ziegelmann, F. (2002), 'Nonparametric Estimation of Volatility Functions: The Local Exponential Estimator', *Econometric Theory* **18**, 985–991.

Revista Colombiana de Estadística
Junio 2010, volumen 33, no. 1, pp. 43 a 61

Appraisal of Several Methods to Model Time to Multiple Events per Subject: Modelling Time to Hospitalizations and Death

Revisión de varios métodos para modelar tiempo a múltiples eventos por sujeto: modelamiento de tiempo a hospitalizaciones y muerte

Javier Castañeda[a], Bart Gerritse[b]

Clinical Outcomes, Research and Biometry, Cardiac Rhythm Disease Management, Medtronic Bakken Research Center, Maastricht, Netherlands

Abstract

During the disease-recovery process of many diseases, such as in Heart Failure (HF), often more than one type of event plays a role. Some clinical trials use the combined endpoint of death and a secondary event; for instance, HF-related hospitalizations. This is often analyzed with time-to-first-event survival analysis which ignores possible subsequent events, such as several HF-related hospitalizations. Accounting for multiple events provides more detailed information on the disease-control process, and allows a more precise understanding of the prognosis of patients.

In this paper we explore and illustrate several modelling strategies to study time to repeated events of disease-related hospitalizations and death per subject. Marginal models are revised in order to account for intra-subject correlation and competing risks. Finally, we recommend a Multi-state model which allows a flexible modelling strategy that incorporates important features in the analysis of HF-related hospitalizations and death, and at the same time extends relevant characteristics of the Andersen & Gill (1982), Wei et al. (1989) and Prentice et al. (1981) models.

***Key words*:** Survival analysis, Competing risks, Correlated observations, Marginal models.

Resumen

Algunos ensayos clínicos para estudiar el efecto de nuevos tratamientos en pacientes con insuficiencia cardiaca (IC) se basan en la evaluación de hospitalizaciones relacionadas con IC y muerte. Frecuentemente el análisis se enfoca en el tiempo a la primera ocurrencia de alguno de estos dos

[a]Biostatistician. E-mail: javier.castaneda@medtronic.com
[b]Principal Statistician. E-mail: bart.gerritse@medtronic.com

desenlaces. Este tipo de análisis ignora importantes eventos como nuevas hospitalizaciones relacionadas con IC, que permiten una mejor descripción y compresión del proceso de recuperación de estos pacientes.

En este trabajo se describen y exploran varias estrategias para el análisis de tiempo a repetidas hospitalizaciones relacionadas con IC y tiempo a la muerte. Se estudian modelos marginales para incorporar la correlación intra-sujeto y riesgos competitivos propios de este tipo de ensayos clínicos. Finalmente, se recomienda un modelo multi-estado como una estrategia sencilla y flexible que incorpora elementos importantes en el análisis de hospitalizaciones relacionadas con IC y muerte, y a la vez extiende características relevantes de los modelos de Andersen & Gill (1982), Wei et al. (1989) and Prentice et al. (1981).

***Palabras clave*:** análisis de sobrevida, riesgos competitivos, observaciones correlacionadas, modelos marginales.

1. Introduction

Some clinical trials use combined endpoints to evaluate the effect of new therapies. For instance, in the treatment of Heart Failure (HF) patients, a common combined endpoint is death and HF-related hospitalizations (Gheorghiade et al. 2005), and this is often analyzed with a time-to-first-event analysis. In the case of a first event being a hospitalization, this analytical approach ignores subsequent hospitalizations or death. Despite the simplicity of time-to-first-event analysis, this strategy has a severe drawback: the waste of information.

As discussed by Gheorghiade et al. (2005) and Solomon et al. (2007), subsequent events provide detailed information on the disease-control process and are worth modelling to get a more precise understanding of patients' prognoses. The objective of this paper is to explore and formulate a simple and flexible modelling strategy for the joint analysis of survival and time to disease-related hospitalizations. Several marginal models are explored in order to illustrate statistical methods that account for intra-subject correlation. Finally, we propose a multistate model as a flexible modelling strategy for the combined analysis of survival and time to disease-related hospitalizations.

Statistical methods for repeated events survival analysis are illustrated using a HF-dataset derived from the PROSPECT study (Predictors of Response to Cardiac Resynchronization Therapy, study described by Yu et al. 2005) and results published by Chung et al. (2008). The HF-dataset incorporates relevant features and information on HF-related hospitalizations and death for 426 patients, randomly assigned to two treatment groups (G1 and G2).

Figure 1 displays the repeated events nature of the dataset. In 18 patients (7 in G1 and 11 in G2) the first event was death, and for 73 patients (33 in G1 and 40 in G2) the first event was hospitalization. Twenty seven patients presented a second hospitalization (15 in G1 and 12 in G2) and only 6 had a third hospitalization (3 in each group). Ten (3 in G1 and 7 in G2), six (3 in each group) and two (1 in each group) patients died after the first, second and third hospitalization, respectively.

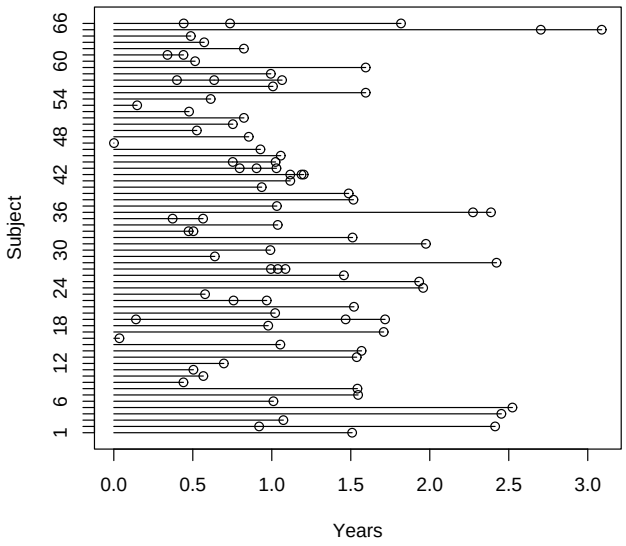

FIGURE 1: Multiple events for 66 patients in the HF-dataset.

In Section 2 several methods and analysis strategies are defined, among them, time to first event Cox proportional hazards model; marginal models for multiple events: Andersen & Gill (1982), Wei et al. (1989) and Prentice et al. (1981) models; and the multi-state model (Andersen & Gill 1982). In Section 3 the pros and cons of these modelling strategies are illustrated using the HF-dataset, analyzing the time to death and/or disease-related hospitalizations. Finally, in Section 4 a discussion is presented and the use of a multi-state model is recommended for the analysis of the HF-dataset. These methods are described next (the descriptions of the Cox and marginal models are based primarily on Therneau & Grambsch (2000)).

2. Methods

Statistical methods for survival analysis, such as the Kaplan-Meier estimator, log-rank test and Cox regression model, can be rewritten as stochastic integrals with respect to counting processes and martingale theory. The counting processes approach is used in this section for the presentation of the different methods (for instance in the description of the Cox model) and the reason for this, as explained by Fleming & Harrington (1991), Therneau & Grambsch (2000) and Andersen et al. (1993), is that counting processes provide a single, elegant, solid basis for survival analysis, which allows flexible ways of modelling (allowing for extensions of the basic survival analysis models to more general multi-state models applicable for event history data) and lead to a unified framework for studying both small sample and asymptotic properties of survival analysis statistics. A complete discussion on

counting processes can be found in the books by Fleming & Harrington (1991) and Andersen et al. (1993).

2.1. The Cox Model

Let $X_{ij}(t)$ be the jth covariate of the ith subject, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$ and X_i denotes the covariate vector for subject i. The hazard for individual i is specified as $\lambda_i(t) = \lambda_0(t)e^{X_i(t)\beta}$, where λ_0 is an unspecified nonnegative function (the baseline hazard), and β is a vector of coefficients. For untied failure time data, Cox (1972) proposed the estimation of β based on the partial likelihood function:

$$PL(\beta) = \prod_{i=1}^{n} \prod_{t \geq 0} \left\{ \frac{Y_i(t) r_i(\beta, t)}{\sum_j Y_j(t) r_j(\beta, t)} \right\}^{dN_i(t)}$$

where $Y_i(t)$ is the indicator function that subject i is still under observation at time t, $N_i(t)$ is the number of observed failures for subject i and $dN_i(t)$ denotes the increment in $N_i(t)$ over the infinitesimal time interval $[t, t + dt)$. $r_i(\beta, t)$ is the risk score for subject i, $r_i(\beta, t) = \exp[X_i(t)\beta] \equiv r_i(t)$. The product integral is defined such that only time points where patient i is at risk ($Y_i(t) = 1$) generate a contribution. Therefore, it is convenient to write the partial likelihood function as:

$$PL(\beta) = \prod_{i=1}^{n} \prod_{t : Y_i(t)=1} \left\{ \frac{r_i(\beta, t)}{\sum_j Y_j(t) r_j(\beta, t)} \right\}^{dN_i(t)}$$

The log partial likelihood is:

$$l(\beta) = \sum_{i=1}^{n} \int_{t:Y_i(t)=1} \left[\log(r_i(\beta, t)) - \log \left(\sum_j Y_j(t) r_j(\beta, t) \right) \right] dN_i(t)$$

$$= \sum_{i=1}^{n} \int_0^{\infty} Y_i(t) \left[X_i(t)\beta - \log \left(\sum_j Y_j(t) r_j(t) \right) \right] dN_i(t)$$

After differentiating the log partial likelihood with respect to β, the $p \times 1$ score vector, $U(\beta)$ is:

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\infty} [X_i(s) - \overline{x}(\beta, s)] dN_i(s)$$

where $\overline{x}(\beta, s)$ is a weighted mean of X, over those observations still at risk at time s,

$$\overline{x}(\beta, s) = \frac{\sum Y_i(s) r_i(s) X_i(s)}{\sum Y_i(s) r_i(s)}$$

$\widehat{\beta}$ is found by solving the partial likelihood equation $U\big(\widehat{\beta}\big) = 0$, and is consistent and asymptotically normally distributed with mean β, and variance $I^{-1}\big(\widehat{\beta}\big)$, the inverse of the observed information matrix.

2.2. Multiple events per subject

A major issue in extending proportional hazards regression models to multiple events per subject is the intra-subject correlation (Therneau & Grambsch 2000). A simple approach that sidesteps this is to take time to first event. Time to first event is simple and easy to interpret, but important information on the disease-recovery process is lost. Other more appropriate approaches are marginal models and multi-state models (with competing risk component). These methods are described below.

2.2.1. Marginal Models

Marginal models offer flexibility in the formation of strata and risk sets, definition of the time scale, and have a well-developed estimator of the variance. Marginal models allow for population average estimation of treatment effect. Therneau & Grambsch (2000) summarize the analysis with these models in three steps:

- Decide on a model (issues such as covariate selection, inclusion of strata, etc.) and structure the data set accordingly.
- Fit the data as an ordinary Cox model, ignoring the possible intra-subject correlation (i.e. treating multiple events from a subject as independent).
- Replace the standard variance estimate with one which is corrected for the possible correlations.

Robust Variance. When a given subject may contribute multiple events, the assumption of independent observations in the standard Cox model does not hold. Lipsitz et al. (1990) showed that marginal models can overcome this assumption for the estimation of the variance of $\widehat{\beta}$ by an appropriate correction based on a grouped jackknife estimate.

Grouped-jackknife values are defined as $J_i = \widehat{\beta} - \widehat{\beta}_{(i)}$, where $\widehat{\beta}_{(i)}$ is the result of the fit that includes all of the individuals except individual i. It is denominated as grouped because in the multiple event case, one individual contributes several observations, and removing a subject implies removing a group of observations. Therneau & Grambsch (2000) describe a way to compute the grouped-jackknife values directly at the Newton-Raphson iteration. The change in the estimated coefficient vector can be expressed in the following way, $\Delta\beta = 1'(UI^{-1}) \equiv 1'D$, where U is the matrix of score residuals. Thus, the change in $\widehat{\beta}$ at each iteration is the column sum of a matrix D, defined as the score residual scaled by I^{-1} (the variance of $\widehat{\beta}$).

This grouped jackknife can be used to derive a robust estimate of the variance for the Cox model. If J is the matrix of grouped-jackknife values (i.e. the ith row of J is $\widehat{\beta} - \widehat{\beta}_{(i)}$), then the grouped jackknife estimate of the variance can be written as the matrix product $V_J = \frac{n-1}{n}\big(J - \overline{J}\big)'\big(J - \overline{J}\big)$, where $\overline{J}$ is the matrix

of column means of J. A natural approximation that is preferred is $D'D$, the matrix product of the approximate jackknife variances (ignoring the $\frac{n-1}{n}$ term). Writing $D'D = I^{-1}(U'U)I^{-1}$, this variance can be viewed as a sandwich estimator ABA, where A is the usual variance and B a correction term. Sandwich estimates are familiar from robust variance estimation in generalized estimating equations (GEE) proposed by Liang & Zeger (1986). Although unbiased, this grouped-jackknife estimate is typically more variable than the ordinary variance of the Cox model, but it is a robust variance that adequately addresses repeat event correlation, and it is expected to be reported when fitting marginal models.

Ordered events. One important issue is to distinguish between data sets where the multiple events have a distinct ordering and those where they do not. In the particular case of the HF-dataset, the outcomes hospitalizations and death are correlated and ordered. Death can happen either as the first event or after hospitalization; there is a specific ordering in this case, obviously after the event of a death it is not possible to have a hospitalization. The most common approaches for correlated ordered outcomes are: the independent increments (Andersen & Gill 1982), marginal (Wei et al. 1989), or PWP (Prentice et al. 1981) models. All three are "marginal" regression models in that $\widehat{\beta}$ is determined from a fit that ignores the correlation between the events followed by a correction of the variance, but differ considerably in their creation of the risk sets.

Andersen and Gill (AG) model. This method is the simplest, but makes the strongest assumptions. Each subject is represented as a series of observations (rows of data) with time intervals as: (entry time, first event], (first event, second event],..., (mth event, last follow up]. The intensity process for subject i is:

$$Y_i(t)\lambda_0(t)\exp(X_i(t)\beta)$$

The difference with the standard Cox model lies in the definition of the at-risk indicator $Y_i(t)$. For survival data, the individual ceases to be at risk when an event occurs and $Y_i(t)$ takes value zero, but for the AG model for recurrent events, $Y_i(t)$ remains one as events occur. Of course the at-risk indicator does not remain one if the event observed is Death. No extra strata or strata by covariate interaction terms are introduced for the multiple events (Therneau & Grambsch 2000). The Andersen-Gill formulation of the Cox proportional hazards model has a number of advantages, including the ability to accommodate left-censored data, time-varying covariates, multiple events, and discontinuous intervals of risks. Some of these practical advantages are discussed in an applied framework by Johnson et al. (2004).

Wei, Lin and Weissfeld (WLW) model. In this model, the ordered outcome dataset is treated as if it were an unordered competing risk case. The number of strata in the analysis will be equal to the maximum number of events a patient reports in the study. Every subject will have one observation in each stratum.

The hazard function for the jth event for subject i is:

$$Y_{ij}(t)\lambda_{0j}(t)\exp(X_i(t)\beta_j)$$

Unlike the AG model, this model allows a separate underlying hazard for each event and for strata by covariate interactions, as shown by the notation β_j. In the WLW model the at-risk indicator for the jth event, $Y_{ij}(t)$, is one until the occurrence of the jth event, unless, of course, some external event causes censoring. When either of those occurs, it becomes zero, indicating that subject is no longer at risk after the last given event (Therneau & Grambsch 2000, Wei et al. 1989). A frequently raised concern is the method's risk set, where each individual is considered to be at risk of all recurrent events from the start of the observation period, and often this method gives estimates that exceed those provided by alternative approaches. By simulation studies, Metcalfe & Thompson (2007) have shown that the WLW model infringes on the proportional hazards assumption when applied to recurrent events data, but the bias this may cause is not behind the distinctive effect estimates. Metcalfe & Thompson (2007) discuss that the analyses of medical data indicate that the infringement of the proportional hazards assumption is not necessarily greater than that experienced with other applications of proportional hazards regression and need not prohibit the application of WLW's method to recurrent events data.

Prentice, Williams and Peterson (PWP) models. This model clearly defines the order of the events. A subject is not at risk for the kth event until he/she has experienced event $k-1$st. Like in the AG model, time intervals are defined as: (entry time, first event], (first event, second event],. . ., (mth event, last follow up], but each event is assigned to a separate stratum (Prentice et al. 1981). The use of time-depending strata means that the underlying hazard function may vary from event to event, unlike the AG model, which assumes that all events are identical. The hazard function for the jth event for subject i is:

$$Y_{ij}(t)\lambda_{0j}(t)\exp(X_i(t)\beta_j)$$

The primary difference between the WLW and PWP models is in the definition of the at-risk indicator and the definition of the strata in the analysis. In the PWP model the at-risk indicator, $Y_{ij}(t)$, is defined as zero until the $j-1$st event and only then becomes one. Once the jth event occurs, $Y_{ij}(t)$ becomes 0 again. The PWP model can be seen as a stratified AG model with event-specific baseline hazards and a restricted risk set. By means of a simulation study, Kelly & Lim (2000) have illustrated that the naive and robust standard errors for the PWP model appear to be similar regardless of the within-subject correlation.

The AG model and the PWP model can be used in the analysis of repeated failure outcomes of the same type, while the approach by the WLW model can be applied to both multiple events of the same type and multiple events of different types as long as there is not a predetermined ordering. The WLW model has a semi-restricted risk set that allows subjects to be at risk for as many events

as the maximum number of events reported per subject in the study, even if most of the subjects only had one event, which (as reported by Kelly & Lim (2000)) leads to overestimation of the treatment effect. For all models, except the PWP model, the robust standard errors become inflated when within-subject events are not independent (Kelly & Lim 2000). When the model is correctly specified (no important covariates are omitted) the PWP model and the AG model estimate unbiased treatment effect and require similar sample size to obtain the same precision in the estimation, while the WLW model estimates biased treatment effect and requires a larger sample size. The PWP model and the AG model are considered to be more efficient than the WLW model, and require less sample size than the time to first event model (Therneau & Grambsch 2000). As noted by Wei & Glidden (1997), the appropriate modelling strategy should be chosen based on the type and nature of the multiple events structure.

2.2.2. Multi-state models

Several of the ideas presented in this section on multi-state models can be found in the 2002 (11) issue of Statistical Methods in Medical Research, entirely devoted to multi-state models. In particular, in the paper by Andersen & Keiding (2002), a multi-state process is defined as a stochastic process $(X(t), t \in T)$ with a finite state space $S = \{1, \ldots, p\}$ and with right-continuous sample paths: $X(t+) = X(t)$. Here $T = [0, \tau]$ or $[0, \tau)$ with $\tau \leq +\infty$. The process has initial distribution $\pi_h(0) = \text{Prob}(X(0) = h), h \in S$. A multi-state process $X(\cdot)$ generates a history consisting of the observation of the process in the interval $[0, t]$. Relative to this history, transition probabilities may be defined by:

$$P_{hj}(s, t) = \text{Prob}(X(t) = j \mid X(s) = h)$$

for $h, j \in S$, and $s, t \in T, s \leq t$ and transition intensities by the derivatives

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{P_{hj}(t, t + \Delta t)}{\Delta t}$$

Graphically, multi-state models may be illustrated using diagrams with boxes representing the states and with arrows between the states representing the possible transitions. A state $h \in S$ is *absorbing* if for all $t \in T, j \in S, j \neq h, \alpha_{hj}(t) = 0$; otherwise h is *transient*. The most simple multi-state model is the two-state model for survival data, which is represented in Figure 2. This model has $p = 2$ states and only one possible transition from state 0 to state 1 (state 0: alive and state 1: dead). The corresponding transition intensity $\alpha_{01}(t)$ is given by the hazard rate function $\alpha(t)$, while $\alpha_{10}(t) = 0$ for all t, that is, state 1 is absorbing. Covariates may be entered into the model using a regression model for $\alpha(\cdot)$. A throughout description of the counting process representation of the multi-state model can be found in the paper by Andersen & Keiding (2002).

In multi-state models, an individual moves from one state to another through time. It is clear that intermediate events, such as disease-related hospitalizations, provide more detailed information on the disease-recovery process and allow for

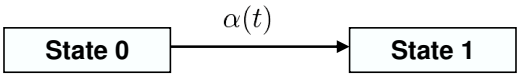

FIGURE 2: The two-state model for survival data.

more precision in predicting the prognosis of patients. These hospitalizations become intermediate events worth modelling in their natural form. These non-fatal events during the course of the disease can be seen as transitions from one state to another. The time origin is characterized by a transition into an initial transient state, such as the start of treatment (entry). Instead of survival data or time-to-event data, data on the history of events is available. Multi-state models provide a framework that allows for the analysis of such event history data (Putter et al. 2007).

If it is assumed that the future depends on the history only through the present, then the process is assumed to be Markovian. In our particular case, given the present state and the event history of a patient, the next state to be visited and the time at which this will occur will only depend on the present state. As explained by Putter et al. (2007), the counting process style of data input with time intervals of (entry time, first event], (first event, second event],..., (mth event, last follow up] can be used by a Markov model.

Andersen & Keiding (2002), Klein & Shu (2002) and Cook & Lawless (2002), discuss the use of multi-state models when the observation plan has been that of cohorts of individuals observed continuously over time. Commenges (2002) treats the situation where individuals are not observed continuously, but only at discrete time points.

3. Statistical Analysis and Results

In this section, each model description indicates how group effect is modelled and what other options are supported by each model. Special attention is given to model assumptions and practical implications for setting up of the dataset.

3.1. Time to first event

The dataset illustrated in Figure 1 was analyzed by a usual "time to first event" analysis. As a first approximation, both outcomes, "hospitalization" and "death", are treated as the same event. The dataset for this analysis has one observation per subject including: Patient identification (ID), time to event in years (TIME), status (event=1 or censoring=0) (STATUS), and the covariates: treatment group (1 or 2)(GROUP), age at enrolment (AGE), and left ventricular ejection fraction (LVEF). Ejection fraction is the percentage of blood pumped out of the left ventricle with each heartbeat. Adamson et al. (2004) have reported

ventricular ejection fraction as a prognostic factor for HF. For patients with ID numbers 35, 36, 37 and 47, the dataset would look as presented in Table 1.

Table 1: Dataset for time to first event analysis (patients with ID 35, 36, 37 and 47).

ID	TIME	STATUS	GROUP	AGE	LVEF
35	0.3723	1	2	63	15
36	2.2735	1	2	50	20
37	1.0322	0	1	52	30
47	0.0010	1	1	79	25

Despite the convenient simplicity of this analysis, there are two important drawbacks. The first one is that information is being wasted-only the first event is considered; the second one is the identical treatment of hospitalization and death, which clinically should be treated as different events.

A more appropriate alternative is an analysis of time to first event, where event can be either hospitalization or death. Since the baseline hazards for the two event types are not expected to be the same, the analysis is stratified by type of event. The dataset for the combined analysis must allow accommodation for competing risks. In this case, the dataset contains one stratum for each outcome type, with each patient appearing in each stratum. The rows in the dataset for patients with ID numbers 35, 36, 37 and 47 are presented in Table 2.

Table 2: Dataset for time to first event analysis with competing risks.

ID	TIME	STATUS	OUTCOME	GROUP	AGE	LVEF
35	0.3723	0	Death	2	63	15
35	0.3723	1	Hospitalization	2	63	15
36	2.2735	0	Death	2	50	20
36	2.2735	1	Hospitalization	2	50	20
37	1.0322	0	Death	1	52	30
37	1.0322	0	Hospitalization	1	52	30
47	0.0010	1	Death	1	79	25
47	0.0010	0	Hospitalization	1	79	25

Note that the first event for patients with ID numbers 35 and 36 is hospitalization, which removes the individuals from being at risk for the first event, Death (therefore, for the first event, Death, Status appears as censor). The lost to follow up of patient with ID=37 makes the status for this individual as censored for both Death and Hospitalization. Finally, the patient with ID=47 has the event Death, therefore removing the patient from being at risk for the first event: Hospitalization.

Several models can be analyzed. However, the interest has been to evaluate the treatment effect adjusted by AGE and LVEF. In the above two model approaches (first one ignoring type of event and second one with competing risks), the estimated treatment effect is not significant (p-value$>$0.6).

3.2. Ordered multiple events

The first alternative to account for multiple events is the AG model. This model treats the two events, hospitalization and death, as if they were the same. Subjects can re-enter the same state multiple times (Figure 3).

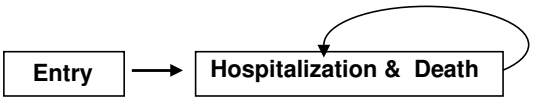

FIGURE 3: Schematic for the AG model approach.

We have the following history of events for patients with ID numbers from 35 to 39 (All patients start at time=0, Table 3). The first event for patient 35 is Hospitalization at 0.372 years after entry; after this the patient has the event Death at time 0.565 years. The first event for the patient with ID=36 is Hospitalization at 2.273 years; after this the patient is loss to follow up at 2.387 years. Patients with ID=37 and 38 do not have any event, they present censoring at times 1.032 and 1.516 years, respectively. Finally, for the patient with ID=39, the first event is hospitalization at 1.117 years after entry, the second event is hospitalization at 1.188 years, and later censoring at time 1.202. Table 3 shows the dataset for the AG model.

TABLE 3: Dataset for multiple events analysis with the AG model (columns: ID, Time1, Time2, Status, Group, Age and LVEF). To adjust a PWP model it is only necessary to add the last column (EventNumber).

ID	TIME1	TIME2	STATUS	GROUP	AGE	LVEF	EventNumber
35	0	0.3723	1	2	63	15	1
35	0.3723	0.5651	1	2	63	15	2
36	0	2.2735	1	2	50	20	1
36	2.2735	2.3874	0	2	50	20	2
37	0	1.0322	0	1	52	30	1
38	0	1.5168	0	1	70	20	1
39	0	1.1170	1	1	80	15	1
39	1.1170	1.1882	1	1	80	15	2
39	1.1882	1.2019	0	1	80	15	3

Note that "type of event" is ignored as one of the assumptions in the AG model. Table 4 shows the parameter estimates for the AG model formulation. As expected in marginal models, robust standard errors (S.E.) are generally larger than model based S.E.

TABLE 4: Parameter estimates for the AG model formulation.

Effect	Parameter estimate	Model based S.E.	Robust S.E.	p-value
GROUP	−0.2787	0.1784	0.2405	0.250
AGE	0.0048	0.0075	0.0089	0.590
LVEF	−0.0562	0.0120	0.0169	0.001

The AG model assumes that all events are identical, which may be too strong an assumption. Figure 4 presents the cumulative hazards for the consecutive events. It clearly suggests that the risk of a new event does not remain constant; on the contrary, it shows that the risk of an event depends on previous events.

Table 5: Parameter estimates for the conditional PWP model formulation.

Effect	Parameter estimate	Model based S.E.	Robust S.E.	p-value
GROUP	−0.3720	0.1809	0.1843	0.044
AGE	0.0028	0.0073	0.0061	0.640
LVEF	−0.0412	0.0121	0.0129	0.001

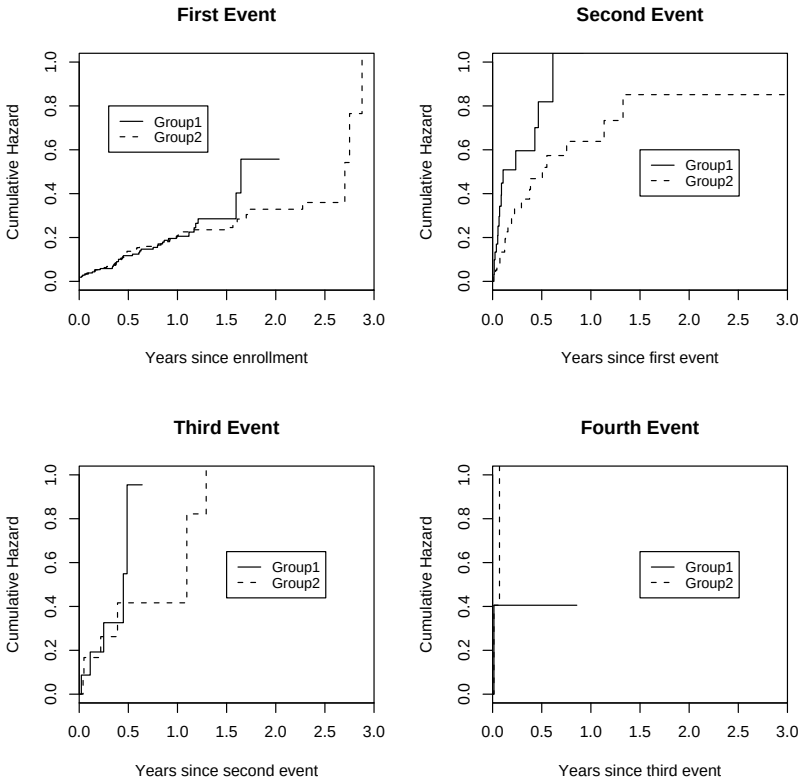

Figure 4: Cumulative hazard for consecutive events.

There is a pre-specified order in the events in HF-patients. A model that encompasses this feature is the PWP model. This model was fitted to incorporate the use of time-depending strata, which means that the underlying hazard function may vary from event to event. The dataset needed to fit a PWP model is the same one used for the AG model, the only difference being that it includes an extra variable, the number of the event (column: EventNumber in Table 3), which is

used for stratification. Table 5 shows the parameter estimates for the conditional PWP model formulation.

The following R code could be used to fit the AG and PWP models.

```
> library(survival)
# AG model #
> coxph(Surv(time1, time2, status) ~ factor(group) + age + lvef
 + cluster(id) , data=data1)
# PWP model #
> coxph(Surv(time1, time2, status) ~ factor(group) + age + lvef
 + cluster(id) + strata(EventNumber), data=data1)
```

3.3. Multi-state models

Two of the characteristics of the HF-dataset are captured by the conditional PWP model presented above, which is able to detect a significant difference between the two treatment groups after controlling for AGE and LVEF (Table 5). First, a patient is not at risk for the kth event until he/she has experienced event $k-1$st. Second, the underlying hazard function may vary from event to event. Unfortunately, the conditional PWP model does not capture the distinction between the two types of events. A very important feature is that Hospitalization and Death, in practice (clinically), cannot be considered equal due to their nature and severity.

The maximum number of hospitalizations for a patient in the HF-dataset is 3. Follow-up after the third hospitalization stops either due to Death or Censoring. Other patients have one or two hospitalizations and afterwards die. A multi-state model where patients transit from one state to another, was proposed to include re-hospitalization and to distinguish the event Death from the event Hospitalization.

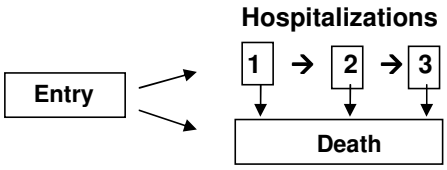

FIGURE 5: Schematic for the full multi-state model.

The multi-state model displayed in Figure 5 does not fit neatly into any of the marginal models described above. The first event can be either Death or Hospitalization (competing risks), a patient is not at risk for a second Hospitalization until he/she experiences Hospitalization as a first event (ordered events), and the underlying risk for the transition from first Hospitalization to Death cannot be assumed to be the same for the transition from the first Hospitalization to a second

Hospitalization. A similar argument holds for the different risks for patients with two or three hospitalizations. In the computer model of the data, there will be seven possible strata-one for each possible transition (Table 6).

Table 6: Strata for the full multi-state model.

Transition	Stratum	Representation
1	Entry - Hospitalization1	E→H1
2	Hospitalization1 - Hospitalization2	H1→H2
3	Hospitalization2 - Hospitalization3	H2→H3
4	Entry - Death	E→D
5	Hospitalization1 - Death	H1→D
6	Hospitalization2 - Death	H2→D
7	Hospitalization3 - Death	H3→D

In setting up the dataset for the multi-state model one must consider the possible transitions that a patient could have at a particular state. For example, the patient with ID=26 has no events, only a censoring at time 1.454 years. This patient has two possible transitions, that is, from Entry to Hospitalization1 (E→H1), or from Entry to Death (E→D); the status in both cases is censoring (0). For the patient with ID=27 the first event is Hospitalization at 0.621 years after entry, his second event is another Hospitalization at time 0.644 years, and finally the patient presents the event Death at 0.672 years after entry. The dataset structure for these two patients is presented in Table 7.

Table 7: Dataset for the full multi-state model.

ID	TIME1	TIME2	TRANSITION	STATUS	GROUP	AGE	LVEF
26	0	1.4543	E→ H1	0	1	73	25
26	0	1.4543	E→D	0	1	73	25
27	0	0.6215	E→H1	1	2	50	20
27	0	0.6215	E→D	0	2	50	20
27	0.6215	0.6439	H1→H2	1	2	50	20
27	0.6215	0.6439	H1→D	0	2	50	20
27	0.6439	0.6720	H2→H3	0	2	50	20
27	0.6439	0.6720	H2→D	1	2	50	20

Note that after the first event (Hospitalization1), the patient with ID=27 has two possible transitions: from Hospitalization1 to Hospitalization2 (H1→H2), or from Hospitalization1 to Death (H1→D). Once the second event is Hospitalization2, the patient is no longer at risk to present transition (H1→D), therefore, the status for this transition is a censoring (0). The same mechanism is applied to the final event.

Table 8 shows parameter estimates for the multi-state model. Due to the assumption of unequal risk for the different transitions, the analysis is stratified by transition.

This multi-state model fully describes the characteristics of the multiple events of the HF-dataset, and is able to detect a significant difference between the two treatment groups after controlling for AGE and LVEF. Subjects in treatment group 2 have a "rate" of state change that is approximately 67% of the rate of those in

Table 8: Parameter estimates for the full multi-state model.

Effect	Parameter estimate	Model based S.E.	Robust S.E.	p-value
GROUP	−0.3880	0.1815	0.1858	0.037
AGE	0.0036	0.0074	0.0061	0.550
LVEF	−0.0414	0.0122	0.0130	0.001

treatment group 1. The assumption for proportional hazards is not rejected (p-value=0.2308). As proposed by Grambsch & Therneau (1994), this proportional hazards test can be seen as a test for assessing the correlation of a scatter plot of the scaled Schoenfeld residuals versus a function of time. This test is implemented and available in R.

The following R code could be used to fit the multi-state model and test for proportional hazards.

```
> library(survival)
> fit1 <- coxph(Surv(time1, time2, status) ~ factor(group) + age
+ lvef + cluster(id) + strata(transition) , data=multistate)
> print(cox.zph(fit1))
```

We can also explore whether treatment affects some transitions more than others by looking at the GROUP by TRANSITION interaction in Table 9. The group effect is strongest with respect to transitions after the first hospitalization (group effect in the H3→D transition is not estimated due to lack of cases in the data). The group by transition interaction is certainly interesting and has to be rigorously proven in order to make conclusions.

Table 9: GROUP by TRANSITION interaction in the full multi-state model.

Effect	Parameter estimate	Exp of parameter estimate
GROUP,E→H1	−0.11310	0.893
GROUP,E→D	−0.18208	0.834
GROUP,H1→H2	−1.03639	0.355
GROUP,H1→D	−0.06922	0.933
GROUP,H2→H3	−1.51823	0.219
GROUP,H2→D	−1.03254	0.356

A simplification of the multi-state model for the HF-dataset is done by assuming the baseline hazards of the H1→H2 and H2→H3 transitions to be proportional, i.e. to have only one Hospitalization state that can be visited more than once. More generally, we may assume a simplification of the multi-state model as shown in Figure 6. Furthermore, this simplification is assuming that baseline hazards for transitions H1→D, H2→D and H3→D are proportional.

To assess whether the assumption of proportional hazards for the transitions is reasonably fulfilled, we explore the estimated cumulative baseline hazards for the multi-state model described in Figure 5. The cumulative baseline hazards for transitions H1→D, H2→D and H3→D are shown in Figure 7.

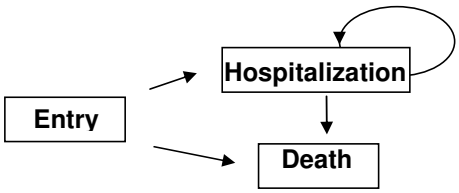

FIGURE 6: Schematic for the simplified multi-state model.

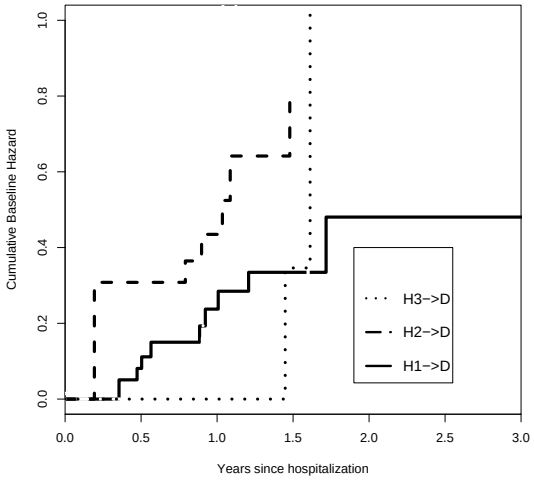

FIGURE 7: Cumulative baseline hazard for transitions H1→D, H2→D and H3→D in the full multi-state model.

The graphical exploration indicates that the baseline hazards are not proportional. This can be checked more rigorously by testing the significance of an interaction between time, and an indicator distinguishing between the three transitions; although this should be considered with caution due to the small sample size. For this case the interaction was significant (p-value <0.01), supporting the graphical check of disproportionate hazards.

The lack of proportionality between the specific transitions suggests not simplifying the multi-state model. The preferred model for the analysis of the HF-dataset is the full multi-state model fitted in Table 8.

4. Conclusion and discussion

Survival analysis has been a common and well-accepted strategy to study treatment effect in a population of patients. During the last few years, there has been an increasing interest in assessing therapy effect not only by using time to Death, but also time to surrogate events; a good example of which is time to hospitalizations. The combined endpoint of time to Death and time to disease-related Hospitaliza-

tions is often analyzed with a time-to-first-event analysis, which has the drawback of waste of information and indistinct handling of two clinically different events.

The analysis of multiple events per subject cannot be approached by a standard Cox model, where the assumption of independence of observations is not valid. In order to account for intra-subject correlation, we have presented the use of marginal and multi-state models using a counting process approach for the joint analysis of survival and time to disease-related Hospitalizations. The characteristics and limitations of the WLW, AG, and PWP models have been illustrated in the modelling of time to HF-related Hospitalizations and Death. All of these models allow for population average estimates of treatment effect and are easily approached using standard statistical software such as SAS, R, and S-Plus. The AG model assumes that all events are identical and can be revisited, and the WLW model only accommodates unordered competing risk. Both models make strong assumptions that are not suitable in the analysis of HF-related Hospitalizations and Death. The PWP model accounts for pre-specified order in the events and competing risks, but has the drawback of assuming Hospitalization and Death as the same type of events, which, given their nature and severity, is clinically unacceptable.

The most appropriate model should be chosen based on the nature of the data. For the HF-dataset we recommend a multi-state model as it allows the incorporation of important features in the analysis of HF-related hospitalizations and death, such as multiple ordered events per subject, event history data, accommodation of competing risks, and the distinction between two different clinical events: death and hospitalization. The proposed simple and flexible multi-state model extends relevant characteristics of the WLW, AG, and PWP models, while capturing important features of time to disease-related Hospitalizations and Death in the HF-dataset, and allows for a more precise understanding of the disease-control process in this particular group of patients.

Acknowledgments

The authors wish to thank the referees for their valuable comments.

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

References

Adamson, P. B., Smith, A. L., Abraham, W. T., Kleckner, K. J., Stadler, R. W., Shih, A. & Rhodes, M. M. (2004), 'Continuous Autonomic Assessment in Patients with Symptomatic Heart Failure. Prognostic Value of Heart Rate Variability Measured by an Implanted Cardiac Resynchronization Device', *Circulation* **110**, 2389–2394.

Andersen, P. K., Borgan, O., Gill, R. D. & Keiding, N. (1993), *Statistical Models Based on Counting Processes*, Springer, New York, United States.

Andersen, P. K. & Gill, R. D. (1982), 'Cox's Regression Model for Counting Pprocesses: A large Sample Study', *Annals of Statistics* **10**, 1100–1120.

Andersen, P. K. & Keiding, N. (2002), 'Multi-State Models for Event History Analysis', *Statistical Methods in Medical Research* **11**, 91–115.

Chung, E. S., Leon, A. R., Tavazzi, L., Sun, J. P., Nihoyannopoulos, P., Merlino, J., Abraham, W. T., Ghio, S., Leclercq, C., Bax, J. J., Yu, C. M., Gorcsan, J., Sutton, S. M., De Sutter, J. & Murillo, J. (2008), 'Results of the Predictors of Response to CRT (PROSPECT) Trial', *Circulation* **117**, 2608–2616.

Commenges, D. (2002), 'Inference for Multi-State Models From Interval-Censored Data', *Statistical Methods in Medical Research* **11**, 167–182.

Cook, R. J. & Lawless, J. F. (2002), 'Analysis of Repeated Events', *Statistical Methods in Medical Research* **11**, 141–166.

Cox, R. D. (1972), 'Regression Models and Life-Tables (with Discussion)', *Journal on the Royal Statistical Society. Series B* .

Fleming, T. R. & Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, John Wiley & Sons, Inc.

Gheorghiade, M., Zannad, F., Sopko, G., Klein, L., Piña, I. L., Konstam, M. A., Massie, B. M., Roland, E., Targum, S., Collins, S. P., Filippatos, G. & Tavazzi, L. (2005), 'Acute Heart Failure Syndromes-Current State and Framework for Future Research', *Circulation* **112**, 3958–3968.

Grambsch, P. M. & Therneau, T. M. (1994), 'Proportional Hazards Tests and Diagnostics Based on Weighted Residuals', *Biometrika* **81**, 515–526.

Johnson, C. J., Boyce, M. S., Schwartz, C. C. & Haroldson, M. A. (2004), 'Modeling Survival: Application of the Andersen-Gill Model to Yellowstone Grizzly Bears', *The Journal of Wildlife Management* **68**, 966–978.

Kelly, P. J. & Lim, L. L. (2000), 'Survival Analysis for Recurrent Event Data: an Application to Childhood Infectious Diseases', *Statistics in Medicine* **19**, 13–33.

Klein, J. P. & Shu, Y. (2002), 'Multi-State Models for Bonemarrow Transplantation Studies', *Statistical Methods in Medical Research* **11**, 117–139.

Liang, K. Y. & Zeger, S. L. (1986), 'Longitudinal Data Analysis Using Generalized Linear Models', *Biometrika* **73**(1), 13–22.

Lipsitz, S. R., Laird, N. M. & Harrington, D. P. (1990), 'Using the Jackknife to Estimate the Variance of Regression Estimators from Repeated Measures Studies', *Communication in Statistics. Theory and Methods* **19**, 821–845.

Metcalfe, C. & Thompson, S. G. (2007), 'Wei, Lin and Weissfeld's Marginal Analysis of Multivariate Failure Time Data', *Statistical Methods in Medical Research* **16**, 103–122.

Prentice, R. L., Williams, B. J. & Peterson, A. V. (1981), 'On the Regression Analysis of Multivariate Failure Time Data', *Biometrika* **68**, 373–379.

Putter, H., Fiocco, M. & Geskus, R. B. (2007), 'Tutorial in Biostatistics: Competing Risks and Multi-State Models', *Statistics in Medicine* **26**, 2389–2430.

Solomon, S. D., Dobson, J., Pocock, S., Skali, H., McMurray, J. J., Granger, C. B., Tusuf, S., Swedberg, K., Young, J. B., Michelson, E. L. & Pfeffer, M. A. (2007), 'Influence of Nonfatal Hospitalization for Heart Failure on Subsequent Mortality in Patients with Chronic Heart Failure', *Circulation* **116**, 1482–1487.

Therneau, T. M. & Grambsch, P. M. (2000), *Modelling Survival Data: Extending the Cox Model*, Springer, New York, United States.

Wei, L. J. & Glidden, D. V. (1997), 'An Overview of Statistical Methods for Multiple Failure Time Data in Clinical Trials', *Statistics in Medicine* **16**, 833–839.

Wei, L. J., Lin, D. Y. & Weissfeld, L. (1989), 'Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions', *Journal of the American Statistical Association* **84**, 1065–1073.

Yu, C. M., Abraham, W. T., Bax, J. J., Chung, E. S., Fedewa, M., Ghio, S., Leclercq, C., Leon, A. R., Merlino, J., Nihoyannopoulos, P., Notabartolo, D., Sun, J. & Tavazzi, L. (2005), 'Predictors of Response to Cardiac Resynchronization Therapy (PROSPECT): study design', *American Heart Journal* **149**, 600–605.

Revista Colombiana de Estadística
Junio 2010, volumen 33, no. 1, pp. 63 a 88

Confidence and Credibility Intervals for the Difference of Two Proportions

Intervalos de confianza y de credibilidad para la diferencia de dos proporciones

Hanwen Zhang[1,a], Hugo Andrés Gutiérrez Rojas[1,b], Edilberto Cepeda Cuervo[2,c]

[1]Centro de Investigaciones y Estudios Estadísticos (CIEES), Facultad de Estadística, Universidad Santo Tomás, Bogotá, Colombia

[2]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

This paper presents a frequentist comparison of the performance of confidence and credibility intervals for the difference of two proportions from two independent samples. The comparison is carried out considering three frequentist criteria. It was found that the intervals with the best performance, in terms of coverage probability, are Bayesians; in terms of expected length and variance of the length, the Newcombe interval shows the best performance. As a final remark, it was found that traditional intervals such as the Wald and adjusted Wald have a poor performance.

***Key words*:** Confidence intervals, Credibility intervals, Difference of two proportions..

Resumen

Este artículo presenta una comparación del comportamiento de intervalos de confianza frecuentistas y de credibilidad bayesianos para la diferencia de dos proporciones provenientes de muestras aleatorias independientes. La comparación se lleva cabo considerando tres criterios frecuentistas con los cuales se concluyó que el mejor comportamiento, en términos de la probabilidad de cobertura, lo tienen los intervalos bayesianos, y en términos de la longitud esperada y varianza de la longitud el mejor comportamiento está dado por el intervalo frecuentista de Newcombe. Como resultado de esta investigación se encontró que los intervalos frecuentistas más populares como Wald y Wald ajustado tienen un comportamiento deficiente.

***Palabras clave*:** intervalos de confianza, intervalos de credibilidad, diferencia de dos proporciones.

[a]Docente investigadora. E-mail: hanwenzhang@usantotomas.edu.co
[b]Director. E-mail: hugogutierrez@usantotomas.edu.co
[c]Profesor asociado. E-mail: ecepedac@unal.edu.co

1. Background

A common problem in practical statistics is estimatig the difference of two proportions by means of interval estimation. This topic is especially important in clinical trials where it is necessary to investigate cure rates of two drugs or treatments. The theoretical background of this research is as follows: suppose that $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are two independent samples such that $X_i \sim Bernoulli(p_1)$ and $Y_j \sim Bernoulli(p_2)$, with $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$. It is necessary to construct a confidence interval or a credibility interval for the difference of the proportions $p_1 - p_2$.

The most popular method for estimating $p_1 - p_2$ by means of frequentist confidence interval is the Wald interval, which is presented in most introductory statistics textbooks in spite of its poor performance. Many modifications have been made to the Wald interval in order to improve it. One of them is the adjusted Wald interval obtained by widening the Wald interval to increase the coverage probability. This improvement is especially meaningful when the sample sizes are small. Another important interval is the score interval (Wilson 1927), obtained by inverting the score test statistics. This interval was first obtained for one proportion, and thereafter was to be extended to deal with the difference of two proportions. However, in that case, the interval lacks a closed form (Pan 2002) and must be computed by numerical approximations. Agresti & Caffo (2000) analyzed the score interval, and derived the Adding-4 method: add 2 successes and 2 failures to sample observation. A considerable number of authors agree that Agresti and Caffo method has a very good performance (Pan 2002, Correa & Sierra 2003, Agresti et al. 2008). Another interval obtained by modifying the score method is the Newcombe interval (Newcombe 1998*a*, 1998*b*), and it seems to have a similar performance to the Agresti and Caffo interval (Correa & Sierra 2003).

In the Bayesian approach, Pham-Gia & Turkkan (1993) used the hypergeometric Appell function and derived the posterior distribution of $p_1 - p_2$ when beta priors are used for each proportion. Given the exact posterior distribution, an exact Bayesian credibility interval for $p_1 - p_2$ can be found. However the computational procedures are somewhat tedious, therefore new computational methods such as the Markov Chain Monte Carlo (MCMC), can be used to make it easier to evaluate posterior distributions for $p_1 - p_2$, as Agresti & Min (2005) argued.

In the literature, many comparisons between confidence intervals have been done (Newcombe 1998*a*, Newcombe 1998*b*, Agresti & Caffo 2000, Pan 2002, Correa & Sierra 2003). The aim of this research is to take into account Bayesian credibility intervals jointly with frequentist confidence intervals. After a brief introduction, Section 2 presents some frequentist and Bayesian intervals for $p_1 - p_2$. Traditional confidence intervals such as the Wald and adjusted Wald are considered, as well as Bayesian credibility intervals with two noninformative priors. Section 3 deals with the comparison criteria for the considered intervals: the coverage probability, the expected length, and the variance of the length are used in order to evaluate the performance of the intervals. Section 4 presents results for the performance of the intervals with varying sample sizes, varying values of a single proportion and, finally, the difference of the two proportions. Other scenarios were analyzed,

but all of them yield similar conclusions. Section 5 provides a survey of other intervals and their performance, and finally Section 6 gives some conclusions and recommendations.

2. Some intervals

In this section we introduce some confidence and credibility intervals that are considered and lead the research through out this paper. We denote $\widehat{p}_1$ as the maximum likelihood estimator of p_1 defined as $\sum_{i=1}^{n_1} \frac{X_i}{n_1}$ and analogously for $\widehat{p}_2$.

2.1. Frequentist intervals

The Wald interval is based on the normal approximation to the distribution of $\widehat{p}_1 - \widehat{p}_2$, when the sample sizes are large, by considering that

$$E(\widehat{p}_1 - \widehat{p}_2) = p_1 - p_2$$
$$Var(\widehat{p}_1 - \widehat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

By the the central limit theorem a $(1-\alpha)100\%$ interval for $p_1 - p_2$ is clearly defined by (L_{low}, L_{upp}), where

$$L_{low} = \widehat{p}_1 - \widehat{p}_2 - z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \tag{1}$$

and

$$L_{upp} = \widehat{p}_1 - \widehat{p}_2 + z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \tag{2}$$

The computation of this interval is very simple, and it is presented in most of the statistical inference textbooks. Despite the fact of its popularity, many authors have shown that the performance of this interval is quite poor (Ghosh 1979, Vollset 1993, Newcombe 1998*a*, Newcombe 1998*b*). Moreover, when the sample sizes are large, the Wald interval still performs poorly (Brown et al. 2001).

Considering that the Wald interval uses a continuous distribution to approximate a discrete distribution, an alternative to for improving the performance of the Wald interval is to incorporate the continuity correction factor by adding a constant term to both the lower and upper limits. The resulting limits of the adjusted Wald interval are:

$$L_{low} = \widehat{p}_1 - \widehat{p}_2 - z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} - \frac{n_1+n_2}{2n_1n_2} \tag{3}$$

and

$$L_{upp} = \widehat{p}_1 - \widehat{p}_2 + z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} + \frac{n_1+n_2}{2n_1n_2} \tag{4}$$

The adjusted Wald interval, by definition, has a wider length than the Wald interval. This leads to an increasing the coverage probability, but at the same time, widening the interval leads to a loss of precision.

Agresti & Caffo (2000) proposed to combine the Wald interval and the score interval, due to Wilson (1927), by adding pseudo observations in order to increase the coverage probability. They found that the optimum number of pseudo observations to add is four: two successes and two failures, and they showed that the performance of the resulting Agresti-Caffo interval is surprisingly high even for small sample sizes. The limits of Agresti-Caffo interval are:

$$L_{low} = \widetilde{p}_1 - \widetilde{p}_2 - z_{1-\alpha/2}\sqrt{V(\widetilde{p}_1,\widetilde{n}_1) + V(\widetilde{p}_2,\widetilde{n}_2)} \tag{5}$$

and

$$L_{upp} = \widetilde{p}_1 - \widetilde{p}_2 + z_{1-\alpha/2}\sqrt{V(\widetilde{p}_1,\widetilde{n}_1) + V(\widetilde{p}_2,\widetilde{n}_2)} \tag{6}$$

with

$$V(\widetilde{p}_i,\widetilde{n}_i) = \frac{1}{\widetilde{n}_i}\left[\widetilde{p}_i - \widetilde{p}_i\frac{n_i}{\widetilde{n}_i} + \frac{1}{2\widetilde{n}_i}\right]$$

where $\widetilde{n}_i = n_i + 2$ for $i = 1, 2$, $\widetilde{p}_1 = \frac{\sum_{j=1}^{n_1} X_j + 1}{\widetilde{n}_1}$ and $\widetilde{p}_2 = \frac{\sum_{j=1}^{n_2} Y_j + 1}{\widetilde{n}_2}$.

Another confidence interval obtained by combining the Wald and the score interval is the Newcombe interval. To compute this interval, the following equation for each p_i should first be solved

$$|\widehat{p}_i - p_i| = z_{1-\alpha/2}\sqrt{\frac{p_i(1-p_i)}{n_i}}$$

Let's denote the solutions by l_i and u_i with $l_i < u_i$, $i = 1, 2$. The limits of the Newcombe interval are

$$L_{low} = \widehat{p}_1 - \widehat{p}_2 - z_{1-\alpha/2}\sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}} \tag{7}$$

and

$$L_{upp} = \widehat{p}_1 - \widehat{p}_2 + z_{1-\alpha/2}\sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{l_2(1-l_2)}{n_2}} \tag{8}$$

Newcombe found that this interval has good coverage and average length properties.

2.2. Bayesian intervals

Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities, such as predictions for new observations (Gelman et al. 2004). This process can be carried out by using Markov Chain Monte Carlo methods that simulate values from the posterior distribution of the parameter of interest[1]. Thus, we appeal to the Gibbs sampling algorithm to simulate values from the posterior distribution.

In order to implement a Gibbs sampling algorithm for the problem of finding a credibility interval for $p_1 - p_2$, we chose the prior distributions of p_1 and p_2 to be $Beta(a_1, b_1)$ and $Beta(a_2, b_2)$, respectively. Once the samples are drawn, the observed information is given by $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$ or equivalently by $S_x = \sum_{j=1}^{n_1} x_j$ and $S_y = \sum_{j=1}^{n_2} y_j$. The posterior marginal distributions of p_1 and p_2 are obtained by Bayes theorem and are given by $Beta(a_1 + S_x, b_1 + n_1 - S_x)$ and $Beta(a_2 + S_y, b_2 + n_2 - S_y)$, respectively (Gelman et al. 2004, p. 34). Since the samples come from two independent populations, the posterior joint distribution of (p_1, p_2) is a product of its marginal distributions and, for this reason, one can get samples from the posterior distribution of $p_1 - p_2$ by simulating N values from the posterior distribution of p_1 and p_2, say $p_1^{(1)}, \ldots, p_1^{(N)}$ and $p_2^{(1)}, \ldots, p_2^{(N)}$, respectively. Then, by computing $p_1^{(1)} - p_2^{(1)}, \ldots, p_1^{(N)} - p_2^{(N)}$, we obtain simulated values from the posterior distribution of $p_1 - p_2$. Note that the algorithm presented here generates independent samples from the posterior, so it is fair to name it as just a Monte Carlo algorithm, rather than a Markov Chain Monte Carlo algorithm.

After that, it is possible to compute the credibility interval[2] of $100 \times (1 - \alpha)\%$ for $p_1 - p_2$ using the percentiles of the values simulated that induce the shortest credible intervals. In this research, we consider two noninformative priors for p_1 and p_2: $Beta(1, 1)$ and $Beta(0.5, 0.5)$ priors. $Beta(1, 1)$ corresponds to the uniform distribution, which provides the same weight along all values in the range $(0, 1)$ for each p_i with $i = 1, 2$. When both priors of p_1 and p_2 are uniform priors, the prior distribution for the difference $p_1 - p_2$ is a triangular distribution with vertices $(-1, 0)$, $(1, 0)$ and $(0, 1)$. That is to say the prior distribution provides greater weight to values of $p_1 - p_2$ close to 0, and small weights to values close to the extremes -1 and 1.

The $Beta(0.5, 0.5)$ is known as the Jeffreys prior, which, according to Carlin & Louis (1998, p. 51), is noninformative in a transformation-invariate sense. However, it provides extra weight to extreme values of p_i, that is, values close to 0 and 1. When both priors of p_1 and p_2 are the Jeffreys prior, the prior distribution of $p_1 - p_2$ is symmetric at the value 0 where it is not defined, increasing for values

[1] In the case of estimating the difference of two proportions, the exact posterior distribution of $p_1 - p_2$ is given by Pham-Gia & Turkkan (1993). However, this exact distribution is somewhat complicated and computationally expensive to obtain.

[2] There are many ways to construct a Bayesian credible interval from the posterior distribution. A naive way to construct it is by using the upper and lower $\alpha/2$ quantiles. However, as the intervals are to be judged by expected length and its variance, it would make more sense to use the highest posterior density intervals which are, by definition, the shortest credible intervals with the given coverage (Carlin & Louis 1998, p. 43).

in $(0, 1)$ and decreasing for values in $(-1, 0)$. The explicit density function of the priori distribution of $p_1 - p_2$ when both priors of p_1 and p_2 are beta is studied in Pham-Gia & Turkkan (1993).

3. Comparison criteria

In this section, we establish some criteria in order to measure the performance of the intervals in a frequentist sense. A good confidence or credibility interval should have the true coverage probability close to or larger than the nominal value. Of course, in most cases, a way to increase the coverage probability is by widening the interval, obtaining intervals with little precision. The comparison of different methods for obtaining confidence intervals for one parameter must take into account their lengths. To accomplish this, mean and variance of those lengths are analyzed in this paper. In conclusion, we use the following criteria:

1. The true coverage probability defined by:

$$CP = E(I(X, Y, p_1, p_2)) \tag{9}$$

where X and Y denote the number of successes in n_1 and n_2 trails, respectively. $I(x, y, p_1, p_2)$ defines an indicator function that is equal to one if the interval contains $p_1 - p_2$ when $X = x$ and $Y = y$, and equal to zero if the interval does not contain $p_1 - p_2$. The coverage probability is given by:

$$CP = \sum_{x=0}^{n_1} \sum_{y=0}^{n_2} I(x, y, p_1, p_2) \binom{n_1}{x} p_1^x (1 - p_1)^{n_1 - x} \binom{n_2}{y} p_2^y (1 - p_2)^{n_2 - y} \tag{10}$$

2. The expected length defined by:

$$l = E(U(X, Y) - L(X, Y)) \tag{11}$$

where $U(X, Y)$ and $L(X, Y)$ are the upper and lower limit of the confidence or credibility interval for $p_1 - p_2$. Note that they are functions of the variables X and Y. The expected length is given by:

$$l = \sum_{x=0}^{n_1} \sum_{y=0}^{n_2} (U(x, y) - L(x, y)) \binom{n_1}{x} p_1^x (1 - p_1)^{n_1 - x} \binom{n_2}{y} p_2^y (1 - p_2)^{n_2 - y} \tag{12}$$

3. Analogously, we define the variance of length by:

$$V = Var(U(X, Y) - L(X, Y)) \tag{13}$$

and it is easy to show that

$$V = \sum_{x=0}^{n_1}\sum_{y=0}^{n_2}(U(x,y)-L(x,y))^2\binom{n_1}{x}p_1^x(1-p_1)^{n_1-x}\binom{n_2}{y}p_2^y(1-p_2)^{n_2-y} - \left(\sum_{x=0}^{n_1}\sum_{y=0}^{n_2}(U(x,y)-L(x,y))\binom{n_1}{x}p_1^x(1-p_1)^{n_1-x}\binom{n_2}{y}p_2^y(1-p_2)^{n_2-y}\right)^2 \tag{14}$$

Notice that these criteria are frequentist, in the sense that in (10), (12) and (14), the proportions p_1 and p_2 are assumed to be fixed values, rather than random variables.

4. Comparison among intervals

In this section, we compare several confidence and Bayesian credibility intervals with respect to coverage probability and mean and variance of their lengths. For confidence intervals (Wald, adjusted Wald, Agresti-Caffo and Newcombe), those values were exactly computed for several combinations of p_1, p_2 and different sample sizes. For Bayesian intervals, the computation was done by means of the simulation of samples of the posterior distributions of p_1 and p_2. These distributions were obtained through the Markov Chain approach, and prior distributions used for p_1 and p_2 were the same: $Beta(1,1)$ and $Beta(0.5,0.5)$. In subsection 4.1, the true coverage probability of 0.95 confidence level or credibility level of intervals are obtained for $p_2 = 0.5$, $p_1 \in (0,1)$ and $n_j \in \{10,50,100\}$, with $j = 1,2$, for the two priors described above. Subsequently, the mean and variance of the intervals were computed. Subsection 4.2 shows the same kind of study, with the same chosen values as in 4.1, except that n_2 is fixed at $n_2 = 30$. In 4.3, $n_1 \in \{1,2,\ldots,500\}$, $n_2 = 30$ and $(p1 - p2) \in \{0, 0.1, 0.5, 0.8\}$.

4.1. Performance of intervals by varying sample sizes

We compare the performance of the confidence and credibility intervals for different sample sizes n_1 and n_2. First, we calculate the true confidence level for the confidence intervals as a function of p_1. The value p_2 is fixed as 0.5, the samples sizes of X and Y are assumed to be the same, and we consider the values $n_1 = n_2 = 10, 50, 100$. The resulting coverage probabilities for the Wald and adjusted Wald intervals are presented in Figure 1. It is seen that the coverage probability of the adjusted Wald interval is always larger than the Wald interval; this fact is intuitive since the adjusted Wald interval is obtained by widening the Wald interval. Additionally, the coverage probability is not affected by the different values of p_1. Also, the poor performance of the Wald interval is noted, especially in small samples.

On the other hand, the coverage probabilities of the Agresti-Caffo and Newcombe intervals are presented in Figure 2. It can be seen that both intervals have

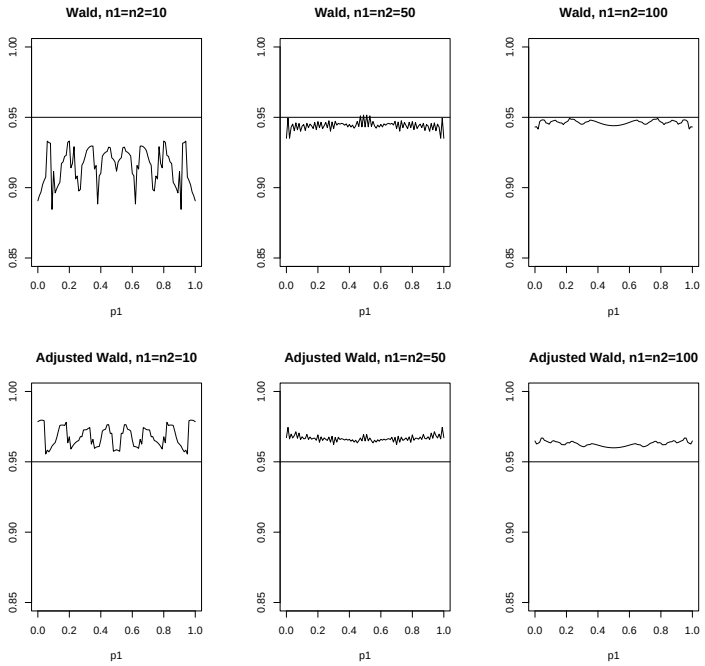

FIGURE 1: True coverage probability of the Wald and Adjusted Wald intervals varying p_1 with $n_1 = n_2 = 10, 50, 100$ with a nominal coverage probability of 0.95

coverage probability quite close to the nominal coverage 0.95, a desirable property that the Wald and adjusted Wald do not have. Although the adjusted Wald interval has coverage probability larger than 0.95, we will see later that its length is the largest. Also, the coverage probability of the Newcombe interval is sean to be affected by different values of p_1, especially when the samples are small.

The coverage probability for Bayesian intervals is presented in Figure 3, where it is seen that the performance of these two intervals are similar, and are quite good in the sense that the coverage probability is stable with respect to p_1, and is close to the nominal 0.95 even when the samples are small. So we can conclude that, in terms of true coverage probability, the Bayesian intervals are better than the frequentist intervals, without ignoring the notable performance of the Agresti-Caffo and Newcombe intervals. As a final remark, the true coverage probabilities of all the intervals considered become more stable with respect to p_1 as the sample sizes increases.

We now compare the intervals in terms of the expected length. The expected lengths of the considered intervals with different sample sizes are presented in Figure 4. It is be seen that the interval with largest length is the adjusted Wald interval. This shows that the high coverage probability is due to the length of the interval, but is not due to its good performance. The shape of the curve for the Wald interval is similar to the adjusted Wald; this is intuitive since the adjusted

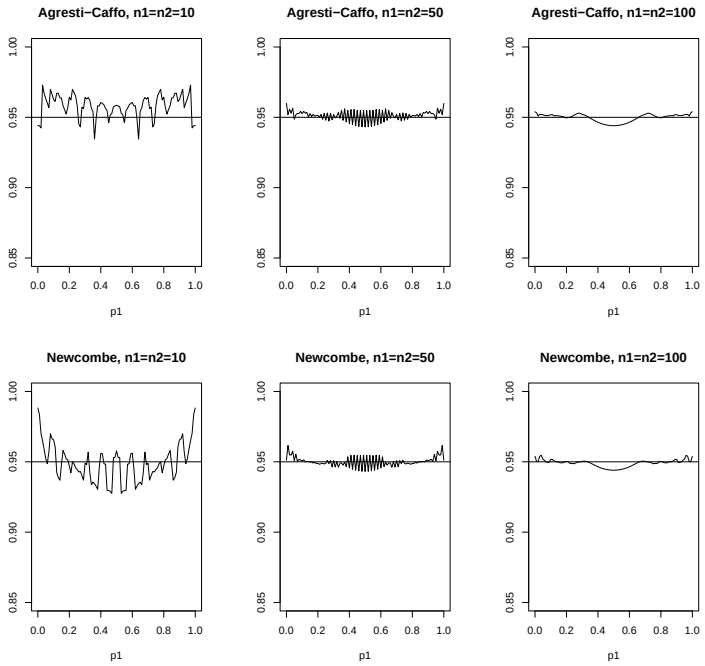

FIGURE 2: True coverage probability of the Agresti-Caffo and Newcombe intervals varying p_1 with $n_1 = n_2 = 10, 50, 100$ with a nominal coverage probability of 0.95

Wald interval is obtained by subtracting and adding a constant to the lower and upper limit of the Wald interval, respectively. As a result then, the following relationship between the lengths of these intervals remains:

$$l_{A.Wald} = l_{Wald} + \frac{n_1 + n_2}{n_1 n_2} \tag{15}$$

The Agresti-Caffo and Newcombe intervals have a more stable expected length with respect to p_1 than the Wald and adjusted Wald intervals. The improvement is noted especially in small samples. In samples with $n_1 = n_2 = 50, 100$, the length of the Agresti-Caffo and Newcombe intervals are smaller than the Wald and adjusted Wald intervals.

The expected lengths of the Bayesian intervals are also presented also in Figure 4, where it is seen that the performance of the intervals with the uniform and Jeffreys prior are similar. However, their expected lengths are larger than the Agresti-Caffo and Newcombe intervals when $n_1 = n_2 = 100$; when $n_1 = n_2 = 50$, the lengths are similar; when $n_1 = n_2 = 10$, the Bayesian intervals show a similar performance to the Newcombe interval while the Agresti-Caffo interval has a slightly larger expected length. In conclusion, the Newcombe interval has the smallest expected length in all sample sizes.

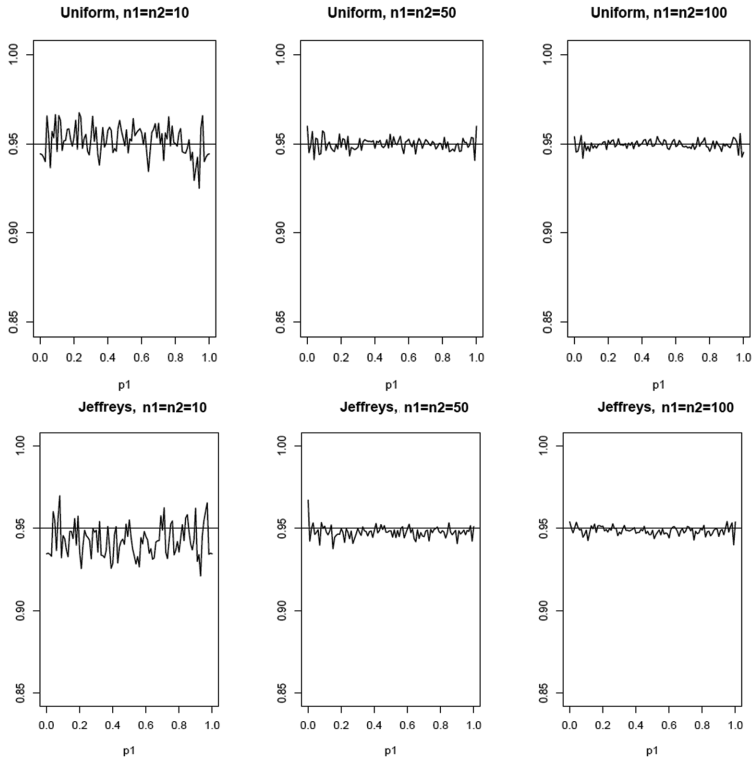

FIGURE 3: True coverage probability of the Bayesian intervals varying p_1 with $n_1 = n_2 = 10, 50, 100$ with a nominal coverage probability of 0.95

Finally, we compare the intervals in terms of variance of the length. Notice that the variance of the length of the adjusted Wald interval is equal to the Wald interval. Recalling (15) and using the property of variance, we have that $Var(l_{A.Wald}) = Var(l_{Wald})$. So in the figures related of the variance of the length, we only plot the variance of length for the Wald interval.

The variances of length for Wald/adjusted Wald, Agresti-Caffo and Newcombe intervals are presented in Figure 5. It is seen that the Newcombe interval has the smallest variance, although very close to the variance of the Agresti-Caffo interval. The huge variance of the Wald and adjusted Wald intervals in small samples is also seen. On the other hand, the variances of the Bayesian intervals are presented in Figure 6, and, the performance of the intervals with the uniform prior and the Jeffreys prior are similar. However, their variance is larger than both the Agresti-Caffo and Newcombe intervals.

In conclusion, in terms of true coverage probability, the best intervals are the Bayesian; in terms of the expected length, the best interval is the Newcombe interval, as well as in terms of variance of the length.

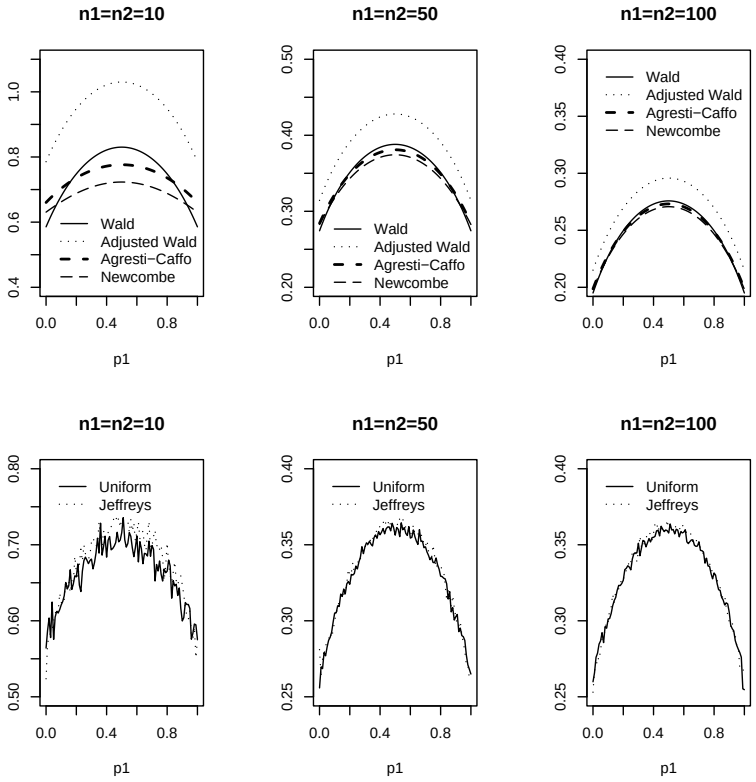

FIGURE 4: Expected length of confidence and Bayesian intervals varying p_1 with $n_1 = n_2 = 10, 50, 100$.

4.2. Performance of intervals varying values of p_1

In this section, we compare the performance of the intervals when different values of p_1 are considered.

First, we compute the true coverage probability as a function of n_1, the value of n_2 is fixed to be 30, the value of p_2 is 0.5, and we consider the values $p_1 = 0.01, 0.1, 0.3, 0.5$. The true coverage probability for the Wald and adjusted Wald intervals are presented in Figure 7. It is seen that, as in the previous section, the coverage probability of the adjusted Wald interval is always larger than the Wald interval. Additionally for the adjusted Wald interval, as the sample size n_1 increases, the coverage probability becomes more stable, while for the Wald interval, the increasing sample size does not improve the coverage probability when $p_1 = 0.01, 0.1, 0.3$.

In Figure 8, the coverage probabilities of the Agresti-Caffo and Newcombe intervals are presented. We see that the performance of the Newcombe interval is better than Agresti-Caffo interval as its coverage probability is more stable; when $p = 0.01, 0.1, 0.3$, it is always larger than the nominal 0.95, and when $p_1 = 0.5$, i

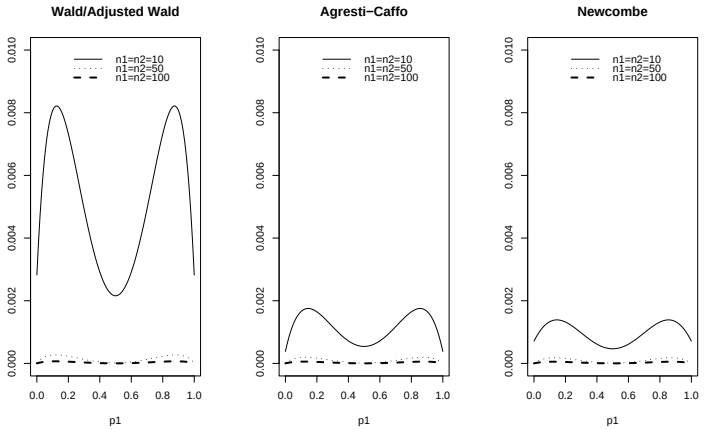

FIGURE 5: Variance of the length of the confidence intervals varying p_1 with $n_1 = n_2 = 10, 50, 100$.

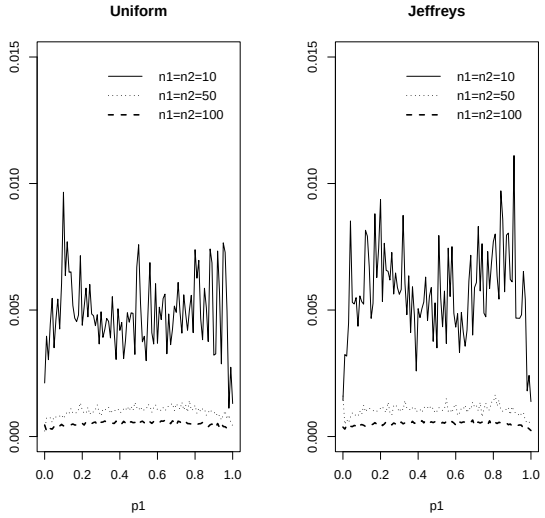

FIGURE 6: Variance of the length of the Bayesian intervals varying p_1 and $n_1 = n_2 = 10, 50, 100$.

tis very close to 0.95. Although the adjusted Wald interval has a larger coverage probability than the Newcombe interval, we will see later that this interval also has a larger expected length.

The results for the Bayesian intervals are those presented in Figure 9, where it is seen that for both intervals, the coverage probability is close to the nominal probability 0.95, and is not affected by different values of p_1; however, it is smaller than the adjusted Wald and Newcombe interval. In conclusion, the best intervals

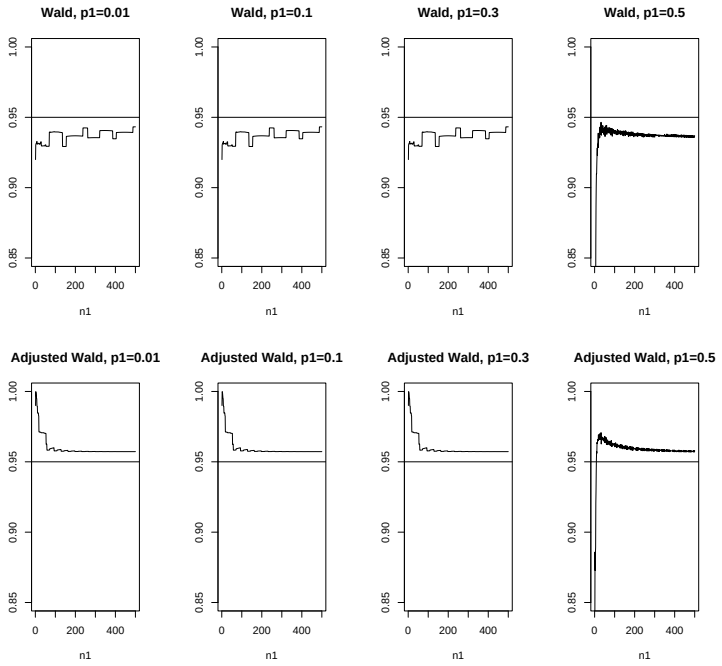

FIGURE 7: True coverage probability of the Wald and Adjusted Wald intervals varying n_1 and p_1 with a nominal coverage probability of 0.95.

in terms of the true coverage probability, are the adjusted Wald and Newcombe intervals.

We compare the intervals in terms of the expected length for different values of p_1. In Figure 10, the expected lengths of the Wald and adjusted Wald intervals are presented. Note that, as in the previous section, the expected length of the adjusted Wald interval is always larger. Thus we do not recommend this interval in spite of its large coverage probability. It is also noted that the lengths get smaller as the value of p_1 decreases and n_1 increases. The expected lengths of theAgresti-Caffo and Newcombe intervals are presented in Figure 11. It can be seen that their performance are very similar, although the length of the Newcombe interval is slightly smaller. In addition it is seen that their lengths are similar to the length of the Wald interval.

In Figure 12, the expected lengths of the Bayesian intervals are presented. It is seen that their performances are almost the same as the Agresti-Caffo and Newcombe intervals. In conclusion, except for the adjusted Wald interval, the performance of the other intervals in terms of the expected length is very similar.

We also compare the intervals considering the variance of the length. The performance of Wald and adjusted Wald intervals is presented in Figure 13. It is seen that for large sample sizes, the variance is almost zero. The variance of the Agresti-Caffo and Newcombe intervals is presented in Figure 14, where it is seen

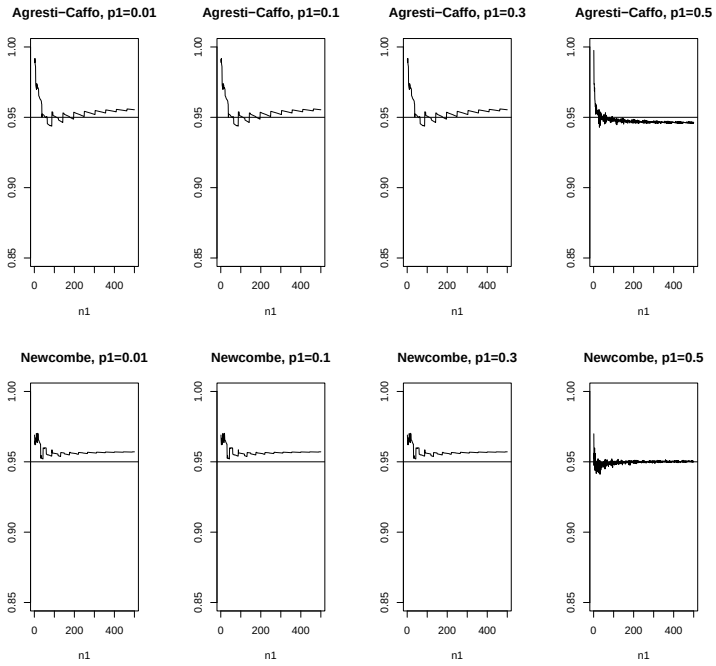

FIGURE 8: True coverage probability of the Agresti-Caffo and Newcombe intervals varying n_1 and p_1 with a nominal coverage probability of 0.95.

that when the sample size n_1 is small, the Newcombe interval always has a smaller variance than the Agresti-Caffo interval; while the difference is negligible when n_1 is large. At any rate, the variance of the Agresti-Caffo and Newcombe intervals is smaller than the Wald and adjusted Wald intervals.

In Figure 15, the variances for the Bayesian intervals are presented. Notice that there is no significant difference between the uniform and Jeffreys prior. However, their variances are smaller than the Wald and adjusted Wald intervals and larger than the Agresti-Caffo and Newcombe intervals. In conclusion, the interval with the smallest variance in length is the Newcombe interval.

4.3. Performance of intervals by varying values of p1 - p2

Since the parameter of interest is the difference between the proportions $p = p_1 - p_2$, it is natural to check the performance of the intervals when this parameter changes. Therefore, we calculate the true coverage probability of the intervals in the case that $p_1 - p_2$=0,0.1,0.5,0.8, the value of n_2 is fixed to be 30, and n_1 takes values $1, 2, \ldots, 500$.

The performance of the Wald and adjusted Wald intervals are presented in Figure 16, where we see that when the difference between p_1 and p_2 is large, the coverage probability of the Wald interval is really small. Further more, in

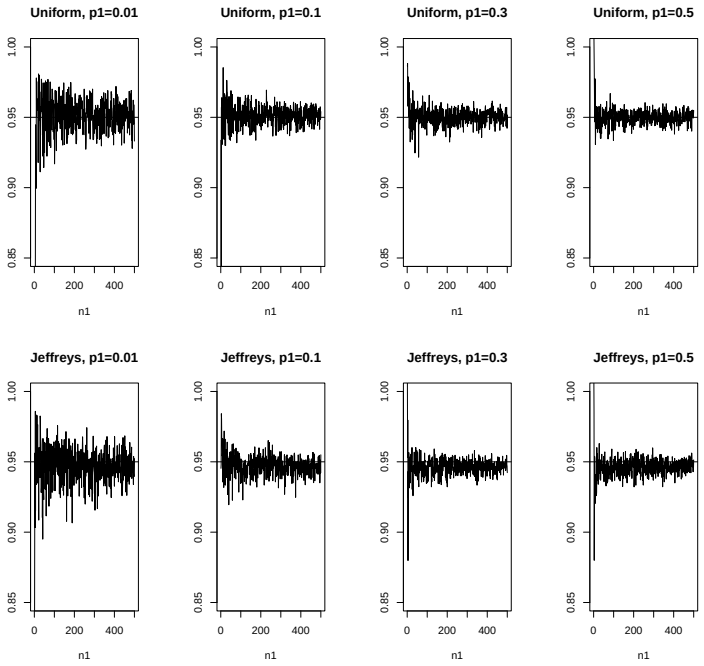

FIGURE 9: True coverage probability of the Bayesian intervals varying n_1 and p_1 with a nominal coverage probability of 0.95.

previous sections, the adjusted Wald always has larger coverage probability than the nominal 0.95, but in the case that $p_1 - p_2 = 0.8$, its coverage probability decreases considerably.

The coverage probabilities of the Agresti-Caffo and Newcombe intervals are presented in Figure 17, where we note that, contrary to the Wald and adjusted Wald intervals, the Agresti-Caffo and Newcombe intervals have larger coverage probability when $p_1 - p_2$ takes larger values. Regarding the Bayesian intervals, whose coverage probabilities are presented in Figure 18, we note that their performance is not affected by the values of $p_1 - p_2$, and that this is an advantage over the confidence intervals.

5. Other intervals

There are many other confidence intervals in statistical literature. Some of them will be briefly presented. Pan (2002) modified the Agresti-Caffo interval using the t distribution instead of the normal distribution to take of the uncertainty in estimating the variance of the observed pseudo proportion into account. It was found that in some situations the proposed method can have a higher coverage probability than the Agresti-Caffo interval. However, the price payed for the Pan interval is the resulting wider length of the intervals. The limits of this interval

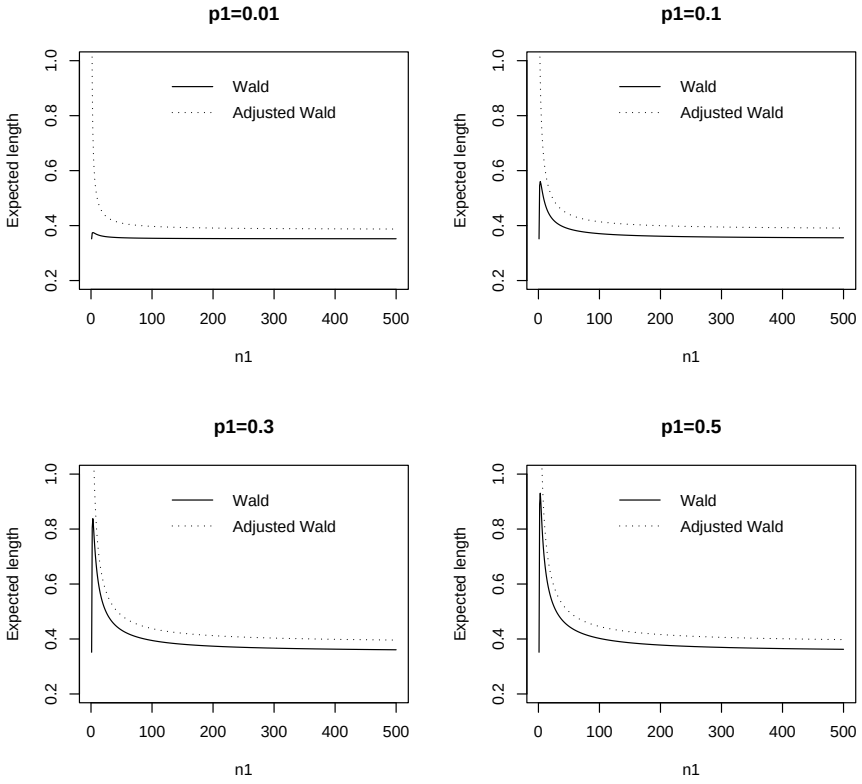

FIGURE 10: Expected length of the Wald and adjusted Wald intervals varying n_1 and p_1.

are:

$$L_{low} = \widetilde{p_1} - \widetilde{p_2} - t_{d,1-\alpha/2}\sqrt{V(\widetilde{p}_1, \widetilde{n}_1) + V(\widetilde{p}_2 + \widetilde{n}_2)} \tag{16}$$

and

$$L_{upp} = \widehat{p_1} - \widehat{p_2} + t_{d,1-\alpha/2}\sqrt{V(\widetilde{p}_1, \widetilde{n}_1) + V(\widetilde{p}_2 + \widetilde{n}_2)} \tag{17}$$

where

$$d \approx \frac{2[V(\widetilde{p}_1, \widetilde{n}_1) + V(\widetilde{p}_2 + \widetilde{n}_2)]}{\Omega(\widetilde{p}_1, \widetilde{n}_1) + \Omega(\widetilde{p}_2 + \widetilde{n}_2)}$$

and

$$\Omega(\widetilde{p}_i, \widetilde{n}_i) = \frac{\widetilde{p}_i - \widetilde{p}_i^2}{\widetilde{n}_i^3} + \widetilde{p}_i + (6\widetilde{n}_i - 7)\widetilde{p}_i^2 + 4(\widetilde{n}_i - 1)(\widetilde{n}_i - 3)\widetilde{p}_i^2 - 2(\widetilde{n}_i - 1)\frac{(2\widetilde{n}_i - 3)\widetilde{p}_i^3}{\widetilde{n}_i^5} - \frac{2\widetilde{p}_i + (2\widetilde{p}_i - 3)\widetilde{p}_i^2 - 2(\widetilde{n}_i - 1)\widetilde{p}_i^3}{\widetilde{n}_i^4}$$

where $\widetilde{p}_i$ and $\widetilde{n}_i$ are similarly defined as in the Agresti-Caffo interval for $i = 1, 2$.

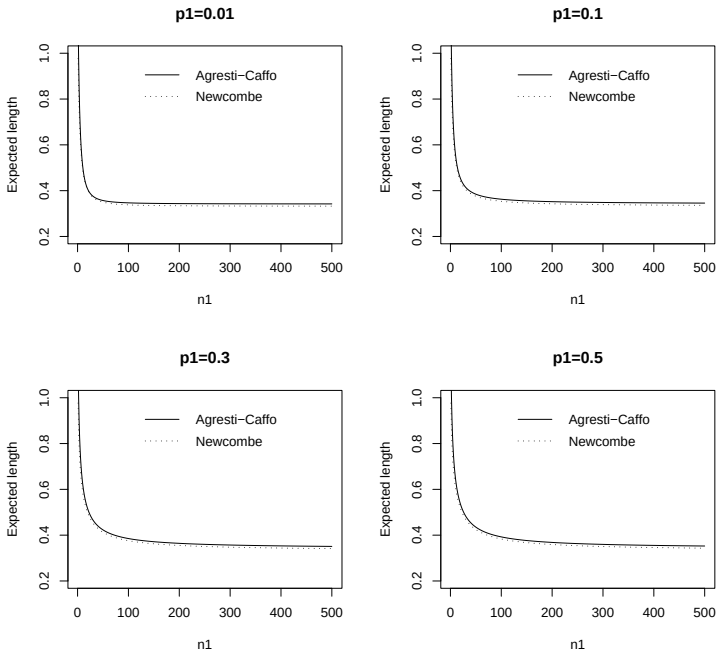

FIGURE 11: Expected length of the Agresti-Caffo and Newcombe intervals varying n_1 and p_1.

Miettinen & Nurminen (1985) proposed an asymptotic method based on the score test statistic, where the following system is considered:

$$H_0 : p_1 - p_2 = p^* \quad versus \quad H_1 : p_1 - p_2 \neq p^*$$

the score test statistic for testing this system is given by

$$S = \frac{\widehat{p}_1 - \widehat{p}_2 - p^*}{\sqrt{\widetilde{p}_1(1-\widetilde{p}_1)/n_1 + \widetilde{p}_2(1-\widetilde{p}_2)/n_2}} \tag{18}$$

where $\widetilde{p}_1$ and $\widetilde{p}_2$ are the maximum likelihood estimates of p_1 and p_2, respectively, under the restriction that $p_1 - p_2 = p^*$. The limits of the score interval L_{low} and L_{upp} are defined to satisfy:

$$\begin{aligned} 1 - \Phi\left(\frac{\widehat{p}_1 - \widehat{p}_2 - L_{low}}{\sqrt{\widetilde{p}_1(1-\widetilde{p}_1)/n_1 + \widetilde{p}_2(1-\widetilde{p}_2)/n_2}}\right) &= \Phi\left(\frac{\widehat{p}_1 - \widehat{p}_2 - L_{upp}}{\sqrt{\widetilde{p}_1(1-\widetilde{p}_1)/n_1 + \widetilde{p}_2(1-\widetilde{p}_2)/n_2}}\right) \\ &= \frac{\alpha}{2} \end{aligned} \tag{19}$$

and the solution of L_{low} and L_{upp} must be found using numerical methods.

In addition, there is the Clopper-Pearson interval, which is strongly associated with the Clopper-Pearson test. However, many authors have criticized this interval for being too conservative in its coverage probability (Vos & Hudson 2008).

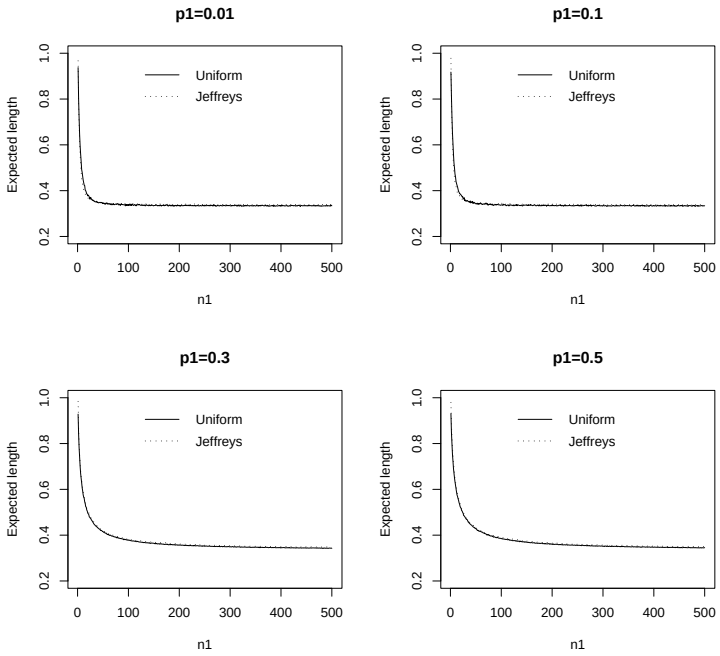

FIGURE 12: Expected length of the Bayesian intervals varying n_1 and p_1.

Another well-known interval is the Blaker interval. This interval has a smaller length than the Clopper-Pearson interval, i.e. it is always contained within the Clopper-Pearson intervals (Blaker 2000).

As we mentioned in Section 2, the exact Bayesian interval for $p_1 - p_2$ can be obtained using the exact posterior distribution. Pham-Gia & Turkkan (1993) established that when the prior distribution for p_i is $Beta(a_i, b_i)$ for $i = 1, 2$, the posterior distribution for $p = p_1 - p_2$ is given by

$$p(p \mid \mathbf{x}, \mathbf{y}) = \begin{cases} \dfrac{1}{k} B(\alpha_2, \beta_1) p^{\beta_1+\beta_2-1}(1-p)^{\alpha_2+\beta_1-1} \\ \qquad F_1(\beta_1, \alpha_1+\beta_1+\alpha_2+\beta_2-2, 1-\alpha_1, \beta_1+\alpha_2, 1-p1-p^2) \\ \qquad\qquad \text{for } \ 0 < p \leq 1 \\ \dfrac{1}{k} B(\alpha_1+\alpha_2-1, \beta_1+\beta_2-1) \qquad \text{for } \ p = 0 \\ \dfrac{1}{k} B(\alpha_1, \beta_2)(-p)^{\beta_1+\beta_2-1}(1+p)^{\alpha_1+\beta_2-1} \\ \qquad F_1(\beta_2, 1-\alpha_1, \alpha_1+\beta_1+\alpha_2+\beta_2-2, \beta_2+\alpha_1, 1-p^2, 1+p) \\ \qquad\qquad \text{for } \ -1 \leq p < 0 \end{cases} \tag{20}$$

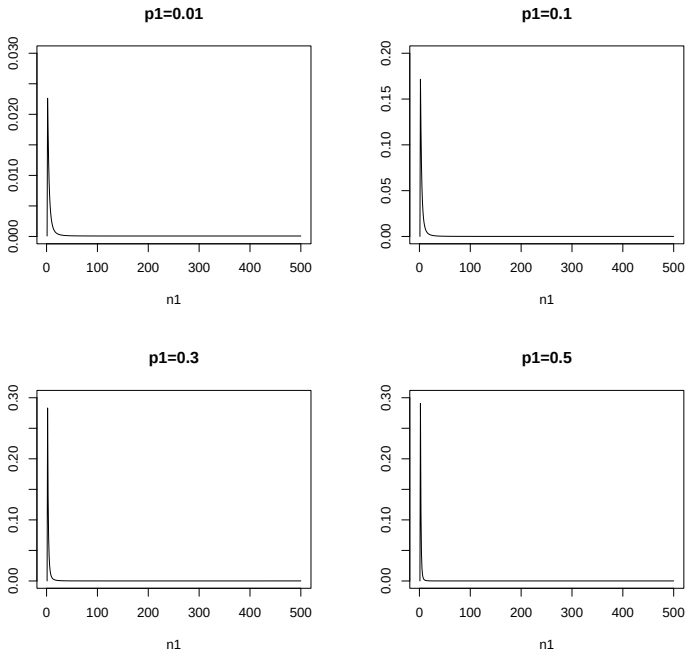

FIGURE 13: Variance of the length of the Wald and Adjusted Wald intervals varying n_1 and p_1.

where $\mathbf{x} = (X_1, \ldots, X_{n_1})$ and $\mathbf{y} = (Y_1, \ldots, Y_{n_2})$.

$k = B(a_1, b_1)B(a_2, b_2)$, with $B(a, b)$ the beta function evaluated in a y b, that is,

$$B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt \tag{21}$$

And $F_1(\varphi, \eta_1, \eta_2, \psi, w_1, w_2)$ is the fourth hypergeometric Appell's function, given by

$$\frac{\Gamma(\psi)}{\Gamma(\varphi)\Gamma(\psi-\varphi)} \int_0^1 u^{\varphi-1}(1-u)^{\psi-\varphi-1}(1-uw_1)^{-\eta_1}(1-uw_2)^{-\eta_2}du \tag{22}$$

when the real part of φ y $\psi - \varphi$ are all positive, for more details, see Bailey (1934).

Given the exact posterior distribution of $p = p_1 - p_2$, a Bayesian interval is defined by the lower limit l and upper limit u such that:

$$Pr(l \leq p \leq u \mid \mathbf{x}, \mathbf{y}) = 1 - \frac{\alpha}{2}$$

l and u are chosen to satisfy $Pr(p < l \mid \mathbf{x}, \mathbf{y}) = Pr(p > u \mid \mathbf{x}, \mathbf{y}) = \alpha/2$. Pham-Gia & Turkkan (1993) considered a numerical example where the prior distribution of p_1 and p_2 are $Beta(3, 5)$ and $Beta(2, 8)$, respectively, and the sampling results are $n_1 = 10$, $s_x = 4$, $n_2 = 6$ and $s_y = 2$. The resulting posterior distribution of $p_1 - p_2$

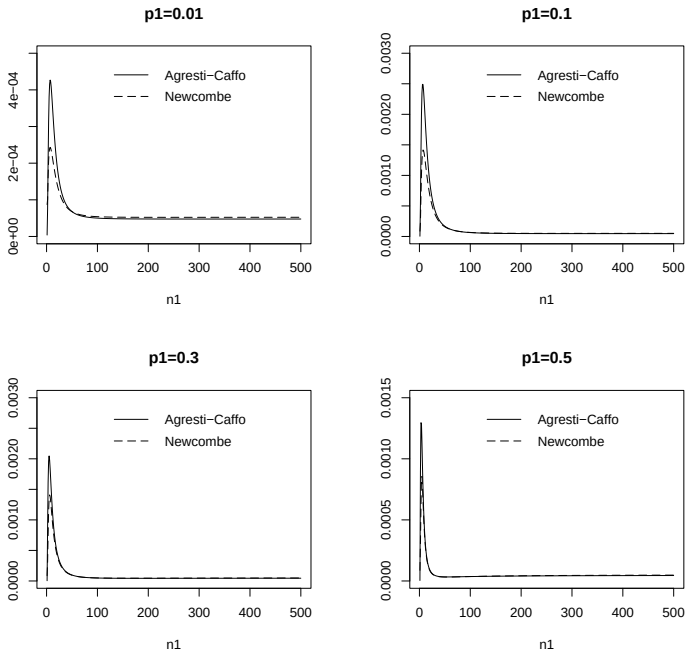

FIGURE 14: Variance of the length of the Agresti-Caffo and Newcombe intervals varying n_1 and p_1.

is bell-shaped, symmetric at the value 0.17, and an exact 90% credibility interval is $(-0.11, 0.39)$.

6. Conclusions

As a first conclusion, we point out that the performance of the Bayesian intervals is not greatly affected by the sample sizes nor by different values of p_1, p_2 or $p_1 - p_2$. In terms of true coverage probability, the best interval is the Bayesian interval, since its coverage probability is always close to the nominal coverage probability and is always stable with respect to different samples sizes. They are followed by the Newcombe and Agresti-Caffo intervals. We discard the use of adjusted Wald interval since its large coverage probability is obtained at the expense of a large length. The Wald interval performs poorly although this poor performance in small samples is a result that is well-known empirically and theoretically (Cepeda 2008). In terms of expected length, the best interval is the Newcombe interval followed by the Agresti-Caffo interval, Bayesian intervals, and the Wald interval. The adjusted Wald interval always has largest length. In terms of the variance of length, the best interval is again the Newcombe interval, followed by the Agresti-Caffo interval, the Wald and adjusted Wald intervals. The intervals with the largest length variance are the Bayesian intervals, there fore the New-

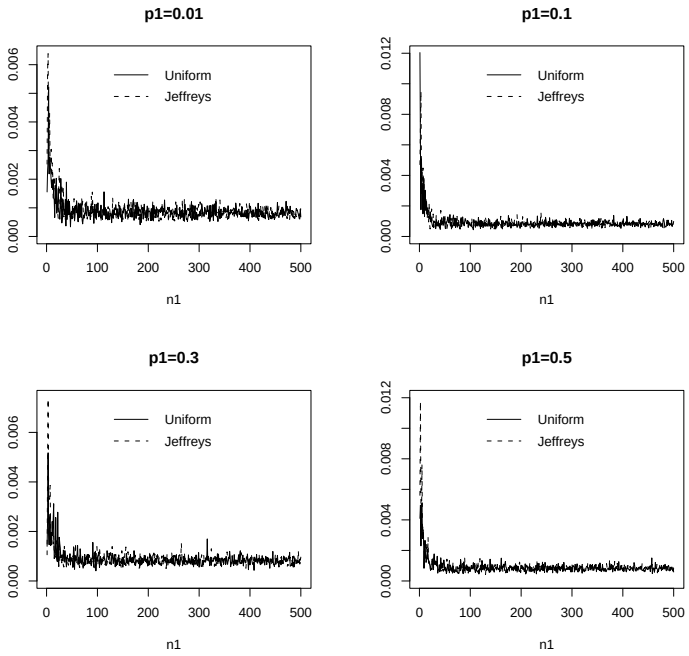

FIGURE 15: Variance of the length of the Bayesian intervals varying n_1 and p_1.

combe interval is strongly recommend. The Wald and adjusted Wald intervals are not recommended.

Acknowledgements

The authors are very grateful to professor Turkkan who kindly answered our inquires and to the anonymous referees for valuable suggestions.

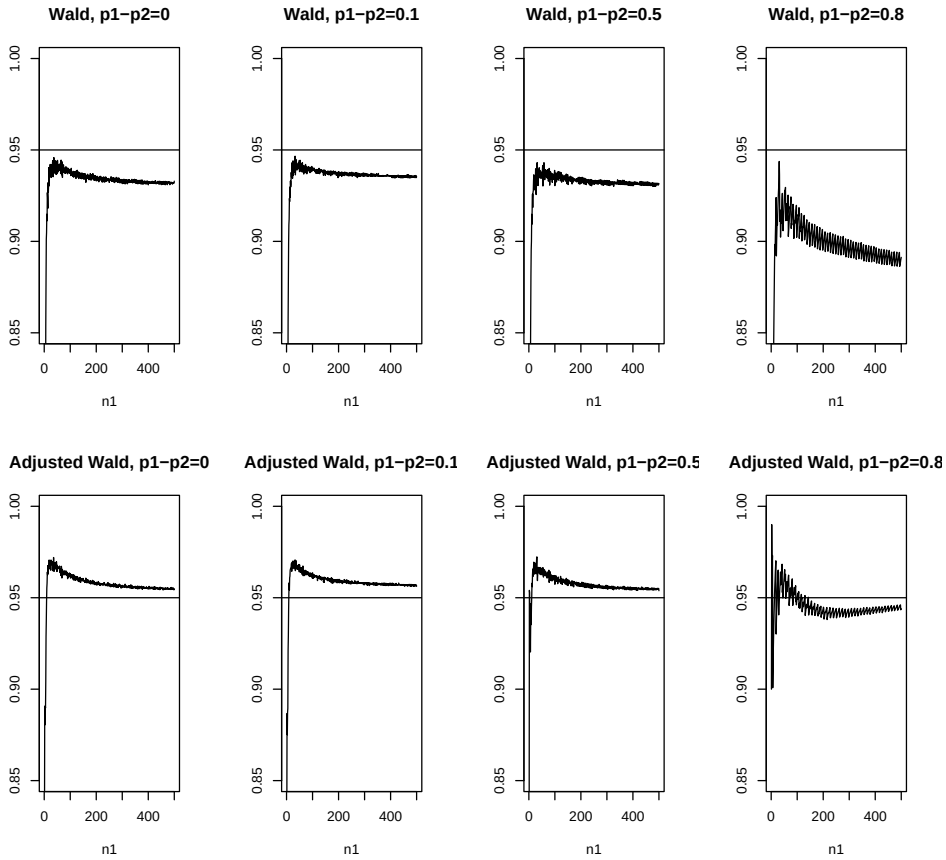

FIGURE 16: True coverage probability of the Wald and Adjusted Wald intervals varying n_1 and $p_1 - p_2$=0,0.1,0.5,0.8 with a nominal coverage probability of 0.95.

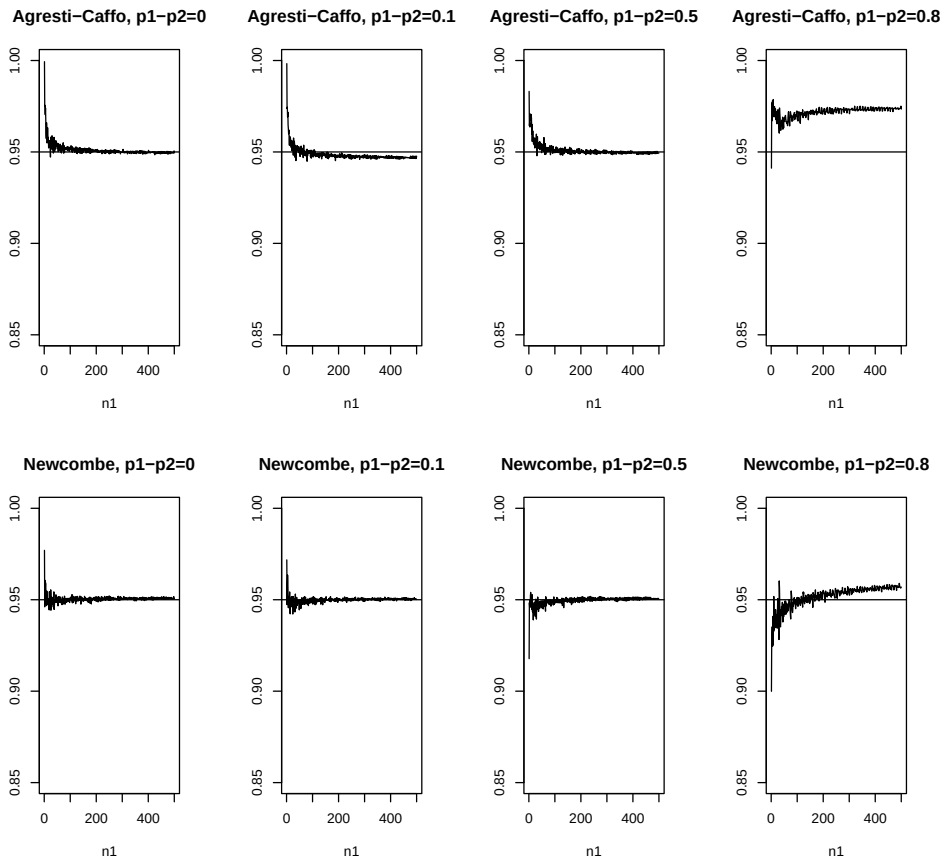

FIGURE 17: True coverage probability of the Agresti-Caffo and Newcombe intervals varying n_1 and $p_1 - p_2$=0,0.1,0.5,0.8 with a nominal coverage probability of 0.95.

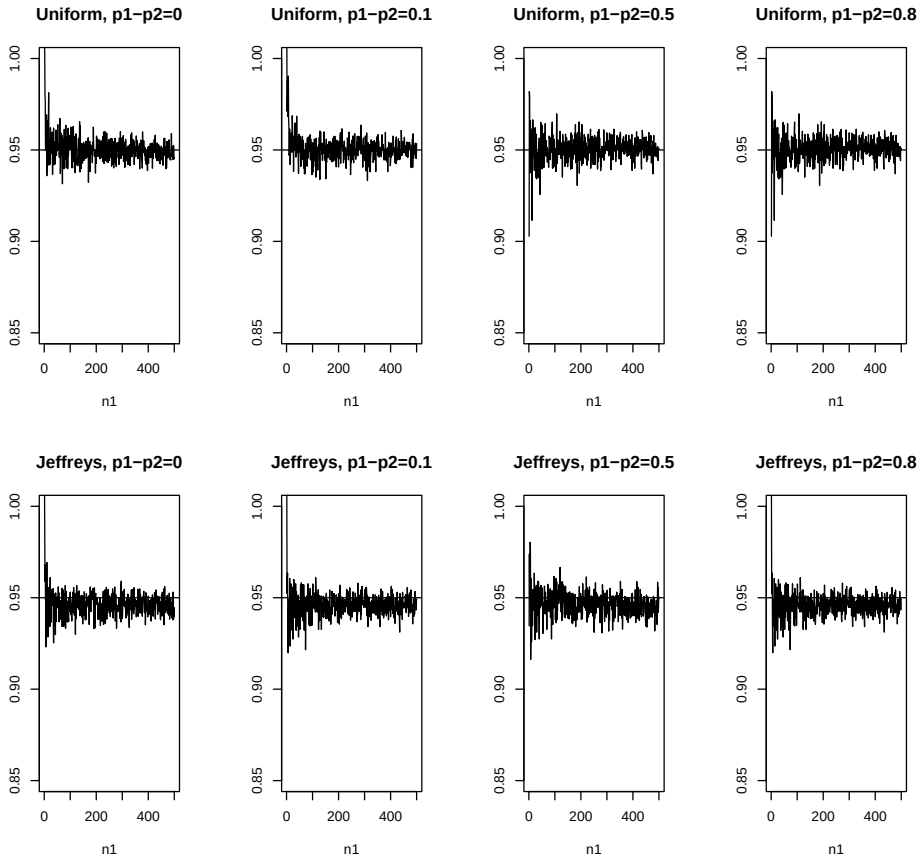

FIGURE 18: True coverage probability of the Bayesian intervals varying n_1 and $p_1 - p_2$=0,0.1,0.5,0.8 with a nominal coverage probability of 0.95.

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

References

Agresti, A., Bini, M., Bertaccini, B. & Ryu, E. (2008), 'Simultaneous Confidence Intervals for Comparing Binomial Parameters', *Biometrics* **64**, 1270–1275.

Agresti, A. & Caffo, B. (2000), 'Simple and Effective Confidence Intervals for Proportions and Differences of Proportions', *American Statistician* **54**(4), 280–288.

Agresti, A. & Min, Y. (2005), 'Frequentist Performance of Bayesian Confidence Intervals for Comparing Proportions in 2×2 Contingency Tables', *Biometrics* **61**, 515–523.

Bailey, W. N. (1934), 'On the reducibility of Appell's Function F4', *The Quarterly Journal of Mathematics* **5**, 291–292.

Blaker, H. (2000), 'Confidence Curves and Improved Exact Confidence Intervals for Discrete Distributions', *The Canadian Journal of Statistics* **28**(4), 783–798.

Brown, L. D., Cai, T. T. & DasGupta, A. (2001), 'Interval estimation of a binomial proportion', *Statistical Science* **16**, 101–133.

Carlin, B. P. & Louis, T. A. (1998), *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall.

Cepeda, E. (2008), 'Intervalos de confianza e intervalos de credibilidad para una proporción', *Revista Colombiana de Estadística* **31**(2), 211–228.

Correa, J. C. & Sierra, E. (2003), 'Intervalos de confianza para la comparación de dos proporciones', *Revista Colombiana de Estadística* **26**(1), 61–75.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004), *Bayesian Data Analysis*, second edn, Chapman & Hall.

Ghosh, B. K. (1979), 'A Comparison of Some Approximate Confidence Intervals for the Binomial Parameter', *Journal of the American Statistical Association* **74**, 894–900.

Miettinen, O. S. & Nurminen, M. (1985), 'Comparative analysis of two rates', *Statistics in Medicine* **4**, 213–226.

Newcombe, R. (1998*a*), 'Interval Estimation for the Difference between Independent Proportions: Comparison of Eleven Methods', *Statistics in Medicine* **17**, 873–890.

Newcombe, R. (1998*b*), 'Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. Statistics in Medicine', *Statistics in Medicine* **17**, 857–872.

Pan, W. (2002), 'Approximate Confidence Intervals for One Proportion and Difference of Two Proportions', *Computational Statistics and Data Analysis* **40**, 143–157.

Pham-Gia, T. & Turkkan, N. (1993), 'Bayesian Analysis of the Difference of Two Proportions', *Communications in Statistics. Theory and Methods* **22**(6), 1755–1771.

Vollset, S. E. (1993), 'Confidence intervals for a binomial proportion', *Statistics in Medicine* **12**, 809–824.

Vos, P. W. & Hudson, S. (2008), 'Problems with Binomial Two-Sided Tests and the Associated Confidence Intervals', *Australian & New Zealand Journal of Statistics* **50**(1), 81–89.

Wilson, E. B. (1927), 'Probable Inference, the Law of Succession, and Statistical Inference', *Journal of the American Statistical Association* **22**, 209–212.

Revista Colombiana de Estadística
Junio 2010, volumen 33, no. 1, pp. 89 a 109

Estimación de las componentes de un modelo de coeficientes dinámicos mediante las ecuaciones de estimación generalizadas

Time-Varying Coefficient Model Component Estimation Through Generalized Estimation Equations

Juan Camilo Sosa[a], Luis Guillermo Díaz[b]

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

Resumen

Se propone una metodología para estimar las componentes de un modelo de coeficientes dinámicos mediante las ecuaciones de estimación generalizadas (Liang & Zeger 1986), con el propósito de incluir directamente en la estimación la posible correlación de las medidas repetidas de cada individuo. La expansión de los coeficientes dinámicos del modelo se lleva a cabo a través de regresión *spline* (Huang et al. 2002). También se propone utilizar el *criterio de información de Akaike en las ecuaciones de estimación generalizadas* (QIC) (Pan 2001) como selector de modelos. Mediante simulación se compara la metodología propuesta con la metodología presentada por Wu & Zhang (2006), donde se estiman las componentes del modelo mediante mínimos cuadrados ponderados y se utiliza el *criterio de información de Akaike* (AIC) como selector de modelos, obteniéndose que en los escenarios simulados la metodología propuesta presenta mejores resultados con relación al error cuadrático medio promedio. Para ilustrar la estrategia de estimación propuesta, se considera el conjunto de datos ACTG 315 (Liang et al. 2003) asociado con un estudio de sida, en el que se modela dinámicamente la relación entre la carga viral y el conteo de células CD4+.

***Palabras clave*:** criterio de información de Akaike, ecuaciones de estimación generalizadas, mínimos cuadrados ponderados, modelo de coeficientes dinámicos, regresión *spline*.

Abstract

[a]Estudiante de doctorado. E-mail: jcsosam@unal.edu.co
[b]Profesor asociado. E-mail: lgdiazm@unal.edu.co

A methodology to estimate time-varying coefficient model's components through generalized estimation equations (Liang & Zeger 1986) is proposed, in order to include directly in the estimation the possible correlation between repeated measurements of each subject. Expansion of the time-varying coefficients is done by means of regression *spline* methods (Huang et al. 2002). Furthermore, is proposed the use of the *Akaike's information criterion in generalized estimating equations* (QIC) proposed by Pan (2001) like model selector. Through simulation are compared the proposed methodology and the methodology presented by Wu & Zhang (2006), where model's components are estimated through weighted least squares and *Akaike's information criterion* (AIC) is used like model selector. It resulted that the proposed methodology gives a better behavior in relation with the average mean square error. In order to illustrate the methodology, is taken into account the data base ACTG 315 (Liang et al. 2003) related to a AIDS study, where it is investigated the relationship between the viral charge and the CD4+ cell count.

***Key words*:** AIC, Generalized estimation equations, Regression spline, Longitudinal data, Time-varying coefficient model, Weighted least squares.

1. Introducción

El análisis de datos longitudinales surge cuando un conjunto de n individuos es observado reiteradamente a través del tiempo, registrando los valores de la respuesta de interés junto con las respectivas covariables que pueden depender o no del instante en que sean medidas. Debido a la naturaleza misma de este tipo de datos, una característica fundamental que los distingue, y que debe tenerse en cuenta en el modelamiento, es la posible correlación dada entre las mediciones repetidas de la variable respuesta en cada individuo, considerando las mediciones entre los individuos independientes. Esto es, las mediciones son correlacionadas dentro e independientes entre individuos. Así, se quiere identificar la evolución de la variable respuesta y determinar cómo es afectada por las covariables. Por ejemplo, en estudios de medicina, interesa evaluar el efecto de la dosis de un medicamento u otros factores asociados, como el género del paciente, sobre el progreso de alguna enfermedad en el tiempo.

Las técnicas paramétricas para el análisis de datos longitudinales han sido estudiadas extensivamente en la literatura (Diggle et al. 1994, Davis 2002, Verbeke & Molenberghs 2000, Fitzmaurice et al. 2009). Aunque el enfoque paramétrico es útil, siempre surgirán dudas sobre la adecuación de los supuestos del modelo y el impacto potencial de la falta de especificaciones del modelo sobre el análisis (Hoover et al. 1998).

Las técnicas no paramétricas han sido introducidas recientemente en el análisis de datos longitudinales ya que permiten una dependencia funcional más flexible de la variable respuesta sobre las covariables. Hart & Wehrly (1986), Altman (1990) y Hart (1991) desarrollaron métodos utilizando funciones *kernel* para la estimación de la esperanza de la variable respuesta, sin la presencia de covariables, y propusieron algunas técnicas de selección de parámetros de suavizamiento a través de

la validación cruzada. Estos métodos consideran solo el posible efecto del tiempo sobre la variable respuesta. Para tener en cuenta la influencia de covariables, Zeger & Diggle (1994) estudiaron un modelo semiparamétrico de la forma

$$y_{ij} = \mu(t_{ij}) + \boldsymbol{x}_{ij}^T(t_{ij})\boldsymbol{\beta} + e_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, n \tag{1}$$

donde n es el número de individuos en estudio, n_i es la cantidad de mediciones del i-ésimo individuo, y t_{ij}, $y_{ij} \equiv y_i(t_{ij})$,

$$\boldsymbol{x}_i(t_{ij}) = [x_{i0}(t_{ij}), x_{i1}(t_{ij}), \ldots, x_{id}(t_{ij})]^T$$

y $e_{ij} \equiv e_i(t_{ij})$ son respectivamente el instante, la variable respuesta de valor real, el vector de covariables en $\mathbb{R}^{d+1}$ y el error aleatorio, asociados con la j-ésima medición del i-ésimo individuo. Además $\mu(t)$ es una función de t, y $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_d]^T$ en $\mathbb{R}^{d+1}$ es un vector de parámetros. En el contexto de los datos longitudinales $e_i(t)$ es asumido como un proceso estocástico con media cero y función de covarianza $\rho_{e_i}(s,t) = Cov(e_i(s), e_i(t))$.

Hoover et al. (1998) proponen una extensión del modelo (1) donde los parámetros pueden variar con el tiempo. La extensión es de la forma

$$y_{ij} = \boldsymbol{x}_{ij}^T(t_{ij})\boldsymbol{\beta}(t_{ij}) + e_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, n \tag{2}$$

donde $\boldsymbol{\beta}(t) = [\beta_{i0}(t_{ij}), \beta_{i1}(t_{ij}), \ldots, \beta_{id}(t_{ij})]^T$ es un vector de funciones de t llamadas coeficientes dinámicos. Este modelo, denominado *modelo de coeficientes dinámicos* (MCD), ha sido estudiado por Wu & Zhang (2006) quienes implementan el método de mínimos cuadrados ponderados junto con la reparametrización de las componentes dinámicas del modelo mediante regresión *spline*. Una de las limitaciones de esta metodología es que no incorpora directamente en el estimador la posible correlación de las mediciones de cada individuo, característica primordial de los datos longitudinales que sí incorpora el estimador propuesto mediante una matriz de correlación de trabajo.

Este artículo está estructurado como sigue. En la sección 2 se revisa brevemente la regresión *spline* y la reparametrización de un MCD utilizando esta técnica. En la sección 3 se revisa la estimación de los parámetros asociados a través del método de mínimos cuadros ponderados y el AIC como selector de modelos. En la sección 4 se propone una alternativa de estimación de los coeficientes mediante las ecuaciones de estimación generalizadas (Liang & Zeger 1986) y el QIC (Pan 2001) como selector de modelos, metodología que no había sido considerada en la estimación de las componentes dinámicas de un MCD. En la sección 5 se muestra un estudio de simulación donde se comparan las metodologías consideradas con base en el error cuadrático medio promedio. En la sección 6 se considera el conjunto de datos ACTG 315 (Liang et al. 2003) asociado con un estudio de sida, en el que se modela dinámicamente la relación entre la carga viral y el conteo de células CD4+. En la sección 7 se discuten los resultados obtenidos y otras alternativas de estimación.

2. Regresión *spline*

En esta sección se presenta la estrategia de estimación de las componentes de un MCD utilizando *regresión spline* (RS). Primero se muestran los conceptos básicos involucrados en la RS y en seguida la estrategia de estimación.

2.1. Conceptos básicos

Polinomio a trozos. Una RS es considerada como un polinomio a trozos. Una función de valor real $f(\cdot)$ definida sobre el intervalo $[a, b]$, es un polinomio a trozos de orden k, o de grado $k-1$, $k \geq 1$, si es obtenida dividiendo el intervalo $[a, b]$ en subintervalos contiguos, de tal modo que se pueda representar la función $f(\cdot)$ mediante un polinomio de orden k en cada subintervalo. Cada subintervalo es de la forma

$$[\tau_l, \tau_{l+1}),\ l = 0, \ldots, K$$

donde

$$a = \tau_0 < \tau_1 < \ldots < \tau_K < \tau_{K+1} = b$$

y $K+1$ es el número de subintervalos. Los puntos τ_l, $l = 0, \ldots, K$, se llaman nodos interiores o simplemente nodos. Note que

$$[a, b) = \bigcup_{l=0}^{K} [\tau_l, \tau_{l+1}).$$

Regresión *spline*. Una RS de orden $k+1$, o de grado k, $k \geq 0$, con nodos interiores τ_l, $l = 0, \ldots, K$, es un polinomio a trozos de orden $k+1$, y tiene hasta $k-1$ derivadas continuas.

Ya que el espacio de funciones de RS es un espacio vectorial de dimensión finita, hay muchas bases para representar dichas funciones, entre otras, se encuentran: la base de polinomios, la base de potencias truncadas y la base de *B-splines*. Una revisión completa y detallada acerca de *B-splines* se encuentra en de Boor (1978).

Base de *B-splines*. Siguiendo a Hastie et al. (1990), antes de definir una *base de B-splines* (BBS), es necesario refinar la secuencia de nodos

$$a = \tau_0 < \tau_1 < \ldots < \tau_K < \tau_{K+1} = b$$

Una secuencia de nodos aumentada se define como una secuencia de nodos tal que

- $\xi_1 \leq \xi_2 \leq \ldots \leq \xi_M \leq \tau_0$;
- $\xi_{j+M} = \tau_j$ para $j = 1, \ldots, K$;
- $\tau_{K+1} \leq \xi_{K+M+1} \leq \xi_{K+M+2} \leq \ldots \leq \tau_{k+2M}$.

Los valores de los nodos adicionales más allá de la frontera del intervalo $[a, b]$ son arbitrarios, y usualmente están dados por

$$\xi_1 = \xi_2 = \ldots \xi_M = \tau_0 \quad \text{y} \quad \tau_{K+1} = \xi_{K+M+1} = \xi_{K+M+2} = \ldots = \tau_{k+2M}$$

Sea $B_{i,m}$ la i-ésima función *B-spline* de orden $k+1$, o grado k, $0 \leq k \leq M-1$, asociado a la secuencia de nodos

$$\{\xi_j : j = 1, \ldots, k + 2M\},$$

para $i = 1, \ldots, K + 2m - k - 1$. Los *B-splines* se definen recursivamente como sigue:

$$B_{i,1}(x) = \begin{cases} 1, & \text{si } x \in [\xi_i, \xi_{i+1}] \\ 0, & \text{si } x \notin [\xi_i, \xi_{i+1}] \end{cases} \tag{3}$$

para $i = 1, \ldots, K + 2M - 1$. Y para $k \geq 1$ se define

$$B_{i,k+1}(x) = \frac{x - \xi_i}{\xi_{i+k} - \xi_i} B_{i,k}(x) + \frac{\xi_{i+k+1} - x}{\xi_{i+k+1} - \xi_{i+1}} B_{i+1,k}(x) \tag{4}$$

para $i = 1, \ldots, K + 2M - k - 1$.

Como puede haber algunos nodos duplicados, se debe tener cuidado con las posibles divisiones por cero en (3) y (4). Por lo tanto se adopta la convención $B_{i,1} = 0$ si $\xi_i = \xi_{i+1}$, o si $\xi_{i+1} = \xi_{i+k+1}$.

Así, una RS de orden $m + 1$, o de grado m, $m \geq 0$, con nodos interiores τ_r, $r = 1, \ldots, K$, puede expresarse usando el siguiente conjunto de *splines* base:

$$\mathcal{B} = \{B_{i,m+1} : i = 1, \ldots, K + 2M - m - 1\} \tag{5}$$

El conjunto (5) de $K + 2M - m - 1$ funciones base se conoce como BBS de orden $m+1$, o grado m, con nodos ξ_j, $j = K + 2M - m - 1$. Usando las funciones de este conjunto, es posible expresar una RS de orden $k + 1$ como

$$f(t) = \sum_{i=1}^{G} \beta_i B_{i,m+1}(t) \tag{6}$$

donde $\beta_1, \ldots, \beta_G$ son los escalares asociados y $G = K + 2M - m - 1$. Una revisión completa y detallada acerca de *B-splines* se encuentra en de Boor (1978).

2.2. Estimación utilizando regresión *spline*

La idea fundamental de la estrategia de estimación a través de RS es expresar cada componente de $\boldsymbol{\beta}(t)$ como una RS, escribiendo cada coeficiente dinámico como una combinación lineal de funciones base, de BBS por ejemplo (Huggins & Loesch 1998, Huang et al. 2002).

La idea básica consiste en expresar cada $\beta_r(t)$, $r = 0, 1, \ldots, d$, como

$$\beta_r(t) = \phi_{r1}(t)\alpha_{r1} + \cdots + \phi_{rp_r}(t)\alpha_{rp_r} = \boldsymbol{\Phi}_{rp_r}(t)^T \boldsymbol{\alpha}_r \tag{7}$$

donde $\boldsymbol{\Phi}_{rp_r}(t) = [\phi_{r1}(t), \ldots, \phi_{rp_r}(t)]^T$ es un vector de $p_r \times 1$ *splines* base, que pueden ser elementos de (5), y $\boldsymbol{\alpha}_r = [\alpha_{r1}, \ldots, \alpha_{rp_r}]^T$ es el vector de $p_r \times 1$ coeficientes asociados.

Remplazando en el MCD (2) cada $\beta_r(t)$ por su expresión equivalente empleando funciones base y acomodando los vectores de una forma apropiada se tiene que:

$$
\begin{aligned}
y_{ij} &= [x_{i0}(t_{ij}), \ldots, x_{id}(t_{ij})] \begin{bmatrix} \boldsymbol{\Phi}_{0p_0}(t_{ij})^T \boldsymbol{\alpha}_0 \\ \vdots \\ \boldsymbol{\Phi}_{dp_d}(t_{ij})^T \boldsymbol{\alpha}_d \end{bmatrix} + e_{ij} \\
\Rightarrow y_{ij} &= [x_{i0}(t_{ij})\boldsymbol{h}_{0ij}^T, \ldots, x_{id}(t_{ij})\boldsymbol{h}_{dij}^T] \begin{bmatrix} \boldsymbol{\alpha}_0 \\ \vdots \\ \boldsymbol{\alpha}_d \end{bmatrix} + e_{ij} \\
\Rightarrow y_{ij} &= [\boldsymbol{x}_{0ij}^T, \ldots, \boldsymbol{x}_{dij}^T] \begin{bmatrix} \boldsymbol{\alpha}_0 \\ \vdots \\ \boldsymbol{\alpha}_d \end{bmatrix} + e_{ij} \\
\Rightarrow y_{ij} &= \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} + e_{ij}
\end{aligned}
$$

para cada $j = 1, \ldots, n_i$ y cada $i = 1, \ldots, n$, donde

$$\boldsymbol{h}_{rij} = \boldsymbol{\Phi}_{rp_r}(t_{ij}),\ \boldsymbol{x}_{ij} = [\boldsymbol{x}_{0ij}^T, \ldots, \boldsymbol{x}_{dij}^T]^T,\ \boldsymbol{x}_{rij} = x_{ri}(t_{ij})\boldsymbol{h}_{rij}$$

para cada $r = 0, 1, \ldots, d$, con $x_{ri}(t)$ la r-ésima componente de $\boldsymbol{x}_i(t)$, y $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_0^T, \ldots, \boldsymbol{\alpha}_d^T]^T$

Note que si el vector de coeficientes $\boldsymbol{\alpha}$ de $p \times 1$, con $p = \sum_{r=1}^{d} p_r$, está completamente especificado en el modelo equivalente

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} + e_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, n \tag{8}$$

entonces cada uno de los parámetros dinámicos del modelo (2) también está determinado. Así, el objetivo es estimar el vector $\boldsymbol{\alpha}$ a través de algún método apropiado.

El modelo (8) escrito en forma vectorial es

$$\boldsymbol{y}_i = \mathbf{X}_i^T \boldsymbol{\alpha} + \boldsymbol{e}_i,\ i = 1, \ldots, n \tag{9}$$

donde

$$\boldsymbol{y}_i = [y_{i1}, \ldots, y_{in_i}]^T$$

es el vector de medidas repetidas,

$$\boldsymbol{e}_i = [e_{i1}, \ldots, e_{in_i}]^T$$

es el vector de errores y

$$\mathbf{X}_i = [\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}]^T$$

es la matriz de diseño, asociados con el i-ésimo individuo. Además, el modelo (9) puede escribirse en forma matricial como

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{e}, \tag{10}$$

donde $\boldsymbol{y} = [\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T]^T$, $\mathbf{X} = [\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T]^T$ y $\boldsymbol{e} = [\boldsymbol{e}_1^T, \ldots, \boldsymbol{e}_n^T]^T$.

En este punto, debido a la expresión de las componentes dinámicas del modelo (2) a través de *B-splines*, este es equivalente a un modelo de regresión lineal en el que se requiere estimar el vector de parámetros $\boldsymbol{\alpha}$.

Una vez obtenida una estimación de los parámetros del modelo (10), denotada por $\widehat{\boldsymbol{\alpha}} = [\widehat{\boldsymbol{\alpha}}_0^T, \ldots, \widehat{\boldsymbol{\alpha}}_d^T]^T$, la estimación de los coeficientes $\beta_r(t)$ está dada por

$$\widehat{\beta}_r(t) = \boldsymbol{\Phi}_{rp_r}(t)^T \widehat{\boldsymbol{\alpha}}_r, \ r = 0, 1, \ldots, d \tag{11}$$

donde $\boldsymbol{\Phi}_{rp_r}(t)$ es el vector de *splines* base asociados con el r-ésimo parámetro dinámico.

La elección de los parámetros de suavizamiento está relacionada directamente con el conjunto de *splines* base con que se expresen los coeficientes dinámicos del MCD (2). Por ejemplo, trabajando con BBS, están dados por

$$p_r = K_r - k_r - 1, \ r = 0, 1, \ldots, d \tag{12}$$

donde K_r es el número de nodos y k_r es el grado de la base asociada con la estimación del r-ésimo coeficiente dinámico. Entonces, seleccionar el valor de los parámetros de suavizamiento es equivalente a seleccionar los K_r y los k_r. Para ello, es costumbre fijar el grado de la base k_r, por ejemplo 1, 2 ó 3, es decir, lineal, cuadrático o cúbico, y en seguida elegir mediante un criterio adecuado el número de nodos K_r para determinar el valor del parámetro de suavizamiento p_r.

Cuando se usa una BBS para expresar los parámetros dinámicos del MCD (2), una vez fijado el número de nodos K_r de cada base a través de algún criterio apropiado, se deben ubicar en el rango de interés, que en el caso de los datos longitudinales es sobre el conjunto de los tiempos de medición

$$\{t_{ij} : i = 1, \ldots, n; j = 1, \ldots, n_i\}$$

En cuanto a la ubicación de los nodos se distinguen dos métodos:

Método 1. Consiste en ubicar los K_r nodos igualmente espaciados en el intervalo de interés $[a, b]$. En el caso de los datos longitudinales, los valores que definen este intervalo son el mínimo y el máximo de todos los tiempos, esto es

$$a = \text{mín}\{t_{ij} : i = 1, \ldots, n; j = 1, \ldots, n_i\}$$
$$b = \text{máx}\{t_{ij} : i = 1, \ldots, n; j = 1, \ldots, n_i\}$$

Método 2. Para ubicar los K_r nodos, según este método, se deben considerar los M tiempos diferentes

$$t_1 < t_2 < \ldots < t_M \tag{13}$$

El método consiste en ubicar los nodos igualmente espaciados sobre los cuantiles de los tiempos (13).

3. Estimación a través del método de mínimos cuadrados ponderados

Una alternativa de estimación de los parámetros del modelo (10) es a través del método de *mínimos cuadrados ponderados* (MCP). Este método consiste en estimar $\boldsymbol{\alpha}$ minimizando la suma de cuadrados

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_i(y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\alpha})^2 \tag{14}$$

donde los w_i son pesos dados usando alguno de los siguientes esquemas de ponderación:

Esquema 1. En el que los pesos están dados por $w_i = 1/N$, $i = 1, \ldots, n$, donde $N = \sum_{i=1}^{n} n_i$.

Esquema 2. En el que los pesos están dados por $w_i = 1/(nn_i)$, $i = 1, \ldots, n$.

El esquema 1 usa el mismo peso para todos los individuos y fue implementado por Hoover et al. (1998). El esquema 2 es considerado por Huang et al. (2002) y usa diferentes pesos para los individuos. Huang et al. (2002) demuestran que el esquema 1 puede llevar a estimaciones inconsistentes de $\boldsymbol{\alpha}$.

Minimizando la función objetivo (14) trabajando con el modelo (10), se obtiene un estimador de $\boldsymbol{\alpha}$ dado por

$$\widehat{\boldsymbol{\alpha}}_{MCP} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{y} \tag{15}$$

donde $\mathbf{W} = diag[\mathbf{W}_1, \ldots, \mathbf{W}_n]$, con $\mathbf{W}_i = w_i\mathbf{I}_{n_i}$ la matriz de pesos del i-ésimo individuo, $i = 1, \ldots, n$, e $\mathbf{I}_{n_i}$ la matriz identidad de $n_i \times n_i$.

Criterio de información de Akaike (AIC). La idea básica del AIC, concebido para la estrategia de estimación que involucra el método de MCP, es encontrar la combinación de parámetros de suavizamiento, determinados por la cantidad de nodos, que minimicen la expresión

$$AIC(\boldsymbol{\rho}) = -2Loglik + 2df, \tag{16}$$

donde $\boldsymbol{\rho} = [p_0, p_1, \ldots, p_d]^T$ es el vector conformado por los parámetros de suavizamiento,

$$Loglik = -\frac{n}{2}log\left(\frac{2\pi e}{n}SCE_\rho\right) \tag{17}$$

con

$$SCE_\rho = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij} - \widehat{y}_{ij})^2$$

y

$$df = tr(\mathbf{A}) \tag{18}$$

donde $\mathbf{A}$ es la matriz de suavizamiento asociada con el estimador, que en el caso del estimador (15) está dada por

$$\mathbf{A} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}$$

Se elige el modelo que minimice la cantidad $AIC(\boldsymbol{\rho})$. Esta medida permite encontrar un equilibrio entre la bondad de ajuste del modelo, representada por $Loglik$, y la complejidad del modelo, representada por df. Es decir, la bondad de ajuste del modelo es penalizada con la complejidad del mismo (Wu & Zhang 2006).

Una de las desventajas de este método es que no tiene en cuenta directamente la posible correlación de las medidas repetidas de cada individuo, una de las características más importantes de la estructura de los datos longitudinales. En la siguiente sección se propone una alternativa de estimación que considera la estructura de correlación de las medidas repetidas.

4. Estimación a través de las ecuaciones de estimación generalizadas

Aquí se propone otra alternativa de estimación de los parámetros del modelo (10) empleando las *ecuaciones de estimación generalizadas* (EEG) de Liang & Zeger (1986), en las que se asume una matriz de correlación de trabajo especifica para la componente del error, con el fin de incluir directamente la posible correlación de las medidas repetidas de cada individuo. Una de las ventajas de este método es que, sin importar la matriz de correlación de trabajo seleccionada, las EEG siempre llevan a estimaciones consistentes y asintóticamente normales, aunque elegir una estructura de correlación adecuada incrementa la eficiencia del método (Davis 2002, pág. 296).

4.1. Metodología

El método de estimación empleando las EEG requiere la selección de una función de enlace, una función de varianza y una matriz de correlación de trabajo. Una vez seleccionadas estas componentes, según lo indique la naturaleza de los datos, la estimación de los parámetros del modelo (10), denotada por $\widehat{\boldsymbol{\alpha}}_{EEG}$, corresponde a la solución para $\boldsymbol{\alpha}$ del sistema de ecuaciones

$$\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\alpha}}\right)^T \mathbf{V}_i^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\mu}\right) = \mathbf{0} \tag{19}$$

donde $\boldsymbol{\mu}_i = [\boldsymbol{\mu}_{i1}, \ldots, \boldsymbol{\mu}_{in_i}]^T$ es el vector de medias asociado con las medidas repetidas del i-ésimo individuo, esto es, $\mu_{ij} = \mathbb{E}(y_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, $\widehat{\boldsymbol{\theta}}$ es un estimador consistente de $\boldsymbol{\theta}$, el vector de parámetros asociado con la matriz de correlación de trabajo del i-ésimo individuo $\mathbf{R}_i \equiv \mathbf{R}_i(\boldsymbol{\theta})$,

$$\mathbf{V}_i \equiv \mathbf{V}_i(\boldsymbol{\theta}) = \phi \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\theta}) \mathbf{A}_i^{1/2},\ i = 1 \ldots, n \tag{20}$$

es la matriz de covarianzas de trabajo, con ϕ un parámetro de escala posiblemente desconocido, y

$$\mathbf{A}_i = diag[V(\mu_{i1}), \ldots, V(\mu_{in_i})],\ i = 1 \ldots, n$$

con $V(\cdot)$ la función de varianza correspondiente.

En Davis (2002, pág. 300) se presenta el proceso correspondiente para resolver el sistema de ecuaciones (19) y también un procedimiento para hallar una estimación consistente de la varianza estimada de $\widehat{\boldsymbol{\alpha}}_{EEG}$.

Criterio de información de Akaike en las ecuaciones de estimación generalizadas (QIC). Considere el conjunto de datos

$$\mathcal{D} = \{(y_{ij}, \boldsymbol{x}_{ij}) : i = 1, \ldots, n; j = 1, \ldots, n_i\}$$

donde y_{ij} y $\boldsymbol{x}_{ij}$ son el valor de la variable respuesta y el vector de covariables asociados con la j-ésima medición del i-ésimo individuo, y $\boldsymbol{\alpha}$ es el vector de parámetros, que puede estimarse a través de las EEG usando cualquier estructura de correlación de trabajo.

Siguiendo a Pan (2001), bajo el modelo de independencia, la cuasiverosimilitud conjunta, basada en el conjunto de datos $\mathcal{D}$, está dada por

$$Q(\boldsymbol{\alpha}, \phi; \mathbf{I}, \mathcal{D}) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} Q(\boldsymbol{\alpha}, \phi; y_{ij}, \boldsymbol{x}_{ij}) \tag{21}$$

donde ϕ es un parámetro de escala posiblemente desconocido y

$$Q(\boldsymbol{\alpha}, \phi; y, \boldsymbol{x}) = Q(g^{-1}(\boldsymbol{x}^T\boldsymbol{\alpha}), \phi; y) = \int_y^{\mu} \frac{y - t}{\phi V(t)} dt \tag{22}$$

con $\mathbb{V}(y) = \phi V(\mu)$, $V(\cdot)$ una función de varianza, $\mu = \mathbb{E}(y) = g^{-1}(\boldsymbol{x}^T\alpha)$, $g(\cdot)$ una función de enlace, y $\boldsymbol{x}$ un vector de covariables.

La idea básica del criterio, concebido para la estimación utilizando las EEG, es minimizar una medida basada en $Q(\widehat{\boldsymbol{\alpha}}, \widehat{\phi}; \mathbf{I}, \mathcal{D})$, la cuasiverosimilitud bajo el modelo de trabajo de independencia con una estimación de $\boldsymbol{\alpha}$, usando cualquier matriz de correlación de trabajo empleada en las EEG.

El modelo de independencia es asumido para la construcción de la cuasiverosimilitud conjunta, aunque en las EEG se puede emplear cualquier estructura de correlación. Este criterio se puede generalizar utilizando cualquier matriz de correlación en la construcción de la cuasiverosimilitud conjunta, pero puede que esta no sea única (Pan 2001).

El QIC consiste en encontrar la combinación de parámetros de suavizamiento, determinados por la cantidad de nodos, que minimicen la expresión

$$QIC = -2Q(\widehat{\boldsymbol{\alpha}}, \widehat{\phi}; \mathbf{I}, \mathcal{D}) + 2tr(\widehat{\boldsymbol{\Omega}}, \widehat{\mathbf{V}}_r) \tag{23}$$

donde $\widehat{\boldsymbol{\alpha}} \equiv \widehat{\boldsymbol{\alpha}}(\mathbf{R})$ es un estimador del vector de parámetros $\boldsymbol{\alpha}$ del modelo (10), obtenido a través del método de las EEG usando cualquier estructura de correlación $\mathbf{R}$, $\widehat{\mathbf{V}}_r$ es cualquier estimación consistente de $\mathbb{V}(\widehat{\boldsymbol{\alpha}})$, como el estimador del *sandwich* (Liang & Zeger 1986) y

$$\widehat{\boldsymbol{\Omega}} = -\partial^2 Q(\boldsymbol{\alpha}, \widehat{\phi}; \mathbf{I}, \mathcal{D})/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}}\, d$$

4.2. Propiedades

Sea $\widehat{\boldsymbol{\beta}}_{EEG}(t)$ el estimador de $\boldsymbol{\beta}(t) = [\beta_0(t), \beta_1(t), \ldots, \beta_d(t)]^T$ utilizando la metodología propuesta apoyada en las EEG. Este estimador está dado por

$$\widehat{\boldsymbol{\beta}}_{EEG}(t) = [\widehat{\beta}_0(t), \widehat{\beta}_1(t), \ldots, \widehat{\beta}_d(t)]^T = \Phi(t)^T \widehat{\boldsymbol{\alpha}}_{EEG}$$

donde $\boldsymbol{\Phi}(t)$ es la matriz de $\sum_{r=0}^{d} p_r \times (d+1)$ dada por

$$\Phi \equiv \Phi(t) = \begin{bmatrix} \boldsymbol{\Phi}_{0p_0}(t) & \mathbf{0}_{p_0} & \cdots & \mathbf{0}_{p_0} \\ \mathbf{0}_{p_1} & \boldsymbol{\Phi}_{1p_1}(t) & \vdots & \mathbf{0}_{p_1} \\ \vdots & \cdots & \ddots & \vdots \\ \mathbf{0}_{p_d} & \mathbf{0}_{p_d} & \cdots & \boldsymbol{\Phi}_{dp_d}(t) \end{bmatrix}$$

con $\mathbf{0}_{p_r}$ un vector columna constituido por p_r ceros, $r = 0, 1, \ldots, d$.

En un tiempo dado, como $\widehat{\boldsymbol{\beta}}_{EEG}(t)$ es una transformación lineal de $\widehat{\boldsymbol{\alpha}}_{EEG}$, el siguiente corolario derivado del teorema 2 de Liang & Zeger (1986, pág. 16) establece la distribución asintótica de $\widehat{\boldsymbol{\beta}}_{EEG}(t)$.

Corolario 1. *Bajo condiciones de regularidad y dado que:*

(i) $\widehat{\boldsymbol{\theta}}$ *es un estimador* $\sqrt{n}$*-consistente de* $\boldsymbol{\theta}$ *dados* $\boldsymbol{\alpha}$ *y* ϕ*;*

(ii) $\widehat{\phi}$ *es un estimador* $\sqrt{n}$*-consistente de* ϕ *dado* $\boldsymbol{\alpha}$*; y*

(iii) $\| \partial\widehat{\boldsymbol{\theta}}(\boldsymbol{\alpha}, \phi)/\partial\phi \| = O_p(1)$*,*

entonces asintóticamente $\sqrt{n}\,(\widehat{\boldsymbol{\beta}}_{EEG}(t) - \boldsymbol{\beta}(t))$ *tiene una distribución normal multivariada con media cero y matriz de covarianza*

$$\lim_{n\to\infty} n\,\Phi\mathbf{M}_0^{-1}\mathbf{M}_1\mathbf{M}_0^{-1}\Phi^T$$

donde

$$\mathbf{M}_0 = \sum_{i=1}^{n} \left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\alpha}}\right)^T \mathbf{V}_i^{-1} \left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\alpha}}\right) \tag{24}$$

y

$$\mathbf{M}_1 = \sum_{i=1}^{n} \left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\alpha}}\right)^T \mathbf{V}_i^{-1}\mathbb{V}(\boldsymbol{y}_i)\mathbf{V}_i^{-1} \left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\alpha}}\right) \tag{25}$$

Así, un estimador de la varianza de $\widehat{\boldsymbol{\beta}}_{EEG}(t)$ está dado por

$$\widehat{\mathbb{V}}(\widehat{\boldsymbol{\beta}}_{EEG}(t)) = \Phi\widehat{\mathbf{M}}_0^{-1}\widehat{\mathbf{M}}_1\widehat{\mathbf{M}}_0^{-1}\Phi^T$$

donde $\widehat{\mathbf{M}}_0$ y $\widehat{\mathbf{M}}_1$ son obtienen como en las expresiones (24) y (25) remplazando $\mathbb{V}(\boldsymbol{y}_i)$ por $(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}})^T$ y $\boldsymbol{\alpha}$, ϕ y $\boldsymbol{\theta}$ por sus respectivos estimadores. Este estimador de la varianza de $\widehat{\boldsymbol{\beta}}_{EEG}(t)$ es un estimador consistente de $\mathbb{V}(\widehat{\boldsymbol{\beta}}_{EEG}(t))$ aunque la matriz de correlación de trabajo $\mathbf{R}_i$ no corresponda a la verdadera matriz de correlación de $\boldsymbol{y}_i$ (ver la prueba en el apéndice A).

5. Simulación

Con el fin de evaluar el desempeño de los métodos de estimación, la comparación entre ellos se hace a través del *error cuadrático medio promedio* (ECMP) dado por

$$ECMP(\kappa) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}[\kappa(t_{ij}) - \widehat{\kappa}(t_{ij})]^2 \tag{26}$$

donde $\kappa(\cdot)$ es una función que corresponde a cualquier coeficiente dinámico del modelo.

La estrategia de simulación utilizada es similar a la presentada en Wu & Liang (2004). El modelo empleado en la simulación es de la forma

$$y_i(t) = \beta_0(t) + x_{i1}(t)\beta_1(t) + e_i(t) \tag{27}$$

donde $\beta_0(t)$ y $\beta_1(t)$ son los coeficientes dinámicos del modelo, $x_i(t)$ es la única covariable del modelo, asociada con $\beta_1(t)$, y $e_i(t)$ es el error en el tiempo t. Note que este modelo corresponde a un MCD dado por (2) donde el vector de parámetros dinámicos es $\boldsymbol{\beta}(t) = [\beta_0(t), \beta_1(t)]^T$ y el vector de covariables es $\boldsymbol{x}_i(t) = [x_{i0}(t), x_{i1}(t)]^T$ con $x_{i0}(t) \equiv 1$.

En el MCD simulado la covariable $x_{i1}(t)$ está dada por

$$x_{i1}(t) = 1 - \exp\left(-0.5t - \frac{i}{n}\right),\ i = 1, \ldots, n \tag{28}$$

y los parámetros dinámicos se definen como

$$\beta_0(t) = 3\exp(t),\ \text{y}\ \beta_1(t) = 1 + cos(2\pi t) + sen(2\pi t) \tag{29}$$

En la figura 1 se muestran los coeficientes dinámicos simulados.

El término de error correspondiente al modelo (27) se simula bajo dos estructuras de correlación:

Estructura de correlación 1. En la que se considera a $e_i(t)$ distribuido normalmente con media 0 y varianza $\sigma_e^2 x_{i1}^2(t)$, donde las medidas repetidas entre individuos y dentro de cada individuo son independientes.

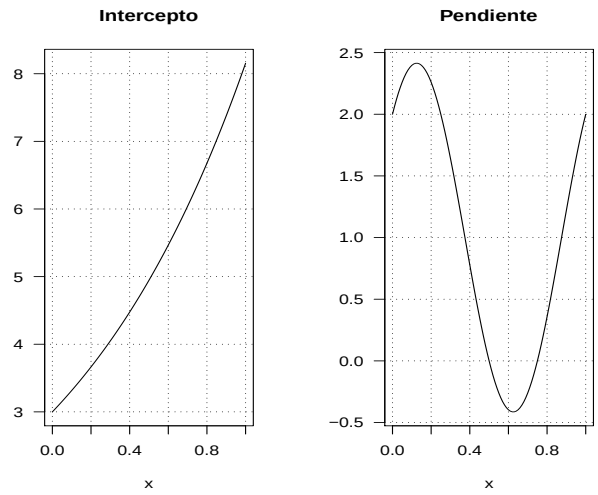

FIGURA 1: Coeficientes dinámicos simulados.

Estructura de correlación 2. En la que se considera al vector de errores asociados con las medidas repetidas de un individuo distribuido normalmente con vector de medias $\mathbf{0}$ y matriz de covarianzas asociada con una estructura de correlación autoregresiva de primer orden, esto es, una matriz de correlación $\mathbf{R} = [R_{kl}]$ donde

$$R_{kl} = \begin{cases} \alpha^{|k-l|}, & \text{si } k \neq l \\ 1, & \text{si } k = l \end{cases}$$

con $0 < \alpha < 1$.

Los tiempos simulados son de la forma

$$t_{ij} = j/(m+1),\ i = 1, \ldots, n,\ j = 1, \ldots, m$$

siendo m un entero positivo, y para simular conjuntos de datos no balanceados[1], una característica propia de la estructura de los datos longitudinales, en cada individuo se remueven aleatoriamente medidas repetidas con una tasa r_m. Así, en promedio hay $m(1 - r_m)$ medidas repetidas por cada individuo y $nm(1 - r_m)$ medidas en total.

Los parámetros de suavizamiento de los métodos respectivos se eligen por medio del AIC y el QIC.

El estudio de simulación se hace en R considerando seis escenarios:

Escenario 1 $n = 25$, $m = 8$, $r_m = 20\,\%$, y $\sigma_e^2 = 0.01$ en la estructura de correlación 1.

[1]El número de medidas repetidas varia de individuo a individuo.

Escenario 2 $n = 25$, $m = 8$, $r_m = 20\,\%$, y $\sigma_e^2 = 0.04$ en la estructura de correlación 1.

Escenario 3 $n = 25$, $m = 8$, $r_m = 20\,\%$, y $\sigma_e^2 = 0.09$ en la estructura de correlación 1.

Escenario 4 $n = 25$, $m = 8$, $r_m = 20\,\%$, y $\phi = 1$ y $\alpha = 0.335$ en la estructura de correlación 2.

Escenario 5 $n = 25$, $m = 8$, $r_m = 20\,\%$, y $\phi = 1$ y $\alpha = 0.665$ en la estructura de correlación 2.

Escenario 6 $n = 25$, $m = 8$, $r_m = 20\,\%$, y $\phi = 1$ y $\alpha = 0.820$ en la estructura de correlación 2.

Cada escenario se replicó $N = 200$ veces y en cada ocasión se calculó $ECMP(\beta_0)$ y $ECMP(\beta_1)$, con el fin de comparar el desempeño relativo de la *estimación a través de las ecuaciones de estimación generalizadas* (EEEG) con la *estimación a través de mínimos cuadrados ponderados* (EMCP). Para ello se definen los indicadores

$$ECMPR = \frac{1}{N}\sum_{k=1}^{N}\frac{ECMP_k(\kappa, EMCP)}{ECMP_k(\kappa, EEEG)} \times 100\,\% \tag{30}$$

y

$$ECMPG = \frac{\sum_{k=1}^{N} I_{\{ECMP_k(\kappa,EMCP) > ECMP_k(\kappa,EEEG)\}}}{N} \times 100\,\% \tag{31}$$

donde $ECMP_k(\kappa, EMCP)$ y $ECMP_k(\kappa, EEEG)$ denotan el valor de $ECMP(\kappa)$ obtenido en la k-ésima réplica de la simulación, $k = 1, \ldots, N$, usando EMCP y EEEG respectivamente, y I_A denota la función indicadora del conjunto A.

Tabla 1: Resultados de la simulación.

Escenario	$ECMPR_0$	$ECMPR_1$	$ECMPG_0$	$ECMPG_1$
Escenario 1	451.6 %	1358.7 %	100.0 %	100.0 %
Escenario 2	388.0 %	979.5 %	100.0 %	100.0 %
Escenario 3	285.1 %	542.1 %	99.0 %	100.0 %
Escenario 4	97.6 %	165.4 %	38.5 %	43.0 %
Escenario 5	170.3 %	401.7 %	62.0 %	64.0 %
Escenario 6	180.8 %	483.4 %	58.5 %	69.5 %

El $ECMPR$ representa la eficiencia relativa promedio asociada con las N réplicas ($N = 200$ en este caso), y $ECMPG$ es el porcentaje de estimadores obtenidos a través de EEG, que son mejores que los obtenidos a través de las MCP en cuanto al ECMP en las N réplicas.

Note que: si $ECMPR \approx 1$ y $ECMPG = 50\,\%$, la EEEG y la EMCP se desempeñan similarmente; si $ECMPR > 1$ y $ECMPG > 50\,\%$, la EEEG se desempeña mejor que la EMCP; y si $ECMPR < 1$ y $ECMPG < 50\,\%$, la EEEG se desempeña peor que la EMCP.

En la tabla 1 se presentan los resultados obtenidos en la simulación. En todos los escenarios considerados, menos el número 4, se obtuvo que $ECMPR > 1$ y

$ECMPG > 50\,\%$, por lo que la EEEG se desempeña mejor que la EMCP con relación al error cuadrático medio promedio. En el escenario 4, como $ECMPR \approx 1$ y $ECMPG \approx 50\,\%$, se tiene que la EEEG y la EMCP se desempeñan similarmente. Esta similitud se debe a que el parámetro de correlación entre las medidas repetidas α no es suficientemente grande para marcar diferencias entre la estructura de correlación autorregresiva de primer orden y la estructura de correlación de independencia. Por ejemplo, note que en el escenario 4 la correlación existente entre la primera medida repetida y la tercera es de apenas R_{13} =0.037.

6. Aplicación

En esta sección se presentan los resultados del análisis del conjunto de datos ACTG 315 (Liang et al. 2003). Este conjunto de datos corresponde a un estudio del *síndrome de inmunodeficiencia adquirida* (sida), llevado a cabo para investigar la relación entre carga viral y número de células CD4+ en individuos infectados con el *virus de inmunodeficiencia humana* (VIH).

6.1. Introducción

La carga viral (*plasma VIH RNA copies/mL*) y el conteo de células CD4+ son actualmente indicadores decisivos para evaluar tratamientos de sida en investigación clínica. Inicialmente se consideró el conteo de células CD4+ como un indicador inmunológico primario de sida, pero últimamente se ha encontrado que la carga viral es más predictiva en los resultados clínicos. Sin embargo, recientemente algunos investigadores han sugerido que la combinación de estos dos indicadores puede ser más apropiada para evaluar los tratamientos de VIH y sida. Por ello es pertinente estudiar la relación entre la carga viral y el conteo de células CD4+ durante el tratamiento (Liang et al. 2003).

En la figura 2 se presentan algunos gráficos de una regresión lineal de la carga viral ($\log_{10}$(RNA)) frente al conteo de células CD4+ en algunos tiempos de un estudio clínico de sida (ACTG 315). En este, se trata de 46 pacientes infectados, con una terapia antiviral que consistía de *ritonavir*, 3TC y AZT. Una vez iniciado el tratamiento, la carga viral y el conteo de células CD4+ fueron observados simultáneamente los días 0, 2, 7, 10, 14, 28, 56, 84, 168 y 336. En total se obtuvieron 361 observaciones, y el número de medidas repetidas por individuo varía de 4 a 10.

En general, se observa que la respuesta virológica (medida por la carga viral) y la respuesta inmunológica (medida por el conteo de células CD4+) del paciente están correlacionadas negativamente, y que su relación es aproximadamente lineal durante el tratamiento antiviral. En la figura 3 se presentan los dispersogramas del conteo de células CD4+ y de la carga viral. Se utiliza el logaritmo de la carga viral para estabilizar la varianza en los procedimientos de estimación de los modelos ajustados en las secciones siguientes.

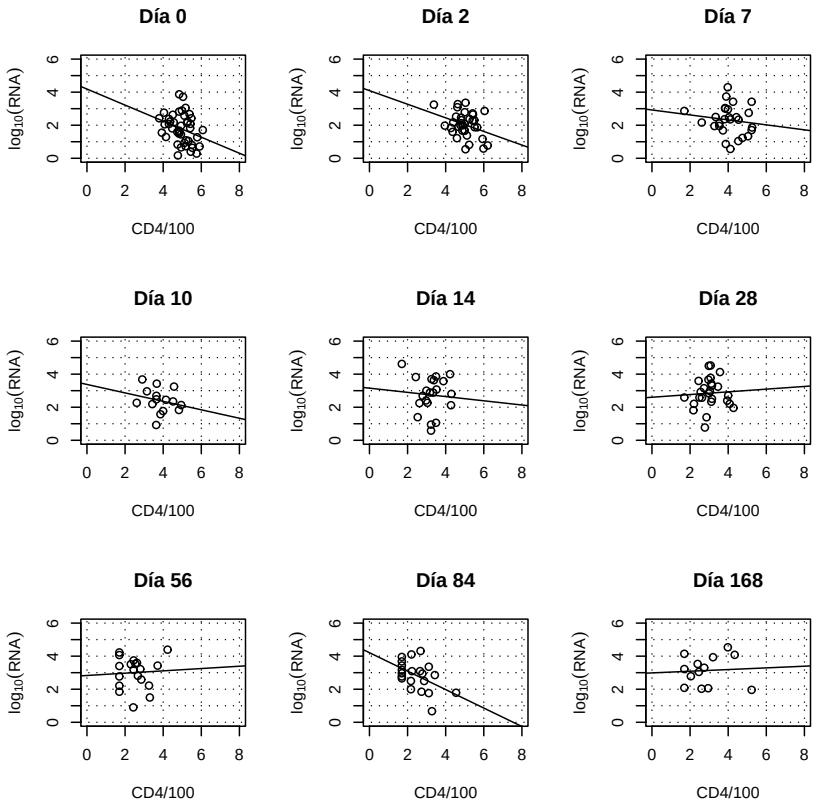

FIGURA 2: Gráficos de una regresión lineal de la carga viral ($\log_{10}(\text{RNA})$) frente al conteo de células CD4+ en algunos tiempos. El modelo ajustado en cada caso es de la forma $\log_{10}(RNA) = \beta_0 + \beta_1(\text{CD4}/100) + e$.

En la figura 2, note que la pendiente de cada regresión lineal cambia con el tiempo. Esto motiva a modelar la carga viral con un MCD.

El conjunto de datos ACTG 315 ya ha sido estudiado ampliamente por Liang et al. (2003), donde se evidencia principalmente una fuerte relación inversa entre la carga viral y el conteo de células CD4+.

6.2. Modelamiento

En esta sección se presentan los resultados del análisis del conjunto de datos ACTG 315 usando un MCD. Se ajusta un MCD para investigar la relación dinámica entre la carga viral (en escala logarítmica) y el conteo de células CD4+.

El MCD ajustado es de la forma

$$y_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij})x_{i1}(t_{ij}) + e_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, 46 \qquad (32)$$

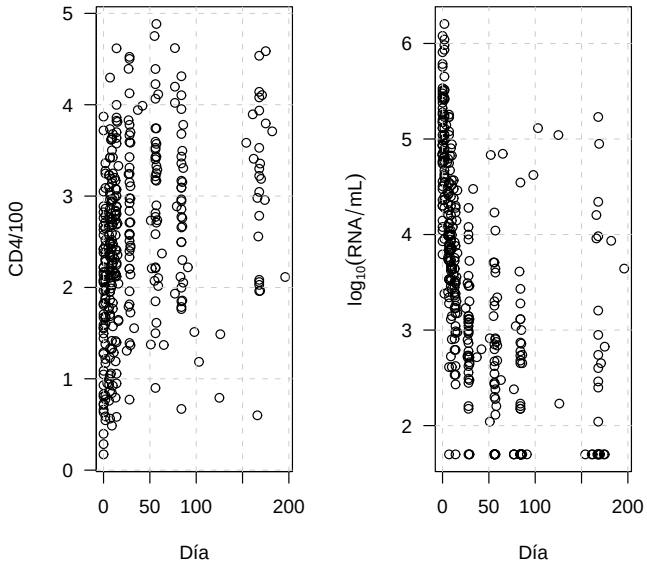

FIGURA 3: Dispersograma del conteo de células CD4+ y de la carga viral.

donde y_{ij}, $x_{i1}(t_{ij})$, y e_{ij} son la carga viral (en escala logarítmica), el conteo de células CD4+ y el error asociados con la j-ésima medición del i-ésimo individuo, respectivamente, y $\beta_0(t)$ y $\beta_1(t)$ son los coeficientes dinámicos del modelo. Note que $\beta_1(t)$ es el coeficiente relacionado con la relación dinámica entre la carga viral y el conteo de células CD4+.

Las componentes dinámicas del modelo se estiman a través de la metodología propuesta utilizando RS y las EEG. La base de *splines* empleada en la RS es una BBS de segundo grado, y en la selección del número de los nodos se emplea el QIC. La ubicación de los nodos se hace según el método 1 en el que se ubican los nodos igualmente espaciados en el intervalo de interés.

TABLA 2: Parámetros de suavizamiento que minimizan el criterio de selección bajo una estructura de correlación de trabajo dada.

p_0	p_1	Estructura	$ECMP(\beta_0)$	$ECMP(\beta_0)$
11	10	Independencia	7.450128e+166	8.706318e-02
4	4	Intercambiable	7.450128e+166	8.701916e-02
8	7	Autorregresiva (1)	7.450128e+166	8.698554e-02

Como se muestra en la tabla 2, la estructura de correlación de trabajo que minimiza el ECMP es la estructura intercambiable. Note que el ECMP de todas las estructuras es similar, así que por simplicidad se eligió la estructura de independencia, donde los parámetros de suavizamiento asociados con $\beta_0(t)$ y $\beta_1(t)$ que

minimizan el QIC son $p_0 = 11$ y $p_1 = 10$, respectivamente. En este caso, la estructura de correlación no tiene mayor incidencia en la estimación de las componentes del modelo, porque la cantidad de medidas repetidas de cada individuo es pequeña comparada con la cantidad de mediciones total.

En la figura 4 se muestra la estimación del coeficiente dinámico asociado con la pendiente del MCD ajustado. Note que al principio del tratamiento la relación entre la carga viral y el conteo de células CD4+ no es tan fuerte hasta aproximadamente el día 90 del tratamiento donde se evidencia una relación directa hasta el día 100. Entre los días 100 y 140 se nota una relación claramente inversa. Entre los días 140 y 150 parece que la relación es directa, pero se nota que vuelve a ser débil. Estos resultados son similares a los obtenidos por Liang et al. (2003).

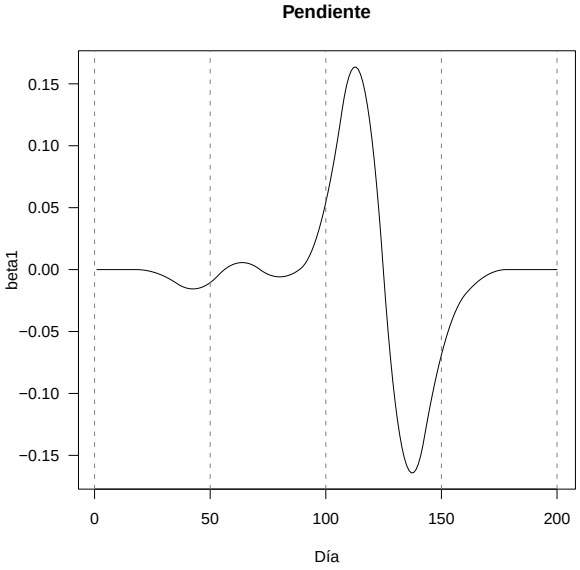

Figura 4: Estimación del coeficiente dinámico asociado con la pendiente del MCD ajustado.

7. Discusión

La metodología propuesta permite involucrar en la estimación de los coeficientes dinámicos y en la selección de los parámetros de suavizamiento la posible correlación de las medidas repetidas de cada individuo en estudio. Por ello en los escenarios simulados donde el grado de correlación es significativo se obtuvo que utilizar las EEG y el QIC da mejores resultados con relación al error cuadrático medio que utilizar el método de MCP y el AIC.

Incluir en el estimador de las componentes del MCD la posible correlación entre las medidas repetidas no tiene un costo computacional importante, puesto que los

procesos computacionales se pueden implementar fácil y eficientemente utilizando software como la función `geese.fit` del paquete geepack de R.

En la metodología propuesta se aproximan los coeficientes dinámicos mediante regresión *spline*, pero cabe notar que no es la única alternativa. También es posible llevar a cabo la aproximación mediante suavizamiento *spline*, polinomios locales *kernel* (Wu & Zhang 2006) o mediante funciones radiales *kernel* (Sosa & Díaz 2009), técnicas a comparar en futuros estudios de simulación.

Utilizando regresión *spline*, también es posible emplear bases diferentes de *B-splines* en la aproximación de los coeficientes dinámicos, por ejemplo bases de potencias truncadas (Ramsay & Silverman 1997). Esta alternativa también es susceptible de investigación en estudios de simulación posteriores.

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

Referencias

Altman, N. S. (1990), 'Kernel Smoothing of Data with Correlated Errors', *Journal of the American Statistical Association* **85**(411), 749–759.

Davis, C. S. (2002), *Statistical Methods for the Analysis of Repeated Measurements*, Springer.

de Boor, C. (1978), *A Practical Guide to Splines*, Springer.

Diggle, P. J., Liang, K. Y. & Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford University Press.

Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (2009), *Longitudinal Data Analysis*, Champman & Hall.

Hart, J. D. (1991), 'Kernel Regression Estimation with Time Series Errors', *Journal of the Royal Statistical Society. Series B (Methodological)* **53**(1), 173–187.

Hart, J. D. & Wehrly, T. E. (1986), 'Kernel Regression Estimation using Repeated Measurements Data', *Journal of the American Statistical Association* **81**(396), 1080–1088.

Hastie, T., Tibshirani, R. & Friedman, J. (1990), *The Elements of Statistical Learning*, Springer.

Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L. P. (1998), 'Nonparametric Smoothing Estimates of Time-Varying Coefficient Models with Longitudinal Data', *Biometrika* **85**(4), 809–822.

Huang, J. Z., Wu, C. O. & Zhou, L. (2002), 'Varying-coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements', *Biometrika* **89**(1), 111–128.

Huggins, R. M. & Loesch, D. Z. (1998), 'On the Analysis of Mixed Longitudinal Growth Data', *Biometrics* **54**(2), 583–595.

Liang, H., Wu, H. & Carroll, R. J. (2003), 'The relationship between Virologic and Immunologic Responses in AIDS Clinical Research using Mixed-Effects Varying-Coefficient Models with Measurement Error', *Biostatistics* **4**(2), 297–312.

Liang, K. Y. & Zeger, S. L. (1986), 'Longitudinal Data Analysis using Generalized Linear Models', *Biometrika* **73**(1), 13–22.

Pan, W. (2001), 'Akaike's Information Criterion in Generalized Estimating Equations', *Biometrics* **57**(1), 120–125.

Ramsay, J. O. & Silverman, B. W. (1997), *Applied Functional Data Analysis*, Springer.

Sosa, J. C. & Díaz, L. G. (2009), Desarrollo de un modelo de coeficientes dinámicos y aleatorios para el análisis longitudinales, Tesis de maestría, Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia.

Verbeke, G. & Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, Springer.

Wu, H. & Liang, H. (2004), 'Backing Random Varying-Coefficient Models with Time-Dependent Smoothing Covariates', *Scandinavian Journal of Statistics* **31**, 3–19.

Wu, H. & Zhang, J. T. (2006), *Nonparametric Regression Methods for Longitudinal Data Analysis*, Wiley.

Zeger, S. L. & Diggle, P. J. (1994), 'Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters', *Biometrics* **50**(3), 689–699.

Apéndice

Sea $\widehat{V}_{EEG}$ un estimador consistente de la varianza de $\widehat{\boldsymbol{\alpha}}_{EEG}$, como el estimador del *sandwich*, que es un estimador consistente aunque la matriz de correlación de trabajo $\mathbf{R}_i(\boldsymbol{\alpha})$ no corresponda a la verdadera matriz de correlación de $\boldsymbol{y}_i$ (Davis 2002, pág. 300).

Como $\widehat{V}_{EEG}$ es un estimador consistente de la varianza de $\widehat{\boldsymbol{\alpha}}_{EEG}$ entonces se tiene que $\widehat{V}_{EEG} \xrightarrow{P} \mathbb{V}(\widehat{\boldsymbol{\alpha}}_{EEG})$, es decir, para cualquier $\epsilon > 0$ y para cualquier $\eta > 0$, existe un entero positivo $n_0 \equiv n_0(\epsilon, \eta)$, tal que

$$P\left[\parallel \widehat{V}_{EEG} - \mathbb{V}(\widehat{\boldsymbol{\alpha}}_{EEG}) \parallel > \epsilon\right] < \eta \;\text{ siempre que }\; n \geq n_0.$$

Sea ϵ^* tal que

$$\epsilon = \frac{\epsilon^*}{\parallel \Phi \parallel \parallel \Phi^T \parallel}$$

por lo que

$$P\left[\parallel \widehat{V}_{EEG} - \mathbb{V}(\widehat{\boldsymbol{\alpha}}_{EEG}) \parallel > \epsilon\right] < \eta \text{ siempre que } n \geq n_0,$$

es equivalente a

$$P\left[\parallel \Phi\widehat{V}_{EEG}\Phi^T - \mathbb{V}(\Phi\widehat{\boldsymbol{\alpha}}_{EEG}) \parallel > \epsilon^*\right] < \eta \text{ siempre que } n \geq n_0,$$

y por tanto $\widehat{\mathbb{V}}(\widehat{\boldsymbol{\beta}}_{EEG}(t)) \xrightarrow{P} \mathbb{V}(\widehat{\boldsymbol{\beta}}_{EEG}(t))$.

Revista Colombiana de Estadística
Junio 2010, volumen 33, no. 1, pp. 111 a 125

Bondad de ajuste empleando la función generadora de momentos

Goodness-of-Fit Employing the Moment Generating Function

Luis Alfonso Muñoz[1,a], Jorge Humberto Mayorga[2,b]

[1]Ciencias Naturales, Institución Educativa Técnico San Juan Bautista, Nariño, Colombia

[2]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Resumen

Proponemos una estadística para evaluar la bondad de ajuste, donde se emplearon las funciones generatrices de momentos muestral y poblacional; a partir de la información considerada se encontró que la estadística G_n tuvo un comportamiento diferente de acuerdo con el modelo usado para el ajuste. Su desempeño fue superior o igual a la estadística de Pearson, pero fue superada por la estadística de K-S; además, para la estadística evaluada no fue influyente el tamaño de la muestra.

***Palabras clave*:** bondad de ajuste, función generadora de momentos, simulación.

Abstract

We propose a statistic to evaluate the goodness-of-fit where we used the empirical moment generating function and the moment generating function, from the considered information it found that the G_n statistic was different behavior according to the used model for the fitting. Its behavior was great or similar to the Pearson statistics, but it was exceeded for the K-S statistic, also for the evaluated statistic was not influential to the sample size.

***Key words*:** Goodness-of-fit, Moment generating function, Simulation.

[a]Profesor. E-mail: lamunozbe@.unal.edu.co
[b]Profesor asociado. E-mail: jhmayorgaa@unal.edu.co

1. Introducción

El objetivo del presente trabajo es proponer una estadística para evaluar la bondad de ajuste desde la perspectiva del cotejo de las funciones generatrices de momentos (fgm) muestral y poblacional. Específicamente se orienta a comparar vía simulación el desempeño en la evaluación del ajuste de la estadística propuesta a un modelo probabilístico de tipo continuo, frente a la evaluación del ajuste con las estadísticas de Pearson y Kolmogorov-Smirnov mediante el uso de un test que se desarrolló para dicho propósito.

2. Función generadora de momentos

La función generadora de momentos de una variable aleatoria X es una función con valores reales $M_X(t) = E(e^{tX})$, siempre que el valor esperado exista para todo $t \in (-h, h), h > 0$; específicamente, $M_X(t) = E(e^{tX}) = \sum_{j=1}^{\infty} e^{tx_j} P_X(x_j)$ si X es discreta o $\int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx$ si X es continua, siendo $P_X(x_j)$ y $f_X(x)$ la función de probabilidad y la función de densidad, respectivamente. A $M_X(t)$ se le denomina fgm porque los momentos ordinarios de X pueden obtenerse derivando esta función y evaluando la derivada en $t = 0$ (Mood et al. 1974).

3. Función generadora de momentos de la muestra

Siendo $X_1, X_2, \ldots, X_n$, una muestra aleatoria de tamaño n proveniente de una población cuya fgm es $M_X(t)$, la función $M_n(t)$ se denomina función generadora de momentos de la muestra y se define según Collander & Chalfant (1985) así: $M_n(t) = \frac{1}{n} \sum_{j=1}^{n} \exp(tX_j)$. La función $M_n(t)$, como estadística, es un estimador insesgado, consistente simple y consistente en media cuadrática para $M_X(t)$.

Quand & Ramsey (1978) fueron los primeros en sugerir la función generadora de momentos de la muestra para la estimación de los parámetros de una mezcla de dos distribuciones normales; la función característica de la muestra, una función con facultades análogas a la fgm, ha sido usada por varios autores para modelar distribuciones desconocidas: Feuerverger & Mureika (1977), Heathcote (1977), Koutrouvelis (1980), Koutrouvelis & Kellermeier (1981), Csörgo (1981), Epps & Pulley (1985), Csörgo (1986) y Csörgo & Heathcote (1987). Trabajos más recientes que emplean la fgm para probar distribuciones específicas son los presentados por Cabaña & Quiroz (2005) y Meintanis (2007).

4. Estructura de la estadística propuesta

Siendo X una variable aleatoria cuya fgm $M_X(t) = E(e^{tX})$ existe, $t \in (-h, h)$ para algún $h > 0$, se define una nueva variable aleatoria $Y = e^{tX}$, entonces $E[Y] = M_X(t)$ y $V[Y] = M_X(2t) - M_X^2(t)$. Considerando una muestra aleatoria $X_1, X_2, \ldots, X_n$, de una población descrita por la variable aleatoria X, donde

$M_X(t)$ es la fgm poblacional y al ser $M_n(t)$ la fgm muestral, el teorema del límite central permite garantizar que: $T_n(t) = \frac{\sqrt{n}(M_n(t)-M_X(t))}{\sqrt{M_X(2t)-M_X^2(t)}} \xrightarrow{d} Z \approx N(0,1)$, para un valor t fijo.

Dada una partición particular del intervalo $(-h', h')$, excluyendo a 0 como punto de subdivisión, se tiene $t_1, t_2, \ldots, t_r$, con $t_1 = -h'$, $t_j = t_{j-1} + \delta$, $j = 1, 2, \ldots, \frac{r}{2}$; $t_r = h'$, $t_{j-1} = t_j - \delta$, $j = \frac{r}{2} + 1, \ldots, r$; donde δ es la amplitud de cada subintervalo, $T_n(t_j) = \frac{\sqrt{n}(M_n(t_j)-M_X(t_j))}{\sqrt{M_X(2t_j)-M_X^2(t_j)}}$ $j = 1, 2, \ldots, r$.

La estadística que este trabajo propone pretende ser un instrumento para determinar la bondad de ajuste y su expresión corresponde a:

$$G_n = \sum_{j=1}^{r} \frac{n(M_n(t_j) - M_X(t_j))^2}{M_X(2t_j) - M_X^2(t_j)}$$

Intuitivamente G_n tenderá a presentar valores pequeños cuando $M_X(t)$ y $M_n(t)$ sean muy similares; por el contrario, tenderá a producir valores grandes cuando las dos funciones difieran ampliamente. La estadística G_n tiene una forma particular de acuerdo con el modelo probabilístico elegido para el ajuste.

4.1. Distribución de la estadística

El hecho de que $T_n(t) \xrightarrow{d} Z \approx N(0,1)$ sugiere que $T_n^2(t) \xrightarrow{d} \chi^2_{(1)}$ y por tanto podría pensarse que la estadística G_n, por su estructura, tendría también distribución asintóticamente ji-cuadrada. Si $T_n^2(t_j)$ y $T_n^2(t_i)$ $i \neq j$, $i, j = 1, 2, \ldots, r$ fueran dos variables aleatorias estadísticamente independientes, $M_{G_n(t_j)} = \prod_{j=1}^{r} M_{T_n^2(t_j)}(t)$ y por tanto la distribución de G_n podría estudiarse analíticamente por medio de esta consecuencia.

Dejando de lado el examen de la independencia anteriormente señalada, para dar paso a la exploración del comportamiento de la estadística propuesta y así tener evidencias iniciales para abordar la manera de proceder, se simuló G_n para el modelo gaussiano y para el modelo uniforme.

4.2. Simulación de la estadística

La distribución de G_n bajo H_0 se simuló empleando el programa IML del paquete estadístico SAS para muestras de tamaño $n = 20, 70, 120, 170$ y 220, con parámetros $\mu = 50, 150, 250$ y 350; además $\sigma^2 = 1, 256$ y 961, en el caso normal. Cuando se emplearon muestras provenientes de la distribución uniforme $(0, 1)$, los tamaños de muestra fueron los mismos que para el caso de la distribución normal (véase Muñoz 1998, p. 10).

Particularmente con el objeto de indagar el efecto de la escogencia de la partición alrededor de cero en el desempeño de la estadística, se simularon valores de ella con base en la partición del intervalo $(-0.1, 0.1)$ con $\delta = 0.001$ y por tanto $r = 20$, la cual se denominó estadística $G_n(20)$; los valores de las estadísticas

$G_n(100)$ y $G_n(500)$ fueron simulados con $\delta = 0.0002$, $r = 100$ y $\delta = 0.00004$, $r = 500$, respectivamente.

La simulación sugirió que la distribución de G_n tendría un valor esperado cercano al número de puntos de la subdivisión del intervalo $(-h', h')$, porque persistentemente mostró promedios alrededor de r, y varianzas que oscilaron cerca de $2r^2$, lo que indujo a considerar no plausible la conjetura de un comportamiento χ^2 de la estadística (tablas 1 y 2).

4.3. Distribución aproximada de la estadística

Los valores simulados se agruparon en histogramas que mostraron siempre sesgo positivo y unimodalidad. Este hecho sugirió que la estadística propuesta tenía las mismas características ya sea usando el modelo normal o el uniforme. Intuitivamente se puede elegir a la familia gamma como modelo de aproximación del comportamiento de G_n, no lejano de la presunción inicial de una distribución χ^2; la función de densidad, valor esperado y varianza de una variable aleatoria X con distribución gamma son respectivamente: $f_X(x) = \frac{1}{\alpha^s \Gamma(s)} x^{s-1} e^{-\frac{1}{\alpha}x} I_{(0,\infty)}(X)$, $E[X] = \alpha s$ y $V[X] = s\alpha^2$.

En las tablas 1 y 2 se ratifican los valores de los parámetros de la distribución aproximada donde el valor esperado es igual a r. Así entonces $r = \alpha s$; por otra parte, el punto de oscilación $2r^2$ sugiere el valor aproximado de la varianza, con lo cual $s\alpha^2 = 2r^2$ y $r\alpha = 2r^2$, de donde $\alpha = 2r$, en consecuencia $2rs = r$ y en síntesis $s = \frac{1}{2}$.

En las tablas 1 y 2 se puede observar la semejanza entre los valores esperados y varianza simuladas con valores esperados y varianzas del modelo gamma, lo cual permite afirmar que la distribución de G_n puede aproximarse por medio de una distribución gamma con parámetros $s = \frac{1}{2}$ y $\alpha = 2r$.

Tabla 1: Media, varianza y percentiles obtenidos por simulación para la estadística G_n con base en 10000 muestras provenientes de la distribución normal.

			PERCENTILES						
ESTADÍSTICA	media	varianza	0.1	0.25	0.5	0.75	0.90	0.95	0.99
$G_n(20)$	20.02	785.48	0.46	2.18	9.23	26.46	53.94	76.37	131.46
percentiles	0.32	2.03	9.10	26.47	54.11	76.83	132.70		
dist. gamma[a]									
$G_n(100)$	100.11	19666.02	2.28	10.85	46.12	132.33	269.85	381.67	657.35
percentiles	1.58	10.15	45.49	132.33	270.55	384.15	663.49		
dist. gamma[a]									
$G_n(500)$	500.73	492477.59	11.31	54.21	230.64	661.69	1350.22	1911.67	3292.77
percentiles	7.90	50.77	227.47	661.65	1352.77	1920.73	3317.45		
dist. gamma[a]									

[a] con parámetros $s = \frac{1}{2}$ y $\alpha = 2r$.

Como consecuencia del examen del comportamiento de los valores simulados de G_n, de sus promedios, varianzas y de sus percentiles que presentan gran similitud con respecto a la distribución gamma, este trabajo propone el siguiente test para evaluar la bondad de ajuste a las distribuciones normal y uniforme, en los siguientes términos:

Tabla 2: Media, varianza y percentiles obtenidos por simulación para la estadística G_n con base en 10000 muestras provenientes de la distribución uniforme.

			PERCENTILES						
ESTADÍSTICA	media	varianza	0.1	0.25	0.5	0.75	0.90	0.95	0.99
$G_n(20)$	19.93	792.48	0.30	1.94	9.00	26.32	54.25	77.62	130.12
percentiles	0.32	2.03	9.10	26.47	54.11	76.83	132.70		
dist. gamma[a]									
$G_n(100)$	99.67	19812.24	1.50	9.74	45.02	131.60	271.24	388.10	650.66
percentiles	1.58	10.15	45.49	132.33	270.55	384.15	663.49		
dist. gamma[a]									
$G_n(500)$	498.41	495354.30	7.51	48.74	225.16	658.03	1356.20	1940.60	3253.40
percentiles	7.90	50.77	227.47	661.65	1352.77	1920.73	3317.45		
dist. gamma[a]									

[a] con parámetros $s = \frac{1}{2}$ y $\alpha = 2r$.

"Rechazar la hipótesis de ajuste al modelo en consideración si $G_n > G_{n,1-\alpha}$". $G_{n,1-\alpha}$ es el percentil $100(1-\alpha)\,\%$ de una variable con distribución gamma con parámetros $s = \frac{1}{2}$ y $\alpha = 2r$.

5. Exploración del ajuste con la estadística propuesta

Para identificar peculiaridades de la estadística propuesta, frente a las estadísticas con las cuales se coteja, se dispuso de cuatro regiones que representaran cercanías o alejamientos entre el promedio de la muestra y el promedio poblacional e igualmente entre la varianza de la muestra y la varianza poblacional. Para ello se utilizó el primer cuadrante del plano cartesiano, tomando como eje de las abscisas la diferencia en valor absoluto entre los promedios y como eje de las ordenadas la diferencia en valor absoluto entre las varianzas. El cuadrante así obtenido se divide en las cuatro regiones como se muestra en la figura 1.

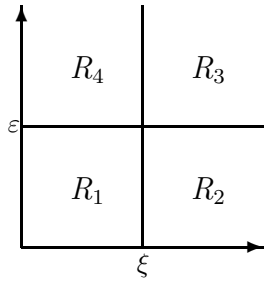

Figura 1: Regiones propuestas para la exploración.

Las cotas ξ y ε, que definen las regiones, se determinaron aplicando la desigualdad de Chebyshev como se describe posteriormente. La región R_1 representa las situaciones donde las diferencias de los promedios y de las varianzas son relativamente pequeñas. La región R_2 representa las ocasiones en las cuales las diferencias entre promedios son grandes y las diferencias entre las varianzas son pequeñas,

la región R_3 representa los casos en los cuales las diferencias entre promedios y varianzas son relativamente grandes y la región R_4 representa los eventos en que las diferencias entre promedios sos pequeñas y las diferencias entre varianzas son grandes.

5.1. Juzgamiento del ajuste a la distribución normal

La unidad de exploración consistió en un conjunto de 1000 muestras aleatorias particulares provenientes de una distribución normal con los mismos parámetros empleados en la sección 4.2.

Con las muestras simuladas se evaluó la bondad del ajuste empleando las estadísticas de Pearson, K-S y la estadística propuesta G_n. La respuesta se cuantificó por medio de la tasa de rechazo de la hipótesis de normalidad, definiéndose como tasa total de rechazos (TTR), tasa de rechazos región uno (TRR_1), tasa de rechazos región dos (TRR_2), tasa de rechazos región tres (TRR_3) y tasa de rechazos región cuatro (TRR_4); las cuatro regiones se definieron como

$$
\begin{aligned}
R_1 &: \text{Si } |\overline{X}_n - \mu| < \xi \quad \text{y} \quad \left|\frac{S_n^2}{\sigma^2} - 1\right| < \varepsilon \\
R_2 &: \text{Si } |\overline{X}_n - \mu| \geq \xi \quad \text{y} \quad \left|\frac{S_n^2}{\sigma^2} - 1\right| < \varepsilon \\
R_3 &: \text{Si } |\overline{X}_n - \mu| \geq \xi \quad \text{y} \quad \left|\frac{S_n^2}{\sigma^2} - 1\right| \geq \varepsilon \\
R_4 &: \text{Si } |\overline{X}_n - \mu| < \xi \quad \text{y} \quad \left|\frac{S_n^2}{\sigma^2} - 1\right| \geq \varepsilon
\end{aligned}
$$

La cota ξ para la diferencia de medias se determinó teniendo en cuenta que $E[\overline{X}_n] = \mu$ y $V[\overline{X}_n] = \frac{\sigma^2}{n}$; luego, utilizando la desigualdad de Chebyshev $P[|\overline{X}_n - \mu_x| < r\sigma_x] \geq 1 - \frac{1}{r^2}$ para $r > 0$ y utilizando el remplazo $\xi = r\frac{\sigma}{\sqrt{n}}$, se tiene $P[|\overline{X}_n - \mu_x| < \xi] \geq 1 - \frac{\sigma^2}{n\xi^2}$ eligiendo $1 - \frac{\sigma^2}{n\xi^2} = 0.05$, entonces $\xi = \frac{\sigma}{\sqrt{n*0.95}}$.

La cota ε para la diferencia de varianzas se determinó denotando $X = \frac{S_n^2}{\sigma^2}$. De esta manera $E[X] = 1$ y $V[X] = \frac{1}{\sigma^4}\, V[S_n^2]$.

Como $V[S_n^2] = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\sigma^4)$, $n > 1$ con $\mu_4 = \frac{4!\sigma^4}{2!2^2} = 3\sigma^4$, entonces $V[S_n^2] = \frac{1}{n}\left(3\sigma^4 - \frac{n-3}{n-1}\sigma^4\right)$ y $V[X] = \frac{1}{n}\left(3 - \frac{n-3}{n-1}\right)$, $n > 1$; empleando la desigualdad de Chebyshev $P[|X - \mu_x| < r\sigma_x] \geq 1 - \frac{1}{r^2}$ para cada $r > 0$, y remplazando $\varepsilon = r\sqrt{\frac{1}{n}(3 - \frac{n-3}{n-1})}$, se tiene que $P[|\frac{S_n^2}{\sigma^2} - 1| < \varepsilon] \geq 1 - \frac{\left(3 - \frac{n-3}{n-1}\right)}{n\varepsilon^2}$. Eligiendo $1 - \frac{\left(3 - \frac{n-3}{n-1}\right)}{n\varepsilon^2} = 0.05$ entonces $\varepsilon = \sqrt{\frac{\left(3 - \frac{n-3}{n-1}\right)}{n*0.95}}$.

Para realizar el ajuste con la estadística de Pearson al modelo normal con valor esperado μ y varianza σ^2 totalmente especificados, se determinó una partición de la recta real de seis subintervalos disjuntos desde $(-\infty, \mu - 2\sigma]$ hasta $(\mu + 2\sigma, \infty]$ y

para efecto de decisión se utilizó el percentil 0.95 de una distribución χ^2 con cinco grados de libertad.

Con el ajuste que empleó la estadística de K-S se utilizó el percentil 0.95 de D_n, calculado con la expresión $1.36/(n+\sqrt{n/10})^{1/2}$ (Conover 1999).

El ajuste a la distribución normal de parámetros específicos mediante la estadística propuesta tuvo en cuenta los percentiles 0.95 que se muestran en las tablas 1 y 2.

5.2. Juzgamiento del ajuste a la distribución uniforme

La unidad de exploración también fueron 1000 muestras aleatorias particulares provenientes de una distribución uniforme de tamaño $n = 20, 70, 120, 170$ y 220; con las muestras simuladas se procedió a realizar la prueba de bondad de ajuste y se cuantificó la tasa de rechazo de igual forma que la descrita en la sección anterior; las regiones para este caso fueron

$$\begin{aligned}
R_1 &: \text{Si } |\overline{X}_n - 0.5| < \xi \quad \text{y} \quad |S_n^2 - \sigma^2| < \varepsilon \\
R_2 &: \text{Si } |\overline{X}_n - 0.5| \geq \xi \quad \text{y} \quad |S_n^2 - \sigma^2| < \varepsilon \\
R_3 &: \text{Si } |\overline{X}_n - 0.5| \geq \xi \quad \text{y} \quad |S_n^2 - \sigma^2| \geq \varepsilon \\
R_4 &: \text{Si } |\overline{X}_n - 0.5| < \xi \quad \text{y} \quad |S_n^2 - \sigma^2| \geq \varepsilon
\end{aligned}$$

Teniendo en cuenta que $E[\overline{X}_n] = 0.5$ y $V\left[\overline{X}_n\right] = \frac{1}{12*n}$, $P\left[\left|\overline{X}_n - 0.5\right| < r\sqrt{\frac{1}{12*n}}\right] \geq 1 - \frac{1}{r^2}$ para $r > 0$, utilizando el remplazo $\xi = r\sqrt{\frac{1}{12*n}}$, se tiene que $P\left[\left|\overline{X}_n - 0.5\right| < \xi\right] \geq 1 - \frac{1}{12*n\xi^2}$ y eligiendo $1 - \frac{1}{12*n\xi^2} = 0.05$, $\xi = \frac{1}{\sqrt{0.95*12*n}}$.

Por otra parte, como $E[S_n^2] = \sigma^2$ y $V[S_n^2] = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\sigma^4), n > 1$, $V[S_n^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\left(\frac{1}{144}\right)\right)$, y $\mu_4 = \frac{1}{5*2^4} = \frac{1}{80}$. Así $V[S_n^2] = \frac{1}{n}\left((\frac{1}{180})\frac{n-3}{n-1}\right), n > 1$ con base en la desigualdad de Chebyshev, $P[|X - \mu_x| < r\sigma_x] \geq 1 - \frac{1}{r^2}$ para cada $r > 0$ y remplazando $\varepsilon = r\sqrt{\frac{1}{n}\left(\left(\frac{1}{180}\right)\frac{n-3}{n-1}\right)}$, se tiene que $P[|S_n^2 - \sigma^2| < \varepsilon] \geq 1 - \frac{\left(\left(\frac{1}{80}\right)\frac{n-3}{n-1}\right)}{n\varepsilon^2}$, eligiendo $1 - \frac{\left(\left(\frac{1}{80}\right)\frac{n-3}{n-1}\right)}{n\varepsilon^2} = 0.05$ entonces $\varepsilon = \sqrt{\frac{\left(\left(\frac{1}{80}\right)\frac{n-3}{n-1}\right)}{n*0.95}}$.

En la evaluación del ajuste de la estadística de Pearson al modelo uniforme en el intervalo $(0, 1)$, se estableció una partición de cinco intervalos de una amplitud de 0.2. Para efecto de decisión se utilizó el percentil 0.95 de una distribución χ^2 con cuatro grados de libertad.

El procedimiento para el ajuste con la estadística de K-S y la estadística G_n fue similar al empleado en el juzgamiento de la distribución normal.

6. Resultados de la exploración con base en la distribución normal

El análisis de los resultados de la exploración se llevó a cabo de manera análoga al análisis de los resultados de un diseño experimental con arreglo factorial de efectos fijos. En este contexto los elementos del análisis fueron:

Respuesta: tasa de rechazo de la hipótesis nula.

Factores:

- Valor esperado μ. Cuatro niveles: 50, 150, 250 y 350.
- Varianza σ^2. Tres niveles: 1, 256 y 961.
- Tamaño de muestra n. Cinco niveles: 20, 70, 120, 170 y 220.
- Para el análisis conjunto, Estadísticas. Cinco niveles: Pearson, K-S, $G_n(20)$, $G_n(100)$ y $G_n(500)$.

Unidad de exploración: 1000 muestras de igual tamaño y generadas con los mismos parámetros.

Réplicas: seis.

6.1. Análisis conjunto

Los resultados de la exploración se examinaron mediante el análisis de varianza para la tasa total de rechazo y la tasa de rechazo en cada una de las regiones. De las tablas de los cinco análisis de varianza se compilaron resultados parciales (véase tabla 3).

Tabla 3: Cuadrados medios para las variables tasa total de rechazos y tasa de rechazos por regiones.

		CUADRADOS MEDIOS				
C de V	gl	TTR	TRR_1	TRR_2	TRR_3	TRR_4
Estadística	4	0.00911823 **	0.00357893 **	0.02928230 **	0.00017162 **	0.01331699 **
μ	3	0.00010199	0.00000069	0.00003909	0.00009976 **	0.00000456
n	4	0.00016816 **	0.00002949 **	0.00009993 **	0.00001706	0.00001728 **
σ^2	2	0.00012652	0.00000208	0.00001031	0.00005176 *	0.00000035
Estadística $*$ μ	2	0.00001374	0.00000115	0.00000632	0.00001433	0.00000160
Estadística $*$ n	6	0.00041797 **	0.00001996 **	0.00013184 **	0.00007314 **	0.00003017 **
Estadística $*$ σ^2	8	0.00003572	0.00000202	0.00003649	0.00000608	0.00000262
$\mu * n$	12	0.00009158 *	0.00000206	0.00004961 *	0.00004758 **	0.00000425
$\mu * \sigma^2$	6	0.00010781 *	0.00000130	0.00013801 **	0.00000613	0.00000226
$n * \sigma^2$	8	0.00005493	0.00000195	0.00004491	0.00007442 **	0.00000498
Error	1719	0.00004300	0.00000164	0.00002485	0.00001557	0.00000284
Total	1799					

* Diferencias estadísticas significativas al 5 %.

** Diferencias estadísticas significativas al 1 %.

Con base en el contenido de la anterior tabla, se puede afirmar que la simulación sugiere que las tasas medias de rechazo son estadísticamente distintas de acuerdo con el tipo de estadística que se utilice, tanto de manera general como vistas las tasas en cada una de las regiones establecidas. Este resultado se confirma en el análisis específico por cada estadística (tabla 4).

Tabla 4: Prueba de comparación de medias de Scheffé al 95 % para las variables tasa total de rechazos y tasa de rechazos por regiones, teniendo en cuenta la estadística.

Estadística	TTR	TRR_1	TRR_2	TRR_3	TRR_4
Pearson	0.0514444 A	0.00721389 A	0.0144278 C	0.0159667 A	0.0138361 A
Kolmogorov-Smirnov	0.0387139 C	0.00084444 B	0.0228667 B	0.0140139 C	0.0009889 B
$G_n(20)$	0.0491444 B	0.00000000 C	0.0341417 A	0.0149750 B	0.0000000 C
$G_n(100)$	0.0492056 B	0.00000000 C	0.0342028 A	0.0149750 B	0.0000000 C
$G_n(500)$	0.0492194 B	0.00000000 C	0.0342139 A	0.0149778 B	0.0000000 C
Diferencia mínima significativa	0.0015	0.0003	0.0011	0.0009	0.0003

Valores con la misma letra no presentan diferencias estadísticas significativas.

El ordenamiento presentado en la tabla 4, producto de la prueba de Scheffé permite concluir que en términos generales la estadística propuesta ocuparía un puesto intermedio entre las estadísticas de Pearson y K-S; igualmente las tres particiones elegidas para la estadística producen resultados muy similares.

Especialmente por región, la estadística propuesta mostró en la simulación su superioridad tanto en la región R_1 como en la región R_4, frente a las dos estadísticas que se comparó; en la región R_3, la estadística tiene el mismo comportamiento que cuando se la analiza de manera global. La debilidad de la estadística propuesta se manifiesta en la región R_2, en la cual es superada por las estadísticas de Pearson y K-S, al presentar las mayores tasas de rechazo.

También se deduce de la tabla 3 que la simulación sugiere que el tamaño de la muestra es un factor que interviene de manera importante en la tasa de rechazo tanto individualmente como de manera colectiva con cada uno de los otros factores. La significancia de la interacción particular con el factor estadística, es la razón que motiva al análisis específico.

Finalmente, de la información acopiada en la tabla 4, se deduce el efecto del valor del parámetro μ y el valor del parámetro σ^2 en la tasa de rechazo en la región R_3, como era lo esperado, puesto que corresponde a la región que bajo normalidad presenta mayor discrepancia entre el promedio de la muestra y el promedio poblacional y entre la varianza de la muestra y la varianza poblacional.

6.2. Análisis específico

Los resultados de la exploración se examinaron mediante el análisis de varianza para la tasa total de rechazo y para la tasa de rechazo en cada una de las regiones, teniendo en cuenta el tipo de estadística empleada para el ajuste. De las tablas de los cinco análisis de varianza se compilaron resultados parciales (véase las tablas 5 a 9).

En las tablas 5 y 6 se pone de manifiesto el efecto que tiene el tamaño de la muestra en la evaluación de la tasa de rechazo en el caso que para el ajuste se emplearon las estadísticas de Pearson y de K-S.

En las tablas 7 a 9 se puede apreciar que la estadística propuesta no depende del tamaño de la muestra ni del tamaño de los parámetros μ y σ^2.

Tabla 5: Cuadrados medios para las variables tasa total de rechazos y tasa de rechazos por regiones empleando para el ajuste la estadística de Pearson.

		CUADRADOS MEDIOS				
C de V	gl	TTR	TRR_1	TRR_2	TRR_3	TRR_4
n	4	0.00029378 **	0.00010506 **	0.00002100	0.00014290 **	0.00013304 **
μ	3	0.00004367	0.00000448	0.00000553	0.00003558	0.00000837
σ^2	2	0.00006738	0.00000882	0.00002712	0.00001383	0.00000914
$n*\mu$	12	0.00008359	0.00000856	0.00002277	0.00000847	0.00002230 *
$n*\sigma^2$	8	0.00002797	0.00000796	0.00001159	0.00000544	0.00001580
$\mu*\sigma^2$	6	0.00000466	0.00000706	0.00001162	0.00001945	0.00000377
Error	319	0.00005029	0.00000721	0.00001377	0.00001645	0.00001197
Total	359					

* Diferencias estadísticas significativas al 5 %.
** Diferencias estadísticas significativas al 1 %.

Tabla 6: Cuadrados medios para las variables tasa total de rechazos y tasa de rechazos por regiones empleando para el ajuste la estadística de K-S.

		CUADRADOS MEDIOS				
C de V	gl	TTR	TRR_1	TRR_2	TRR_3	TRR_4
n	4	0.00147881 **	0.00000429 **	0.00052488 **	0.00014165 **	0.00000408 *
μ	3	0.00000394	0.00000081	0.00002627	0.00001424	0.00000176
σ^2	2	0.00001059	0.00000135	0.00004157	0.00002788	0.00000084
$n*\mu$	12	0.00003926	0.00000089	0.00002072	0.00001893	0.00000161
$n*\sigma^2$	8	0.00003075	0.00000118	0.00002075	0.00003500 *	0.00000129
$\mu*\sigma^2$	6	0.00001383	0.00000056	0.00001963	0.00000738	0.00000181
Error	319	0.00003705	0.00000092	0.00002352	0.00001625	0.00000123
Total	359					

* Diferencias estadísticas significativas al 5 %.
** Diferencias estadísticas significativas al 1 %.

Tabla 7: Cuadrados medios para las variables tasa total de rechazos y tasa de rechazos por regiones empleando para el ajuste la estadística $G_n(20)$.

		CUADRADOS MEDIOS				
C de V	gl	TTR	TRR_1	TRR_2	TRR_3	TRR_4
n	4	0.00002307	0.0	0.00002694	0.00000836	0.0
μ	3	0.00003724	0.0	0.00001124	0.00003603	0.0
σ^2	2	0.00006984	0.0	0.00003370	0.00001186	0.0
$n*\mu$	12	0.00002902	0.0	0.00001889	0.00001675	0.0
$n*\sigma^2$	8	0.00002474	0.0	0.00002464	0.00002343	0.0
$\mu*\sigma^2$	6	0.00005346	0.0	0.00004731	0.00000175	0.0
Error	319	0.00004528	0.0	0.00003095	0.00001610	0.0
Total	359					

Tabla 8: Cuadrados medios para las variables tasa total de rechazos y tasa de rechazos por regiones empleando para el ajuste la estadística $G_n(100)$.

		CUADRADOS MEDIOS				
C de V	gl	TTR	TRR_1	TRR_2	TRR_3	TRR_4
n	4	0.00002211	0.0	0.00002691	0.00000836	0.0
μ	3	0.00003587	0.0	0.00001034	0.00003594	0.0
σ^2	2	0.00006180	0.0	0.00002762	0.00001143	0.0
$n * \mu$	12	0.00002948	0.0	0.00001901	0.00001685	0.0
$n * \sigma^2$	8	0.00002266	0.0	0.00002381	0.00002345	0.0
$\mu * \sigma^2$	6	0.00005367	0.0	0.00004635	0.00000186	0.0
Error	319	0.00004501	0.0	0.00003069	0.00001612	0.0
Total	359					

Tabla 9: Cuadrados medios para las variables tasa total de rechazos y tasa de rechazos por regiones empleando para el ajuste la estadística $G_n(500)$.

		CUADRADOS MEDIOS				
C de V	gl	TTR	TRR_1	TRR_2	TRR_3	TRR_4
n	4	0.00002225	0.0	0.00002755	0.00000835	0.0
μ	3	0.00003622	0.0	0.00001100	0.00003531	0.0
σ^2	2	0.00005979	0.0	0.00002625	0.00001109	0.0
$n * \mu$	12	0.00002994	0.0	0.00001908	0.00001677	0.0
$n * \sigma^2$	8	0.00002308	0.0	0.00002420	0.00002382	0.0
$\mu * \sigma^2$	6	0.00005300	0.0	0.00004566	0.00000184	0.0
Error	319	0.00004502	0.0	0.00003070	0.00001613	0.0
Total	359					

7. Resultados de la exploración con base en la distribución uniforme

Cuando se empleó la distribución uniforme, e independientemente de la estadística usada para el ajuste, solo se presentaron tasas de rechazo en las regiones uno y cuatro descritas anteriormente.

El análisis de los resultados de la exploración se llevó a cabo de manera análoga al análisis de los resultados de un diseño experimental con arreglo factorial de efectos fijos. Los elementos del análisis fueron:

Respuesta: tasa de rechazo de la hipótesis nula.

Factores:

- Estadísticas. Cinco niveles: Pearson, K-S, $G_n(20)$, $G_n(100)$ y $G_n(500)$.
- Tamaño de muestra n. Cinco niveles: 20, 70, 120, 170 y 220.

Unidad de exploración: 1000 muestras de igual tamaño.

Réplicas: seis.

7.1. Análisis de los resultados

Los resultados de la exploración se examinaron mediante el análisis de varianza para la tasa total de rechazo, tasa de rechazo en la región uno y tasa de rechazo en la región cuatro. De las tablas de los tres análisis de varianza se compilaron resultados parciales (tabla 10).

Tabla 10: Cuadrados medios para las variables tasa total de rechazos, tasa de rechazos región uno y tasa de rechazos región cuatro.

		CUADRADOS MEDIOS		
C de V	gl	TTR	TRR_1	TRR_4
Estadística	4	0.00074934 **	0.00133998 **	0.00087633 **
n	4	0.00009902 **	0.00029941 **	0.00005594 **
Estadística $*$ n	16	0.00002924	0.00003099 *	0.00002563 *
Error	120	0.00332488	0.00001627	0.00001275
Total	149			

* Diferencias estadísticas significativas al 5 %.
** Diferencias estadísticas significativas al 1 %.

Con base en la anterior tabla, se puede afirmar que la simulación sugiere que las tasas medias de rechazo son estadísticamente distintas de acuerdo con el tipo de estadística que se utilice, tanto de manera general como para las regiones uno y cuatro.

El resultado del análisis de varianza se confirma en la tabla 11, donde el ordenamiento presentado producto de la prueba de Scheffé permite concluir que la estadística de Pearson y la estadística propuesta en cuanto a la tasa total de rechazos no presentan diferencias estadísticas entre sí, siendo estas tasas mayores que para la estadística de K-S.

La estadística propuesta en la simulación presenta peor desempeño frente a las dos estadísticas con que se comparó en la región R_1, mientras que en la región R_4 tiene un desempeño similar a la estadística de K-S, ambas estadísticas con menores tasas de rechazo que la estadística de Pearson. Además las tres particiones elegidas para la estadística propuesta producen resultados similares (tabla 11).

Tabla 11: Prueba de comparación de medias de Scheffé al 95 % para las variables tasa total de rechazo, tasa de rechazos región uno y tasa de rechazos región cuatro, teniendo en cuenta la estadística.

Estadística	TTR	TRR_1	TRR_4
Pearson	0.049233 A	0.019533 B	0.0297000 A
Kolmogorov-Smirnov	0.038133 B	0.019867 B	0.0182667 B
$G_n(20)$	0.049333 A	0.031900 A	0.0174333 B
$G_n(100)$	0.049333 A	0.031900 A	0.0174333 B
$G_n(500)$	0.049333 A	0.031900 A	0.0174333 B
Diferencia mínima significativa	0.0043	0.0033	0.0029

Valores con la misma letra no presentan diferencias estadísticas significativas.

En la tabla 10 se puede apreciar también que la simulación sugiere que el tamaño de la muestra es un factor que interviene de manera importante en la tasa de rechazo tanto para la tasa total de rechazo como en cada una de las regiones. La interacción de los factores estadística y tamaño de la muestra presenta diferencias estadísticas únicamente para las tasas de rechazo en las regiones uno y cuatro.

En general en la tabla 12 se aprecia la prueba de Scheffé, donde se revela que el efecto del tamaño de la muestra no influye en las estadísticas de Pearson, K-S y la propuesta para la tasa de rechazo en la región R_4. Las estadísticas de Pearson y la propuesta no tienen influencia del tamaño de muestra en la tasa total de rechazos.

La estadística de K-S presenta influencia del tamaño de la muestra en cuanto a la tasa total de rechazo. Además esta estadística y la propuesta presentan influencia del tamaño de la muestra para la tasa de rechazos en la región R_1.

Tabla 12: Prueba de comparación de medias de Scheffé al 95 % para las variables tasa total de rechazos y tasa de rechazos por regiones, teniendo en cuenta la estadística y el tamaño de la muestra.

	TASA TOTAL DE RECHAZOS				
Tamaño de muestra	Pearson	K-S	$G_n(20)$	$G_n(100)$	$G_n(500)$
20	0.049833 A	0.030167 B	0.046500 A	0.046500 A	0.046500 A
70	0.047500 A	0.038000 AB	0.051667 A	0.051667 A	0.051667 A
120	0.050000 A	0.040667 A	0.048500 A	0.048500 A	0.048500 A
170	0.051167 A	0.039667 A	0.050167 A	0.050167 A	0.050167 A
220	0.047667 A	0.042167 A	0.049833 A	0.049833 A	0.049833 A
Diferencia mínima significativa	0.0133	0.0095	0.0103	0.0103	0.0103
	TASA DE RECHAZOS REGIÓN UNO				
Tamaño de muestra	Pearson	K-S	$G_n(20)$	$G_n(100)$	$G_n(500)$
20	0.020333 A	0.011833 B	0.025333 B	0.025333 B	0.025333 B
70	0.022667 A	0.020833 A	0.033667 AB	0.033667 AB	0.033667 AB
120	0.017333 A	0.021333 A	0.031667 AB	0.031667 AB	0.031667 B
170	0.019167 A	0.021500 A	0.035000 A	0.035000 A	0.035000 A
220	0.018167 A	0.023833 A	0.033833 AB	0.033833 AB	0.033833 AB
Diferencia mínima significativa	0.0081	0.0068	0.0086	0.0086	0.0086
	TASA DE RECHAZOS REGIÓN CUATRO				
Tamaño de muestra	Pearson	K-S	$G_n(20)$	$G_n(100)$	$G_n(500)$
20	0.029500 A	0.018333 A	0.021167 A	0.021167 A	0.021167 A
70	0.024833 A	0.017167 A	0.018000 A	0.018000 A	0.018000 A
120	0.032667 A	0.019333 A	0.016833 A	0.016833 A	0.016833 A
170	0.032000 A	0.018167 A	0.015167 A	0.015167 A	0.015167 A
220	0.029500 A	0.018333 A	0.016000 A	0.016000 A	0.016000 A
Diferencia mínima significativa	0.0112	0.0056	0.0062	0.0062	0.0062

8. Conclusiones

Este trabajo recurrió a la simulación como medio para explorar el comportamiento de la estadística propuesta. En tal sentido, se derivaron de dicha simulación y por tanto su alcance está limitado a las condiciones particulares elegidas.

Se encontró que la estadística propuesta tuvo un desempeño intermedio para la evaluación del ajuste a la distribución normal, superando a la estadística de Pearson, pero superada por la estadística de K-S.

Al realizar la evaluación del ajuste al modelo normal se encontró que la estadística propuesta no depende del tamaño de la muestra como sí sucede cuando se emplean las estadísticas de Pearson y K-S.

La estadística propuesta y la estadística de Pearson en la evaluación del ajuste al modelo uniforme, presentaron mayor tasa total de rechazos que la estadística de K-S.

En el ajuste a la distribución uniforme al emplear las estadísticas de Pearson y la propuesta, el tamaño de la muestra no influyó en la tasa total de rechazos como sí sucedió con la estadística de K-S.

La estadística propuesta presentó un comportamiento diferente de acuerdo con el modelo empleado para el ajuste. Además en los dos modelos usados las particiones elegidas para la estadística producen resultados similares o iguales.

Finalmente este trabajo propone realizar el estudio teórico de la distribución de la estadística. También realizar la valoración del comportamiento de la estadística cuando la hipótesis de nulidad no se cumple.

[Recibido: febrero de 2009 — Aceptado: mayo de 2010]

Referencias

Cabaña, A. & Quiroz, J. (2005), 'Using the Empirical Moment Generating Function in Testing for the Weibull an the Type I Extreme Distributions', *Test* **12**(2), 417–431.

Collander, R. & Chalfant, J. (1985), An alternative Approach to Decisions under Uncertainty: The Empirical Moment Generating Function, Work Paper, Universiy of California, Department of Agricultural and Resources Economics.

Conover, W. (1999), *Practical Nonparametric Statistics*, third edn, John Wiley, United States of America.

Csörgo, S. (1981), 'Multivariate Empirical Characteristic Function. Zeitschrift für Wahrscheintichkeits-Theoric und Verwandte Gabiete', *Annals of Probability* (55), 203–229.

Csörgo, S. (1986), 'Testing for Normality in Arbitrary Dimension', *Annals of Statistics* **14**(2), 708–723.

Csörgo, S. & Heathcote, C. (1987), 'Testing for Symetry', *Biometrika* **1**(74), 177–184.

Epps, T. & Pulley, L. (1985), Two Test of fit Based on the Sample Characteristic Function with Applications to Exponentiality, Paper presented at the annual meeting of the American Statistical Association, Las Vegas, NE.

Feuerverger, A. & Mureika, A. (1977), 'The Empirical Characteristic Function and its Applications', *Annals of Statistics* (5), 88–97.

Heathcote, C. (1977), 'The Integrated Squared Error Estimation of Parameters', *Biometrika* **2**(64), 255–264.

Koutrouvelis, I. (1980), 'A goodness-of-fit Test of Simple Hypotheses Based on the Empirical Characteristic Function', *Biometrika* **1**(67), 238–240.

Koutrouvelis, I. & Kellermeier, J. (1981), 'A goodness-of-fit Test Based on the Empirical Characteristic Function when Parameters must be Estimated', *Journal of the American Statistical Association* **2**(43), 173–176.

Meintanis, S. (2007), 'A Kolmogorov-Smirnov Type Test for Skew Normal Distributions Based on the Empirical Moment Generating Function', *Journal of Statistical Planning and Inference* **137**(8), 2681–2688.

Mood, A., Graybill, F. & Boes, D. (1974), *Introduction to the Theory of Statistics*, third edn, McGraw-Hill, Singapore.

Muñoz, L. (1998), Bondad de ajuste empleando la función generadora de momentos, Tesis de maestría, Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia.

Quand, R. & Ramsey, J. (1978), 'Estimating Mixtures of Normal Distributions and Switching Regressions', *Journal of the American Statistical Association* (73), 730–741.

Revista Colombiana de Estadística
Junio 2010, volumen 33, no. 1, pp. 127 a 147

Synthesizing the Ability in Multidimensional Item Response Theory Models

Habilidad sintética en modelos multidimensionales de teoría de respuesta al ítem

Álvaro Mauricio Montenegro Díaz[a], Edilberto Cepeda[b]

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

Abstract

A central problem associated with Multidimensional Item Response Theory (MIRT) Models is the impossibility of ordering the examinees. In this paper, we obtain two unidimensional synthetic indices that are optimal linear combinations of the ability vector. These synthetic indices are similar to the reference composite commonly used in MIRT models, but they are easier to calculate and interpret. The synthetic indices are compared with the unidimensional ability obtained when a multidimensional data is fitted with an unidimensional IRT (UIRT) model.

***Key words*:** Binary response, Item response theory, Index, Multidimensional data, Synthetic estimator, Latent trait.

Resumen

Un problema central asociado con los Modelos Multidimensionales de Teoría de Respuesta al Item (TRIM) es la imposibilidad de ordenar a los examinados. En este artículo, se obtienen dos índices sintéticos unidimensionales que son combinaciones lineales óptimas del vector de habilidades. Estos índices sintéticos son semejantes a la composición de referencia comúnmente usada en los modelos TRIM, pero son más fáciles de calcular. Los índices sintéticos se comparan con el parámetro de habilidad obtenido cuando un conjunto de datos multidimensionales es ajustado con un modelo TRI unidimensional.

***Palabras clave*:** respuesta binaria, teoría de respuesta al ítem, índice, datos multidimensionales, estimador sintético, trazo latente.

[a]Assistant professor. E-mail: ammontenegrod@unal.edu.co
[b]Associate professor. E-mail: ecepedac@unal.edu.co

1. Introduction

This research originated in recent results obtained by Levine (2003) and Carroll & Levine (2007) in the context of Multidimensional Item Response Theory. They proved that any multidimensional model has unidimensional submodels that are equivalent to the original model.

The unidimensional item response theory models (UIRT) consist of models according to which the interactions of persons with items can be adequately represented by a unique parameter describing the characteristics of the person (Reckase 2009).

The multidimensional item response theory (MIRT) models are based on the assumption that people require more than one basic ability to response correctly to an item in a test. There are two major types of MIRT models-the compensatory models Reckase (1985, 1997, 2007) and the non-compensatory or partial compensatory models (Sympson 1978). In this research, we only refer to the compensatory MIRT models, that we will call them simply MIRT models.

Stout (1990) introduced the concept of essential unidimensionality. The central idea of Stout is that even though the ability space is multidimensional, the set of items used in a test may be sensitive, mainly to differences along one of the dimensions. The statistical tests to assess unidimensionalidad can reject the unidimensional assumption. Stout et al. (1999) developed DIMTEST, a procedure to test the assumption of essential unidimensionality of the person's ability.

Several authors tried to determine the relationship between the ability vector $\boldsymbol{\theta}$ and the unidimensional ability denoted θ, obtained by fitting a unidimensional model to data that were generated from multidimensional models. Ansley & Forsyth (1985) examined the unidimensional estimates for two dimensional data using a noncompensatory model. They studied situations in which the θ's were correlated with correlations values of 0.0, 0.3, 0.6, 0.9 and 0.95. Way et al. (1988) also compared the effects of using a UIRT model to estimate two dimensional data for both the noncompensatory and the compensatory MIRT model. Reckase (1990, 1986) reported that, in some situations, where a multidimensional data matrix was fitted with a UIRT model, the dimensionality and the difficulty were confused.

Ackerman (1989) reported that, in his simulations, the unidimensional estimate of θ was highly correlated with $(\theta_1 + \theta_2)/2$; this correlation was better when the correlation of the abilities was increased. Reckase & Ackerman (1988) suggested to build unidimensional tests from multidimensional items by grouping the items that measure more similar the linear combinations of abilities.

Folk & Green (1989) stated that $\widehat{\theta}$ is strongly related to some optimal combination of θ_1 and θ_2, even for simulate samples with low correlations. Doody (1985) reported studies about the robustness of unidimensional fitting applied to two dimensional data. Zhao et al. (2002), in a simulated study of computerized adaptive tests, founded similar results. As Ackerman, they compared the ability $\widehat{\theta}$ with $(\theta_1 + \theta_2)/2$. Walker & Beretvas (2003) compared multidimensional and uni-

dimensional proficiency using real data from a large-scale mathematics test and obtained similar results.

Recently, Sheng (2007) showed that when each one of the items measures essentially only one ability , a multi-unidimensional model fits better the data.

In this paper, we review the previous works about the important issue of synthesizing the latent ability vector in MIRT models. We derive two optimal linear combinations of the components of the ability vector, which are synthetic indices of the abilities. Through a simulation study, we compared the proposed indices with the others proposed previously, and we infer that all the synthetic indices are similar. Our indices are easier to compute and interpret by the experts. The synthetic indices obtained are also estimations of the linear combination of the latent ability vector that is best measured by a test. We state how the covariance of the latent ability vector affects the synthetic index. Finally, we infer through a second simulation study that when the multidimensional data is fitted with a unidimensional model, the unidimensional latent ability is precisely the synthetic index of the ability vector. In the paper, the terms latent ability and latent trait are used as synonyms.

2. The geometrical facts

When a UIRT model is used to fit data set, it is usual to assume a normal standard distribution for the abilities of the individuals. Clearly, if the data is multidimensional, the correlation matrix of the ability vector is the identity matrix. But, if really the correlation matrix is not the identity, the obvious question is what happens with the item and the ability parameters when this information about the correlation matrix of the abilities is omitted?

The works reviewed in Section 1 suggested us that when a data set is generated from a MIRT model and the correlation matrix of the ability vector is not the identity, a unidimensional model can fit well the data. This lead us to conjecture that if the unidimensional model is used with the assumption that the abilities have a normal standard distribution, the correlation matrix of the abilities transforms the direction of the items in such a way that in the extreme case all of them must be aligned. The direction of an item is discussed in Section 3. Also, the results reported in Section 1 seems to suggest that in the extreme case the unique direction of the items is just $\frac{1}{\sqrt{d}}\mathbf{1}_d$, where d is the dimension of the ability space. This conjecture lead us to propose and prove the results of this Section. The required facts from d-dimensional geometry can be consulted in the Appendix.

Theorem 1. *Let Σ be a $d \times d$ symmetric and positive definite matrix, such that all its diagonal elements are 1 and the off-diagonal elements are nonnegative. Let $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ be unitary vectors of $\mathbb{R}^d$, such that all their elements are nonnegative. Let $|\boldsymbol{\Sigma}|$ be the determinant of $\boldsymbol{\Sigma}$, then*

$$\left[\frac{\boldsymbol{\beta_1^t\Sigma\beta_2}}{\sqrt{(\boldsymbol{\beta_1^t\Sigma\beta_1})(\boldsymbol{\beta_2^t\Sigma\beta_2})}}\right]^2 \geq 1-|\boldsymbol{\Sigma}|(1-(\boldsymbol{\beta_1^t\beta_2})^2) \tag{1}$$

Proof. Let $\Sigma^{1/2}$ be the squared root of Σ. Let $\gamma_i = (\Sigma^{1/2}\beta_i)/\sqrt{\beta_i^t\Sigma\beta_i}$, $i = 1, 2$. Then, the vectors γ_1 and γ_2 are unitary. Let $vol(\gamma_1, \gamma_2)$ be the volume of the parallelotope determined by the vectors γ_1 and γ_2. From equations (34), (36) and (37) in the Appendix, it follows that

$$vol^2(\gamma_1,\gamma_2) = 1 - \left[\frac{\boldsymbol{\beta_1^t\Sigma\beta_2}}{\sqrt{(\boldsymbol{\beta_1^t\Sigma\beta_1})(\boldsymbol{\beta_2^t\Sigma\beta_2})}}\right]^2 \tag{2}$$

and

$$vol^2(\gamma_1,\gamma_2) = \frac{|\boldsymbol{\Sigma}|vol^2(\boldsymbol{\beta_1},\boldsymbol{\beta_2})}{(\boldsymbol{\beta_1^t\Sigma\beta_1})(\boldsymbol{\beta_2^t\Sigma\beta_2})} \tag{3}$$

The properties of matrix Σ permit us to conclude that $\beta_i^t\Sigma\beta_i \geq 1,\ i = 1, 2$. The result follows from this fact and also from the previous two equations and Lemma 3 in the Appendix. □

Corollary 1. *Under the conditions of Theorem 1, we have that*

$$\left[\frac{\boldsymbol{\beta_1^t\Sigma\beta_2}}{\sqrt{(\boldsymbol{\beta_1^t\Sigma\beta_1})(\boldsymbol{\beta_2^t\Sigma\beta_2})}}\right] \geq (\boldsymbol{\beta_1^t\beta_2}) \tag{4}$$

Proof. The result follows from the fact that $|\boldsymbol{\Sigma}| \leq 1$. □

In the next result, we assume that $\boldsymbol{\Sigma_m^{1/2}}$ is the squared root of $\boldsymbol{\Sigma_m}$.

Theorem 2. *Let $\boldsymbol{\Sigma_m}$ be a sequence of $d \times d$ matrices that have the same properties than $\boldsymbol{\Sigma}$ in Theorem 1, and such that their determinants are decreasing and that $|\boldsymbol{\Sigma_m}| \to 0$ as $m \to \infty$. Let $\boldsymbol{\beta_m} = \boldsymbol{\Sigma_m^{1/2}\beta}$, where $\boldsymbol{\beta}$ is any not-zero vector, where all of its components are nonnegative. Thus, $\boldsymbol{\beta_m}/||\boldsymbol{\beta_m}|| \to \frac{1}{\sqrt{d}}\mathbf{1}_d$, where $\mathbf{1}_d$ is the vector with 1's at all its components.*

Proof. It is easy to see that $|\boldsymbol{\Sigma}| = 0$, if and only if $\boldsymbol{\Sigma} = J_d$, where J_d is the matrix with 1's in all of its components. Thus, $\boldsymbol{\Sigma_m^{1/2}} \to \frac{1}{\sqrt{d}}J_d$. □

Suppose that $\boldsymbol{\Sigma}$ is a correlation matrix. It can be shown that if the off-diagonal elements of the matrix $\boldsymbol{\Sigma}$ become large, then the determinant of the matrix $\boldsymbol{\Sigma}$ decreases due to the relationship

$$|\ \boldsymbol{\Sigma}\ | = (1 - R^2_{p.1\cdots p-1})(1 - R^2_{p-1.1\cdots p-2})\cdots(1 - R^2_{2.1})$$

where $R^2_{k\cdots d}$ is the squared multiple correlation coefficient between the variable k and the following variables. See, for example (Peña 2002), (Peña & Rodríguez 2003).

From theorems 1 and 2 we conclude that, if the off-diagonal elements of the matrix $\mathbf{\Sigma}$ are increased, all the transformed vectors $\mathbf{\Sigma_m^{1/2}\beta}$ have a smaller angle between them than the original vectors, and the respective transformed normalized vectors have a greater orthogonal projection between them. Also, all the transformed vectors are conducted toward the unitary vector $\frac{1}{\sqrt{d}}\mathbf{1}_d$. In the limit case, all the transform vectors align with that unitary vector.

3. The nature of the items in the MIRT model

In this Section, we show that any item in a compensatory MIRT model is essentially unidimensional and prove that the item response hypersurface of an item in a MIRT model is monotonic along any direction. This property allows exchanging the item response function (IRF) and the item response hypersurface (IRHS) as in the unidimensional case, but also permits us to determine what an item really measures in a MIRT model.

In the logistic two parameter model (Baker & Seok-Ho 2004), (Bock 1972), (Bock & Jones 1968), (Hambleton & Rogers 1991), the probability of a correct response for the unidimensional case is given by

$$p_j(\theta_i) = P(X_{ij} = 1 \mid \theta_i, a_j, b_j) = \frac{1}{1 + e^{-a_j(\theta_i - b_j)}} \tag{5}$$

where X_{ij} is the response of person i to item j; $X_{ij} = 1$ if the examinee i responses correctly to item j, and $X_{ij} = 0$ otherwise; θ_i is the unidimensional ability parameter for person i. The scale parameter a_j is called the discrimination parameter of item j, and b_j is the difficulty or position parameter of item j.

The function $f_j(\theta) = p_j(\theta)$ is called the item response function (IRF) and its graph is the item response curve (IRC). Note that

$$f_j(b_j) = \frac{1}{2} \tag{6}$$

and,

$$f'(b_j) = \frac{1}{4}a_j \tag{7}$$

so, except by the term 1/4, a_j represents the slope of the IRC at the point b_j.

In the classical compensatory MIRT model, there is more that one ability measured by a test. Let $\boldsymbol{\theta}_i$ be a vector of $\mathbb{R}^d$ that represents the ability vector of the examinee i. The parameters of item j in this case are: $\boldsymbol{a}_j$, a vector of $\mathbb{R}^d$ related with the discrimination of the item and γ_j, a scalar related with the difficulty of the item. The probability that an examinee with ability vector $\boldsymbol{\theta}_i$ responses correctly to item j is given by

$$P(X_{ij} = 1 \mid \boldsymbol{\theta}_i, \boldsymbol{a}_j, \gamma_j) = \frac{1}{1 + e^{-(\boldsymbol{a}_j^t\boldsymbol{\theta}_i + \gamma_j)}} \tag{8}$$

The component θ_{ik} of $\boldsymbol{\theta_i}$ represents the ability of the person i in the k-th dimension. The interpretations of $\boldsymbol{a}_j$'s and γ_j's parameters are a little different from those in the unidimensional case. Reckase (1985, 1997, 2007) states that the MIRT model does not provide a direct interpretation about the parameters $\boldsymbol{a}_j$ and γ_j. In this case, the item response function $f_j(\boldsymbol{\theta}) = p_j(\boldsymbol{\theta})$ is a multivariate function and its graph is a hypersurface. Let α_j be the norm of the vector $\boldsymbol{a}_j$, that is,

$$\alpha_j = \sqrt{\sum_{k=1}^{d} a_{jk}^2}$$

where the a_{jk}'s are the components of vector $\boldsymbol{a}_j$. Then, the vector $\boldsymbol{a}_j$ can be rewritten as

$$\boldsymbol{a}_j = \alpha_j \boldsymbol{\beta}_j \tag{9}$$

where $\boldsymbol{\beta_j} = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jd})^t$, $\beta_{jk} = a_{jk}/\alpha_j$. Clearly, $\boldsymbol{\beta}_j$ is a unitary vector of $\mathbb{R}^d$. Thus, the model given by Equation (8) can be rewritten as

$$P(X_{ij} = 1 \mid \boldsymbol{\theta}_i, \alpha_j, \boldsymbol{\beta}_j, b_j) = \frac{1}{1 + e^{-\alpha_j(\boldsymbol{\beta_j^t \theta_i} - b_j)}} \tag{10}$$

where $b_j = -\gamma_j/\alpha_j$. Reckase (1985) defined the value α_j as the multidimensional discrimination (MDISC) parameter and the value b_j as the multidimensional difficulty (MDIFF) parameter. He showed that α_j is the slope at the point of the steepest slope in the direction specified by the vector $\boldsymbol{\beta}_j$, called the *direction of item j*.

Additionally, he proved that b_j is the distance from the origin to the point of the steepest slope. We will show in this Section why the MDISC and MDIFF names are justified.

At this point, we introduce the concept of item response hypersurface. In the UIRT models, one may use the item response function (IRF) and its geometrical representation-the item response curve (IRC)-almost interchangeable. In the multidimensional case, however, the matter is not so straightforward.

First, we fix some notations. For any $v \in \mathbb{R}^d$, the ray of v is defined to be the line $\mathbb{R} \cdot v$ in $\mathbb{R}^d$ determined by $\mathbb{R} \cdot v = \{tv \in \mathbb{R}^d \mid t \in \mathbb{R}\}$. Similarly, for $v, w \in \mathbb{R}^d$ the directed line going through w is defined by

$$w + \mathbb{R} \cdot v = \{w + tv \in \mathbb{R}^d \mid t \in \mathbb{R}\}$$

Definition 1. A dichotomous item response hypersurface is a d-dimensional smooth submanifold M of $\mathbb{R}^d \times [0,1]$, so that for any two vectors $v, w \in R^d$ the intersection of $(w + \mathbb{R} \cdot v) \times [0,1]$ and M is the graph of a monotonic function $f_{v,w} : w + R \cdot v \to [0,1]$.

We shall use the notation $f_v = f_{v,0}$. Definition 1 and the notation were taken from Antal's paper (Antal 2007).

Lemma 1. *The graph of the item response function given by,*

$$f(\boldsymbol{\theta}) = \frac{1}{1 + e^{-\alpha_j(\boldsymbol{\beta}_j^t\boldsymbol{\theta}+\gamma_j)}} \tag{11}$$

is a dichotomous item response hypersurface.

Proof. Let $\boldsymbol{v}, \boldsymbol{w}$ be two arbitrary vectors of $\mathbb{R}^d$ and consider the line given by $\boldsymbol{\eta}(t) = \boldsymbol{w} + t\boldsymbol{v}, t \in \mathbb{R}$. Clearly, $\boldsymbol{\beta}_j^t\boldsymbol{\eta}(t) = \boldsymbol{\beta}_j^t w + (\boldsymbol{\beta}_j^t\boldsymbol{v})t$ is a monotonic function of t and then $f(\boldsymbol{\eta}(t))$ is a monotonic function along the direction $\boldsymbol{v}$ through $\boldsymbol{w}$. □

As a consequence of Lemma 1, the item response function (11) defines a dichotomous item response hypersurface and the MIRT model is completely determined by these hypersurfaces.

Lemma 2. *The item response function $f_j(\boldsymbol{\theta})$ of a MIRT model is constant in the orthogonal complement of vector $\boldsymbol{\beta}_j$.*

Proof. For any vector $\boldsymbol{\eta}$ in the orthogonal subspace of $\boldsymbol{\beta}_j$, $\boldsymbol{\beta}_j^t\boldsymbol{\eta} = 0$, so, $f_j(\boldsymbol{\eta}) = 1/(1 + e^{-\alpha_j\gamma_j})$. □

The next Corollary can be directly proven.

Corollary 2. *Given $\boldsymbol{w} \in \mathbb{R}^d$, the item response function $f_j(\boldsymbol{\theta})$ is constant in the hyperplane parallel to the orthogonal complement of vector $\boldsymbol{\beta_j}$ that contains $\boldsymbol{w}$.*

This Corollary is well-known. It states that the contours of equiprobability are hyperplanes, and that all of them are parallel. However, the important fact is that they are orthogonal to the vector $\boldsymbol{\beta_j}$. Theorem 3 is the main result of this Section. It establishes that the item response function $f_j(\boldsymbol{\theta})$ is a trivial extension of a unidimensional item response function (UIRF). According to Equation (9) we will use the expression $\boldsymbol{a_j} = \alpha_j\boldsymbol{\beta_j}$ in the Proof. It is not necessary, but is useful to understand the result.

Theorem 3. *The multidimensional IRF $f_j(\boldsymbol{\theta})$ of a MIRT model is a trivial extension of a classical UIRF.*

Proof. Let $\boldsymbol{\theta}$ be a vector in $\mathbb{R}^d$, and let $\{\boldsymbol{\beta}_j, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{d-1}\}$ be a normed orthogonal basis of $\mathbb{R}^d$ that contains the vector $\boldsymbol{\beta}_j$. Then, there exist real numbers $t, t_1, \ldots, t_{d-1}$ such that

$$\boldsymbol{\theta} = t\boldsymbol{\beta}_j + t_1\boldsymbol{v}_1 + \cdots + t_{d-1}\boldsymbol{v}_{d-1}$$

then,

$$\boldsymbol{\beta}_j^t\boldsymbol{\theta} = (\boldsymbol{\beta}_j^t\boldsymbol{\beta}_j)t = t \tag{12}$$

Hence,

$$f_j(\boldsymbol{\theta}) = \frac{1}{1 + e^{-\alpha_j\boldsymbol{\beta}_j^t\boldsymbol{\theta}-\gamma_j}} = \frac{1}{1 + e^{-\alpha_j t-\gamma_j}} = \frac{1}{1 + e^{-\alpha_j(t-b_j)}} = p_{\boldsymbol{\beta}_j}(t) \tag{13}$$

□

The notation $p_{\boldsymbol{\beta}_j}$ is used to emphasize the direction $\boldsymbol{\beta}_j$, and that $f_j(\boldsymbol{\theta})$ is an extension of a UIRF. Theorem 3 shows an explicit way to construct the hypersurface defined by $f_j(\boldsymbol{\theta})$ from a unidimensional IRC. Let $p_j(t)$ be the UIRF defined by

$$p_j(t) = \frac{1}{1 + e^{-\alpha_j(t-b_j)}} \tag{14}$$

The function $p_j(t)$ can be trivially extended to a multivariate function by $p_j(t_1, \ldots, t_d) = p_j(t_1)$. The original hypersurface is obtained by a rigid rotation of the hypersurface defined by $p_j(t_1, t_2, \ldots, t_d)$ on the hyperplane defined by the canonical vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$, which aligns vector $\boldsymbol{e}_1$ with vector $\boldsymbol{\beta}_j$. This is a general result, since any rotation in $\mathbb{R}^d$ can be done in this way. The theory of rigid rotations in d-dimensional spaces can be found in Aguilera & Pérez-Aguila (2004) and Mortari (2001). A direct and important consequence of Theorem 3 is stated in the next Corollary.

Corollary 3. *Let's suppose that the directions of all items in a MIRT model coincide, that is, $\boldsymbol{\beta}_i = \boldsymbol{\beta}$, for all i. Then, the model is essentially unidimensional. In other words, the MIRT model is a trivial extension of a UIRT model.*

The result of Corollary 3 was first proven by Stout and Reckase in a paper presented at a meeting of the Psychometric Society (Reckase & Stout 1995). Reckase (2009) reproduced the result (Theorem 1, page 197).

Other useful properties of the MIRT model follow. On the hyperplane $\boldsymbol{\beta}_j^t\boldsymbol{\theta} - b_j = 0$ we have that

$$f_j(\boldsymbol{\theta}) = 1/2 \tag{15}$$

It is straightforward to verify that for all $\boldsymbol{\theta}$ in that hyperplane

$$\frac{\partial f_j}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{1}{4}\alpha_j\boldsymbol{\beta}_j \tag{16}$$

So, as in Equation (7), the parameter α_j, except by the constant $1/4$, is the slope of the IRHS for all $\boldsymbol{\theta}$ in the hyperplane $\boldsymbol{\beta}_j^t\boldsymbol{\theta} - b_j = 0$. The slope in the direction $\boldsymbol{\beta}_j$ is maximum when the IRHS crosses the hyperplane (Reckase 1985).

From equations (10), (15) and (16), we can conclude that IRHS of item j in the MIRT model is a trivial extension of a unidimensional IRC whose parameters of discrimination and difficulty are respectively α_j and $b_j = -\gamma_j/\alpha_j$. Also, it is clear that item j measures the linear combination of the abilities given by $\boldsymbol{\beta}_j^t\boldsymbol{\theta}$.

4. Synthesizing the latent ability

A unidimensional synthetic index of the latent trait vector in a MIRT model is usually called a composite. The formal concept is given in Definition 2.

Definition 2. A *composite* Θ_β of the complete latent trait vector $\boldsymbol{\Theta}$ is a linear combination of $\boldsymbol{\Theta}$, that is $\Theta_\beta = \boldsymbol{\beta^t\Theta} = \sum_{k=1}^{d}\beta_k\Theta_k$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_d)^t$ is a constant vector called the direction of the composite Θ_β. If $Var(\Theta_\beta) = 1$, Θ_β will be called a normalized composite.

Some authors have done theoretically developments to construct a unidimensional synthetic index of the latent trait vector. Yen (1985) considered an approximation of a MIRT model by a UIRT, using a least squares (LS) approach. She used the objective function

$$G\left[\widehat{a},\widehat{b},\widehat{\theta}\right]=\sum_{i}\sum_{j}\left[\widehat{a_j}(\widehat{\theta_i}-\widehat{b_j})-\alpha_j\boldsymbol{\beta}_j^t\boldsymbol{\theta}_i-\gamma_j\right]^2 \tag{17}$$

where $\widehat{a}=(\widehat{a_1},\ldots,\widehat{a_p})^t$, $\widehat{b}=\left(\widehat{b_1},\ldots,\widehat{b_p}\right)^t$, $\widehat{\theta}=\left(\widehat{\theta_1},\ldots,\widehat{\theta_N}\right)^t$ are the corresponding parameters in an approximate UIRT model, and p is the number of items. The respective LS equations do not have a closed solution. Then, she assumed the particular case where $\boldsymbol{\beta}_i=\boldsymbol{\beta}, i=1,\ldots,p$, to obtain the solution

$$\widehat{\theta_i}=\frac{\boldsymbol{\beta}^t\boldsymbol{\theta}_i}{\sqrt{\boldsymbol{\beta^t\Sigma\beta}}} \tag{18}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the latent trait $\boldsymbol{\theta}$. This result can be obtained as a direct consequence of Theorem 3, since in this particular case all directions of the items coincide, and then we have essentially a UIRT along the direction $\boldsymbol{\beta}$.

Let $\{X_1,\ldots,X_m\}$ be a subtest, and let $Y=\sum_{j=1}^m X_j$ be the subtest number correct score, let $\xi(\boldsymbol{\theta})=\sum_{j=1}^m p_j(\theta)$ be the true subtest score. Zhang & Stout (1999) defined the direction of score Y as the vector $\boldsymbol{\xi}$ that maximizes the *expected multidimensional critical ratio* (EMCR) defined as

$$EMCR(\boldsymbol{\xi},\boldsymbol{\theta};Y)=E\left[\frac{\nabla_{\boldsymbol{\xi}}\xi(\boldsymbol{\theta})}{[Var(Y\mid\boldsymbol{\theta})]^{\frac{1}{2}}}\right] \tag{19}$$

where $\nabla_{\boldsymbol{\xi}}\xi(\boldsymbol{\theta})$ is the directional derivative of the true score $\xi(\boldsymbol{\theta})$ in the direction $\boldsymbol{\xi}$. The EMCR function gives the average discrimination power of the observed score Y in the direction $\boldsymbol{\xi}$. They showed that vector $\boldsymbol{\xi}$ is given by

$$\boldsymbol{\xi}=\sum_{j=1}^{m}\omega_i\boldsymbol{\beta}_j \tag{20}$$

where $\omega_i=cE\left[H_i'(\alpha_j\boldsymbol{\beta}_j^t\boldsymbol{\theta}+\gamma_j)\,/\,\sqrt{Var(Y\mid\boldsymbol{\theta})}\right]$. $H_i(\cdot)$ represents the item response function. Clearly, the direction $\boldsymbol{\xi}$ in Equation (20) depends on the response function, and it is an average on the latent trait population. In this case, $\boldsymbol{\xi}^t\boldsymbol{\theta}$ is the composite that is best measured by the subtest. The reference direction $\boldsymbol{\xi}$ was called the direction of the subtest.

Wang (1985, 1986) constructed a unidimensional approximation to a multidimensional data matrix that he called the *reference composite trait*. He used the transformation $y=ln[p/(1-p)]$, the item logistic score, and rewrote the logistic MIRT model as

$$\boldsymbol{Y}=\boldsymbol{\theta}\boldsymbol{A}^t+\mathbf{1}_K\boldsymbol{\gamma}^t \tag{21}$$

where $\boldsymbol{\theta}$ is the matrix of the latent traits, $\boldsymbol{A}$ is the $K\times d$ matrix of the discrimination parameters in the MIRT model, K is the number of items, $\mathbf{1}_K$ is the K-vector of

ones and $\boldsymbol{\gamma}$ is the vector associated with the difficulty. The objective function in this case is the trace of $\left(\boldsymbol{Y}-\widehat{\boldsymbol{Y}}\right)^t\left(\boldsymbol{Y}-\widehat{\boldsymbol{Y}}\right)$, where $\widehat{\boldsymbol{Y}}=\boldsymbol{G}\boldsymbol{H}^t+\boldsymbol{1}_K\boldsymbol{\gamma}^t$. Here $\boldsymbol{G}$ is the unidimensional latent trait in the approximate model and $\boldsymbol{H}$ is the vector of discrimination item parameters in that model. Observe that it is assumed that the difficult parameters do not change. Wang showed that in this case

$$\boldsymbol{G}=\boldsymbol{\theta}\boldsymbol{\omega}, \tag{22}$$

where $\boldsymbol{\omega}$ is the (unit-length) eigenvector associated with the largest eigenvalue of the matrix $\boldsymbol{A^tA}$.

Theorem 3 states that all items in a compensatory MIRT model are essentially unidimensional. Then, the multidimensional nature of a MIRT model can only be attributed to the item directions $\boldsymbol{\beta_j}$. Corollary 3 states that when the directions of all items coincide, the model is a trivial extension of a UIRT model. These results encouraged us to derive a unidimensional synthetic ability in a different way than Yen, Wang, and Zhang and Stout.

We observed that if all $\boldsymbol{\beta_j}$'s are the same, and $\boldsymbol{\beta_j}=\boldsymbol{\beta}$, $j=1,\ldots,K$, where K is the number of items in the test, then Equation (10) reduces to

$$P(X_{ij}=1\mid\boldsymbol{\theta_i},\alpha_j,\boldsymbol{\beta},b_j)=\frac{1}{1+e^{-\alpha_j(\boldsymbol{\beta^t\theta_i}-b_j)}} \tag{23}$$

that is a trivial extension of an UIRT model, where each one of the items measures the same composite of the abilities given by $\boldsymbol{\beta}^t\boldsymbol{\theta}_i$. This observation suggests looking for a vector $\boldsymbol{\beta}$ that summarizes the $\boldsymbol{\beta}_j$'s. Since these vectors are all unitary, they are in the unitary hypersphere of $\mathbb{R}^d$. Also, we can assume that the components of the vectors $\boldsymbol{\beta_j}$ are all non-negative, then all the vectors are in the same hyper-quadrant. Therefore, it is reasonable to expect that the vector that summarizes all the $\boldsymbol{\beta_j}$'s is the same hyper-quadrant of the unitary hypersphere. This leads us to search the vector $\boldsymbol{\beta}$ by optimizing the objective function given by

$$h(\beta_1,\ldots,\beta_d)=\sum_{l=1}^{d}\sum_{k=1}^{K}(\beta_{kl}^2-\beta_l^2)^2 \tag{24}$$

whose solution is the unitary vector given by

$$\beta_l=\sqrt{\left(\frac{1}{K}\sum_{j=1}^{K}\beta_{kl}^2\right)} \tag{25}$$

We will denote the solution vector in this case as $\boldsymbol{\beta}_h$.

Alternatively, it is also reasonable to optimize the objective function

$$g(\beta_1,\ldots,\beta_d)=\sum_{l=1}^{d}\sum_{k=1}^{K}(\beta_{kl}-\beta_l)^2 \tag{26}$$

whose solution, considering a unitary vector is given by

$$\beta_l = \frac{\sum_{k=1}^{K} \beta_{kl}}{||\sum_{k=1}^{K} \beta_{kl}||} \quad l = 1, \ldots, d \tag{27}$$

The solution vector in this case will be denoted as $\boldsymbol{\beta}_g$.

We finish this Section with an approach about the role of the latent trait correlation matrix. It is usual to assume that the abilities of the examinees in a test constitute a sample drawn from a normal d-dimensional distribution $N(0, \boldsymbol{\Sigma})$. The marginal EM estimation is based on this assumption (Bock & Aitkin 1981).

To obtain an identifiable model, most of the programs written to estimate MIRT models assume that $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, where $\boldsymbol{I}_d$ is the identity matrix. Examples are TESTFACT (Wilson et al. 1987) and recently the ltm package (Rizopoulos 2006). Those are examples of programs that use this assumption. In general, this is not a realistic situation. Software NOHARM (Fraser 1988) estimates the item parameters and the correlation matrix, but it does not estimate the latent abilities. Bégin & Glass (2001) and De la Torre & Patz (2005) proposed MCMC algorithms that simultaneously estimate the item parameters, the latent abilities and the matrix $\boldsymbol{\Sigma}$. In this work, we assume only that the diagonal elements are all 1. This assumption defines a common scale along the canonical axis of the ability space. Ackerman (1989) stated that, in the case where the matrix $\boldsymbol{\Sigma}$ is not the identity, the difficulty and the dimensionality can be confused.

The usual assumption that the correlation matrix is the identity probably resulting the problem mentioned by Ackerman. Let's assume that $\boldsymbol{\theta}$, the latent ability of the examinees, is a sample from a normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. Then $\boldsymbol{\Sigma}$ has the stochastic representation $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Upsilon}$, where $\boldsymbol{\Upsilon}$ has a multivariate normal standard distribution, and $\boldsymbol{\Sigma}^{\mathbf{1/2}}$ is the squared root of $\boldsymbol{\Sigma}$. Then, we have that

$$\boldsymbol{\beta}^t\boldsymbol{\theta} = \left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}\right)^t \boldsymbol{\Upsilon} \tag{28}$$

Hence, when in the estimation process it is assumed that the correlation matrix is the identity matrix, the direction of each item is estimated in a transformed space determined by $\boldsymbol{\Sigma}^{1/2}$. Equation (28) shows a procedure to compute the reference direction when the correlation matrix is available.

It is clear that if $\boldsymbol{\theta}$ has a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, any composite $\boldsymbol{\beta}^t\boldsymbol{\theta}$ has a different scale, since $Var(\boldsymbol{\beta}^t\boldsymbol{\theta}) = \boldsymbol{\beta}^t\boldsymbol{\Sigma}\boldsymbol{\beta}$. In this case, the reference direction must be computed from the transformed vectors $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}$, and the synthetic ability must be computed using the transformed ability $\boldsymbol{\Upsilon} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta}$.

5. Simulation study

Two simulations were developed to evaluate and compare the synthetic indices $\boldsymbol{\beta}_h^t\boldsymbol{\theta}$ and $\boldsymbol{\beta}_g^t\boldsymbol{\theta}$. This indices are compared with the synthetic indices $\boldsymbol{\xi}^t\boldsymbol{\theta}$ and $\boldsymbol{\omega}^t\boldsymbol{\theta}$, where $\boldsymbol{\xi}$ is the reference direction obtained by Zhang and Stout and $\boldsymbol{\omega}$ is the reference direction obtained by Wang.

5.1. Comparison of the reference directions

Conceptually, the construction of the reference direction of Wang and the reference directions proposed in this paper are very similar. The construction of the reference direction proposed by Zhang and Stout is different.

The vector $\boldsymbol{\xi}$ is the direction in which the total score Y has maximum discriminating power (Zhang & Stout 1999). The vector $\boldsymbol{\omega}$ maximizes the projection of the direction of the items along of it. The vector $\boldsymbol{\beta}_h$ essentially minimizes the angle between this reference direction and the direction of the items. The vector $\boldsymbol{\beta}_g$ minimizes the distance between this reference direction and the direction of the items as points of the latent space. However, all the directions are very similar as we show in this Section.

To review this fact, we generate a set of 60 vector directions in the 3-dimensional latent space. We generate 4 clusters, each one with fifteen directions. To do that, we fixed four directions: $\boldsymbol{b}_1 = (1.0, 1.0, 1.0)^t$, $\boldsymbol{b}_2 = (1.0, 0.2, 0.1)^t$, $\boldsymbol{b}_3 = (0.3, 1.0, 0.1)^t$ and $\boldsymbol{b}_4 = (0.25, 0.25, 1.0)^t$. Then, we generate the vectors of each cluster by adding random noise to each component of the vectors $\boldsymbol{b}$. The noise is smaller in cluster 1 and is augmented progressively until cluster 4.

In a second step, we compute the reference directions $\boldsymbol{\omega}$, $\boldsymbol{\beta}_h$ and $\boldsymbol{\beta}_g$, from all the item directions and from the item directions in each cluster. Additionally, we simulate values of MDISC and MDIFF parameters to generate all the item parameters for 60 items, and then we also computed the reference direction $\boldsymbol{\xi}$ from all the items and from the items in each cluster. We used a logistic response function, and Equation (20).

We considered two different distributions for the latent ability. First, we assumed a 3-variate normal standard distribution and then a 3-variate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & 0.3 & 0.6 \\ 0.3 & 1.0 & 0.4 \\ 0.6 & 0.4 & 1.0 \end{pmatrix}$$

Tables 1 and 2 show the results. In Table 1 columns 3, 4 and 5 correspond to the components of the reference directions for the first distribution of the latent abilities and columns 7, 8 and 9 are the components of the reference directions for the second distribution. Finally, we evaluate the synthetic indices comparing them with the original composites. We computed the quantity

$$\boldsymbol{\Delta}_v = \frac{1}{K_v} \sum_{j=1}^{Kv} E\left[\left|\boldsymbol{\beta}^t \boldsymbol{\theta}_v - \boldsymbol{\beta}_{vj}^t \boldsymbol{\theta}\right|\right] \tag{29}$$

where v is respective the cluster, and K_v the size of the cluster.

Table 2 shows the scalar product between the four reference directions.

TABLE 1: Reference directions for each cluster. Columns 3, 4 and 5 are the components of the reference directions for the distribution $N(\mathbf{0}, \boldsymbol{I})$ and columns 7, 8 and 9 for the distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$.

cluster	vector	comp.1	comp.2	comp.3	Δ	comp.1	comp.2	comp.3	Δ
all	ξ	0.5614	0.6035	0.5663	0.4236	0.4029	0.5666	0.7188	0.2805
	ω	0.5887	0.6071	0.5337	0.4269	0.4254	0.5679	0.7047	0.2784
	β_h	0.5825	0.5843	0.5651	0.4237	0.4544	0.5758	0.6797	0.2822
	β_g	0.5870	0.6025	0.5408	0.4260	0.4187	0.5686	0.7081	0.2787
1	ξ	0.5637	0.5730	0.5949	0.0472	0.4335	0.5509	0.7131	0.0267
	ω	0.5618	0.5735	0.5962	0.0470	0.4324	0.5512	0.7136	0.0266
	β_h	0.5621	0.5731	0.5964	0.0470	0.4328	0.5512	0.7133	0.0267
	β_g	0.5870	0.6025	0.5408	0.0689	0.4187	0.5686	0.7081	0.0303
2	ξ	0.9667	0.2079	0.1495	0.0687	0.7367	0.3718	0.5649	0.0380
	ω	0.9675	0.2157	0.1318	0.0694	0.7388	0.3786	0.5576	0.0385
	β_h	0.9631	0.2206	0.1542	0.0700	0.7380	0.3797	0.5578	0.0386
	β_g	0.9675	0.2157	0.1317	0.0694	0.7388	0.3786	0.5576	0.0385
3	ξ	0.2500	0.9605	0.1225	0.0994	0.2193	0.8837	0.4134	0.0718
	ω	0.2488	0.9619	0.1133	0.0986	0.2195	0.8863	0.4078	0.0716
	β_h	0.2634	0.9534	0.1475	0.1046	0.2312	0.8827	0.4092	0.0711
	β_g	0.2488	0.9619	0.1135	0.0986	0.2195	0.8863	0.4078	0.0716
4	ξ	0.1317	0.2341	0.9632	0.1677	0.1301	0.2750	0.9526	0.1595
	ω	0.1571	0.2412	0.9577	0.1683	0.1533	0.2846	0.9463	0.1591
	β_h	0.2102	0.2825	0.9360	0.1757	0.2008	0.3148	0.9277	0.1628
	β_g	0.1581	0.2412	0.9575	0.1683	0.1537	0.2843	0.9464	0.1592

TABLE 2: Scalar product between the reference vectors.

cluster	$< \boldsymbol{\xi}.\boldsymbol{\beta}_h >$	$< \boldsymbol{\omega}.\boldsymbol{\beta}_h >$	$< \boldsymbol{\beta}_g.\boldsymbol{\beta}_h >$	$< \boldsymbol{\omega}.\boldsymbol{\beta}_g >$	$< \boldsymbol{\xi}.\boldsymbol{\beta}_g >$	$< \boldsymbol{\xi}.\boldsymbol{\omega} >$
all	0.9979	0.9992	0.9989	1.0000	0.9998	0.9997
1	1.0000	1.0000	0.9997	0.9997	0.9997	1.0000
2	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
3	0.9999	0.9999	0.9999	1.0000	1.0000	1.0000
4	0.9964	0.9982	0.9983	1.0000	0.9997	0.9997
mean	0.9988	0.9995	0.9994	0.9999	0.9998	0.9999

5.2. Comparison of the synthetic ability $\boldsymbol{\beta}_h^t\boldsymbol{\theta}$ with the unidimensional ability of a UIRT model

To evaluate $\boldsymbol{\beta}_h^t\boldsymbol{\theta}$ as a synthetic index of the latent trait vector, we used the following strategy. It is reasonable to expect that the synthetic index of the ability is a good unidimensional summary of the ability vector. Then, if a multidimensional data set is fitted with a unidimensional model, the unidimensional estimative of the ability parameter must be also an estimative of the synthetic index.

In this Section, we evaluate the synthetic index $\boldsymbol{\beta}_h^t\boldsymbol{\theta}$ in forty simulated examples. For clarity, the subscript h will be omitted. All examples are based on 2-dimensional models. One hundred item parameters were simulated as follows. First, the MDISC (the α_j's) parameters were generated from a uniform distribution in the range $[.4, 2]$. Second, the parameters b_j were generated from a normal distribution $N(0, 1)$. Third, the angles that determine the direction of the vectors

$\boldsymbol{\beta_j}$ were generated from a uniform distribution in the range $[5, 50]$. The MDISC parameters were generated in the range $[.4, 2]$ because, this is the usual range of this parameter in real tests. Different prior distributions are assumed for these parameters as a log-normal or a non-informative positive flat distribution (Sheng 2008). We used the last option. The range of angles was chosen to yield a more disperse set of angles as possible. In the simulation of the previous Section, the simulated angles were less dispersed in each cluster.

A sample of 4000 examinees was drawn from the normal bivariate distribution $N(0, \boldsymbol{I_d})$. To examine the impact of the correlation between the $\boldsymbol{\theta}$'s, we respectively introduced correlations of $0, .3, .6$ and $.9$. In all cases, the diagonal elements were 1, thus $\boldsymbol{\Sigma}$ is always a correlation matrix. Also, in all cases a normal standard distribution is assumed for the ability vectors in the estimation process.

Finally, for each correlation matrix a set of binary responses were generated as follows: for each ability vector and each parameter set, the probability of a correct response was computed using Equation (8). Then, a random number u was obtained, from the uniform distribution in the range $[0, 1]$. If the probability of correct response was greater or equal than u the value 1 was assigned to the response. Otherwise, the 0 value was assigned (Kromrey et al. 1999).

We fitted 10 unidimensional models for each set of responses using the ltm package (Rizopoulos 2006). First, we took the first 10 items; then, we took the first 20 items and so until all items were taken. Table 3 shows the main results.

A number of statistical indices were calculated at the simulate sample level to evaluate the synthetic index $\boldsymbol{\beta^t\theta}$. Let $\boldsymbol{\beta_k}$, $k = 1, \ldots, 40$ be the vector $\boldsymbol{\beta}$ in each one of the 40 simulations. Let $\widehat{\theta}_i$ be the estimation of the ability parameter obtained, when the multidimensional data was fitted with the unidimensional model. The bias index can be expressed as

$$bias_k = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{\beta_k^t\theta_i} - \widehat{\theta}_i\right) \quad (k = 1, \ldots, 40) \tag{30}$$

The error index included is the mean absolute error (mae) defined as

$$mae_k = \frac{1}{N}\sum_{i=1}^{N}\left|\boldsymbol{\beta_k^t\theta_i} - \widehat{\theta}_i\right| \quad (k = 1, \ldots, 40) \tag{31}$$

To evaluate the precision of the *mae* index, we included the standard deviation sd_k of values $\left|\boldsymbol{\beta_k^t\theta_i} - \widehat{\theta}_i\right|$. A fidelity index was computed, the Pearson product-moment rho correlation, denoted by ρ. Additionally, we computed the least squares (LS) - fitting between the values $\boldsymbol{\beta_k^t\theta_i}$ and $\widehat{\theta}_i$. We took the synthetic index as the explanatory variable. The c-values were the coefficients and the R^2-values the corresponding R^2 statistics of the fitting in each simulation.

Also, we compared the estimations $\widehat{\theta}_i$ with $(\theta_1 + \theta_2)/2$. The indices mae^1 and c^1 were computed by replacing the values $\boldsymbol{\beta_k^t\theta_i}$ with the values $(\theta_1 + \theta_2)/2$ in the previous respective indices.

TABLE 3: Statistical indices to evaluate the synthetic index $\boldsymbol{\beta^t\theta}$. The value p is the number of items, σ is the correlation between the θ's, β_1 and β_2 are the components of vector $\boldsymbol{\beta}$, γ_1 and γ_2 are the minimum and maximum angles of the vectors $\boldsymbol{\beta_j}$ with respect to the horizontal in each simulation.

p	σ	β_1	β_2	γ_1	γ_2	bias	mae	sd	ρ	c	R^2	mae^1	c^1	mae^α
10	0.0	0.94	0.34	5.2	34.3	0.022	0.38	0.30	0.88	0.73	0.77	0.40	0.93	0.10
10	0.3	0.90	0.44	13.6	37.1	0.024	0.35	0.28	0.89	0.74	0.80	0.36	0.99	0.06
10	0.6	0.85	0.53	22.4	39.6	0.026	0.35	0.27	0.90	0.75	0.81	0.32	1.03	0.08
10	0.9	0.78	0.63	34.2	42.5	0.026	0.33	0.28	0.90	0.76	0.82	0.29	1.07	0.07
20	0.0	0.87	0.49	5.2	49.3	0.006	0.30	0.24	0.92	0.83	0.85	0.35	1.11	0.12
20	0.3	0.84	0.55	13.6	48.1	0.005	0.28	0.22	0.93	0.84	0.87	0.31	1.16	0.08
20	0.6	0.80	0.60	22.4	47.1	0.007	0.27	0.21	0.94	0.86	0.89	0.29	1.20	0.07
20	0.9	0.75	0.66	34.2	46.0	0.004	0.26	0.21	0.94	0.86	0.89	0.27	1.22	0.08
30	0.0	0.88	0.47	5.2	49.3	0.002	0.27	0.21	0.94	0.87	0.88	0.36	1.16	0.10
30	0.3	0.84	0.54	13.6	48.1	0.007	0.25	0.19	0.95	0.88	0.90	0.32	1.21	0.06
30	0.6	0.80	0.59	22.4	47.1	0.000	0.23	0.18	0.96	0.89	0.91	0.29	1.25	0.07
30	0.9	0.75	0.66	34.2	46.0	0.022	0.23	0.18	0.96	0.90	0.91	0.27	1.28	0.08
40	0.0	0.88	0.47	5.2	49.3	0.003	0.25	0.19	0.95	0.90	0.90	0.35	1.20	0.10
40	0.3	0.84	0.54	13.6	48.1	−0.008	0.23	0.18	0.96	0.91	0.92	0.31	1.25	0.07
40	0.6	0.80	0.59	22.4	47.1	−0.011	0.21	0.16	0.96	0.92	0.93	0.29	1.29	0.08
40	0.9	0.75	0.66	34.2	46.0	0.033	0.21	0.16	0.96	0.93	0.93	0.27	1.32	0.09
50	0.0	0.89	0.46	5.2	49.3	−0.002	0.22	0.17	0.96	0.92	0.92	0.35	1.24	0.10
50	0.3	0.85	0.53	13.6	48.1	−0.008	0.21	0.16	0.97	0.94	0.93	0.32	1.30	0.08
50	0.6	0.81	0.59	22.4	47.1	0.001	0.19	0.15	0.97	0.96	0.94	0.30	1.35	0.10
50	0.9	0.76	0.65	34.2	46.0	0.038	0.19	0.15	0.97	0.97	0.94	0.29	1.37	0.12
60	0.0	0.88	0.48	5.2	50.0	−0.007	0.20	0.16	0.97	0.95	0.93	0.35	1.28	0.11
60	0.3	0.84	0.55	13.6	48.7	−0.014	0.19	0.15	0.97	0.97	0.94	0.32	1.34	0.11
60	0.6	0.80	0.60	22.4	47.5	0.014	0.18	0.14	0.97	1.00	0.95	0.32	1.40	0.15
60	0.9	0.75	0.66	34.2	46.1	0.040	0.18	0.14	0.98	1.01	0.95	0.30	1.42	0.16
70	0.0	0.87	0.49	5.2	50.0	0.006	0.19	0.15	0.97	0.97	0.94	0.35	1.33	0.13
70	0.3	0.84	0.55	13.6	48.7	−0.034	0.18	0.14	0.98	1.00	0.95	0.33	1.38	0.14
70	0.6	0.80	0.60	22.4	47.5	−0.012	0.17	0.14	0.98	1.03	0.96	0.33	1.45	0.18
70	0.9	0.75	0.66	34.2	46.1	−0.001	0.17	0.14	0.98	1.05	0.96	0.33	1.48	0.22
80	0.0	0.88	0.48	5.2	50.0	0.020	0.18	0.14	0.97	0.99	0.95	0.36	1.35	0.14
80	0.3	0.84	0.54	13.6	48.7	−0.045	0.17	0.14	0.98	1.02	0.96	0.34	1.41	0.15
80	0.6	0.80	0.60	22.4	47.5	0.074	0.18	0.14	0.98	1.05	0.96	0.34	1.48	0.20
80	0.9	0.75	0.66	34.2	46.1	0.011	0.17	0.13	0.98	1.06	0.96	0.33	1.50	0.22
90	0.0	0.88	0.48	5.2	50.0	0.002	0.18	0.14	0.98	1.01	0.95	0.37	1.37	0.14
90	0.3	0.84	0.54	13.6	48.7	−0.040	0.17	0.14	0.98	1.04	0.96	0.35	1.44	0.17
90	0.6	0.80	0.60	22.4	47.5	0.076	0.18	0.14	0.98	1.07	0.96	0.35	1.50	0.21
90	0.9	0.75	0.66	34.2	46.1	−0.068	0.18	0.14	0.98	1.09	0.96	0.36	1.55	0.26
100	0.0	0.88	0.48	5.2	50.0	0.009	0.17	0.13	0.98	1.02	0.96	0.37	1.39	0.15
100	0.3	0.84	0.55	13.6	48.7	−0.054	0.17	0.14	0.98	1.07	0.96	0.36	1.47	0.19
100	0.6	0.80	0.60	22.4	47.5	0.079	0.18	0.15	0.98	1.10	0.96	0.37	1.54	0.25
100	0.9	0.75	0.66	34.2	46.1	−0.076	0.19	0.15	0.99	1.12	0.97	0.38	1.59	0.29

Finally, in Table 3 we included the mae^{α} index for the α-parameters. This index was computed as

$$mae_k^{\alpha} = \frac{1}{p}\sum_{j=1}^{p} |\alpha_{jk} - \widehat{\alpha}_{jk}| \tag{32}$$

for each simulation k. The value $\widehat{\alpha}_{jk}$ is the slope parameter of the unidimensional model estimated in simulation k.

6. Discussion

Carroll & Levine (2007) and Levine (2003) proved that any MIRT model can be approximated by unidimensional models. However, their approximate models are non-parametric and the response functions are not necessarily monotone.

In this paper, we reviewed the main aspects concerning to synthesize the latent ability vector in compensatory MIRT models. We used composites, that are linear combinations of the latent trait vector.

Theorem 3 shows that each item j in a MIRT model is essentially unidimensional along the direction given by the vector $\boldsymbol{\beta_j}$. Item j measures the composite ${\boldsymbol{\beta_j}}^t\boldsymbol{\theta_i}$. Then, each item measures a different linear combination of the $\boldsymbol{\theta_i}$, unless all the vectors $\boldsymbol{\beta_j}$ have the same direction.

In realistic problems, where a test measures more than one latent trait, the components of the latent trait vector are correlated. However, Equation (28) shows that if the latent trait random vector $\boldsymbol{\theta}$ has multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, then any composite $\boldsymbol{\beta}^t\boldsymbol{\theta}$ can be rewritten as $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^t\boldsymbol{\Upsilon}$, where $\boldsymbol{\Upsilon}$ has a normal standard distribution. This transformation has two important consequences. First, according to Corollary 1, the transformation induced by $\boldsymbol{\Sigma^{1/2}}$ shrinks the direction vectors $\boldsymbol{\beta}_j$. Second, if a vector $\boldsymbol{\beta}$ is unitary, the composite $\boldsymbol{\beta}^t\boldsymbol{\Upsilon}$ is normalized, and any normalized composite has a normal standard distribution.

In Section 3, we stated that each item is essentially unidimensional along the direction of the item. In Corollary 3 we proved that if all the directions of the items coincide, the test is essentially unidimensional along the unique direction of the items.

The important issue about how to obtain a unidimensional synthetic index of the multidimensional latent trait was discussed in Section 4. Previous works of Yen, Wang and Zhang and Stout was reviewed. Wang and Zhang and Stout proposed two alternative synthetic indices, called respectively reference composite ($\boldsymbol{\omega}^t\boldsymbol{\theta}$) and the direction of the test ($\boldsymbol{\xi}^t\boldsymbol{\theta}$). We proposed two new synthetic indices: $\boldsymbol{\beta}_h^t\boldsymbol{\theta}$ and $\boldsymbol{\beta}_g^t\boldsymbol{\theta}$. Computing these alternative indices can be easier that Computing the previous indices, and they are more natural and easy to use by the experts.

Tables 1 and 2 of the first simulation study (Section 5.1) show that all the reference directions are very similar. This is not surprising, because although the constructions are different, the objective in all cases is the same: to obtain a synthetic index of the multidimensional latent trait. However, if we joint all the

results, we can conclude additionally that each one of the reference composites is an approximation to the best composite that is measured by a subtest. This fact, is illustrated in Section 4, where we compared the theoretical synthetic index $\boldsymbol{\beta}_h^t\boldsymbol{\theta}$ with the unidimensional latent trait index obtained by fitting a multidimensional data set with a UIRT model.

7. Conclusions and future work

From a geometrical point of view, we showed in this paper how in tests that measure more than one latent trait the multidimensional latent trait vectors can be synthesized to obtain unidimensional measures of the examinees. The approach can be applied to subtests obtained from clusters of the items, or to the full test.

In the paper, nothing was stated about the item parameters that are estimated when a unidimensional model is used to fit a multidimensional data set. We showed that the correlation in the latent trait vector must be considered when a synthetic latent trait must be computed. In this case, a right computation implies to transform the direction of the items by a non orthogonal projection. But, in this scenery, the open question is: how must be modified the MDISC and MDIFF parameters to conserve approximately the same probability of response?. In other words, what is the relationship between the item parameters of the MIRT model and the item parameters of the UIRT when a unidimensional model is used to fit a multidimensional data set?.

Acknowledgements

This research is based on the doctoral dissertation in Statistics at the Universidad Nacional de Colombia of Álvaro Mauricio Montenegro Díaz, under the advice of Edilberto Cepeda. The research was supported by Universidad Nacional de Colombia through a study grant. The calculations were computed with R (Team 2008). The estimations were run with the ltm-package (Rizopoulos 2006) written for R. The code can be obtained by e-mailing the authors. The authors thank Professor Mark Reckase for sending the manuscripts of Min Wang to us, and thank the referees by their suggestions to improve the paper.

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

References

Ackerman, T. (1989), 'Unidimensional IRT Calibration of Compensatory and Non-compensatory Multidimensional Items', *Applied Psychological Measurement* **13**, 113–127.

Aguilera, A. & Pérez-Aguila, R. (2004), General n-Dimensional Rotations, *in* 'WSCG SHORT Communications papers proceedings', Union Agency - Science Press, Czech Republic.

Ansley, T. & Forsyth, R. (1985), 'An Examination of the Charateristics of Unidimensional IRT Parameter Estimates Derived from Two Dimensional Data', *Applied Psychological Measurement* **9**, 27–48.

Antal, T. (2007), 'On multidimensional item response theory: a coordinate free approach', *Electronic Journal of Statistics* **1**, 290–306.

Baker, F. B. & Seok-Ho, K. (2004), *Item Response Theory*, 2 edn, Marcel Decker Inc.

Bégin, A. & Glass, C. A. (2001), 'MCMC estimation and some Model-Fit Analysis of Multidimensional IRT Models', *Psychometrika* **66**(4), 541–562.

Bock, R. D. (1972), 'Estimating Item Parameters and Latent Ability when Responses are Scored in Two or more Nominal Categories', *Psychometrika* **37**, 29–51.

Bock, R. D. & Aitkin, M. (1981), 'Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm', *Psychometrika* **46**, 443–459.

Bock, R. D. & Jones, L. V. (1968), *The Measurement and Prediction of the Judge and Choice*, San Francisco: Holden-Day.

Carroll, J. Williams, B. & Levine, M. (2007), 'Multidimensional Modeling with Unidimensional Approximations', *Journal of Mathematical Psychology* **51**.

De la Torre, J. & Patz, R. (2005), 'Making the Most of what we have: A Practical Application of Multidimensional Item Response Theory in Test Scoring', *Journal of Educational and Behavioral Statistics* **30**(3), 295–311.

Doody, E. (1985), Examining the Effects of Multidimensional Data on Ability and Item Parameter Estimation using the Three-Parameter Logistic Model, *in* 'The 2002 annual meeting of American Educational Research Association', Chicago.

Folk, V. & Green, V. (1989), 'Adaptive Estimation when the Unidimensionality Assumption of IRT is Violated', *Applied Psychological Measurement* **13**, 373–389.

Fraser, C. (1988), 'NOHARM II: A Fortran Program for Fitting Unidimensional and Multidimensional Normal Ogive Models of Latent Trait Theory', The University of New England, Armidale, Australia.

Hambleton, R. K. Swaminathan, H. & Rogers, H. J. (1991), *Fundamentals of Item Response Theory*, Sage Publications, Newbury Park, United States.

Kendall, M. (1961), *A Course in the Geometry of n Dimensions*, Charles Griffin and Company Limited, London.

Kromrey, D., Parshall, C. & Chason, W. (1999), Generating Item Responses Bsed on Multidimensional item Response Theory, *in* 'SUGI 24', SAS.

Levine, M. (2003), 'Dimension in Latent Variable Models', *Journal of Mathematical Psychology* **47**, 450–466.

Mathai, M. (1999), 'Random p-Content of a p-Parallelotope in Euclideann-Space', *Advances in Applied Probability* **31**(2), 343–354.

Mortari, D. (2001), 'On the Rigid Rotation Concept in n-Dimensional Spaces', *Journal of the Astronautical Sciences* **49**(3), 401–420.

Peña, D. (2002), *Análisis de Datos Multivariantes*, McGraw Hill.

Peña, D. & Rodríguez, J. (2003), 'Descriptive Measures of Multivariate Scatter and Linear Dependence', *Journal of Multivariate Analysis* **85**(2), 361–374.

Reckase, M. (1985), 'The Difficulty of Test Items that Measure more than one Ability', *Applied Psycological Measurement* **9**(9), 401–412.

Reckase, M. (1990), Unidimensional Data from Multidimensional Data from Unidimensional Tests, *in* 'Paper presented at the annual meeting of American Educational Research Association', Boston.

Reckase, M. (1997), 'The Past and the Future of Multidimensional Item Response Theory', *Applied Psychological Measurement* **21**(1), 25–36.

Reckase, M. (2007), 'Multidimensional Item Response Theory', *Handbook of Statistics* **26**, 607–642.

Reckase, M. (2009), *Multidimensional Item Response Theory*, Statistics for Social and Behavior Sciences, Springer.

Reckase, M. & Ackerman, T. (1988), 'Building a Unidimensional Test Using Multidimensional Items', *Journal of Educational Measurement* **25**(3), 193–203.

Reckase, M., Carlson, J. & Ackerman, T. (1986), The Interpretation of the Unidimensional IRT Parameters when Estimate from Multidimensional Data, *in* 'Annual Meeting of Psychometrics Society', Toronto.

Reckase, M. & Stout, W. (1995), Conditions under which Items that Assess Multiple Ablilities will be fit by Unidimensional IRT Models, *in* 'The European meeting of Psychometric Society', Leyden, Holanda.

Rizopoulos, D. (2006), 'ltm: An R Package for Latent Variable Modeling and Item Response Theory Models', *Journal of Statistical Software* **17**(5), 1–25.

Sheng, Y. (2007), 'Comparing Multiunidimensional and Unidimensional Item Response Theory Models', *Educational and Psychological Measurement* **67**(6), 899–919.

Sheng, Y. (2008), 'Bayesian Multidimensional IRT Models with a Hierarchical Structure', *Educational and Psychological Measurement* **68**(3), 413–430.

Stout, W. (1990), 'A new Item Response Theory Modeling Approach with Applications to Unidimensionality Assessment and Ability Estimation', *Psychometrika* **55**, 293–325.

Stout, W., Douglas, B., Junker, B. & Roussos, L. (1999), DIMTEST, Computer software, The William Stout Institute for Measurement, Champaing, IL.

Sympson, J. (1978), A Model for Testing with Multidimensional Items, *in* 'Proceedings of the 1977 Computarized Adaptive Testing Conference', Minneapolis: University of Minnesota, Department of Psychology, pp. 82–98.

Team, R. D. C. (2008), *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

Walker, C. & Beretvas, S. (2003), 'Comparing Multidimensional and Unidimensional Proficiency Calssifications: Multidimensional IRT as Diagnostic Aid', *Journal of Educational Measurement* **40**(3), 255–275.

Wang, M. (1985), Fitting a Unidimensional Model to Multidimensional Item Response Data: The effect of latent space misspecification on the application of IRT, Research Report MW: 6-24-85, University of Iowa, Iowa City.

Wang, M. (1986), Fitting a Unidimensional Model to Multidimensional Item Response Data, The Office of Naval Research Contractors Meeting, Gartlingburg.

Way, W., Ansley, T. & Forsyth, R. (1988), 'The Comparative Effects of Compensatory and Noncompensatory Two-dimensional Data Items on Unidimensional IRT Estimates', *Applied Psychological Measurement* **12**, 239–252.

Wilson, D., Wood, R. & Gibbons, R. (1987), TESTFACT [Computer program], *in* 'Scientific Software', Mooresville IN.

Yen, W. (1985), 'Increasing Item Complexity: A Possible Cause of Scale Shrinkage for Unidimensional Item Response Theory', *Psychometrika* **50**(4), 399–410.

Zhang, J. & Stout, W. (1999), 'Conditional Covariance Structure of Generalized Compensatory Multidimensional Items', *Psychometrika* **64**, 129–152.

Zhao, J., McMorris, R. & Pruzek, R. (2002), The Robusnetss of the Unidimensional 3PL IRT when Applied to Two-Dimensional Data in Computarized Adaptive Testing, *in* 'The 2002 annual meeting of American Educational Research Association', New Orleans.

Appendix

For the concepts of n-dimensional geometry, see for example (Kendall 1961). Let $v_1, \ldots, v_d$ be an ordered set of vectors in $\mathbb{R}^n, n \geq d$. The parallelotope [1] with sides $v_1, \ldots, v_d$ is the convex hull created by this vectors. This parallelotope is denoted by $P(v_1, \ldots, v_d)$. It is well known that the volume or content of $P(v_1, \ldots, v_d)$ is

$$vol(v_1, \ldots, v_d) = |V^t V|^{1/2} \tag{33}$$

where $V = (v_1, \ldots, v_d)$, see for example (Mathai 1999). It is immediate that

$$vol(\lambda v_1, \ldots, v_d) = \lambda \cdot vol(v_1, \ldots, v_d) \tag{34}$$

Also, if S is region of $\mathbb{R}^n$ and Σ a $n \times n$ matrix, then

$$vol(\Sigma S) = |\Sigma| vol(S) \tag{35}$$

From Equation (35) it is straightforward that

$$vol(\Sigma v_1, \ldots, \Sigma v_d) = |\Sigma| \cdot vol(v_1, \ldots, v_d) \tag{36}$$

Lemma 3. *Let $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ be unitary vectors of $\mathbb{R}^n$, then*

$$vol^2(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}) = 1 - \boldsymbol{\beta_1^t \beta_2} \tag{37}$$

Proof. It follows directly from Equation (33). □

[1] The parallelotope is the generalization of a parallelepiped to $\mathbb{R}^d$

Revista Colombiana de Estadística
Junio 2010, volumen 33, no. 1, pp. 149 a 166

Un test de similitud entre dos secuencias dicotómicas ordenadas

A Similarity Test between Two Dichotomic Ordered Sequences

Ramón Giraldo Henao[a], Jimmy Corzo Salamanca[b]

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

Resumen

Se propone una prueba para la hipótesis de similitud de dos secuencias dicotómicas ordenadas. Un estudio de potencia basado en simulación indica que la prueba propuesta mantiene su tamaño bajo la hipótesis nula y que su potencia crece adecuadamente con el tamaño de muestra. Además la prueba tiene la misma potencia que la prueba del signo y supera en potencia a las pruebas basadas en las estadísticas de antirrachas y de Wilcoxon.

***Palabras clave*:** similitud, datos dicotómicos, potencia de una prueba.

Abstract

We propose a test for the hypothesis of similarity between dichotomic ordered sequences. A simulation study was carried out to estimate the power of the proposed test. It is shown that the test maintain its size under the null hypothesis and that its power increase with the considered alternative hypothesis. In addition the proposed test is as powerful as the sign test and overtakes the antiruns test and the Wilcoxon test.

***Key words*:** Similarity, Dichotomous data, Power of a test.

1. Introducción

En muchas áreas de la ciencia se llevan a cabo experimentos en los que se mide una variable respuesta de tipo dicotómico bajo dos tratamientos, en los que dicha respuesta tiene un orden específico asociado a bloques o covariables. Por ejemplo en farmacia cuando se realiza un bioensayo con el propósito de establecer diferencias entre machos y hembras respecto a su respuesta a un fármaco (vive/muere,

[a]Profesor Asociado. E-mail: rgiraldoh@unal.edu.co
[b]Profesor Asociado. E-mail: jacorzos@unal.edu.co

mejora/no mejora) y se considera como covariable la dosis del fármaco o la edad de los individuos. En psicología cuando se aplican dos test a los mismos individuos que aprueban o no aprueban el test y estos son de diferente edad o tienen diferente nivel de educación. En la industria cuando dos catadores califican como buenas o malas las muestras de algún producto y estas tienen diferente nivel de calidad. Es importante resaltar que en las tres situaciones descritas, aunque hay un orden en las respuestas, las observaciones pueden considerarse independientes puesto que estas mismas son evaluadas en individuos distintos. En este trabajo se presenta una estadística que permite realizar pruebas de hipótesis con información obtenida en experimentos análogos a los arriba mencionados, es decir en aquellos en los que la variable respuesta es binaria y tiene un orden implícito dado por una covariable, hay dos tratamientos y se asume independencia entre las observaciones.

El artículo se organiza como sigue: en la sección 2 se plantea el tipo de hipótesis de interés, se define la estadística de prueba y su distribución bajo la hipótesis nula. En la sección 3 se presentan teoremas respecto a los dos primeros momentos y se calcula la distancia entre la distribución de la estadística de prueba y la normal estándar para varios tamaños de sucesión. En la sección 4 se muestran los resultados de un estudio de potencia para la prueba propuesta y se compara esta con la potencia de otras pruebas útiles para la misma hipótesis . En la sección 5 se presentan aplicaciones de la metodología propuesta a dos conjuntos de datos reales correspondientes a mediciones de oxígeno disuelto en dos niveles de la columna de agua en la Ciénaga Grande de Santa Marta (CGSM)(IGAC 1973). Algunas conclusiones y propuestas de trabajo futuro se dan en la sección 6. El artículo finaliza con un apéndice en el que se presentan tablas de valores críticos y el código R (R Development Core Team 2005) usado.

2. Hipótesis y estadística de prueba

Sean $\eta_{11}, \ldots, \eta_{1n}$ y $\eta_{21}, \ldots, \eta_{2n}$ dos sucesiones dicotómicas observadas bajo los tratamientos 1, 2 y que están ordenadas según una covariable que tiene niveles $1, \ldots, n$. Si se define

$$\tau_k = \begin{cases} 0 & \text{si } \eta_{1k} = \eta_{2k} \\ 1 & \text{si } \eta_{1k} \neq \eta_{2k} \end{cases}, k = 1, \ldots, n \tag{1}$$

entonces $\tau_1, \ldots, \tau_n$ conforma una nueva sucesión dicotómica que en caso de contener muchos ceros indicará que las dos secuencias son símiles (concordantes) o por el contrario que las respuestas bajo los dos tratamientos son disímiles (discordantes) cuando esté compuesta por muchos unos. En términos de probabilidad se dirá que hay similitud entre las dos secuencias siempre que en la sucesión $\tau_1, \ldots, \tau_n$ la probabilidad del valor cero sea mayor o igual que la del valor uno. De otro lado, se estará bajo la hipótesis alterna (no similitud entre las dos secuencias) cada vez que la probabilidad del valor uno en la sucesión sea mayor que la del valor cero. De acuerdo con lo anterior, las hipótesis de interés pueden plantearsen de la siguiente

forma:

$$\begin{aligned} H_0 &: P(\tau_k = 0) \geq P(\tau_k = 1) \\ H_{\mathrm{a}} &: P(\tau_k = 0) < P(\tau_k = 1) \end{aligned} \tag{2}$$

A continuación se define la estadística de prueba propuesta para la hipótesis dada en (2). Esta se basa en el conteo del número de discordancias entre las sucesiones a comparar y en la posición de estas dentro de la sucesión de los τ_i. Adicionalmente incluye dos términos que facilitan su interpretación. La expresión de la estadística

$$GC(n) = \tau_\bullet + K + n(n-2) - h(n, \tau_\bullet) \tag{3}$$

donde n es el tamaño de la sucesión y los otros términos se definieron como

$$\tau_\bullet = \sum_{k=1}^{n} \tau_k, \tag{4}$$

$$K = \sum_{k=1}^{n} \delta_k, \ \text{ con } \delta_k = \begin{cases} k & \text{si } \tau_k = 1 \\ \tau_\bullet - n & \text{si } \tau_k = 0 \end{cases} \tag{5}$$

y

$$h(n, \tau_\bullet) = \begin{cases} -2n & \text{si } \tau_\bullet = 0 \\ 0 & \text{si } \tau_\bullet = 1 \\ a_m - b_m n & \text{si } \tau_\bullet = 2, \ldots, n, \ m = \tau_\bullet - 2 \end{cases} \tag{6}$$

En (6) $a_0 = 0$, $b_0 = -1$ y para $m \geq 1$ se utilizan las expresiones recursivas $a_m = a_{m-1} + (m^2 + m)$ y $b_m = b_{m-1} + (m - 1)$.

La variable $\tau_\bullet$ definida en la ecuación (4) cuenta el número total de discordancias entre las dos sucesiones que están siendo comparadas, es decir, cuenta el número de unos de la sucesión $\tau_1, \ldots, \tau_n$. Esta variable tiene la misma expresión de la estadística usada en la prueba del signo (Conover 1999). En la ecuación (5) K indica la posición de los unos dentro de la secuencia $\tau_1, \ldots, \tau_n$ y su posición dentro de la misma. K es menor cuando las discordancias están al inicio de la secuencia que cuando se presentan hacia el final de la misma. Su valor mínimo se da cuando las dos secuencias originales son totalmente símiles y su valor máximo se obtiene cuando $\tau_k = 1$, para todo k, $k = 1, \ldots, n$, es decir cuando las dos secuencias son totalmente disímiles. Para facilitar la interpretación del indicador dado en (3) se incluyen los términos $n(n-2)$ y $h(n, \tau_\bullet)$. Con estos se logra, que independientemente del tamaño de la sucesión, la variable $GC(n)$ tome su mínimo en cero y aumente consecutivamente en la escala de los enteros positivos. En resumen, la estadística de prueba planteada detecta las diferencias entre dos secuencias binarias ordenadas y permite describir en qué posición del orden considerado es que estas mismas se presentan.

Tabla 1: Valores de $GC(4)$ en las posibles sucesiones dicotómicas de tamaño 4.

τ_1	τ_2	τ_3	τ_4	$\tau_\bullet$	δ_1	δ_2	δ_3	δ_4	K	$GC(4)$
0	0	0	0	0	−4	−4	−4	−4	−16	0
1	0	0	0	1	1	−3	−3	−3	−8	1
0	1	0	0	1	−3	2	−3	−3	−7	2
0	0	1	0	1	−3	−3	3	−3	−6	3
0	0	0	1	1	−3	−3	−3	4	−5	4
1	1	0	0	2	1	2	−2	−2	−1	5
1	0	1	0	2	1	−2	3	−2	0	6
0	1	1	0	2	−2	2	3	−2	1	7
1	0	0	1	2	1	−2	−2	4	1	7
0	1	0	1	2	−2	2	−2	4	2	8
0	0	1	1	2	−2	−2	3	4	3	9
1	1	1	0	3	1	2	3	−1	5	10
1	1	0	1	3	1	2	−1	4	6	11
1	0	1	1	3	1	−1	3	4	7	12
0	1	1	1	3	−1	2	3	4	8	13
1	1	1	1	4	1	2	3	4	10	14

A manera de ilustración del efecto de las expresiones (4), (5) y (6) en el valor del indicador propuesto en la ecuación (3), se presenta el cálculo de $GC(4)$ con todos los posibles arreglos de ceros y unos de una secuencia binaria de tamaño cuatro (tabla 1). Se observa que $GC(4)$ es una variable discreta, monótona creciente, con valores entre 0 y 14. Para este caso los valores del indicador definen puntualmente lo ocurrido respecto al número de unos y a la posición de los mismos dentro de las sucesiones, excepto cuando $GC(4)$ es igual a 7. Valores de $GC(4)$ entre 1 y 4 indican que hubo un solo uno en la sucesión y cada número revela la posición que este ocupa en la misma (1 si el uno está en la primera posición, 2 si está en la segunda, etc.). Valores entre 5 y 9 corresponden a sucesiones en las que hubo dos unos, con $GC(4)$ igual a 5 cuando los dos unos están en las primeras dos posiciones de la sucesión ordenada y a 9 cuando están en las dos últimas. Los valores 6, 7 y 8 reflejan la transición de los dos unos de las dos primeras a las dos últimas posiciones. Valores entre 10 y 13 indican que hubo tres unos y cada uno de estos valores corresponde a una única sucesión (no hay empates y por consiguiente definen explícitamente lo sucedido en la sucesión respecto a la posición de los unos). El valor de $GC(4)$ será 10 cuando los unos estén en las tres primeras posiciones e igual a 13 cuando estén en las tres últimas. Los valores 0 y 14 se obtendrán cuando en la sucesión no haya unos (las dos sucesiones originales son totalmente símiles) o todos los valores sean iguales a uno (las dos sucesiones originales son totalmente disímiles), respectivamente.

Los valores grandes de $GC(n)$ conducen a rechazar la hipótesis de similitud porque estos se presentan cuando la sucesión dicotómica $\tau_1, \ldots, \tau_n$ está compuesta por muchos unos, lo que indica, de acuerdo con el criterio de dicotomización dado en (1), que hay poca semejanza entre las dos sucesiones binarias originales. De lo anterior, la prueba basada en $GC(n)$ rechaza H_0 a favor de H_1 a un nivel de significancia α dado cuando $GC(n) \geq g_{1-\alpha}$, donde $g_{1-\alpha}$ es tal que $P_{H_0}(GC(n) \geq g_{1-\alpha}) = \alpha$. En la sección 6 se presentan los valores críticos para varios niveles

de significancia con tamaños de muestra entre 3 y 15. Usando el programa R (R Development Core Team 2005) dado en la sección 6, se puede hacer estimación de dichos valores críticos para cualquier tamaño de sucesión n. Los valores críticos también pueden obtenerse mediante una aproximación a la distribución normal, como se describe en la siguiente sección.

3. Propiedades de la distribución de $GC(n)$

En esta sección se enuncian dos teoremas referentes a los momentos de primer y segundo orden de la estadística de prueba y se estudia la aproximación de su distribución a una normal. La demostración de los teoremas puede consultarse en Giraldo (2003), disponible en http://www.docentes.unal.edu.co/rgiraldoh/docs/.

En el caso de hipótesis compuestas, del tipo

$$H_0 : \theta \in \Theta_0$$
$$H_a : \theta \in \Theta_1$$

con $\Theta_1 = \Theta - \Theta_0$ y Θ el espacio de parámetros, el nivel de significancia se define por $\alpha = \underset{\theta \in \Theta_0}{\text{máx}} P$ (rechazar H_0) (Dudewicz & Mishra 1988). La hipótesis nula de interés dada en (2) es compuesta y la distribución de la estadística de prueba $GC(n)$ bajo H_0 depende de la probabilidad que se asuma para $P(\tau_k = 0)$ en la sucesión $\tau_1, \ldots, \tau_n$. Bajo H_0 se tiene que $1/2 \leq P(\tau_k = 0) \leq 1$ y el α deseado se obtiene cuando $P(\tau_k = 0) = 1/2$ (Giraldo 2003). Por lo anterior, para el cálculo del valor esperado y la varianza de $GC(n)$ se supone que $P(\tau = 0) = P(\tau = 1) = 1/2$.

Teorema 1. *Sea $\tau_1, \ldots, \tau_n$ una sucesión dicotómica de tamaño n con $P(\tau_k = 1) = 0.5$ y $GC(n)$ definido como en (3); entonces:*

$$E(GC(n)) = \frac{n^3 + 5n}{12} \tag{7}$$

Teorema 2. *Sea $\tau_1, \ldots, \tau_n$ una sucesión dicotómica de tamaño n con $P(\tau_k = 1) = 0 : 5$ y $GC(n)$ definido como en (3); entonces:*

$$\begin{aligned} V(GC(n)) = & \frac{n(n-1)(n-2)(n-3)(4n^2 + 45n - 4)}{576} \\ & + \frac{n(n-1)(90 - 303n + 444n^2 - 27n^3)}{144} \\ & + \frac{(2n^5 - 34n^4 + 54n^3 - 26n^2 + 10n)}{12} - \left(\frac{n^3 + 5n}{12}\right)^2 \end{aligned} \tag{8}$$

En este trabajo no se realiza un estudio teórico de la distribución asintótica de la estadística $GC(n)$. Sin embargo a manera de exploración se estudia la convergencia a la normal calculando, para diferentes tamaños de sucesión n, la diferencia

máxima entre la frecuencia acumulada exacta y esta misma bajo la normal estándar. Para cada n se evalúa:

$$d_n = \max_x \left| F_n(x) - \Phi\left(\frac{x-\mu_n}{\sigma_n}\right) \right| \tag{9}$$

donde x corresponde a un valor de $GC(n)$, $F_n(x)$ es la distribución exacta de $GC(n)$, μ_n y σ_n^2 corresponden al valor esperado y a la varianza dados en (7) y (8), respectivamente, y $\Phi(\cdot)$ es la función de distribución de la normal estándar. Los cálculos de (9) en el total de la distribución y en las colas de la misma, con n entre 3 y 100, se dan en la tabla 2.

Tabla 2: Diferencias máximas entre la distribución de $GC(n)$ y la normal estándar en el total de la distribución y en las colas (al 5 %) de la misma, para tamaños de sucesión n entre 3 y 100.

n	d_n total	$F_n(x) \leq 0.05$	$F_n(x) \geq 0.95$
3	0.120		
4	0.081		
5	0.078	0.004	0.036
6	0.071	0.012	0.027
7	0.059	0.013	0.021
8	0.063	0.012	0.016
9	0.054	0.011	0.013
10	0.058	0.009	0.011
11	0.052	0.009	0.010
12	0.055	0.009	0.010
13	0.051	0.009	0.009
14	0.053	0.009	0.010
15	0.049	0.009	0.007
20	0.047	0.008	0.009
50	0.037	0.009	0.009
100	0.034	0.007	0.007

En la tabla 2 se puede observar que las diferencias en la distribución completa tienen un alto decrecimiento hasta $n = 7$ y que a partir de ahí las diferencias toman valores parecidos que varían entre 6.3 % y 4.9 %. En las colas se presentan comportamientos distintos. En la cola superior las diferencias toman su máximo para un tamaño de sucesión 5 (3.6 %) y de ahí en adelante disminuyen hasta alcanzar, para tamaños de sucesión mayores de 10, diferencias cercanas al 1 %. En la cola inferior cuando el tamaño de sucesión es pequeño ($n = 3$) se obtiene el valor mínimo (0.4 %), posteriormente las diferencias máximas se incrementan hasta 1.3 %, con $n = 7$, y cuando n aumenta se estabilizan en valores cercanos a 0.7 %. Los resultados muestran que en el total de la distribución el ajuste a la normal no es muy bueno, pero que en las colas las diferencias máximas definidas en (3) son relativamente pequeñas (menores del 1 % para tamaños de sucesión mayores de 13). En la figura 1 se presenta una comparación entre las funciones de distribución de la variable $GC(n)$ y la de la normal para varios tamaños de sucesión. Se observa que la distribución de $GC(n)$ es multimodal y que por ello esta se distancia de la normal y además que hacia las colas las diferencias son mucho menores. Para

propósitos de inferencia (prueba de hipótesis) solo se requiere que haya buen ajuste en las colas de la distribución, por ello el uso de una normal en este caso puede ser razonable. De acuerdo con lo anterior, el test dado en (2) podría realizarse de manera aproximada basado en la estadística:

$$Z_c = \frac{GC(n) - E(GC(n))}{\sqrt{V(GC(n))}}$$

donde el valor esperado y la varianza se obtienen de (7) y (8), respectivamente. La hipótesis nula se rechazaría al nivel α si $Z_c > z_{(1-\alpha)}$, con $z_{(1-\alpha)}$ el percentil $(1-\alpha)$ de la distribución normal estándar.

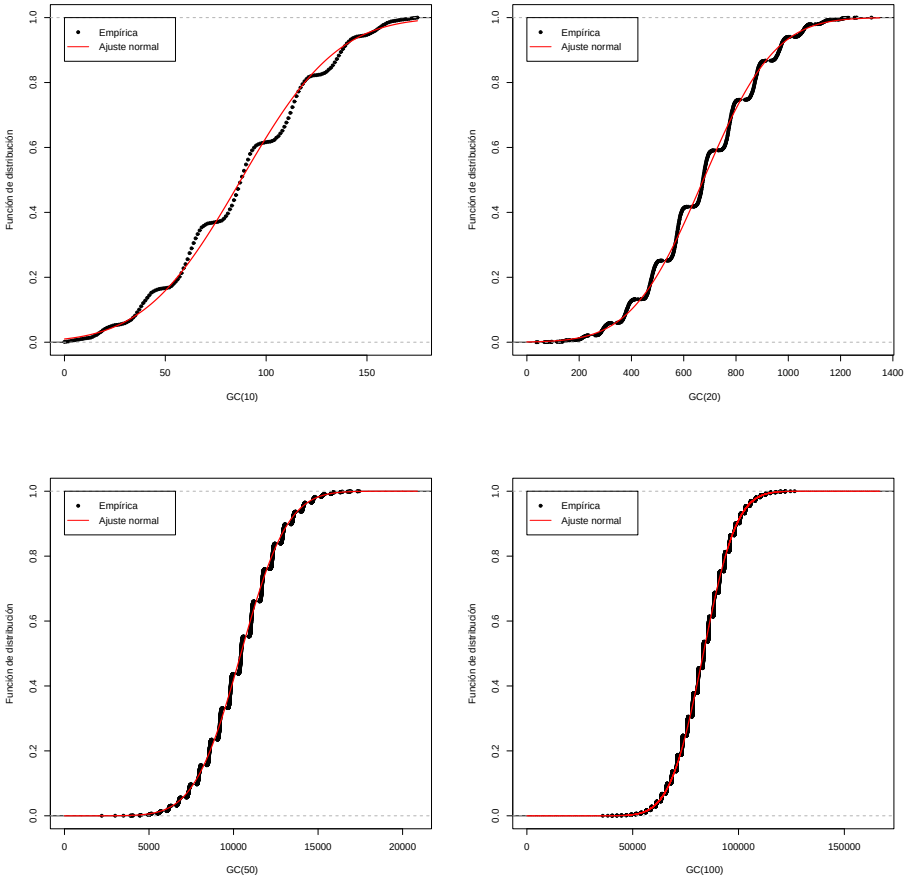

FIGURA 1: Comparación entre la función de distribución de $GC(n)$ y la función de distribución normal con parámetros $E(GC(n))$ y $V(GC(n))$. $n = 10$ (arriba izquierda); $n = 20$ (arriba derecha); $n = 50$ (abajo izquierda); $n = 100$ (abajo derecha). La función de distribución $F_n(x)$ de la variable $GC(n)$ se estima por simulación en el caso de n = 50, 100.

4. Potencia

El estudio de potencia tuvo dos enfoques: en el primero se consideró solo el indicador $GC(n)$ y se calculó la potencia de este para diferentes tamaños de la sucesión (n = 5, 10, 15, 20 y 30). Se usó en cada caso un α del 5 %. Para los valores de $n \leq 15$ se tomaron los valores críticos exactos, por exceso (ver sección 6). Con tamaños de sucesión 20 y 30 se estimaron los valores críticos a través de simulaciones de tamaño 10000 (con $P(\tau_k = 0) = 0.5$). Para estimar la potencia de la prueba se hicieron nuevamente simulaciones de tamaño 10000 bajo la hipótesis alterna, generando sucesiones dicotómicas de tamaño 5, 10, 15, 20 y 30, en las que las probabilidades fueron mayores de 0.5 ($P(\tau_k = 0)$ = 0.6, 0.7, 0.8, 0.9 y 0.99). En cada caso la potencia de la prueba se estimó calculando el número de veces que la estadística $GC(n)$ superó el correspondiente valor crítico antes calculado y dividiendo este sobre el número de simulaciones.

En el segundo enfoque se comparó la potencia de la estadística $GC(n)$ en la prueba de similitud con las obtenidas al usar adaptaciones, para este mismo fin, de las estadísticas $-C$ (Corzo 1990), del signo y de Wilcoxon (Conover 1999). Se emplearon tamaños de sucesión 5, 10 y 15 y en todos los casos se usó α del 5 %. Los valores críticos para $GC(n)$ y para la estadística $-C$ se presentan en la sección 6. Los de las estadísticas del signo y Wilcoxon se tomaron de Conover (1999) y Hollander & Wolfe (1999), respectivamente. Debido a que las estadísticas son discretas y no se consiguen valores críticos exactos para el nivel de significancia usado, se emplearon pruebas aleatorizadas (Dudewicz & Mishra 1988). Las correspondientes funciones críticas para los tres tamaños de sucesión considerados aparecen en Giraldo (2003). Para estimar las correspondientes potencias se simularon 10000 sucesiones dicotómicas de tamaño 5, 10 y 15, respectivamente, para valores de $P(\tau_k = 1)$ = 0.6, 0.7, 0.8, 0.9 y 0.99. Con cada sucesión se calcularon los valores de las cuatro estadísticas $GC(n)$, $-C$, del signo (S) y Wilcoxon (W) y se evaluaron las correspondientes funciones críticas. La potencia en cada caso resultó del cociente entre la suma de los valores obtenidos en las funciones críticas sobre 10000 (tamaño de la simulación).

4.1. Potencia de $GC(n)$

Una propiedad deseable de una prueba es que sea insesgada (Dudewicz & Mishra 1988), es decir que, para un tamaño de muestra fijo, la potencia de la prueba aumente en la medida en que haya alejamiento de la hipótesis nula y que bajo H_0 la probabilidad de rechazo sea pequeña. En la figura 2 se muestran las curvas de potencias estimadas para los 5 tamaños de sucesión considerados. Se observa que, para cada valor de n, la potencia tiene una tendencia creciente en la medida en que se incrementa (es decir cuando hay alejamiento de la hipótesis nula) y que en ningún caso la potencia estimada bajo H_0 excede el valor 0.05, lo cual permite concluir, desde el punto de vista de la simulación, que la prueba basada en $GC(n)$ es insesgada.

Por otra parte, en la figura 3 se muestra el comportamiento de la potencia como función del tamaño de la sucesión. Se puede comprobar allí, gráficamente,

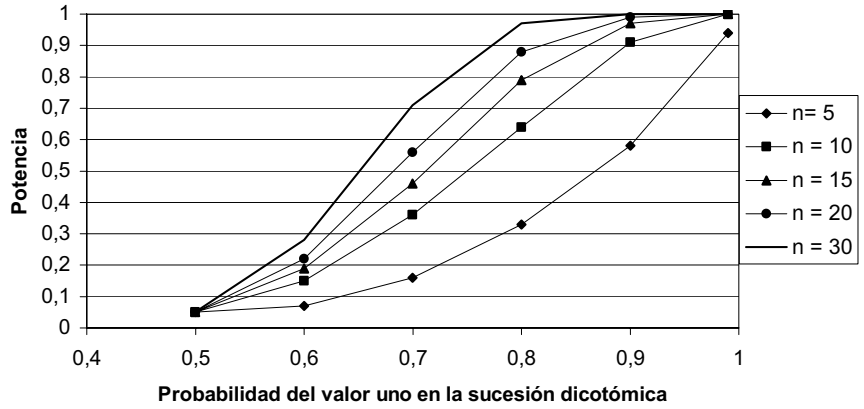

FIGURA 2: Potencia en la prueba de similitud de dos sucesiones dicotómicas con base en $GC(n)$ para varias alternativas (n corresponde al tamaño de la sucesión).

que la potencia es una función creciente en términos del tamaño de la sucesión (a mayor tamaño de sucesión, mayor potencia), lo que indica, con base en resultados de simulación, que esta es consistente (Hollander & Wolfe 1999).

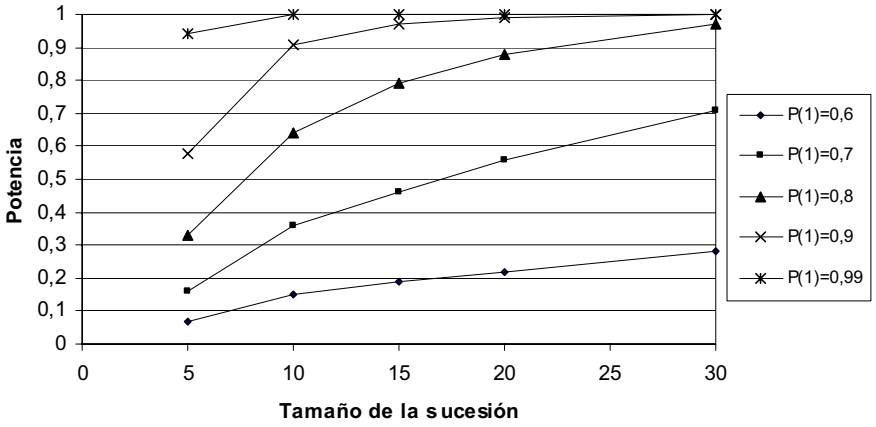

FIGURA 3: Potencia en la prueba de similitud de sucesiones dicotómicas con la estadística $GC(n)$, según el tamaño de la muestra para varias alternativas. P(1) corresponde a $P(\tau_k = 1)$.

4.2. Estudio comparativo de potencia

Con base en el procedimiento de simulación antes mencionado, se obtuvieron los valores de potencia dados en las tablas 3 a 5. Los resultados muestran que para el

tamaño de sucesión más pequeño (n= 5) las cuatro estadísticas consideradas tienen igual potencia (tabla 3); que para los otros dos tamaños de sucesión estudiados (n = 10 y 15) las pruebas basadas en la estadística $GC(n)$ y en la del signo tienen mayor potencia que las dos restantes; y que la prueba basada en la estadística de Wilcoxon resulta ser la menos potente entre las cuatro consideradas.

Tabla 3: Potencia de las pruebas de similitud de dos sucesiones dicotómicas basada en $GC(n)$, y las estadísticas $-C$, del signo (S) y de Wilcoxon (W), para tamaño de sucesión 5 y nivel de significancia del 5 %.

$P(\tau_i = 1)$	$GC(n)$	$-C$	S	W
0.5	0.05	0.05	0.05	0.05
0.6	0.11	0.11	0.11	0.11
0.7	0.21	0.21	0.21	0.21
0.8	0.38	0.38	0.38	0.38
0.9	0.63	0.63	0.63	0.63
0.99	0.95	0.95	0.95	0.95

Los valores de potencia estimados se redondearon a dos cifras significativas.

Tabla 4: Potencia de las pruebas de similitud de dos sucesiones dicotómicas basada en $GC(n)$, y las estadísticas $-C$, del signo (S) y de Wilcoxon (W), para tamaño de sucesión 10 y nivel de significancia del 5 %.

$P(\tau_i = 1)$	$GC(n)$	$-C$	S	W
0.5	0.05	0.05	0.05	0.05
0.6	0.16	0.16	0.16	0.14
0.7	0.36	0.33	0.35	0.29
0.8	0.65	0.61	0.65	0.58
0.9	0.91	0.88	0.91	0.85
0.99	0.99	0.99	0.99	0.99

Los valores de potencia estimados se redondearon a dos cifras significativas.

Al comparar los resultados de en las tablas 2 a 4 se puede establecer que las cuatro estadísticas producen pruebas insesgadas (en la subsección anterior se estableció esto solamente para la estadística $GCij(n)$). Los resultados descritos, aunque no permiten la deducción de conclusiones desde un punto de vista formal, puesto que se basaron en simulación y se obtuvieron solo para tres tamaños de sucesión, sí hacen posible intuir que la prueba basada en la estadística $GCij(n)$ puede tener la misma potencia que la realizada con base en la estadística del signo (tablas 3 a 5). Esto resulta muy relevante teniendo en cuenta que esta última es la más potente para hipótesis de similitud como la planteada en la ecuación (2) (Randles & Wolfe 1979), cuando la información original es dicotómica (mínima escala de medida).

Tabla 5: Potencia de las pruebas de similitud de dos sucesiones dicotómicas basada en $GC(n)$, y las estadísticas $-C$, del signo (S) y de Wilcoxon (W), para tamaño de sucesión 15 y nivel de significancia del 5 %.

$P(\tau_i = 1)$	$GC(n)$	$-C$	S	W
0.5	0.05	0.05	0.05	0.05
0.6	0.19	0.18	0.19	0.17
0.7	0.47	0.42	0.46	0.38
0.8	0.80	0.75	0.79	0.70
0.9	0.98	0.96	0.98	0.94
0.99	0.99	0.99	0.99	0.99

Los valores de potencia estimados se redondearon a dos cifras significativas.

5. Aplicación: riesgo de muerte de aerobios en la CGSM de acuerdo con el nivel de oxígeno

Una función para medir la influencia del oxígeno disuelto en el riesgo de muerte de aerobios en sistemas tropicales costeros se presentan en la figura 4 (Mancera et al. 1996).

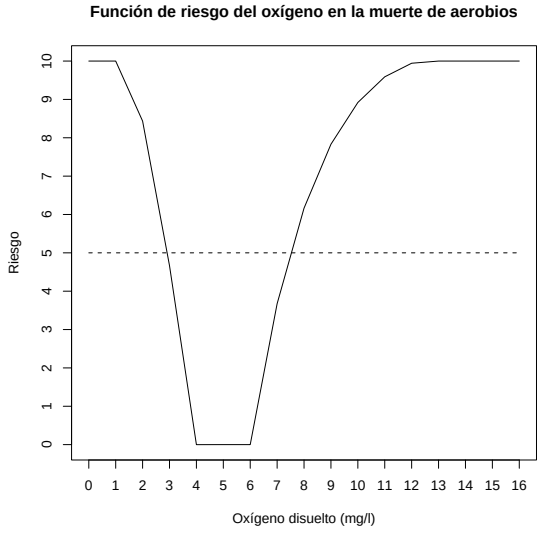

Figura 4: Función de riesgo del oxígeno disuelto en el cálculo del indicador de riesgo de mortandad de aerobios (IRMA). La línea punteada corresponde a riesgo igual a 5.

Magnitudes de oxígeno por debajo de 3 (mg/l) o por encima de 7.5 (mg/l) corresponden a valores de la función de riesgo superiores a 5 (figura 4) e implican un riesgo moderado o alto de muerte de peces y otros organismos. En este trabajo se usa dicha función para convertir en datos binarios los registros de oxígeno disuelto obtenidos en 114 sitios de la Ciénaga Grande de Santa Marta en un muestreo llevado a cabo el 8 de marzo de 1997 (Giraldo et al. 2000) y para probar con base

en estos si el oxígeno del sistema se encontraba en un nivel de riesgo normal el día del muestreo. Las observaciones se tomaron en dos niveles de la columna de agua (superficie y fondo).

Un test de igualdad de medias basado en los datos originales permitió establecer que el día de la muestra había diferencias significativas ($P < 0.05$) entre los dos niveles de la columna de agua. En la figura 5 se muestran los correspondientes intervalos de confianza del 95 % para las medias. De acuerdo con esta figura se concluye, como era de esperarse, que la media de oxígeno en superficie es mayor que la media de oxígeno en el fondo de la columna. Este resultado, aunque muy relevante desde el punto de vista biológico, no permite establecer en términos globales si la variable en cuestión está en un nivel normal o de riesgo para la vida de los organismos dentro del sistema. La comparación de las medias de los dos niveles respecto a los puntos críticos 3.5 mg/l y 7 mg/l, aunque útil desde un punto de vista descriptivo, tampoco permite hacer una prueba formal de esta hipótesis. Por ello el uso de la prueba propuesta en este trabajo resulta de interés con la información considerada.

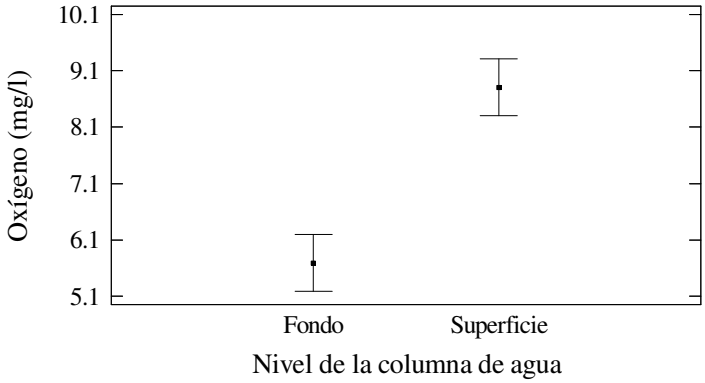

Figura 5: Intervalos de confianza del 95 % para la media de oxígeno en dos niveles de la columna de agua en la Ciénaga Grande de Santa Marta. Datos medidos en marzo de 1997 en 114 sitios del sistema.

En todos los casos los valores en superficie fueron mayores que los del fondo. Valores de oxígeno en fondo menores de 3 mg/l y de oxígeno superficial mayores de 7.5 mg/l se codificaron con el valor 1. Aunque hubo diferencias entre superficie y fondo mayores de 4.5 mg/l, en ningún sitio se dio el caso de que simultáneamente el oxígeno de fondo fuera menor de 3 mg/l y el de superficie mayor de 7.5 mg/l, es decir las concordancias en 1 no se presentaron. Con esta dicotomización las concordancias en ceros implican ausencia de riesgo (oxígeno de fondo mayor de 3 mg/l y oxígeno de superficie menor de 7.5 mg/l). De acuerdo con esto, las hipótesis de interés son

$$H_0 : P(\tau_k = 0) = P(\tau_k = 1) = 1/2$$
$$H_{\mathrm{a}} : P(\tau_k = 0) < P(\tau_k = 1)$$

En la tabla 6 se muestran los tres primeros y los tres últimos datos de las secuencias binarias obtenidas. Las sucesiones dicotómicas están ordenadas de acuerdo con la batimetría (m), teniendo en cuenta que a mayor profundidad menor nivel de oxígeno.

Tabla 6: Esquema de las sucesiones dicotómicas.

Profundidad	OSB (η_{in})	OFB (η_{jn})	Diferencia (τ_{ijn})
0.25	0	0	0
0.40	1	0	1
0.50	0	0	0
⋮	⋮	⋮	⋮
2.00	0	0	0
2.10	1	0	1
2.50	1	0	1

Obtenidas con los siguientes criterios:
Oxígeno superficial (1: mayor de 7.5 mg/l; 0: menor de 7.5 mg/l)
Oxígeno en el fondo (1: menor de 3 mg/l; 0: mayor de 3 mg/l)
Las secuencias están ordenadas según la batimetría.
Los datos originales de oxígeno se tomaron en 114 sitios de
la Ciénaga Grande de Santa Marta en marzo de 1997.
OSB: oxígeno en superficie convertido en binario
OFB: oxígeno en fondo convertido en binario

En 70 de los 114 sitios se obtuvo un 1 para la sucesión τ_n, es decir que en un 61 % de los sitios de la muestra hay un nivel de riesgo por encima de 5. De los 70 sitios con valores uno en la secuencia dicotómica τ_n, solo dos corresponden a sitios con oxígeno de fondo menor de 3 mg/l. Los restantes unos de dicha sucesión se deben a niveles de oxígeno mayores de 7.5 mg/l en superficie. Esto indica de manera descriptiva que el nivel de oxígeno en el sistema estaba el día de la toma de la muestra en condiciones no favorables para la vida de organismos aerobios, específicamente porque los niveles de oxígeno eran mayores de 7.5 mg/l (en 37 de los 114 sitios hubo valores mayores de 7.5 mg/l tanto en superficie como en fondo).

Usando el programa R del apéndice se obtuvo el valor de la estadística de prueba $GC(114)$ y una estimación del correspondiente valor crítico para dos colas con un α del 10 %. Estos fueron respectivamente 164836 y 152479. El valor de la estadística de prueba estandarizado es 2.4027, el cual es mayor que el valor crítico estimado para la prueba de dos colas 1.6843 (ligeramente mayor que el de la normal para el mismo nivel de significancia). Con el mismo programa se obtuvo un error del 0.00717 en la aproximación por la normal. Por lo tanto se rechaza con un nivel de significancia del 10 % la hipótesis de que el nivel de oxígeno del sistema se encontraba en un nivel de riesgo normal. Desde un punto de vista práctico puede concluirse que el 8 de marzo de 1997 los organismos aerobios de la CGSM estaban expuestos a niveles de oxígeno de riesgo moderado o alto (de acuerdo con la función de riesgo dada en la figura 4) especialmente en los sitios de mayor profundidad, puesto que el valor muestral de la estadística está a la derecha de la media de la distribución (123509). El conjunto total de datos y el código R para el análisis de los datos puede obtenerse en la página http://www.docentes.unal.edu.co/rgiraldoh/docs/.

6. Conclusión

La estadística propuesta para probar la hipótesis de similitud de dos secuencias binarias ordenadas es insesgada y consistente y tiene según los resultados de simulación la misma potencia de la prueba basada en la estadística del signo. Su ventaja radica en las posibilidades de interpretación en los casos en los que se rechaza la hipótesis de interés. Con la estadística propuesta es posible establecer si las diferencias tienden a estar al comienzo o al final de la sucesión. El análisis de datos realizado muestra la utilidad práctica de la prueba planteada.

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

Referencias

Conover, W. (1999), *Practical Nonparametric Statistics*, John Wiley & Sons, New York.

Corzo, J. (1990), 'Teoría de rachas', *Revista Colombiana de Estadística* **19-20**, 80–93.

Dudewicz, E. & Mishra, N. (1988), *Modern Mathematical Statistics*, John Wiley & Sons, New York.

Giraldo, R. (2003), Construcción de un indicador para el estudio conjunto de la distribución espacial de múltiples variables binarias, Tesis de maestría, Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia.

Giraldo, R., Méndez, N. & Troncoso, W. (2000), 'Geoestadística: una herramienta para la modelación en estuarios', *Revista Académica Colombiana Ciencia* **24**(90), 57–92.

Hollander, T. & Wolfe, D. (1999), *Nonparametric Statistical Methods*, John Wiley & Sons, New York.

Mancera, E., Giraldo, R. & Salazar, J. (1996), IRMA: indicador de riesgo de mortandad de aerobios. Instituto de Investigaciones Marinas (INVEMAR).

R Development Core Team (2005), *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.

Randles, R. & Wolfe, D. (1979), *Introduction to the Theory of Non-parametric Statistics*, John Wiley & Sons, New York.

Apéndice

Valores críticos de la estadística $GC(n)$

Valores críticos $GC(n)$, de una cola superior, para muestras dicotómicas de tamaños 3 a 15. Se reportan los valores de $GC(n)$ cuyo nivel de significancia exacto, α^*, es más cercano al teórico por defecto (línea superior) o por exceso (línea inferior). El asterisco significa que el estadístico $GC(n)$ no tiene valores con probabilidad inferior al nivel de significancia dado.

	α			
Tamaño de la s ucesión	**0.05**	**0.025**	**0.01**	**0.005**
3	*	*	*	*
	7 (0.125)	7 (0.125)	7 (0.125)	7 (0.125)
4	*	*	*	*
	14 (0.0625)	14 (0.0625)	14 (0.0625)	14 (0.0625)
5	25 (0.0313)	*	*	*
	24 (0.0625)	25 (0.0313)	14 (0.0625)	14 (0.0625)
6	39 (0.0469)	41 (0.0156)	*	*
	38 (0.0625)	40 (0.0313)	41 (0.0156)	41 (0.0156)
7	58 (0.0469)	61 (0.0234)	63 (0.0078)	*
	57 (0.0547)	60 (0.0313)	62 (0.0156)	63 (0.0078)
8	82 (0.0430)	87 (0.0234)	91 (0.0078)	92 (0.0039)
	81 (0.0508)	86 (0.0273)	90 (0.0117)	91 (0.0078)
9	114 (0.0430)	118 (0.0234)	125 (0.0098)	128 (0.0039)
	113 (0.0508)	117 (0.0273)	124 (0.0117)	127 (0.0059)
10	152 (0.0488)	159 (0.0025)	166 (0.0098)	171 (0.0049)
	151 (0.0508)	158 (0.0264)	135 (0.0107)	170 (0.0059)
11	193 (0.0479)	208 (0.0249)	216 (0.0088)	222 (0.0049)
	192 (0.0527)	207 (0.0269)	215 (0.0103)	221 (0.0054)
12	250 (0.0498)	258 (0.0249)	277 (0.0093)	282 (0.0046)
	249 (0.0535)	257 (0.0269)	276 (0.0105)	281 (0.0054)
13	303 (0.0494)	328 (0.0233)	347 (0.0098)	355 (0.0048)
	302 (0.0508)	327 (0.0254)	346 (0.0101)	354 (0.0054)
14	378 (0.0481)	407 (0.0246)	420 (0.0097)	439 (0.0049)
	377 (0.0511)	406 (0.0255)	419 (0.0106)	438 (0.0052)
15	468 (0.0473)	481 (0.0247)	516 (0.0095)	524 (0.0049)
	467 (0.0507)	480 (0.0262)	515 (0.0103)	523 (0.0053)

Los valores entre paréntesis son los valores de significancia exactos, es decir $\alpha* = P(GC(n) \geq x)$, con x un valor de $GC(n)$.

Valores críticos de la estadística de antirrachas

Valores críticos de la estadística de antirrachas, de una cola superior, para muestras dicotómicas de tamaños 3 a 15. Se reportan los valores de $-C$ cuyo nivel de significancia exacto, α^*, es más cercano al teórico por defecto (línea superior) o por exceso (línea inferior). El asterisco significa que el estadístico $-C$ no tiene valores con probabilidad inferior al nivel de significancia dado.

	α			
Tamaño de la sucesión	**0.005**	**0.01**	**0.025**	**0.05**
3	*	*	*	*
	2.00 (0.125)	2.00 (0.125)	2.00 (0.125)	2.00 (0.125)
4	*	*	*	*
	2.50 (0.0625)	2.5 (0.0625)	2.5 (0.0625)	2.5 (0.0625)
5	*	*	*	3.00 (0.0313)
	3.00 (0.0313)	3.00 (0.0313)	3.00 (0.0313)	2.25 (0.0625)
6	*	*	3.50 (0.0156)	2.80 (0.0313)
	3.50 (0.0156)	3.50 (0.0156)	2.80 (0.0313)	2.50 (0.0938)
7	*	4.00 (0.0078)	3.33 (0.0156)	3.33 (0.0156)
	4.00 (0.0078)	3.33 (0.0156)	3.00 (0.0547)	3.00 (0.0547)
8	4.50 (0.0039)	3.86 (0.0078)	3.86 (0.0078)	3.00 (0.0430)
	3.86 (0.0078)	3.50 (0.0313)	3.50 (0.0313	2.80 (0.0625)
9	4.37 (0.0039)	4.37 (0.0039)	3.43 (0.0234)	3.12 (0.0391)
	4.00 (0.0195)	4.00 (0.0195)	3.37 (0.0254)	3.00 (0.0703)
10	4.56 (0.0029)	4.56 (0.0029)	3.60 (0.0225)	3.50 (0.0439)
	4.50 (0.0107)	4.50 (0.0107)	3.50 (0.0439)	3.43 (0.0508)
11	5.10 (0.0015)	4.50 (0.0073)	4.11 (0.0127)	3.37 (0.0449)
	5.00 (0.0059)	4.31 (0.0117)	4.00 (0.0308)	3.36 (0.0571)
12	5.00 (0.0042)	4.55 (0.0095)	4.10 (0.0220)	3.56 (0.0398)
	4.89 (0.0063)	4.50 (0.0183)	4.00 (0.0269)	3.50 (0.0562)
13	5.17 (0.0039)	5.10 (0.0051)	4.09 (0.0248)	3.57 (0.0480)
	5.10 (0.0051)	5.00 (0.0107)	4.00 (0.0408)	3.56 (0.0514)
14	5.58 (0.0028)	5.00 (0.0095)	4.55 (0.0184)	3.82 (0.0494)
	5.50 (0.0062)	4.89 (0.0105)	4.50 (0.0259)	3.80 (0.0502)
15	5.54	5.17 (0.0088)	4.46	4.09 (0.0378)
		5.10 (0.0105)		4.00 (0.0504)

Los valores entre paréntesis son los valores de significancia exactos, es decir $\alpha* = P(-C \geq x)$, con x un valor de $-C$.

Código R para cálculo de valores críticos de la estadística $GC(n)$

En esta sección se muestra el código R usado en la aplicación y en la estimación de potencias de la sección 2. El código puede obtenerse en http://www.docentes.unal.edu.co/rgiraldoh/docs/.

```
################################################################
# 1. Programa general. Este usa las funciones GC, GCCritico y
#    decisión definidas de 2 a 4. Para compilar este programa
#    con sus datos cambie oxigeno.txt por su archivo de datos
#    el cual debe contener una columna con la secuencia dico-
#    tómica de diferencias nombrada como suc.
################################################################

rm(list=ls())
source("GC.R")
source("GCCritico.R")
source("Decision.R")
```

```
suc<-read.table("oxigeno.txt", head=T)
attach(suc) suc<-suc$suc
Estadistica<-GC(suc)
Valor.Critico<-GCCritico(length(suc))
Decision(Estadistica,Valor.Critico)

#########################################################
2. Función para el cálculo del indicador
#########################################################
GC<-function(suc)
{
suc<-as.vector(suc)
tau.punto<-sum(suc)
n<-length(suc)

############################
# 2.1. Cálculo de K
############################
delta<-rep(0,n)
for (i in 1:n)
     {
     if(suc[i]==1) delta[i]<-i else (delta[i]<-tau.punto-n)
     }
K<-sum(delta)
##############################
# 2.2. Cálculo de h(n, tau)
##############################
a<-0
b<--1
am<-0
bm<--1
for (i in 1:(n-2))
     {
     a<-a+((i^2)+i)
     am<-rbind(am,a)
     b<-b+(i-1)
     bm<-rbind(bm,b)
     }
ab<-as.matrix(cbind(am,bm))
m<-tau.punto-2
if(tau.punto==0) h<--2*n
if(tau.punto==1) h<--0
if(tau.punto>1 ) h<-ab[m+1,1]-(ab[m+1,2]*n)
##############################
# 2.3 Valor de la estadística
##############################
GC<- tau.punto+K+(n*(n-2))-h
```

```
return(list(tau.punto=tau.punto, K=K,h=h, GC=GC))
}

################################################
# 3. Función para el cálculo del valor crítico
################################################
GCCritico<-function(n)
{
 critico<-NULL
 for(i in 1:10000)
    {
    suc<-rbinom(n,1,0.5) # Cambiar 0.5 por 0.6,...,.99
                         # para calcular potencia
    GCal<-GC(suc)
    critico<-rbind(critico,GCal$GC)
    }
 valor.critico<-quantile(critico,0.95)
 return(valor.critico)
}

#########################
# 4. Función de decisión
#########################

Decision<-function(Estadistica, Valor.Critico)
{
 rechazo<-paste("SE", "RECHAZA", "H0")
 no.rechazo<-paste("NO","HAY", "EVIDENCIA",
                   "PARA", "RECHAZAR", "H0")
 decision<-ifelse (Estadistica$GC>Valor.Critico,
                   rechazo, no.rechazo)
 return(list(Estadistica=Estadistica$GC,ValorCritico=Valor.Critico,
            Decision=decision))
}
```

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 Statistcs* (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

La Revista Colombiana de Estadística agradece a las personas, que no integran los Comités Editorial y Científico, por su colaboración en el volumen 32 (2009).

Aracelis Hernández, Ph.D.
Purificación Vicente, Ph.D.
Carlos Walter Robledo, Ph.D.
Carlos Noguera, Ph.D.
Marcelo Kittlein, Ph.D.
Manuel Galea, Ph.D.
Luis Mauricio Castro Cepero, Ph.D.
Eliseo Martínez, Ph.D.
Ramón Ferreiro García, Ph.D.
Inmaculada Arostegui Madariaga, Ph.D.
Jairo Alberto Fúquene Patiño, M.Sc.
Jairo Angel Guzmán, M.Sc.
David Ospina Botero, Ph.D.
Enrique M. Cabaña Pérez, Pregrado
Holger Dette, Ph.D.
Luis Fernando Melo Velandia, M.Sc.
Lourdes Contreras Montenegro, Ph.D.
Oscar Andrés Nupia Martínez, Ph.D.
Manuel Miguel Ramos Álvarez, Ph.D.
Pablo Pincheira, Ph.D.
José Antonio Gutiérrez Gallego, Ph.D.
Francisco J. Pérez Arredondo, Ph.D.
Gregorio Saravia Atuncar, Ph.D.
Fernando Quintana, Ph.D.
Arturo T. De Zan, Ph.D.
Alfredo García Hiernaux, Ph.D.
Manuel Vargas, Ph.D.
Edilberto Cepeda Cuervo, Ph.D.