Revista Colombiana de Estadística

Volumen 34. Número 1 - junio - 2011 **ISSN 0120 - 1751**

UNIVERSIDAD
NACIONAL
DE COLOMBIA
SEDE BOGOTÁ
FACULTAD DE CIENCIAS
DEPARTAMENTO DE ESTADÍSTICA

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. E-mail: apchavezr@gmail.com
Impresión: Editorial Universidad Nacional de Colombia, Tel. 57-1-3165000 Ext. 19645, Bogotá.

Revista Colombiana de Estadística	Bogotá	Vol. 34	Nº 1
ISSN 0120 - 1751	COLOMBIA	junio-2011	Págs. 1-209

Contenido

Editorial

Saludo del editor entrante

Leonardo Trujillo[a]

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

---

Es un gran honor y placer asumir el papel de editor general de la *Revista Colombiana de Estadística* (RCE). Mi colaboración en los últimos años había sido como árbitro y coeditor de algunos artículos, y posteriormente como miembro del Comité Científico, lo cual me dejó una grata experiencia y la motivación para asumir tan importante responsabilidad. Aprovecho la oportunidad para agradecer al profesor Ramón Giraldo, director del Departamento de Estadística y a los comités Científico y Editorial por su confianza en mí y por su activa participación en los últimos años para reducir los tiempos de arbitraje, pero mejorando el nivel de calidad de las publicaciones.

La RCE ha vivido un proceso de ascenso, convirtiéndose en una de las revistas de Estadística más prestigiosas en Latinoamérica y el resto del mundo. Este ascenso se ha logrado bajo el liderazgo de los últimos editores generales: profesores Campo Elías Pardo, Jorge Ortíz y B. Piedad Urdinola. La RCE ha publicado 33 volúmenes anuales con dos números por año. En el primer volumen de 1968, hubo 3 artículos publicados. En el volumen 33 de 2010, contamos con 52 artículos sometidos y 18 publicados. En lo corrido del año hasta mayo de 2011, contamos ya con 35 artículos sometidos.

En este primer número del volumen 34 de 2011, que hoy presentamos, hemos puesto a disposición de nuestros lectores los 10 primeros que han sido recientemente aprobados, sin contar los artículos del número especial de la Revista, que se publican en paralelo con este número. Un reto que se debe enfrentar es el número creciente de artículos sometidos, lo cual requiere un excelente y riguroso proceso de arbitraje en áreas de la Estadística cada vez más especializadas. Cada artículo requiere entre dos y tres árbitros como mínimo para su revisión, lo cual implica implica un trabajo bastante cuidadoso y elaborado en la búsqueda de árbitros idóneos. Sin embargo, el número creciente de artículos sometidos es al mismo tiempo un buen signo de que vamos por un buen camino.

Mis mayores objetivos como editor general para los próximos años serán 1) mejorar la disponibilidad y visibilidad de la RCE manteniendo las políticas de mis predecesores y las propias de los organismos editoriales y las instituciones de indexación internacionales, 2) aumentar su exogenidad en lo concerniente a la afiliación de autores, árbitros y coeditores; y 3) mantener el nivel de calidad alcanzado hasta ahora, que nos ha llevado a obtener la máxima calificación A1 bajo el índice de Publindex de revistas científicas nacionales.

[a]Editor de la Revista Colombiana de Estadística, Profesor asistente.
E-mail: ltrujilloo@bt.unal.edu.co

Con respecto a la exogenidad de la RCE, en el volumen 33 de 2010, se publicaron 18 artículos. 6 de ellos corresponden a publicaciones del Departamento de Estadística de la Universidad Nacional de Colombia, -sede Bogotá- (33 %), pertenecen a publicaciones nacionales de investigadores de la Universidad Nacional de Colombia, sede Medellín-, la Universidad de Córdoba, la Universidad de la Guajira y la Universidad Santo Tomas (22 %) 8 restantes a investigadores internacionales de Bélgica, Brasil, España, India, México y Venezuela (44 %). El comité editorial está conformado por investigadores de Brasil, Colombia, España, México, Puerto Rico, Estados Unidos, Venezuela. Los artículos en proceso de evaluación provienen de países tan disímiles como Brasil, Chile, China, Colombia, Cuba, Emiratos Árabes Unidos, Eritrea, España, India, Irán, Japón, México, Pakistán, Turquía, USA Estados Unidos y Venezuela. La participación de colegas internacionales en los diferentes roles de la Revista ha llegado a su número más alto y esperamos siga en alza.

Es muy grato para mí presentarles el primer número de 2011 (volumen 34, número 1), donde hemos escogido 10 artículos para publicación. En paralelo con este número, se está publicando un número especial de la RCE, que se refiere a artículos sobre aplicaciones de la Estadística en la industria (8 artículos adicionales cuyos editores invitados fueron los profesores Jorge Romeu y Piedad Urdinola). Por lo regular, las organizaciones industriales requieren el uso de métodos estadísticos en cuanto al aseguramiento de la calidad, compras, desarrollo de productos, finanzas, manufactura, mercadotecnia, recursos humanos y el respaldo posterior a la venta o garantía. Por tanto, este número especial quiso hacer énfasis en los métodos estadísticos usados para cada uno de estos fines.

Desde ya, hacemos un llamado para el envío de artículos a ser publicados en julio de 2012 en un número especial en Bioestadística. El objetivo de este número especial es difundir desarrollos teóricos y aplicaciones en diversos temas que cubre la Bioestadística, como: biología, demografía, ecología, epidemiología y medicina, entre muchos otros. Las editoras invitadas de este número especial son las profesoras Liliana López-Kleine y B. Piedad Urdinola.

Finalmente, quiero extender las gracias a autores, árbitros, coeditores y, editores por su trabajo arduo, así como a nuestros lectores por su apoyo. Todos ellos nos llevan a mejorar día a día. Esperamos que todo nuestro esfuerzo se vea recompensado algún día en un mayor índice de impacto de nuestras publicaciones.

Editorial

A Message from the Incoming Editor

Leonardo Trujillo[a]

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

---

It is a great honor and pleasure to assume the role of General Editor of the Colombian Journal of Statistics (CoJS). My involvement over the last years as a reviewer and coeditor of some papers and after, as a member of the Scientific Committee of the Journal, was a great experience and the main motivation in taking up such an important responsibility. I would like to thank to Professor Ramon Giraldo, Director of the Department of Statistics and to the Editorial and Scientific Committees for their trust on me and for their active participation during the last years to further reduce the time from submission to final decision but keeping, and even improving, at the same time, the top-quality of our published papers.

The CoJS had a rapid growing process, becoming one of the most prestigious journals in Statistics in Latin America and at the international standard. This was possible under the leadership of the former editors: Professors Campo E. Pardo, Jorge E. Ortiz y B. Piedad Urdinola. In all its history, the CoJS has published 33 annual volumes with two numbers per year. The first volume of 1968 had got 3 published papers; the 33rd volume last year (2010) had 52 submitted papers with 18 of them finally published in our Journal. So far, on May 2011, we can already count 35 submitted papers.

In this first number of the 34th volume of 2011, I am happy to present the 10 first papers approved for publication. Here, we are not taking into account those papers belonging to the special issue of the Journal, which have been published in parallel with this issue. Therefore, to face a relevant challenge concerning to the steadily growing number of submitted papers, an excellent and rigorous peer review process in more specialized areas of Statistics is necessary. Each paper requires typically two or three reviews as a minimum, implying a very careful and elaborated search of reviewers. However, the growing number of submitted papers indicates we are going in a good way.

The main aims in my work as editor for the next years will be to 1) improve the availability and visibility of the CoJS keeping the policies of my pre-decessors and those proper of the editorial and indexing international institutions, 2) increase the participation of authors, reviewers and associate editors in an exogenously way and 3) keep the standard top-quality level achieved so far, which has permitted us to reach the maximum level A1 at the national Publindex classification of scientific journals.

[a]General Editor of the Colombian Journal of Statistics, Assistant Professor.
E-mail: ltrujilloo@bt.unal.edu.co

Regarding to the second point in the last paragraph, for the 33rd volume on 2010, we published 18 papers. Six of them corresponding to the Statistics Department of the National University of Colombia in Bogota (33 %), four of them from national publications from academic institutions throughout the country, such as the National University of Colombia in Medellin, the Santo Tomas University, the University of Cordoba and the University of Guajira (22 %) and the other eight publications (44 %) come from researchers around the world, from countries such as Belgium, Brazil, India, Mexico, Spain and Venezuela. The Editorial Committee is formed by recognized international researchers in Statistics from Brazil, Colombia, Mexico, Puerto Rico, Spain, USA and Venezuela. The submitted papers still waiting for approval come from countries such as Brazil, Chile, China, Colombia, Cuba, Eritrea, India, Iran, Japan, Mexico, Pakistan, Spain, Turkey, United Arab Emirates, USA and Venezuela. The participation from international researchers in the CoJS has come to the highest number in the history of the journal and we expect this number continues to rise.

I am very glad to edit the first issue in 2011 (volume 34, number 1), where we have chosen ten papers for publishing. In parallel with this issue, we had launched a special issue. This one refers to papers about Statistical Applications in the Industry (eight additional papers that guest editors were Professors Jorge Romeu and B. Piedad Urdinola). Industrial organizations commonly require the use of statistical methods in terms of development of products, finances, human resources, manufacture, marketing, purchasing, quality assessment, and after-sales warranty. Then, this special issue emphasizes on the different statistical methods used for these purposes.

We are making a special call for the submission of papers to be published on July 2012 in a special issue in Biostatistics. The aim of this special issue is to diffuse theoretical developments and applications in several topics about Biostatistics such as: biology, demography, ecology, epidemiology and medicine, among many others. The guest editors for this special issue are Professors Liliana López-Kleine and B. Piedad Urdinola.

Finally, I would like to extend my acknowledgements to the authors, coeditors, editors, reviewers for their very hard work, and readers that make us to improve day by day. We hope all this effort will be compensated one day with a higher impact index of all our publications.

Revista Colombiana de Estadística
Junio 2011, volumen 34, no. 1, pp. 1 a 14

Validación de la escala de actitudes hacia la estadística en estudiantes españoles de ciencias de la actividad física y del deporte

Validation of the Scale of Attitudes toward Statistics in Spanish Students of Physical Activity and Sport Sciences

Carlos M. Tejero-González[1,a], María Castro-Morera[2,b]

[1]Departamento de Educación Física, Deporte y Motricidad Humana, Facultad de Formación de Profesorado y Educación, Universidad Autónoma de Madrid, Madrid, España

[2]Departamento de Métodos de Investigación y Diagnóstico en Educación, Facultad de Educación, Universidad Complutense de Madrid, Madrid, España

Resumen

Este trabajo analiza la estructura dimensional de la Escala de Actitudes hacia la Estadística en su aplicación a estudiantes de Ciencias de la Actividad Física y del Deporte. En virtud de los datos obtenidos con una muestra de 145 participantes de ambos sexos que fueron seleccionados por muestreo incidental en dos universidades públicas españolas, se concluye que no son plausibles las estructuras dimensionales propuestas por otros autores. Al mismo tiempo, se defiende una solución factorial basada en tres dimensiones y doce ítems, con capacidad para explicar el 68 % de la varianza del instrumento y con una fiabilidad alfa de Cronbach igual a 0,87. Los estudiantes universitarios de ciencias del deporte declaran niveles medios de ansiedad hacia la estadística, consideran que la utilidad o importancia de esta asignatura es media-baja, y declaran baja predisposición hacia dicha materia.

***Palabras clave*:** análisis factorial, escala, estadística, medición de actitud, psicometría, validación.

Abstract

This article analyses the dimensional structure of the Scale of Attitudes toward Statistics in its implementation for Physical Activity and Sport Science students. On the data obtained with a sample of 145 participants of both sexes who were selected by incidental sampling in two Spanish universities, it is concluded that dimensional structures proposed by other autores

[a]Profesor. E-mail: carlos.tejero@uam.es
[b]Profesora titular. E-mail: maria.castro@edu.ucm.es

are not plausible. At the same time, this study presents a factorial solution based on three dimensions and twelve items, with capacity to explain 68 % of the variance and with a Cronbach's alpha reliability of 0.87 from Sport Sciences declare average levels of anxiety, consider low usefulness or importance and declare low predisposition to statistics.

***Key words*:** Attitude measurement, Factor analysis, Psychometrics, Scale, Statistics, Validation.

1. Introducción

De acuerdo con el Real Decreto 1393/2007 por el que se establece la ordenación de las enseñanzas universitarias oficiales del Estado Español (Ministerio de Educación y Ciencia 2007), todas las titulaciones de grado deben estar asignadas a una rama de conocimiento, existiendo cinco áreas: artes y humanidades, ciencias, ciencias de la salud, ciencias sociales y jurídicas, e ingeniería y arquitectura. Asimismo, independientemente de la titulación, los planes de estudios deberán contener 60 créditos de formación básica, estableciéndose una relación de materias básicas que son específicas para cada rama de conocimiento, de tal forma que las titulaciones tienen que ofertar al menos 36 créditos de materias básicas asignadas a su rama de conocimiento, mientras que el resto de créditos básicos hasta llegar a los 60 podrán pertenecer a la misma u otra rama de conocimiento, o bien a otras materias si se justifica su carácter transversal. En este contexto legislativo, la estadística es una materia básica asignada a dos ramas de conocimiento: la de ciencias sociales y jurídicas y la de ciencias de la salud.

Si a ello añadimos que para la titulación de grado en ciencias de la actividad física y del deporte, las ramas de asignación preferentes son la de ciencias sociales y jurídicas y la de ciencias de la salud (Agencia Nacional de Evaluación de la Calidad y Acreditación ANECA 2009, Pleno de la Conferencia Conferencia Española de Institutos y Facultades de Ciencias de la Actividad Física y del Deporte 2007), cabe pensar que con la implantación del nuevo espacio universitario europeo, la estadística es una asignatura que ha ganado extensión y protagonismo en los estudios universitarios de ciencias de la actividad física.

En virtud de este encadenamiento de valoración de la estadística como materia básica en los estudios universitarios vinculados a las ciencias del deporte, y a tenor de la importancia que tiene la actitud en el proceso de enseñanza-aprendizaje de esta asignatura (Gómez 2000, Blanco 2004, Bazán & Aparicio 2006, Mondéjar, Vargas & Bayot 2008, Estrada 2009, Mondéjar & Vargas 2010), una pertinente línea de investigación es aquella que analice qué instrumentos son válidos y fiables para medir las actitudes hacia la estadística de los estudiantes universitarios de ciencias de la actividad física y del deporte.

Casi de forma exclusiva, el tipo de instrumento que se ha utilizado para medir las actitudes hacia la estadística ha sido el cuestionario o escala (Carmona 2004, Blanco 2008). Sin el ánimo de ser exhaustivos, algunos ejemplos son los siguientes: Statistics Attitudes Survey (Roberts & Bilderback 1980), Attitudes Toward Statistics (Wise 1985), Statistics Attitude Inventory (Zeidner 1991), Attitude Toward

Statistics (Miller, Behrens, Green & Newman 2007), Survey of Attitudes Toward Statistics (Schau, Stevens, Dauphinee & Del Vecchio 1995) y Quantitative Attitudes Questionnaire (Chang 1996). Por otra parte, diversos autores han diseñado instrumentos en idioma español, entre otros: Auzmendi (1992), Velandrino & Parodi (1999), Muñóz (2002), Mondéjar et al. (2008), Estrada, Batanero & Fortuny (2004).

De los instrumentos que se han diseñado en idioma español, este estudio se centrará en la Escala de Actitudes hacia la Estadística EAE de Auzmendi (1992), por ser una de las escalas más investigadas y replicadas con publicación de resultados psicométricos en revistas científicas, siendo un instrumento de 25 ítems y cinco dimensiones: utilidad (ítems 1, 6, 11, 16 y 21), ansiedad (ítems 2, 7, 12, 17 y 22), confianza (ítems 3, 8, 13, 18, 23), agrado (ítems 4, 9, 14, 19 y 24), y motivación (ítems 5, 10, 15, 20 y 25) (Apéndice). En el protocolo del instrumento se solicita a los estudiantes que expresen su grado de acuerdo con los diferentes enunciados, valiéndose de una escala Likert de cinco puntos, donde 1 significa total desacuerdo y 5 total acuerdo. Según los datos de la propia autora, el instrumento tiene capacidad para explicar el 60,7 % de las puntuaciones, extrayendo las dimensiones con método de Componentes Principales y rotación Varimax.

Como ya se ha señalado, la calidad técnica de la EAE se ha analizado en diferentes ocasiones, a saber: Sánchez-López (1996), Darías (2000) y Méndez & Macía (2007). Puesto que el estudio de Darías (2000) es una extensión de la investigación de Sánchez-López (1996), la atención de este trabajo recaerá en los estudios de Darías (2000) y Méndez & Macía (2007).

Darías (2000) validó el instrumento con una muestra de 188 estudiantes de primeros cursos de psicología, procediendo con extracción de Componentes Principales y rotación Varimax. La estructura factorial explicó el 53 % de la dispersión del instrumento a partir de cuatro factores y no de cinco: seguridad (ítems 2, 3, 7, 8, 12, 13, 17, 18 y 22), importancia (ítems 4, 9, 14, 19, 20 y 24), utilidad (ítems 6, 10, 11, 16 y 21) y deseo de saber (ítems 1, 5, 15 y 23). El ítem 25 no se consideró por saturar en diferentes dimensiones con bajo peso en cada una.

Méndez & Macía (2007) hicieron lo propio con una muestra de 168 estudiantes universitarios chilenos, encontrando como en el caso anterior una estructura de cuatro dimensiones: factor I (ítems 4, 9, 14, 19, 21, 24), factor II (ítems 2, 7, 12, 15, 17, 22), factor III (ítems 1, 5, 6, 10, 11, 16, 23, 25) y factor IV (ítems 3, 8, 13, 18). Esta validación excluyó el ítem 20 por no saturar en ninguno de los factores. La varianza explicada fue del 49 % con extracción por Factorización de Ejes Principales y rotación Equamax.

Con estos antecedentes, el presente trabajo tiene como objetivo analizar la estructura factorial de la escala de actitudes hacia la Estadística de Auzmendi (1992), en su aplicación a estudiantes universitarios de ciencias de la actividad física y del deporte.

2. Método

2.1. Participantes

La muestra está formada por 145 estudiantes de ambos sexos (25 % de mujeres y 75 % de hombres), con una edad promedio de 22 años (media = 22, 55; desviación estándar = 3, 78; mínimo = 20; máximo = 39). Los participantes se seleccionaron mediante muestreo no aleatorio incidental en dos universidades públicas del Estado español, de acuerdo con un criterio de viabilidad de acceso y en virtud de la idoneidad de perfil: ser estudiante que cursa la materia de estadística dentro de un plan de estudios de formación universitaria en ciencias de la actividad física y del deporte.

2.2. Diseño y recolección de datos

El presente es un estudio *ex post facto*. El instrumento utilizado ha sido la escala de actitudes hacia la estadística de Auzmendi (1992), con rango de respuesta Likert de cinco puntos. Los datos fueron recogidos por el investigador principal de forma masiva en el aula habitual de los estudiantes, antes de una clase de estadística del último mes de curso. Se resolvieron las escasas dudas que surgieron y no se recompensó la participación, siendo ésta anónima y voluntaria.

2.3. Análisis de los datos

Los datos se analizaron en dos fases consecutivas. En la primera, se contrastó la idoneidad empírica de las estructuras dimensionales propuestas por Auzmendi (1992), Darías (2000) y Méndez & Macía (2007) con la muestra descrita, para lo que se procedió con Análisis Factorial Confirmatorio. Posteriormente, se llevó a cabo un proceso de selección de ítems con el objeto de postular un instrumento válido y fiable para medir las actitudes hacia la estadística de los estudiantes de ciencias de la actividad física y del deporte, procediéndose con Análisis Factorial Exploratorio y Análisis Factorial Confirmatorio. También se procedió a un estudio descriptivo de los resultados para observar el comportamiento de estos estudiantes ante la estadística como objeto de estudio. Se utilizaron las aplicaciones informáticas IBM SPSS Statistics 18 y AMOS 17.

3. Resultados

Después de comprobar por una parte la pertinencia de analizar factorialmente la EAE (prueba de Kaiser-Meyer-Olkin (KMO) =, 903. Prueba de esfericidad de Barlett: $\chi^2 = 1895, 84$; g.l.= 300; $p <, 001$), y por otra parte que el tamaño de la muestra es adecuado, pues supera las cinco unidades muestrales por ítem -umbral mínimo en el caso de instrumentos de autoinforme para la medición de actitudes (Morales, Urosa & Blanco 2003)-, se analizó el ajuste de las estructuras dimensionales propuestas por Auzmendi (1992), Darías (2000) y Méndez & Macía

(2007). Se procedió con el Análisis Factorial Confirmatorio y con la estimación de parámetros por método de máxima verosimilitud, ya que una muestra entre 100 y 150 participantes, como es el caso de este estudio, "asegura el uso apropiado de MLE (estimación por máxima verosimilitud)" (Hair, Anderson, Tatham & Black 2004, p. 632).

Los resultados del Análisis Factorial Confirmatorio señalaron que los datos obtenidos en este estudio no ajustan adecuadamente a ninguna de las estructuras dimensionales. Las propuestas de Auzmendi (1992) y Méndez & Macía (2007) mostraron índices de ajuste no aceptables, y la de Darías (2000) manifestó problemas de identificación que exigía añadir diferentes restricciones (tabla 1).

Tabla 1: Análisis factorial confirmatorio de trabajos previos.

Trabajo	Modelo	Índices de bondad de ajuste					
		GFI	NFI	TLI	CFI	RMSEA	χ^2/gl
Auzmendi (1992)	5 factores independientes, 1 factor superior	,74	,71	,79	,81	,096	2,31
Méndez y Macía (2007)	Cuatro factores relacionados	,75	,73	,80	,82	,094	2,26
Darías (2000)	4 factores independientes, 1 factor superior	Modelo con problemas de identificación que exige añadir restricciones					

A partir de aquí se procedió con múltiples análisis factoriales exploratorios en aras de encontrar una estructura de factores robustos, consistentes y unipolares, para lo que se llevó a cabo un exigente proceso de selección de ítems de acuerdo con tres criterios teóricos y metodológicos establecidos desde un principio. El primer criterio parte de la premisa de que una escala no tiene mayor calidad técnica por tener más ítems, sino por garantizar la máxima explicación de varianza sin perder validez de contenido. En este sentido, es posible una estructura de cinco factores, como propone Auzmendi (1992), o de cuatro factores, como defienden Vargas & Mondéjar (2009); si bien, según otras perspectivas teóricas (Gil 1999, Gómez 2000, Bazán & Aparicio 2006), también es pertinente una estructura de tres factores correlacionados con vinculación a las esferas fisiológica, cognitiva y conductual. El segundo criterio exigía que el modelo factorial garantizara al menos el 60 % de la varianza de las puntuaciones, que los ítems saturasen en su factor de pertenencia por encima de 0,50, sin cargar de forma estadísticamente significativa en otros factores, y que todos los ítems alcanzaran una comunalidad mínima de 0,50 (Hair et al. 2004). Finamente, el tercer criterio establecía que el instrumento, sea cual fuera su estructura factorial, debía alcanzar una fiabilidad de 0,80, donde todos y cada uno de los ítems covariasen entre ellos con correlaciones superiores a 0,30, sin implicar un crecimiento de la fiabilidad global de la escala en el caso de que alguno de los ítems fuera eliminado (Martínez, Hernández & Hernández 2006). Así, después de transformar o invertir la escala de los ítems negativos (Apéndice), se obtuvo una estructura dimensional de tres factores y doce

ítems con capacidad para explicar el 68 % de la varianza del instrumento, mediante método de extracción de componentes principales. Posteriormente la solución factorial fue rotada mediante rotación Promax, al ser un método oblicuo adecuado para factores que estén correlacionados (Pardo & Ruiz 2002). Se alcanzaron pesos factoriales entre 0,52 y 0,92, comunalidades entre 0,51 y 0,80, y correlaciones interelementos entre 0,40 y 0,73. La escala mostró alta consistencia interna con un coeficiente alfa de 0,87, sin que la eliminación de ningún ítem elevara la fiabilidad global. Se muestra en tabla 2 la matriz factorial rotada, las saturaciones superiores a 0,35, y los valores de homogeneidad y de comunalidad.

Tabla 2: Análisis factorial exploratorio y análisis de fiabilidad.

ítems	Factores			Comunalidad	Correlación	Alfa sin
	1	2	3		ítem-total	ítem
12	,92			,80	,53	,86
13	,85			,73	,55	,86
22	,82			,73	,59	,85
7	,75			,64	,58	,85
24		,86		,67	,53	,86
19		,85		,70	,61	,85
4		,77		,67	,64	,85
14		,65		,71	,73	,84
10			,97	,79	,44	,86
5			,70	,56	,51	,86
1			,60	,63	,54	,86
16			,52	,51	,40	,86
Varianza	42 %	17 %	8 %			
Fiabilidad	,87	,83	,76			

Varianza global de la escala: 68 %

Fiabilidad global de la escala; $\alpha =, 87$

La primera y principal dimensión (42 % de varianza, $\alpha = 0, 87$) indica la respuesta fisiológica al aprendizaje de la estadística en su rasgo de calma/ansiedad. La segunda dimensión (17 % de la varianza, $\alpha = 0, 83$) recoge ítems relacionados con una predisposición activa o tendencia positiva hacia la estadística. Y la tercera dimensión (8 % de varianza, $\alpha = 0, 76$), de naturaleza cognitiva, hace referencia a qué es lo que piensan los estudiantes de la estadística y qué percepción tienen de su utilidad o importancia. Se muestra en tabla 3 la configuración de dimensiones e ítems, acompañados de estadísticos descriptivos de centralidad, dispersión y normalidad.

Tabla 3: Dimensiones e ítems. Estadísticos descriptivos.

ítems	Dimensiones	Media aritmética	Desviación típica	Asimetría	Curtosis
	Respuesta fisiológica de no ansiedad	2,99	1,00	,14	−,65
7	*La estadística es una de las asignaturas que más temo	2,50	1,33	,52	−,98
12	*Cuando me enfrento a un problema de estadística, me siento incapaz de pensar con claridad	3,23	1,11	−,28	−,65
13	Estoy calmado/a y tranquilo/a cuando me enfrento a un problema de estadística	3,01	1,11	,01	−,70
22	*La estadística hace que me sienta incómodo/a y nervioso/a	3,23	1,14	−,10	−,72
	Predisposición positiva y activa	2,23	,85	,40	−,44
4	El utilizar la estadística es una diversión para mí	2,07	,97	,67	−,09
14	La estadística es agradable y estimulante para mí	2,35	1,01	,32	−,65
19	Me gustaría tener una ocupación en la cual tuviera que utilizar la estadística	2,14	1,03	,63	−,23
24	Si tuviera oportunidad, me inscribiría en más cursos de estadística de los que son necesarios	2,39	1,13	,44	−,65
	Percepción de utilidad e importancia	2,85	,78	−,10	−,39
1	Considero la estadística como una materia muy necesaria en la carrera	3,13	,98	−,40	−,25
5	*La estadística es demasiado teórica como para ser de utilidad práctica para el profesional medio	3,50	1,08	−,41	−,40
10	*La estadística puede ser útil para el que se dedique a la investigación pero no para el profesional medio	2,90	1,14	−,059	−,75
16	*Para el desarrollo profesional de nuestra carrera considero que existen otras asignaturas más importantes que la estadística	1,88	,91	,91	,62
	*ítem negativos en los que se ha invertido la escala con el fin de positivar el significado del enunciado				

Posteriormente, se procedió con Análisis Factorial Confirmatorio. Pese al incumplimiento mayoritario de la normalidad de los ítems, se utilizó el método de extracción de máxima verosimilitud, ya que el método de mínimos cuadrados generalizados es "impracticable a medida que el modelo aumenta en tamaño y complejidad" (Hair et al. 2004, p. 630). El modelo es plausible de acuerdo con los índices de ajuste alcanzados: GFI=0,91, NFI= 0,898, TLI=0,94, CFI= 0,95, RMESA=0,069 y χ^2/gl= 1,68 (figura 1).

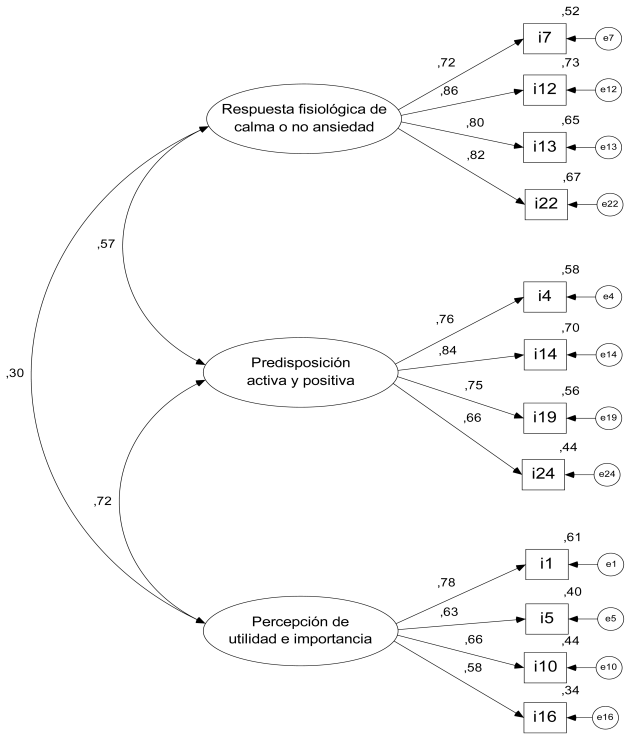

Figura 1: Análisis factorial confirmatorio.

GFI=,91, NFI= ,898, TLI=,94, CFI= ,95, RMESA=,069, χ^2/gl= 1,68

No se procedió a analizar la validez cruzada del modelo o la invarianza de los residuos debido al tamaño de la muestra.

4. Discusión y conclusiones

Una vez argumentada la importancia de la estadística en la nuevas titulaciones de grado en ciencias de la actividad física y del deporte, y en sintonía con la "creciente actitud positiva para abordar la problemática del aprendizaje de la estadística" (Behar & Grima 2001, p. 192), este trabajo ha analizado la estructura dimensional de la Escala de Actitudes hacia la Estadística (EAE)

(Auzmendi 1992) y también la estructura dimensional de otras propuestas basadas en la EAE (Darías 2000, Méndez & Macía 2007), utilizando una muestra representativa de estudiantes universitarios de ciencias del deporte.

Al efectuar la revisión bibliográfica, la EAE no muestra indicios de estabilidad factorial, pues la estructura dimensional propuesta por Auzmendi (1992) sólo fue parcialmente validada por Darías (2000), y ambas difieren sustancialmente de la encontrada por Méndez & Macía (2007). Al respecto, procediendo con análisis factorial confirmatorio y con los datos de este estudio, una primera conclusión es que la EAE –tal y como se había postulado hasta el momento: cinco o cuatro dimensiones y 25 ó 24 ítems–, no es un instrumento válido en la actualidad para estudiantes de ciencias de la actividad física y del deporte.

En relación con esta primera conclusión, más allá de que el estudio de Méndez & Macía (2007) utilizara un método de extracción y rotación distinto al empleado por Auzmendi (1992) y Darías (2000), y reconociendo la posible singularidad de la muestra, es decir, que los estudiantes universitarios españoles de ciencias del deporte pudieran diferenciarse de otro tipo de estudiantes, como por ejemplo los estudiantes de psicología o los estudiantes universitarios chilenos, a juicio de los autores de este trabajo la razón primordial de la no plausibilidad de la EAE se debe a la propia evolución de la enseñanza de la estadística desde 1992, momento del primer estudio de Auzmendi, a la actualidad, donde el desarrollo de aplicaciones informáticas estadísticas permite impulsar procesos de enseñanza-aprendizaje de mayor lógica aplicada, de tal forma que "los paquetes de software estadístico no solo sirven para ayudar en los cálculos, también se pueden utilizar para ver los conceptos clave" (Grima 2009, p. 25).

Así, se ha llevado a cabo una exigente selección de ítems de la EAE, hasta alcanzar un instrumento actual, válido y fiable, con dimensiones unipolares, consistentes y robustas. Se propone un instrumento de doce ítems y tres dimensiones, bien justificado no sólo psicométricamente sino también desde una perspectiva teórica y de congruencia de significado o contenido.

La primera dimensión, con un relevante componente fisiológico, hace referencia a la respuesta de calma/ansiedad que se produce en el aprendizaje de la estadística, siendo el factor con mayor capacidad explicativa y fiabilidad (42 % de la varianza y consistencia interna igual a 0,87). Datos que ratifican la idea de que la ansiedad es la dimensión fundamental del constructo actitud hacia la estadística (Baloglu & Zelhart 2003, Onwuegbuzie & Wilson 2003). Los ítems de esta dimensión son los siguientes: "La estadística es una de las asignaturas que más temo", "Cuando me enfrento a un problema de estadística, me siento incapaz de pensar con claridad", "Estoy calmado/a y tranquilo/a cuando me enfrento a un problema de estadística", y "La estadística hace que me sienta incómodo/a y nervioso/a".

La segunda dimensión, pudiendo interpretarse como un indicador de la esfera conductal, tiene que ver con la predisposición activa y tendencia positiva hacia la estadística. Explica el 17 % de la varianza, con un coeficiente α de 0,83. Queda formado por los siguientes elementos: "El utilizar la estadística es una diversión para mí", "La estadística es agradable y estimulante para mí", "Me gustaría tener una

ocupación en la cual tuviera que utilizar la estadística", y "Si tuviera oportunidad, me inscribiría en más cursos de estadística de los que son necesarios".

Y la tercera dimensión, siendo la dimensión más débil desde una perspectiva metodológica de medición: 8 % de varianza y α=0,76, tiene un claro componente cognitivo, haciendo referencia a qué es lo piensan los estudiantes de la estadística y qué percepción tienen de su utilidad o importancia. La relación de ítems de esta dimensión es la siguiente: "Considero la estadística como una materia muy necesaria en la carrera", "La estadística es demasiado teórica como para ser de utilidad práctica para el profesional medio", "La Estadística puede ser útil para el que se dedique a la investigación pero no para el profesional medio", y "Para el desarrollo profesional de nuestra carrera considero que existen otras asignaturas más importantes que la estadística".

De acuerdo con los parámetros estandarizados obtenidos en el análisis factorial confirmatorio, el valor máximo de correlación se encontró entre las dimensiones predisposición activa y percepción de utilidad ($r =, 72$), seguida de la asociación entre no ansiedad y predisposición activa ($r =, 57$) y, en menor medida, la relación entre percepción de utilidad y respuesta de calma o no ansiedad ($r =, 30$). En todo caso, existe una covariación lineal positiva estadísticamente significativa en todas las ocasiones ($p <, 001$), de tal forma que a valores altos en una dimensión corresponden valores altos en las otras dimensiones. Valores de asociación que son lógicos y coherentes con el planteamiento teórico: un estudiante con buena actitud hacia la estadística mostrará bajos niveles de ansiedad y altos valores de predisposición activa y percepción de utilidad, ocurriendo lo contrario para el caso de un estudiante con baja actitud hacia la estadística.

Finalmente, de acuerdo con una escala entre 1 y 5, los datos del estudio permiten inferir que los estudiantes de ciencias de la actividad física declaran niveles medios de ansiedad (M= $2, 99$; D$T = 1$), consideran que la utilidad o importancia de la estadística es media-baja (M= $2, 85$, DT= $0, 78$), y su predisposición hacia la materia es baja (M= $2, 23$; DT= $0, 85$). Resultados similares a los de Mondéjar et al. (2008), y menos optimistas que los defendidos por Gil (1999) y Estrada et al. (2004).

Éste ha sido el primer estudio en idioma español sobre medición de actitudes hacia la estadística en estudiantes de Ciencias de la Actividad Física y del Deporte.

[Recibido: septiembre de 2010 — Aceptado: enero de 2011]

Referencias

ANECA (2009), *Libro Blanco de Título de Grado en Ciencias de la Actividad Física y Del Deporte*, Agencia Nacional de Evaluación de la Calidad y Acreditación.

Auzmendi, E. (1992), *Las Actitudes Hacia la Matemática-Estadística en las Enseñanzas Medias y Universitarias*, segunda edn, Mensajero, Bilbao, España.

Baloglu, M. & Zelhart, P. F. (2003), 'Statistical anxiety: a detailed review', *Psychology and Education* **40**, 27–37.

Bazán, J. L. & Aparicio, A. S. (2006), 'Las actitudes hacia la matemática-estadística dentro de un modelo de aprendizaje', *Revista de Educación de la Pontificia Universidad Católica del Perú* **15**(28), 7–20.

Behar, R. & Grima, P. (2001), 'Mil y una dimensiones del aprendizaje de la estadística', *Estadística Española* **43**(148), 189–207.

Blanco, A. (2004), Enseñar y aprender estadística en las titulaciones universitarias de ciencias sociales: apuntes sobre el problema desde una perspectiva pedagógica, *in* J. C. Torres & J. Gil, eds, 'Hacia una Enseñanza Universitaria Centrada en el Aprendizaje', Universidad Pontificia de Comillas, Madrid, España.

Blanco, A. (2008), 'Una revisión crítica de la investigación sobre las actitudes de los estudiantes universitarios hacia la estadística', *Revista Complutense de Educación* **19**(2), 311–320.

Carmona, J. (2004), 'Una revisión de las evidencias de fiabilidad y validez de los cuestionarios de actitudes y ansiedad hacia la estadística', *Statistics Education Research Journal* **3**(1), 5–28.

Chang, L. (1996), 'Quantitative attitudes questionnaire: instrument development and validation', *Educational and Psychological Measurement* **56**(6), 1037–1042.

Darías, E. J. (2000), 'Escala de actitudes hacia la estadística', *Psicothema* **12**(2), 175–178.

Estrada, A. (2009), *Las Actitudes Hacia la Estadística en la Formación de los Profesores*, Milenio, Lleida, España.

Estrada, A., Batanero, C. & Fortuny, J. M. (2004), 'Un estudio comparado de las actitudes hacia la estadística en profesores en formación y en ejercicio', *Enseñanza de las Ciencias* **22**(2), 263–274.

Gil, J. (1999), 'Actitudes hacia la estadística. incidencia de las variables sexo y formación previa', *Revista Española de Pedagogía* **214**, 567–590.

Gómez, I. M. (2000), *Matemática Emocional. Los Afectos en el Aprendizaje Matemático*, Narcea, Madrid, España.

Grima, P. (2009), Ideas y experiencias acerca de la enseñanza de la estadística, *in* 'II Encuentro Iberoamericano de Biometría', Universidad Veracruzana, Veracruz, México.

Hair, J. F., Anderson, R. E., Tatham, R. L. & Black, W. C. (2004), *Análisis Multivariante*, Pearson, Madrid, España.

Martínez, M. R., Hernández, M. J. & Hernández, M. V. (2006), *Psicometría*, Alianza Editorial, Madrid, España.

Méndez, D. & Macía, F. (2007), 'Análisis factorial confirmatorio de la escala de actitudes hacia la estadística', *Cuadernos de Neuropsicología* **3**(1), 174–371.

Miller, R. B., Behrens, J. T., Green, B. A. & Newman, D. (2007), 'Goals and perceived ability: impact on student valuing, self-regulation and persistence', *Contemporary Educational Psychology* **18**, 2–18.

Ministerio de Educación y Ciencia (2007), *REAL DECRETO 1393/2007, de 29 de octubre, por el que se establece la ordenación de las enseñanzas universitarias oficiales*, Boletín Oficial de Estado de 3 de octubre de 2007, Madrid, España.

Mondéjar, J. & Vargas, M. (2010), 'Determinant factors of attitude towards quantitative subjects: differences between sexes', *Teaching and Teacher Education* **3**(26), 688–693.

Mondéjar, J., Vargas, M. & Bayot, A. (2008), 'Medición de la actitud hacia la estadística. Influencia en los procesos de estudio', *Revista Electrónica de Investigación Psicoeducativa* **3**(16), 729–748.

Morales, P., Urosa, B. & Blanco, A. (2003), *Construcción de Escalas de Actitud Tipo Likert*, La Muralla, Madrid, España.

Muñóz, I. (2002), Actitudes hacia la estadística y su relación con otras variables en alumnos universitarios del área de las ciencias sociales, PhD thesis, Universidad Pontificia de Comillas.

Onwuegbuzie, A. J. & Wilson, V. A. (2003), 'Statistics anxiety: nature, etiology, antecedents, effects, and treatments. a comprehensive review of the literature', *Teaching in Higher Education* **2**(8), 195–209.

Pardo, A. & Ruiz, M. A. (2002), *SPSS 11: Guía para el Análisis de Datos*, McGrawHill, Madrid, España.

Pleno de la Conferencia (2007), celebrado el 14 de diciembre de 2007, Conferencia Española de Institutos y Facultades de Ciencias de la Actividad Física y del Deporte, Instituto Nacional de Educación Física, Madrid, España.

Roberts, D. M. & Bilderback, E. W. (1980), 'Reliability and validity of a statistics attitude survey', *Educational and Psychological Measurement* **40**, 235–238.

Sánchez-López, C. R. (1996), 'Validación y análisis ipsativo de la escala de actitudes hacia la estadística', *Análisis y modificación de conducta* **86**(22), 799–819.

Schau, C., Stevens, J., Dauphinee, T. L. & Del Vecchio, A. (1995), 'The development and validation of the survey attitudes toward statistics', *Educational and Psychological Measurement* **5**(55), 868–875.

Vargas, M. & Mondéjar, J. (2009), 'Structure of latent factors in the learning of statistics', *Journal of College Teaching and Learning* **3**(6), 49–54.

Velandrino, A. P. & Parodi, L. H. (1999), La escala de actitudes hacia la estadística (eae): Desarrollo y propiedades psicométricas, *in* 'Conferencia Internacional Experiencias e Expectativas do Ensino da Estatística: Desafíos para o Século XXI', Florianópolis, Brasil.

Wise, S. L. (1985), 'The development and validation of a scale measurement attitudes toward statistics', *Educational and Psychological Measurement* **45**, 401–405.

Zeidner, M. (1991), 'Statistics and mathematics anxiety in social science students: some interesting parallels', *British Journal of Educational Psychology* **61**, 319–328.

Apéndice
Escala de Actitudes hacia la estadística (Auzmendi 1992)

1. **Considero la estadística como una materia muy necesaria en la carrera**
2. *La asignatura de Estadística se me da bastante mal*
3. El estudiar o trabajar con la estadística no me asusta en absoluto
4. **El utilizar la estadística es una diversión para mí**
5. ***La estadística es demasiado teórica como para ser de utilidad práctica para el profesional medio***
6. Quiero llegar a tener un conocimiento más profundo de la estadística
7. ***La estadística es una de las asignaturas que más temo***
8. Tengo confianza en mí mismo/a cuando me enfrento a un problema de estadística
9. Me divierte el hablar con otros de estadística
10. ***La estadística puede ser útil para el que se dedique a la investigación pero no para el profesional medio***
11. Saber utilizar la estadística incrementaría mis posibilidades de trabajo
12. ***Cuando me enfrento a un problema de Estadística, me siento incapaz de pensar con claridad***
13. **Estoy calmado/a y tranquilo/a cuando me enfrento a un problema de estadística**
14. **La Estadística es agradable y estimulante para mí**

15. *Espero tener que utilizar poco la estadística en mi vida profesional*
16. ***Para el desarrollo profesional de nuestra carrera considero que existen otras asignaturas más importantes que la estadística***
17. *Trabajar con la estadística hace que me sienta muy nervioso/a*
18. No me altero cuando tengo que trabajar en problemas de estadística
19. **Me gustaría tener una ocupación en la cual tuviera que utilizar la Estadística**
20. Me provoca una gran satisfacción el llegar a resolver problemas de estadística
21. Para el desarrollo profesional de mi carrera una de las asignaturas más importantes que ha de estudiarse es la estadística
22. ***La estadística hace que me sienta incómodo/a y nervioso/a***
23. Si me lo propusiera creo que llegaría a dominar bien la estadística
24. ***Si tuviera oportunidad me inscribiría en más cursos de estadística de los que son necesarios***
25. La materia que se imparte en las clases de estadística es muy poco interesante

Nota.
Los ítems seleccionados se presentan en negrita.
Los ítems negativos con escala a invertir se presentan en cursiva

Revista Colombiana de Estadística
Junio 2011, volumen 34, no. 1, pp. 15 a 37

Donde se muestran algunos resultados de atribución de autor en torno a la obra cervantina

Wherein are Shown some Results of Autorship Attribution to Cervantes' Work

Freddy López[a]

Departamento de Matemáticas, Instituto Venezolano de Investigaciones Científicas, Estado Miranda, Venezuela

Resumen

En este artículo se aplican algunos métodos de clasificación a un conjunto de textos con el objetivo de estudiar la probabilidad que el libro *Novela de la tía fingida* haya sido escrita por Miguel de Cervantes. Esta novela se le ha atribuido históricamente, pero existen algunas posiciones encontradas al respecto. Los métodos usados en este artículo contemplan: regresión logística, regresión logística aditiva, análisis discriminante lineal, cuadrático, regularizado, de mezclas y flexible, árboles de clasificación, método de los k-ésimos vecinos más cercanos, método de Bayes ingenuo y máquinas de soporte vectorial.

Los métodos fueron calibrados y aplicados utilizando un corpus de autores contemporáneos a Cervantes (Lope de Vega, Jerónimo de Pasamonte, Alonso Fernández de Avellaneda, Mateo Alemán y Francisco de Quevedo) junto con más de cuarenta variables, principalmente palabras y signos de puntuación, medidas sobre muestras de los textos escritos por estos autores.

Con respecto a estos métodos, la mayoría clasifica la obra como cervantina; sin embargo, es recomendable ampliar el corpus utilizado para el estudio e incluir más autores para la comparación.

***Palabras clave*:** análisis discriminante, árboles de clasificación, máquinas de aprendizaje, regla de Bayes, regresión logística, validación cruzada.

Abstract

In this paper, some classification methods are applied to a set of texts with the aim of studying the probability that the book *Novela de la tía fingida* has been written by Miguel de Cervantes. This novel has been historically attributed to him but there are some encountered positions about this. The methods used in this paper range from: logistic regression, additive logistic

[a]Estudiante de postgrado. E-mail: freddy.vate01@gmail.com

regression, linear, quadratic, regularized, mixture and flexible discriminant analysis, classification tree, k-nearest neighbour, Naive Bayes method and support vector machines.

Methods were trained and applied using a corpus of authors contemporary to Cervantes as Lope de Vega, Jerónimo de Pasamonte, Alonso Fernández de Avellaneda, Mateo Alemán, and Francisco de Quevedo and more than forty variables, mainly words and punctuation marks, measured over written texts by these authors.

Respect to these methods, most of them classify the novel as another Cervantes' work; however, is our recommendation to include more texts from these authors and more authors.

***Key words*:** Bayes rule, Classification tree, Cross validation, Discriminant analysis, Logistic regression, Machine learning.

1. Introducción

El problema de atribución de autor hace referencia a la asignación de un autor a un texto cuya autoría es desconocida (anónima) y es un problema que puede abordarse de diversas maneras, como por ejemplo, atribuir un texto a determinado autor por el lugar donde fue encontrado, por semejanzas y giros del lenguaje propios de un autor, por el estilo, por el tema tratado, por el metro y ritmo (en textos poéticos), etc.

La atribución *cuantitativa* consiste en realizar mediciones al texto anónimo, y compararlas con textos de autores de la época y asignar la autoría del texto a aquel autor al que esté estadísticamente más cercano.

Los problemas de atribución pueden pensarse como un problema de clasificación bajo el supuesto de que se conocen con certeza los posibles autores del texto. Muchas veces no se puede estar plenamente seguro de que los autores postulados sean efectivamente posibles autores, y es un problema que puede no tener solución.

Si se tiene un grupo de candidatos a autores que sea confiable, entonces es razonable aplicar técnicas estadísticas (multivariantes), reducción de dimensiones y técnicas de clasificación. En general, las variables que se utilizan en este tipo de trabajo son el resultado del conteo de palabras más frecuentes (ver sección 2).

En la literatura reciente se encuentra que Jockers, Witten & Criddle (2008), comparan dos métodos de clasificación para determinar la autoría del *Libro del Mormón* entre un grupo de posibles autores (Salomón Spalding, Sidney Rigdon y Oliver Cowdery). Al final, el estudio concluye que el libro es autoría de Rigdon y Spalding.

Hoover (2002) investiga el análisis de conglomerados dentro de los textos de un mismo autor y logra clasificar correctamente todas las secciones de todas las novelas investigadas. Cada una de estas secciones consta de 2500 palabras. Lamentablemente, reporta que cuando el número de palabras más frecuentes se hace muy grande, la clasificación puede fallar. Luego propone utilizar el orden de aparición de las palabras más frecuentes dentro de la sección estudiada.

Binongo (2003) utiliza los dos primeros componentes principales para la distinción de las características de los dos autores que estudia: L.F. Baum y R.P. Thompson. El problema abordado por Binongo fue la autoría del libro *The Royal Book of Oz* (el libro número 15 de la saga; ver Baum 2001), y contó con esos dos autores potenciales. En su trabajo, el primer componente separó claramente a los dos autores y parte de la validación de sus resultados la logró incluyendo, de forma ilustrativa (Lebart, Morineau & Warwick 1984) otros trabajos de Baum. El trabajo concluye gratamente con la inclusión del libro de Martin Gardner *Visitors from Oz* (Gardner 1998) y notando que la forma de escribir de Gardner está más próxima al estilo de Thompson que de Baum, creador original de la historia.

Koppel, Schler & Argamon (2009) dan un resumen sobre las técnicas estadísticas empleadas hasta hoy y aborda especialmente el tema de las máquinas de soporte vectorial. Proponen un método denominado *desenmascaramiento*, cuya idea principal consiste en remover, por etapas, las variables que son más útiles al momento de separar un texto de un autor y de otro; de esta forma, la precisión en la clasificación se deteriorará conforme se vayan removiendo variables. La idea es que cuando un texto pertenece a un autor, sus características propias permanecerán en él a pesar del tema o género tratado. Cuando las variables que mejor separan han sido removidas, los textos de un mismo autor serán prácticamente indistinguibles.

Dos excelentes trabajos de recopilación y fuentes bibliográficas son el trabajo de Joula (2006) y el trabajo de Grieve (2007). Este último compara las distintas variables frecuentemente utilizadas en la atribución de autor.

Este trabajo está organizado como sigue: La sección 2 describe las variables a ser utilizadas, cómo se trabajan y cuáles serán los textos a utilizar; la sección 3 lista los métodos que se emplearán dando una cortísima descripción de cada uno; la sección 4 presenta los resultados de entrenar y aplicar los métodos con los textos bajo examen, incluyendo la aplicación a la *Novela de la tía fingida*, atribuida a Cervantes y, por último, se presentan algunas conclusiones.

2. Procesamiento de los textos

2.1. Las variables

El procedimiento inicial consistió en tomar un libro y considerar sus divisiones naturales, los capítulos, como individuos. En algunos casos no fue posible dado que el tamaño del capítulo no era razonable o el libro no presentaba divisiones. Estos puntos serán abordados más adelante en la sección 4. Así, a estos individuos les serán medidas las siguientes variables:

1. **Conteo de las palabras más utilizadas.** Una palabra es una cadena de caracteres delimitada por un signo de puntuación o espacio. Estas se buscan de la siguiente manera: se colapsan todos textos en estudio, se contabilizan todas las palabras y se ordenan. Las palabras más frecuentes son utilizadas entonces como variables (en términos de estadística multivariante).

2. **Conteo de las frases de dos y tres palabras más utilizadas.** Estas frases de dos y tres palabras se encuentran de la misma manera como se describe en el ítem anterior, con la diferencia de que no son palabras aisladas. Estas no pueden ser muchas porque no se repiten mucho en todos los textos.

3. **Conteo de signos de puntuación más utilizados.** La cantidad de signos de puntuación utilizados en una novela es amplia. Para utilizar los signos de puntuación como variables o características, son necesarias aquellas que más se repiten.

4. **Medida de riqueza verbal.** Existen muchas medidas de riqueza verbal (un listado de ellas puede ser consultado en Grieve 2007). Una de ellas, la que es utilizada aquí, consiste en el número de palabras distintas divididas entre el número palabras utilizadas en el texto. Una pequeña introducción de esta medida puede revisarse en Bird, Klein & Loper (2009, p. 8).

Así, para todos los textos utilizados (ver sección 2.4),

```
"que",    "de",     "y",     "la",    "a",      "en",   "el",  "no",
"con",    "los",    "se",    "por",   "lo",     "las",  "su",  "le",
"me",     "como",   "del",   "un",    "si",     "mi",   "es",  "yo",
"para",   "al",     "una",   "dijo",  "porque", "ni"
```

son las 30 palabras más repetidas de un total de 47.560 palabras, mientras que

```
",", ".", ";", "-", ":", "?", "£", "ą", "!"
```

son los signos de puntuación que más se repiten, de un total de 28 signos de puntuación.

Además de ellos,

```
"de la"  ,  "lo que",  "que no"
```

son las 3 frases de dos palabras más usadas, y

```
"de lo que"
```

es la frase de tres palabras más repetida.

Estas características son contadas en cada capítulo y guardadas en un registro, junto a la medida de riqueza verbal (`RT`). Frecuentemente, a estas características (signos de puntuación y palabras) se les denomina *tokens*. Es importante observar que los *tokens* anteriores no se encuentran necesariamente en todos los textos utilizados. En efecto, los *tokens* definidos arriba son los que más se repiten en todos los textos a ser utilizados y no hay garantía que cada uno de ellos se encuentre en cada capítulo.

Además, se cuentan cuántos *tokens* tiene cada fragmento, y luego se divide por él cada una de las mediciones anteriores; de esta manera, se tienen registros estandarizados por el número de *tokens* que tiene cada fragmento a ser estudiado.

En la siguiente sección se hará una pequeña demostración de cómo se trabaja con cada capítulo.

2.2. Ejemplo

Supóngase que es de interés obtener las medidas anteriormente descritas al siguiente fragmento del prólogo a las *Novelas Ejemplares* de Cervantes.

> A esto se aplicó mi ingenio, por aquí me lleva mi inclinación, y más, que me doy a entender, y es así, que yo soy el primero que he novelado en lengua castellana, que las muchas novelas que en ella andan impresas todas son traducidas de lenguas estranjeras (sic), y éstas son mías propias, no imitadas ni hurtadas: mi ingenio las engendró, y las parió mi pluma, y van creciendo en los brazos de la estampa. Tras ellas, si la vida no me deja, te ofrezco los *Trabajos de Persiles*, libro que se atreve a competir con Heliodoro, si ya por atrevido no sale con las manos en la cabeza; y primero verás, y con brevedad dilatadas, las hazañas de don Quijote y donaires de Sancho Panza, y luego las *Semanas del jardín*.

Así, este fragmento, contiene seis veces la palabra `que`, cinco veces la palabra `de`, nueve veces la palabra `y`, etc. Nótese que no contiene las frases `lo que`, `que no`, `de lo que` mientras la frase `de la` aparece una vez. Estos valores serán divididos por la cantidad de *tokens* que haya en el texto.

El fragmento contiene además diecisiete comas (`,`), dos puntos (`.`), un punto y coma (`;`), un dos puntos (`:`) y no hay signos de admiración ni interrogación.

Este fragmento muestra 87 *tokens* distintos de un total de 153, así, su riqueza verbal, según se ha definido, es $\frac{87}{153} = 0,568627451$, y la proporción de la palabra `que`, por ejemplo, es $\frac{6}{153} = 0,03921569$.

De esta forma, para este fragmento, se cuenta con los siguientes valores observados para las variables estudiadas:

que	de	y	la	a
0,039215686	0,032679739	0,058823529	0,019607843	0,019607843
en	el	no	con	los
0,026143791	0,006535948	0,019607843	0,019607843	0,013071895
se	por	lo	las	su
0,013071895	0,013071895	0,000000000	0,039215686	0,000000000
le	me	como	del	un
0,000000000	0,019607843	0,000000000	0,006535948	0,000000000
si	mi	es	yo	para
0,013071895	0,026143791	0,006535948	0,006535948	0,000000000
al	una	dijo	porque	ni
0,000000000	0,000000000	0,000000000	0,000000000	0,006535948
s.coma	s.punto	s.puntoycoma	s.guión	s.dospuntos
0,111111111	0,013071895	0,006535948	0,000000000	0,006535948
s.intcerrado	s.intabierto	s.exclabierto	s.exclcerrado	RT
0,000000000	0,000000000	0,000000000	0,000000000	0,568627451
de_la	lo_que	que_no	de_lo_que	
0,006535948	0,000000000	0,000000000	0,000000000	

2.3. Reducción de dimensiones

Una vez se cuente con todas estas medidas para cada uno de los textos en estudio, se realizará un análisis de componentes principales (Jolliffe 2002) y se retendrán algunos componentes. Se compararán los resultados de clasificación con el uso de las variables originales y los componentes retenidos.

2.4. Textos a ser procesados

Se consideran textos de Cervantes así como de otros autores de la época. A continuación se listarán los autores y libros utilizados:

- De Cervantes: Las dos partes de Don Quijote: *El ingenioso hidalgo don Quijote de la Mancha* y *El ingenioso caballero don Quijote de la Mancha*; sus novelas ejemplares: *La gitanilla*, *El amante liberal*, *Rinconete y Cortadillo*, *La española inglesa*, *El licenciado Vidriera*, *La fuerza de la sangre*, *El celoso extremeño*, *La ilustre fregona*, *Las dos doncellas*, *La señora Cornelia*, *El casamiento engañoso* y *Los trabajos de Persiles y Segismunda*.
- De Lope de Vega: *La Dorotea* y *Novelas a Marcia Leonarda*.
- De Jerónimo de Pasamonte: *Vida y trabajos de Jerónimo de Pasamonte*.
- De Alonso Fernández de Avellaneda: *Segundo tomo del Ingenioso Hidalgo don Quijote de la Mancha*.
- De Mateo Alemán y de Enero: Los dos libros de Guzmán de Alfarache: *Primera parte de Guzmán de Alfarache*, *Segunda parte de la vida de Guzmán de Alfarache, atalaya de la vida humana*.
- Francisco de Quevedo y Villegas: El Buscón: *Historia de la vida del Buscón, llamado Don Pablos, ejemplo de vagabundos y espejo de tacaños*.

3. Métodos de clasificación

En esta sección se hará una muy corta explicación de los métodos que se emplearán aquí. Para un estudio en profundidad de cada técnica se recomienda revisar la bibliografía recomendada en cada aparte. En lo sucesivo, Y representará una variable dicotómica, tomando valor 1 cuando el texto evaluado pertenece a Cervantes y 0 en caso contrario. $\mathbf{x}$ representa, por su parte, las variables predictoras (que están en función de las variables descritas en la sección 2) con las que se busca explicar Y.

3.1. Regresión logística

La regresión logística es un tipo de regresión en donde la variable respuesta es categórica. En el caso que nos ocupa, es específicamente dicotómica. Siguiendo a Hosmer & Lemeshow (2000), la esperanza condicional que el resultado

(`1=Cervantes, 0=otro autor`) esté presente se denotará por $E(Y = 1 \mid \mathbf{x}) = \pi(\mathbf{x})$, donde $E(\cdot)$ es el operador esperanza. La forma específica del modelo de regresión logística es

$$\pi(\mathbf{x}) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} \tag{1}$$

que, luego de aplicar la 'transformación logit', toma la forma

$$g(\mathbf{x}) = \ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{2}$$

donde los parámetros a ser estimados son $\beta_0, \ldots, \beta_p$.

3.2. Regresión logística aditiva

Cuando cada término lineal es reemplazado por una función suavizada más general, digamos f_j, se obtiene la denominada regresión logística aditiva, cuya forma es

$$g(\mathbf{x}) = \beta_0 + f_1(x_1) + \cdots + f_p(x_p)$$

Para trabajar con esta ecuación se debe minimizar

$$\sum_{i=1}^{n}(y_i - \alpha - \sum_{j=1}^{p} f_j(x_{ij}))^2 + \sum_{j=1}^{p} \lambda_j \int f_j''(t_j)^2 dt_j \tag{3}$$

donde λ_j son parámetros que deben ser calibrados y, además, se puede demostrar que la minimización de (3) es un modelo aditivo de splines cúbicos. La forma más popular de maximizar esta ecuación y hallar a las funciones f_j, es conocida como 'algoritmo de *backfitting*'. Consiste básicamente en fijar un valor para α (por ejemplo el promedio de los valores de Y) y luego aplicar un suavizado de splines cúbicos a $\{y_i + \widehat{\alpha} + \sum_{k\neq j} \widehat{f}_k(x_{ik})\}$, $i = 1, \ldots, n$, para obtener un nuevo valor de $\widehat{f}_j$ (para mayores detalles consultar Hastie, Tibshirani & Friedman 2009).

3.3. Análisis discriminante lineal y cuadrático

Asúmase que se tienen dos poblaciones normales con distintos vectores de medias $\boldsymbol{\mu}_0$ y $\boldsymbol{\mu}_1$ e igual matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$. Asúmase que la de dimensión de esas poblaciones es p.

La función discriminante es la combinación lineal de las p variables que forman el conjunto de datos tal que se maximice la distancia entre los dos grupos de vectores de medias.

La combinación lineal es de la forma $z = \mathbf{a}'\mathbf{y}$, donde el vector de parámetros a estimar es $\mathbf{a}$. Se puede demostrar $\mathbf{a}$ está en función de $\mathbf{S}_i$, n_i, $\mathbf{S}_{\text{man}} = \frac{(n_0-1)\mathbf{S}_0+(n_1-1)\mathbf{S}_1}{n_0+n_1-2}$, los estimadores de la matriz de covarianza de cada población, los tamaños de muestra y la matriz mancomunada de varianzas y covarianzas con $\overline{\mathbf{y}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{y}_{ij}$, para $i = 0, 1$. Se utiliza $\mathbf{S}_{\text{man}}$ bajo el supuesto de que las dos

poblaciones involucradas tengan iguales matrices de varianzas y covarianzas. En este caso, una nueva observación, $\mathbf{x}_0$, se clasificará en una de las poblaciones de acuerdo con la cercanía de $\mathbf{a}'\mathbf{x}_0$ al punto medio $m = \frac{1}{2}(\overline{x}_0 + \overline{x}_1)$, donde $\overline{x}_i = \mathbf{a}\overline{\mathbf{y}}_i$ (ver Rencher (2002), capítulo 8 y Johnson & Wichern (1998), capítulo 11).

Por otra parte, si la igualdad en las matrices de varianzas y covarianzas no se puede sostener, se puede usar la regla de asignar $\mathbf{x}_0$ a la población que haga máximo el valor $Q_i(\mathbf{x}_0) = \log p_i - \frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}_0 - \overline{\mathbf{y}})^t\mathbf{S}_i^{-1}(\mathbf{x}_0 - \overline{\mathbf{y}})$, $i = 0, 1$, donde p_i son las probabilidades previas de cada grupo. Este cómputo, $Q_i(\mathbf{x}_0)$, es conocido como *puntaje de discriminación cuadrático*.

En general, los términos p_i son desconocidos y se suele trabajar asignándoles valores proporcionales al número de individuos que presenta cada población según su aparición en el conjunto de datos. En este trabajo se adoptará este enfoque (algunos autores, sin embargo, sugieren el uso de p_i completamente equilibrados. Ver por ejemplo Johnson & Wichern 1998, pp. 670-672).

3.4. Análisis discriminante regularizado

El análisis discriminante regularizado busca un equilibrio entre el análisis discriminante lineal y el análisis discriminante cuadrático; obligando a la matriz de covarianza muestral de cada población, $\mathbf{S}_k$, $k = 0, 1$, acercarse a la matriz mancomunada de covarianza $\mathbf{S}_{\text{man}}$, en un intento de reducir el sesgo en la estimación de los autovalores (Jolliffe 2002, p. 207). La regularización tiene la forma

$$\mathbf{S}_k(\alpha) = \alpha\mathbf{S}_k + (1 - \alpha)\mathbf{S}_{\text{man}}$$

con $k = 0, 1$ y $\alpha \in [0, 1]$; en la práctica se suele hallar el valor de α por validación cruzada.

3.5. Análisis discriminante de mezclas

Supóngase que los datos pueden ser expresados como una distribución de mezclas. Este supuesto puede establecerse cuando los grupos no son homogéneos o se sospeche que no lo son. Entonces, un modelo de mezclas normal para la k-ésima clase tiene una densidad expresada por

$$P(\mathbf{x} \mid Y = k) = \sum_{r=1}^{n_k} \pi_{kr}\phi(\mathbf{x}; \mu_{kr}, \mathbf{\Sigma})$$

donde la notación $\phi(\mathbf{x}; \mathbf{m}, \mathbf{s})$ representa la densidad normal de la variable $\mathbf{x}$, de media $\mathbf{m}$ y matriz de varianzas y covarianzas $\mathbf{s}$, los valores π_{kr} suman 1, n_k es el tamaño de cada población y $k \in \{0, 1\}$, las dos poblaciones. Adicionalmente, se asume que cada uno de los grupos tiene la misma matriz de varianzas y covarianzas $\mathbf{\Sigma}$. Ahora, dado el modelo normal anterior para cada clase, las probabilidades de clase posterior están dadas por

$$P(Y = k \mid \mathbf{x} = x) = \frac{\sum_{r=1}^{n_k} \pi_{kr}\phi(\mathbf{x}; \mu_{kr}, \Sigma)p_k}{\sum_{l=1}^{K}\sum_{r=1}^{n_l} \pi_{lr}\phi(\mathbf{x}; \mu_{lr}, \Sigma)p_l}$$

donde p_k son las probabilidades previas de cada clase, tal que $p_0 + p_1 = 1$. Estas últimas pueden ser vistas como la proporción de elementos pertenecientes a cada clase.

Los parámetros de los modelos normales se encuentran maximizando el logaritmo de la función de verosimilitud conjunta sobre $P(Y, \mathbf{x})$ gracias al algoritmo `EM` (Dempster, Laird & Rubin 1977). Este método y su forma de cómputo puede consultarse en detalle en Hastie et al. (2009, ver sección 12.7).

3.6. Análisis discriminante flexible

Supóngase que se cuenta con datos de la forma $(y_j, \mathbf{x}_j) = (y_j, x_1, \ldots, x_p)$, $j = 1, \ldots, n$, donde y_j puede tomar valores en $\{0, 1\}$ y x_j, $j = 1, \ldots, n$, es un conjunto de variables métricas.

Entonces se define una función $\theta : \{0, 1\} \longmapsto \mathbb{R}$ que asigna puntajes reales a las categorías de la variable respuesta, de manera que las clases así transformadas sean óptimamente predichas por una regresión lineal cuyas variables predictoras son $x_1, \ldots, x_p$, con parámetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$.

De esta forma el problema se reduce a resolver

$$\text{mín}_{\boldsymbol{\beta},\theta} \sum_{j=1}^{n} \left[\theta(y_j) - x_j^t \boldsymbol{\beta}\right]^2$$

Es decir, hallar aquellos valores de $\boldsymbol{\beta}$ y θ para los que la predicción sea mejor en términos del error cuadrado medio. En general, suponiendo que se cuente con K categorías para la variable respuesta, se pueden escoger hasta $L \leq K - 1$ conjuntos de puntajes independientes para las etiquetas de las clases, $\theta_1, \ldots, \theta_L$, con L correspondiendo a una función lineal tal que $\eta_l(\mathbf{x}) = \mathbf{x}^t \boldsymbol{\beta}_l$, $l = 1, \ldots, L$ escogidos para que sean óptimos para la regresión lineal en $\mathbb{R}^p$. Los puntajes $\theta_l(\cdot)$ y los coeficientes $\boldsymbol{\beta}_l$ son escogidos de manera que minimicen el error cuadrado promedio

$$ECM = \frac{1}{n} \sum_{l=1}^{L} \left[\sum_{j=1}^{n} (\theta_l(y_j) - \mathbf{x}_j^t \boldsymbol{\beta}_l)^2 \right]$$

Este tema puede ser consultado en Hastie et al. (2009, sección 12.5).

3.7. Árboles de clasificación

Los árboles de clasificación son básicamente objetos gráficos. Se construyen particionando el espacio de posibles observaciones dentro de subregiones que se corresponden con las *hojas*. Cada observación será clasificada en una hoja del árbol.

Asimismo, la forma de construir el árbol se diferencia de otras técnicas computacionales (más que estadísticas) principalmente en la estrategia de *poda* y la

estrategia al dividir y formar nodos. Algunos métodos y algoritmos son reseñados en Ripley (1996).

En principio, se considera un *atributo* A, el cual puede dividirse con el objetivo de tomar una decisión y la atención se centra entonces en hallar aquel valor que divide el atributo de la mejor manera. Este razonamiento se aplica a los demás atributos del problema.

La estrategia de poda entra en juego debido a que, como en cualquier problema estadístico multivariante, hay variables que no contribuyen al análisis y se hace importante seleccionar aquellas variables que sean determinantes y no tener *ramas* inútiles. Seleccionar aquellas ramas que son de verdadera utilidad y deshacerse de las que no es lo que se conoce como poda. Este trabajo se logra por medio de algoritmos computacionales y existen variantes de ello. El lector interesado puede consultar Ripley (1996).

La poda puede realizarse de dos formas principalmente: de forma manual o de forma automática. La forma manual supone una revisión por parte del investigador de aquella cantidad de hojas que son importantes en la discriminación (esto se logra, en general, por validación cruzada) mientras que la forma automática supone entregarle al software la libertad de encontrar aquel número de ramas adecuado. Ambas estrategias se explican en detalle en Venables & Ripley (2002, pp. 251-256).

3.8. Método de los k-ésimos vecinos más cercanos

Este método consiste en asignar la categoría de un individuo, dependiendo cómo estén distribuidos sus vecinos según las variables que lo caracterizan. Así, supóngase que se desea estudiar si el individuo $\mathbf{x}_j$ pertenece a uno de dos grupos, $y_j = \{0, 1\}$ (el conjunto de datos puede representarse nuevamente como $(y_j, \mathbf{x}_j) = (y_j, x_1, \ldots, x_p)$, $j = 1, \ldots, n$). Entonces se define a $N_k(\mathbf{x}_j)$ como la vecindad a $\mathbf{x}_j$ (según alguna métrica) teniendo en cuenta un entorno que considere solo k vecinos: los k vecinos más cercanos a $\mathbf{x}_j$. Con esta información, se calcula $\widehat{Y}(\mathbf{x}_j)$, que se define como

$$\widehat{Y}(\mathbf{x}_j) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x}_j)} y_i$$

Como las observaciones $y_j \in \{0, 1\}$, entonces una nueva observación será asignada a un grupo u otro si $\widehat{Y}(\mathbf{x}_j)$ es mayor o menor que 0,5. La métrica utilizada al definir la vecindad más cercana es comúnmente la distancia euclídea (para otras métricas, revisar pp. 197-198 de Ripley, 1996).

Para la elección del número de vecinos, k, se acostumbra utilizar validación cruzada.

3.9. Método de Bayes ingenuo

El método de Bayes ingenuo asume que, dada una clase $Y = j$, con $j \in \{0, 1\}$, las variables x_k son independientes y, por tanto, $P(\mathbf{x} \mid Y = j) = P_j(\mathbf{x}) =$

$\prod_{k=1}^{p} P_{jk}(x_k)$. Además, asume distribuciones normales para las variables predictoras que son métricas. En nuestro caso, todas las variables predictoras son de este tipo. Luego, trabajando con el Teorema de Bayes, se tiene la regla siguiente

$$P(Y = j \mid \mathbf{x}) = \frac{P_j(\mathbf{x})\pi_j}{\sum_{l=1}^{K} P_l(\mathbf{x})\pi_l} = \frac{\prod_{k=1}^{p} P_{jk}(x_k)\pi_j}{\sum_{l=1}^{K}\prod_{k=1}^{p} P_{jk}(x_k)\pi_l}$$

donde π_j son las probabilidades previas de cada población. Este cálculo sencillo, en general, para grandes volúmenes de datos, reduce significativamente los cálculos.

Una introducción a esta técnica puede ser estudiada en Witten & Frank (2005, pp. 94-97).

3.10. Máquinas de soporte vectorial

Supóngase que se está en el caso en que ningún punto de los dos grupos se superpone; esto es, que cada punto de cada clase está, digamos, *separado* del otro grupo. A esto se le conoce como *caso separable*, y supóngase que se cuenta con datos del tipo $(\mathbf{x}_j, y_j)$, $j = 1, \ldots, n$, con $y_j \in \{-1, 1\}$. Se define un hiperplano por

$$\left\{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^t\boldsymbol{\beta} + \beta_0 = 0\right\}$$

donde $\|\boldsymbol{\beta}\| = 1$, y la regla de clasificación inducida por $f(\mathbf{x})$ es $G(x) = \text{signo}(\mathbf{x}^t\boldsymbol{\beta} + \beta_0)$. De manera que se busca una línea que divide a los dos grupos. Además, Hastie et al. (2009) muestran que $f(\mathbf{x})$ es la distancia con signo de un punto $\mathbf{x}$ al hiperplano $f(\mathbf{x}) = \mathbf{x}^t\boldsymbol{\beta} + \beta_0 = 0$. Como las categorías son separables es posible hallar una función $f(\mathbf{x}) = \mathbf{x}^t\boldsymbol{\beta} + \beta_0$ con $y_i f(x_i) > 0\ \forall i$ lo que implica que se pueden crear los 'márgenes' más grandes entre los puntos de las clases -1 y 1. Un problema de optimización equivalente es

$$\text{mín}_{\boldsymbol{\beta},\beta_0,\|\beta\|=1}\, M \text{ sujeto a } y_i(\mathbf{x}^t\boldsymbol{\beta} + \beta_0) \geq M, i = 1, \ldots, n$$

donde M es la distancia mínima que existe de cada grupo a la recta $f(\mathbf{x}) = \mathbf{x}^t\boldsymbol{\beta} + \beta_0$; así, la distancia que existe de un grupo a otro es de $2M$.

En el caso que los puntos en las clases no sean separables, una de las formas de abordar el problema, es aún maximizar M pero permitir que algunos puntos estén en el lado incorrecto de los márgenes. Esto se logra incluyendo las variables de holgura $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ en el problema de optimización anterior (Hastie et al. 2009, capítulo 12).

Las soluciones consideradas de esta manera incluyen de entrada las variables originales $(x_1, \ldots, x_p)$ con el objetivo de hallar 'márgenes lineales'. Sin embargo, la mayoría de las veces resulta útil modificar estas variables para obtener en este espacio clases que estén más separadas; esto se logra aplicando diversas transformaciones, digamos $h_l(\cdot)$, a las variables de entrada x_j.

Los métodos más populares para transformar variables son los métodos de expansión de bases, especialmente polinomios. Éste consiste en tomar una variable, digamos x_k y representarla como $x_k = \sum_{m=1}^{L} \beta_l h_l(x_k)$, donde $h_l(\cdot)$ son polinomios.

4. Resultados del entrenamiento y clasificación

4.1. Preliminares. Análisis de componentes principales

Para el caso de los textos listados en la sección 2.4, se cuenta con una matriz de datos de dimensión 393×45, cuyas columnas se corresponden con las 44 variables descritas en la sección 2, junto con una variable respuesta que indica si el texto pertenece a Cervantes o no. El número de filas se corresponde principalmente con el número de capítulos que existen en los libros utilizados en el entrenamiento (ver sección 2.4). Sin embargo, para evitar distorsiones en algunas medidas, se limitaron las dimensiones de los capítulos a no menos de 1000 y a no más de 10000 *tokens*. Esto debido a que algunos capítulos son muy cortos (por ejemplo los primeros capítulos del Pasamonte: el primero con apenas 143 *tokens*) o son muy largos, como las novelas ejemplares que no presentan divisiones (por ejemplo La Gitanilla que cuenta con 28006 *tokens*). Cada capítulo de dimensión inferior a 1000 *tokens* fue unido al capítulo que le precedía o sucedía dependiendo de la situación. Cada capítulo de más de 10000 *tokens*, fue dividido en varias partes de 4000 *tokens*, cada una aproximadamente. Nótese que en esta etapa no se esta trabajando aún con la *Novela de la tía fingida*.

Ahora, la forma de las variables estudiadas no presenta en general asimetrías fuertes y presentan patrones como los que se pueden apreciar en la figura1. Los histogramas representan la cantidad de veces que se repitió en el conjunto de datos cada proporción de cada palabra. Así, por ejemplo, se encuentran más de cien fragmentos donde la palabra `que` tuvo una proporción entre 0,045 y 0,05. Tampoco hubo ningún texto donde la proporción de esta palabra estuviera por encima de 0,7.

Asimismo, es importante notar que los datos están siendo considerados como una matriz rectangular sin tomar en cuenta el posible efecto serial que pueda haber entre capítulos o individuos consecutivos. Esto debido a que se está trabajando con varios autores y varios libros. En aquellos casos donde se trabaje con un solo autor con una cantidad suficientemente grande de capítulos es recomenable hacer un estudio previo de la posible correlación serial existente en la secuencia de estos y utilizar las técnicas apropiadas. Nótese que la correlación serial no tiene por qué ser semejante para un mismo autor en diferentes libros. En la figura 2 se muestra la riqueza verbal para las tres series de datos más largas dentro del conjunto de datos en estudio (I Quijote: *El ingenioso hidalgo don Quijote de la Mancha*; II Quijote: *El ingenioso caballero don Quijote de la Mancha* y Persiles: *Los trabajos de Persiles y Segismunda*). Aparentemente no existe un patrón fuerte presente, a excepción, quizá, de la primera parte de don Quijote.

Ahora, a esta matriz de datos se le aplicó la técnica de componentes principales con el objetivo de trabajar con menos cantidad de variables que con el conjunto original (sin embargo, muchas veces es difícil saber qué componente, si lo hay, es el que discrimina mejor. Ver especialmente sección 9.1 de Jolliffe (2002), y las referencias allí citadas).

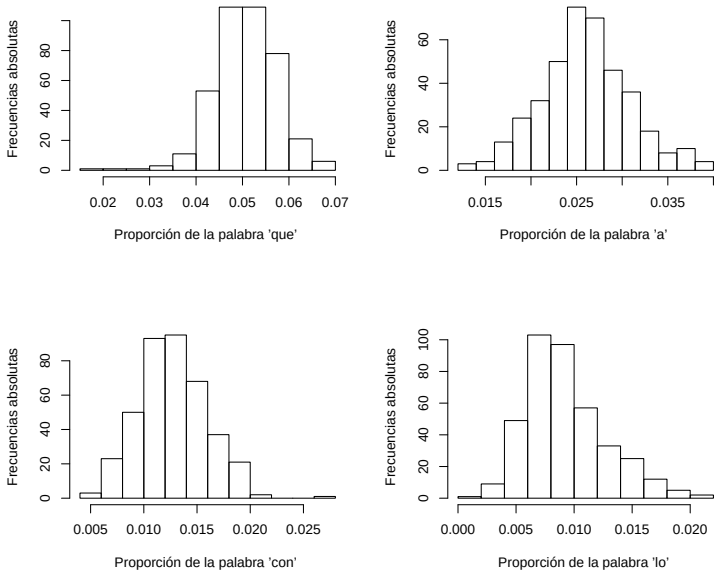

FIGURA 1: Histograma de la proporción de algunas palabras en los textos, a saber: `que, a, con, lo`.

La ejecución del método de componentes principales arroja los autovalores que se observan en la figura 3. Nótese que el criterio clásico de retener aquellos componentes mayores a la unidad se ve satisfecho con al menos 13 componentes. No obstante, en este trabajo se decidió trabajar con aquellos componentes que contribuyeron más en la discriminación según el método de regresión logística univariado de la forma que es sugerida por Hosmer & Lemeshow (2000, cap. 4 y 5). Esto es, se estimaron 44 modelos de la forma

$$g(\mathbf{x}) = \beta_0 + \beta_1 \text{COMPONENTE}_j, j = 1, \ldots, 44$$

donde $g(\mathbf{x})$ es definida como en la ecuación (2). De estos modelos se retuvieron aquellos componentes cuyo β_1 fuera significativo al 15 % puesto que cuando se utilizan valores p tradicionales (como 0,05) frecuentemente se falla al identificar variables que pueden ser importantes. Este procedimiento produjo la retención de 14 componentes. La cantidad de varianza explicada por cada uno de los componentes retenidos puede estudiarse en la tabla 1. Es interesante notar que existen componentes que explican muy poca cantidad de varianza ($<$2 %) pero sin embargo tienen alto poder de clasificación como el componente 38.

Una muestra de los dos primeros componentes principales puede apreciarse en las figuras 4 y 5. Se puede observar cómo estos dos componentes separan perfectamente la obra de Cervantes de Lope de Vega, Jerónimo de Pasamonte, Mateo Alemán y Francisco de Quevedo; mientras que, curiosamente, la separación no queda clara para Alonso Fernández de Avellaneda. Se puede observar asimismo

que el segundo componente en su lado negativo representa al relato autobiográfico por las variables `mi`, `yo` y `me` que allí se agrupan.

También se trabajó con las variables originales o un subconjunto de ellas. Este subconjunto se escogió utilizando el criterio de Akaike en un algoritmo por pasos aplicado a un modelo de regresión logística con todas las variables en estudio (Venables & Ripley 2002, p. 175). Esto es, se comenzó con el modelo inicial (con 44 variables predictoras)

$$g(\mathbf{x}) = \beta_0 + \beta_1 \text{que} + \beta_2 \text{de} + \cdots + \beta_{43} \text{que\_no} + \beta_{44} \text{de\_lo\_que}$$

donde $g(\mathbf{x})$ se define como en la ecuación (2) y utilizando las variables que fueron establecidas en la sección 2.2. En los pasos sucesivos de minimización del AIC se retuvieron en el modelo conjunto las variables siguientes: `que`, `y`, `a`, `el`, `no`, `se`, `por`, `lo`, `las`, `le`, `mi`, `yo`, `para`, `s.punto`, `s.guión` y `que_no`.

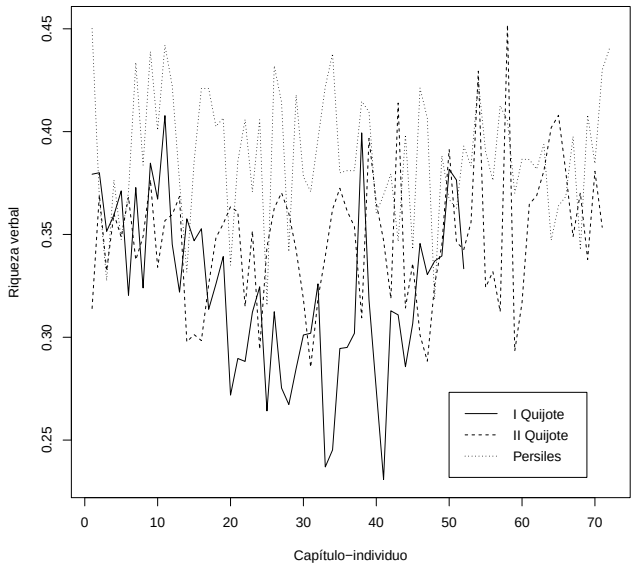

Figura 2: Tres libros más extensos vistos como una serie temporal.

Así, para la mayoría de las técnicas, se trabajó con dos conjuntos de datos principalmente: aquellos derivados de los componentes principales y aquellos que trabajan directamente sobre los datos originales.

4.2. Aplicación de las técnicas

Para el método de los vecinos más cercanos, se utilizaron cien repeticiones para cada número de vecinos para encontrar el número k más adecuado. El resultado puede observarse en la figura 6 y de allí se desprende que se haya escogido como

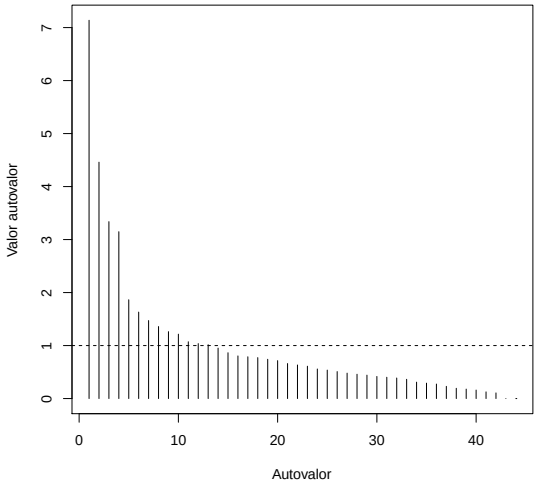

FIGURA 3: Autovalores: Método de componentes principales aplicado a las 44 variables en estudio

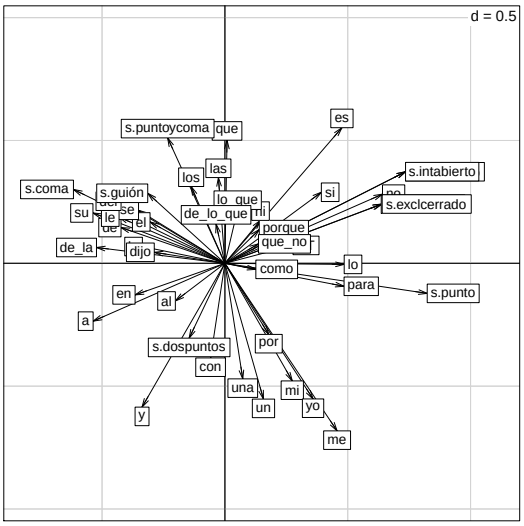

FIGURA 4: Gráfico de los dos primeros componentes principales (variables).

número óptimo 10 vecinos. Se puede notar que el promedio de desaciertos estuvo por debajo del 4 %, siendo un valor considerablemente bueno.

Para tener una idea de la efectividad del clasificador empleado, el conjunto de datos se dividió en dos partes seleccionadas completamente al azar: un 70 % de entrenamiento y 30 % como muestra de prueba. Con el conjunto de datos de entrenamiento se estimó el modelo correspondiente y se probó con el conjunto de

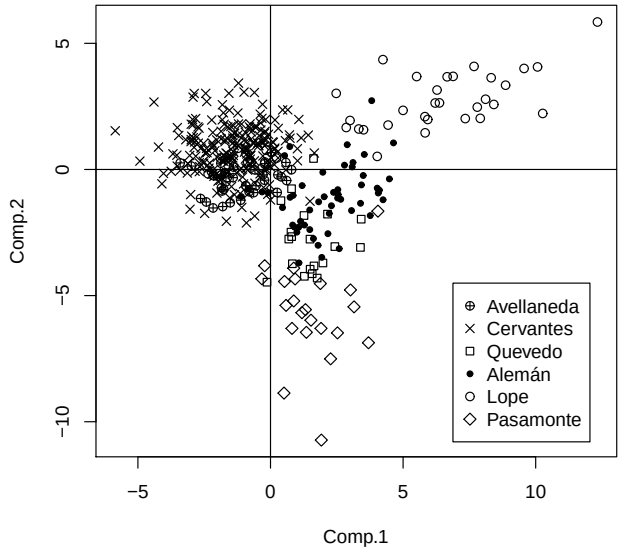

Figura 5: Gráfico de los dos primeros componentes principales (individuos).

Tabla 1: Cantidad de varianza explicada por los componentes retenidos.

No. de componente	1	2	3	4	5	6	7
% de var. explicada	16.22	10.14	7.59	7.16	4.24	3.71	3.35
% de var. acum.	16.22	26.36	33.95	41.10	45.34	49.05	52.40
No. de componente	8	9	12	13	20	31	38
% de var. explicada	3.09	2.87	2.36	2.31	1.63	0.92	0.44
% de var. acum.	55.49	58.36	60.71	63.02	64.65	65.57	66.01

datos para tal fin. El porcentaje de desaciertos fue guardado para dar una idea de lo efectivo que puede ser el método y este procedimiento se repitió 300 veces para cada clasificador (ver figura 7). En esta figura se han puesto, además, en color gris los métodos basados en los datos originales y en blanco los métodos que utilizaron como matriz de entrada algunos de los componentes principales.

Es importante notar que los cuatro primeros métodos con menor error de clasificación (máquinas de soporte vectorial, regresión logística, regresión logística generalizada y análisis discriminante de mezclas) se basan en datos originales y no en derivados del análisis de componentes principales. También es notable que más del 70 % de los métodos ostenten un error promedio de clasificación menor a 10 %, cuestión que pone de manifiesto la alta efectividad de la mayoría de ellos.

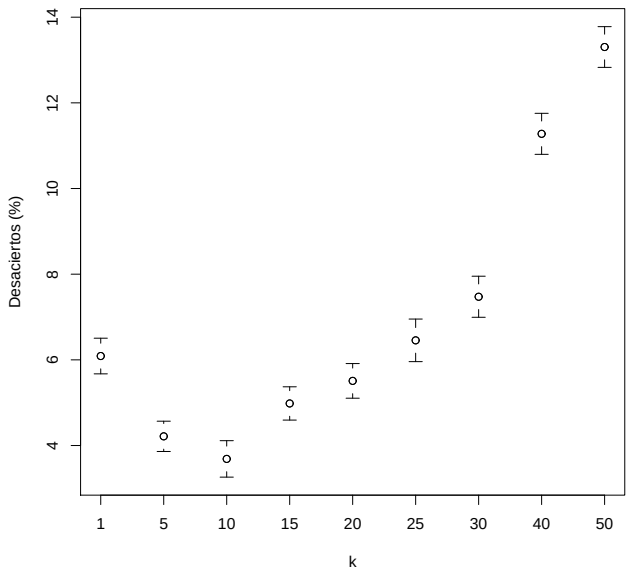

FIGURA 6: Escogencia del número de vecinos en el método de vecinos más cercanos. Nótese que el error de clasificación se hace menor con $k = 10$.

Por otro lado, los métodos basados en árboles de clasificación presentaron las tasas más altas de desaciertos para todas las variaciones ensayadas: todos en torno a 15 %, lo que sugiere la imposibilidad de esta técnica para afinar su calidad luego de cierto punto para este conjunto de datos (ver figura 8 y el detalle de un árbol en la figura 9).

4.3. El caso de la *Novela de la tía fingida*

El caso de la *Novela de la tía fingida* ha enigmado por mucho tiempo a los cervantistas. Hay posiciones encontradas con respecto a la autoría de esta novela. Hablando de los que apoyaban la tesis que Cervantes era su original autor, Andrés Bello, por ejemplo, decía (Aylward 1982, p. 27):

> ...se me acusará de temerario en poner este asunto otra vez en tela de juicio, mayormente después de lo que ha escrito, en modo incisivo i perentorio que acostumbra, don Bartolomé José Gallardo en el número 1ž de *El Criticón*. Pero, después de haber leído cuanto sobre esta materia me ha venido a las manos, que a la verdad no es mucho, no acabo de asegurarme...

y alega posteriormente razones de estilo y lenguaje entre las obras que sabe cervantinas y las que no.

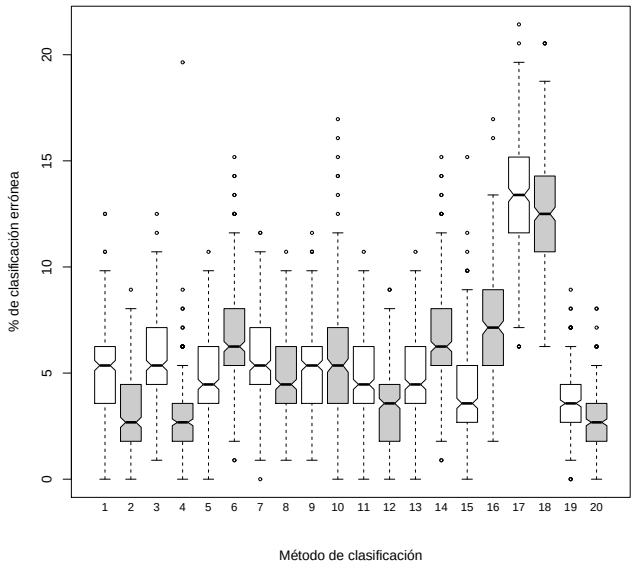

FIGURA 7: Métodos de clasificación: 1 y 2: Regresión logística. 3 y 4: Regresión logística aditiva generalizada. 5 y 6: Análisis discriminante lineal. 7 y 8: Análisis discriminante cuadrático. 9 y 10: Análisis discriminante regularizado. 11 y 12: Análisis discriminante mixto. 13 y 14: Análisis discriminante flexible. 15 y 16: Método de los vecinos más cercanos (10). 17 y 18: Método de Bayes ingenuo. 19 y 20: Máquinas de soporte vectorial (kernel radial). Se comparan los métodos utilizando los componentes principales y las variables originales.

Madrigal (2003), en un trabajo reciente, da razones para creer que el autor de esta novela es el propio Cervantes; y en su opinión, la atribución de una obra podría darse con frases como 'pasando por una calle' (que da inicio a la novela y que es usada sólo por Cervantes en algunas obras citadas por él). Posteriormente encuentra frases en novelas cervantinas y las compara con sus pares en la *Novela de la tía fingida* y concluye así que esta obra es de Cervantes.

Un resumen de la aplicación de las técnicas antes descritas se presenta en las tablas 2 y 3. En ellas se muestran dos columnas con probabilidades. La primera de ellas (con un signo ✠) muestra la probabilidad que la obra sea de Cervantes de acuerdo al método evaluado y curiosamente se obtiene que cerca del 70 % de ellos asignan la novela al grupo de Cervantes, no sin ciertas peculiaridades dignas de mención.

Se observa, por ejemplo, que todos los métodos que utilizan los componentes principales clasifican la novela como cervantina (ver tabla 2). Esto levanta la sospecha que los métodos que utilizan los componentes principales tienden a asignar indiscriminadamente la obra a Cervantes sin ser un resultado justo. La comprobación de esta sospecha es difícil de discutir y de desentrañar puesto que no es sencillo

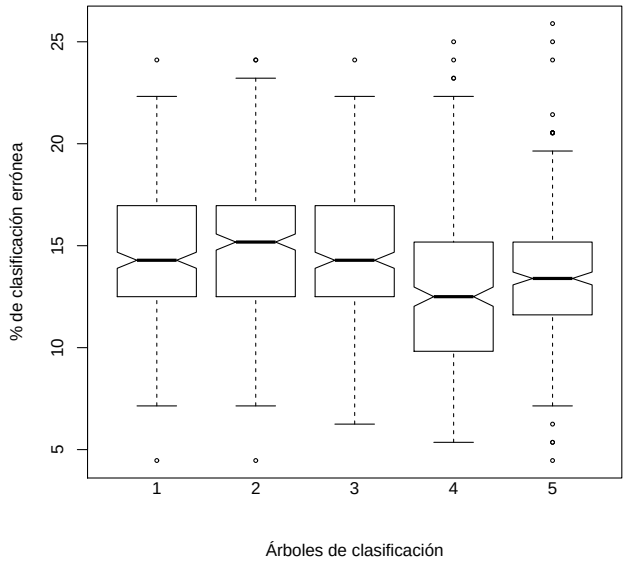

FIGURA 8: Árboles de clasificación: 1: Resultado con poda automática con un subconjunto de los componentes principales. 2: Resultado con poda automática con todos los componentes principales. 3: Resultado con poda por validación cruzada con un subconjunto de los componentes principales. 4: Resultado con poda por validación cruzada con los datos originales. 5: Resultado con poda automática con los datos originales.

saber qué componente pudiera estar causando este efecto (si el efecto, ciertamente, existe). Es probable también que la configuración de los componentes escogidos, conjuntamente, produzca una especie de enmascaramiento. Es notable, además, que las probabilidades que la obra sea una creación cervantina obtenidas con las variables originales, en general, son bastante bajas. La controversia pudiera aún continuar al revisar los resultados de los árboles de clasificación (ver tabla 3) donde la probabilidad más baja (0,82) se corresponde con el uso de un subconjunto de los componentes principales.

La columna de probabilidades de las tablas 2 y 3, cuyo signo es ¶, muestra los resultados de aplicar las extensiones naturales de los métodos expuestos en la sección 3 al caso de varias categorías (donde cada categoría es uno de los autores listados en la sección 2.4).

En el resultado de esta aplicación se confirma la curiosa sospecha de la semejanza entre Alonso Fernández de Avellaneda y Miguel de Cervantes, puesto que únicamente hacen aparición estos dos autores[1].

[1]Los resultados para el análisis discriminante cuadrático y el modelo de regresión multinomial aditivo producen errores de estimación y convergencia. Por estas razones esos resultados particulares no muestran.

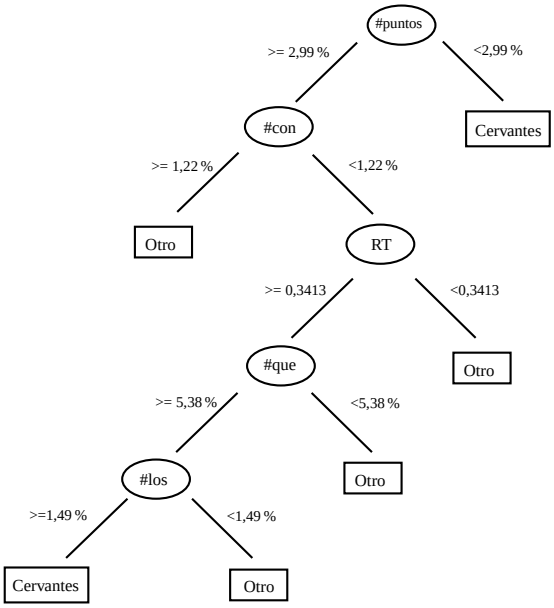

FIGURA 9: Árbol de clasificación: se muestra el árbol que utiliza las variables originales bajo el esquema de poda manual (se corresponde con el método 4 de la figura 8).

Otro dato curioso es el resultado obtenido con el método de las máquinas de soporte vectorial cuando se utilizan todas las variables originales: en el modelo dicotómico se asigna la obra al grupo donde no está Cervantes y al aumentar las clases asigna a la obra como cervantina.

5. Conclusiones

Los métodos de clasificación en general mostraron en el entrenamiento un muy buen desempeño estando la mayoría por debajo del 10 % de desaciertos y aún por debajo del 5 %. El método de Bayes ingenuo presentó una tasa de desacierto superior al 10 % (Yu (2008) obtuvo un resultado donde el método de Bayes ingenuo fue tan competitivo como el método de las máquinas de soporte vectorial).

En ciertos casos, trabajar con algunos componentes principales produjo menor error de clasificación (tal es el caso de las máquinas de soporte vectorial o el análisis discriminante cuadrático); en otros, trabajar con las variables originales produjo mejores resultados (como el método de los vecinos más cercanos o el análisis discriminante lineal) mientras que en otros, como en el análisis discriminante regularizado o el método de Bayes ingenuo, el trabajar con un conjunto de datos u otro no representó mayor diferencia. Esta situación es más palpable cuando se utiliza el método de árboles de clasificación (ver figura 8), donde los resultados fueron prácticamente los mismos para las diferentes estrategias utilizadas. Note

Tabla 2: Resultados de aplicar diferentes métodos de clasificación a la *Novela de la tía fingida*, históricamente atribuida a Cervantes.

Método	Prob.✠	Prob.¶	Método	Prob.✠	Prob.¶
Reg. logística†	0,78	$1{,}00^{C}$	An. Disc. cuadrático†	0,94	$1{,}00^{C}$
Reg. logística‡	0,00	$1{,}00^{A}$	An. Disc. cuadrático‡	0,00	-
Reg. Log. aditiva†	1,00	-	An. Disc. regularizado†	0,80	$0{,}99^{C}$
Reg. Log. aditiva‡	0,00	-	An. Disc. regularizado‡	0,00	$0{,}60^{A}$
An. Disc. lineal†	0,80	$0{,}99^{C}$	10 vecinos más cercanos†*	1,00	$1{,}00^{C}$
An. Disc. lineal‡	0,33	$0{,}60^{A}$	10 vecinos más cercanos◇*	1,00	$1{,}00^{C}$
An. Disc. de mezclas†	0,74	$0{,}99^{C}$	Bayes ingenuo†	0,96	$0{,}90^{C}$
An. Disc. de mezclas‡	0,30	$0{,}86^{A}$	Bayes ingenuo◇	1,00	$1{,}00^{C}$
An. Disc. flexible†	0,80	$0{,}99^{C}$	Máq. de soporte vectorial†	0,99	$0{,}99^{C}$
An. Disc. flexible‡	0,33	$0{,}61^{A}$	Máq. de soporte vectorial◇	0,12	$0{,}71^{C}$

✠: Probabilidades del modelo binario (Cervantes vs Otro).
¶: Probabilidades del modelo que considera cada autor como un grupo.
†: Considerando el subconjunto de los componentes principales establecido en la tabla 1.
‡: Considerando el subconjunto de las variables originales reseñado en la sección 4.1.
◇: Considerando todas las variables originales.
*: Este valor no es una probabilidad. Este método asigna como cervantina a esta novela.
^{A}: Significa que el modelo de múltiple respuesta asigna la obra a Alonso Fernández de Avellaneda.
^{C}: Significa que el modelo de múltiple respuesta asigna la obra a Miguel de Cervantes.

Tabla 3: Resultados de aplicar árboles de clasificación a la *Novela de la tía fingida*, históricamente atribuida a Cervantes.

Método	Prob.✠	Prob.¶
Árboles de clasificación†	0,97	0,92
Árboles de clasificación‡	0,97	0,92
Árboles de clasificación◇	0,82	0,62
Árboles de clasificación♣	0,97	0,97
Árboles de clasificación♠	0,97	0,94

✠: Probabilidades del modelo binario (Cervantes vs Otro). ¶: Probabilidades del modelo que considera cada autor como un grupo (todos asignaron la obra como Cervantina). †: Considerando poda automática con un subconjunto de los componentes principales. ‡: Considerando poda automática con todos los componentes principales. ◇: Considerando poda por validación cruzada con un subconjunto de los componentes principales. ♣: Considerando poda por validación cruzada con los datos originales. ♠: Considerando poda automática con los datos originales.

el lector que se está haciendo referencia a los resultados generales de clasificación (ver figura 7) y no al resultado particular motivo del estudio.

A este respecto resulta, a la luz de los resultados encontrados en este trabajo, extremadamente difícil dar una conclusión concreta. Por ejemplo, si se decidiera considerar solo aquellos clasificadores cuyo error promedio es menor al 5 % (10 en total) se encontraría que 5 de ellos asignan la obra a Cervantes y los restantes 5

a otro autor. Si se decidiera utilizar los clasificadores con un error promedio de menos del 10 %, se encontraría que 10 de ellos clasifican la obra como cervantina y 8 no. Si se consideran todos los métodos, el 68 % de ellos asigna la obra como cervantina.

Por lo pronto, parte de las ampliaciones y mejoras directas que tiene este trabajo comienzan por la extensión del corpus utilizado para autores diferentes a Cervantes, ampliar el panorama de autores considerados, utilizar otras técnicas de clasificación (por ejemplo, Tibshirani, Hastie, Narashimhan & Chu 2003), variar los métodos de calibración empleados (cambiar, por ejemplo, las probabilidades previas a cada grupo; cambiar el punto de corte para los modelos binarios, etc.).

Por último, merece un comentario la sorpresa de la aparente semejanza estilográfica entre Miguel de Cervantes y Alonso Fernández de Avellaneda. Es probable que 'descubrir' que el autor de *La novela de la tía fingida* es estadísticamente muy parecido a Cervantes no dé muchas luces al problema puesto que se ignora qué escritor, *encubriendo su nombre, fingiendo su patria*, estuvo detrás de este pseudónimo.

[Recibido: abril de 2010 — Aceptado: enero de 2011]

Referencias

Aylward, E. T. (1982), *Cervantes: Pioneer and Plagiarist*, Tamesis Books Limited, Londres, UK.

Baum, L. F. (2001), *The Royal Book of Oz*, Dover Publications, New York, States United. Escrito con 'colaboración' de R. Thompson.

Binongo, J. (2003), 'Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution', *Chance* **16**(2), 9–17.

Bird, S., Klein, E. & Loper, E. (2009), *Natural Language Processing with Python*, O'Really, Sebastopol, States United.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Pattern Recognition* **39**, 1–38.

Gardner, M. (1998), *Visitors from Oz: The Wild Adventures of Dorothy, the Scarecrow, and the Tin Woodman*, St Martins Press, New York, States United.

Grieve, J. (2007), 'Quantitative Authorship Attribution: An Evaluation of Techniques', *Literacy and Linguistic Computing* **22**(3), 251–270.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2 edn, Springer, New York, States United.

Hoover, D. L. (2002), 'Multivariate Analysis and Study of Style Variation', *Literacy and Linguistic Computing* **18**(4), 341–360.

Hosmer, D. & Lemeshow, S. (2000), *Applied Logistic Regression*, 2 edn, Wiley, New York, States United.

Jockers, M., Witten, D. & Criddle, C. (2008), 'Reassessing authorship of the Book of Mormon using delta and nearest shrunken centroid classification', *Literacy and Linguistic Computing* **23**(4), 465–491.

Johnson, R. & Wichern, D. (1998), *Applied Multivariate Statistical Analysis*, fourth edn, Prentice Hall, New York, States United.

Jolliffe, I. T. (2002), *Principal Component Analysis*, 2 edn, Springer, New York, States United.

Joula, P. (2006), 'Authorship Attribution', *Foundations and Trends in Information Retrieval* **1**(3), 233–334.

Koppel, M., Schler, J. & Argamon, S. (2009), 'Computational methods in authorship attribution', *Journal of the American Society for Information Science and Technology* **60**(1), 9–26.

Lebart, L., Morineau, A. & Warwick, K. (1984), *Multivariate Descriptive Statistical Analysis*, John Wiley & Sons, New York, States United.

Madrigal, J. L. (2003), 'De cómo y por qué La tía fingida es de Cervantes', *Artifara* (2).

Rencher, A. (2002), *Methods of Multivariate Analysis*, second edn, Wiley, New York, States United.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, UK.

Tibshirani, R., Hastie, T., Narashimhan, B. & Chu, G. (2003), 'Class prediction by nearest shrunken centroids with applications to DNA microarrays', *Statistical Science* **18**(1), 104–117.

Venables, W. N. & Ripley, B. D. (2002), *Modern Applied Statistics with S*, fourth edn, Springer, New York, States United.
*http://www.stats.ox.ac.uk/pub/MASS4

Witten, I. H. & Frank, E. (2005), *Data Mining: Practical Machine Learning Tools and Techniques*, 2 edn, Elsevier, San Francisco, States United.

Yu, B. (2008), 'An evaluation of text classification methods for literacy studies', *Literacy and Linguistic Computing* **23**(3), 327–343.

Revista Colombiana de Estadística
Junio 2011, volumen 34, no. 1, pp. 39 a 72

Determinantes socioeconómicos de la mortalidad infantil en Colombia, 1993

Socioeconomic Determinants of Infant Mortality in Colombia, 1993

B. Piedad Urdinola[a]

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

Resumen

Este artículo estima los determinantes socioeconómicos de la mortalidad infantil en Colombia, haciendo uso de los datos censales más recientes y disponibles al público en el país (1993). Para tal fin, se estiman las tasas de mortalidad infantil (TMI) de manera indirecta siguiendo el método Brass-Trussell, dadas las altas tasas de subregistro, que pueden alcanzar el 30 %. Estas estimaciones permiten una mejor medición de la TMI por diferentes características socioeconómicas, nunca antes medidas en el país, así como plantear un modelo estadístico que mide paramétricamente los principales determinantes socioeconómicos de la TMI en el país. Los resultados destacan la educación materna, la calidad de la vivienda, el acceso a servicios públicos y a servicios sanitarios como los principales determinantes de la TMI en Colombia.

***Palabras clave*:** análisis multivariado, demografía, modelos lineales, mortalidad infantil.

Abstract

This article considers the socioeconomic determinants of infant mortality in Colombia, by using the most recent and available census to the public data in Colombia (1993). For such aim, the Infant Mortality Rate (IMR) is calculated by indirect estimation techniques following the Brass-Trussell method, given the high rates of sub-registry, which can reach 30 %. These estimations allow better measurements of IMR by different socioeconomic characteristics, never measured before in Colombia, as well as to apply a statistical model that parametrically measures the main socioeconomic determinants of the IMR. The results prove maternal education, predominant housing materials, access to public and sanitary services to be the main determinants of IMR in Colombia.

***Key words*:** Demography, Infant mortality, Linear models, Multivariate analysis.

[a]Profesora asociada. E-mail: bpurdinolac@unal.edu.co

1. Introducción

La tasa de mortalidad infantil (TMI) es uno de los indicadores demográficos que mejor refleja el contexto socioeconómico de un país. Se define como la razón de defunciones a la edad de 0 a 1 año, frente a los nacimientos del mismo período. Además de ser un indicador efectivo en describir las condiciones de mortalidad, la TMI es muy eficiente en capturar diferentes problemas de bienestar social y de desarrollo socioeconómico de cualquier población, que se asocia a las mejoras en capital físico (por ejemplo infraestructura y hospitales) y humano (como la educación de los padres) que debe hacer una sociedad por mejorar sus condiciones de vida. Este trabajo busca estimar puntualmente los determinantes socioeconómicos de la mortalidad infantil en Colombia, aprovechando la riqueza de la información censal de 1993, con el fin de dar luces en la focalización de esfuerzos si la meta es reducir la TMI, tal como lo considera el cuarto objetivo de desarrollo del milenio de Naciones Unidas.

Este trabajo estima indirectamente la TMI en Colombia utilizando la información del último censo de población disponible, 1993, y aplicando métodos indirectos de estimación demográfica (Brass 1975), dado que las cifras de defunciones del país se encuentran subregistradas y en mayor proporción para los menores de un año (Somoza 1980, Pabón 1993, Flórez & Méndez 1997, Medina & Martínez 1999, PAHO 1999, Urdinola 2004). Además, los trabajos realizados hasta el momento dedicados a la medición y corrección de subregistro de la mortalidad infantil se limitan a generar estimaciones directas o indirectas por zonas, departamentos o ciertas características socioeconómicas de las madres, pero no evalúan todas estas características en conjunto bajo un modelo estadístico.

Luego de esta introducción, la segunda sección hace una revisión al estado del arte con énfasis en el contexto colombiano y latinoamericano. La tercera sección contiene la metodología de la estimación indirecta de la TMI a partir de la información censal de 1993 y el modelo estadístico escogido para hallar los determinantes socioeconómicos. Luego se exponen los resultados obtenidos de estas mediciones y por último las conclusiones y recomendaciones.

2. Antecedentes

A partir de los años 50, Colombia experimenta importantes cambios socioeconómicos que han influido positivamente en el descenso de indicadores demográficos como la fecundidad y la mortalidad general, e incluso la mortalidad infantil. En particular, el aumento en los niveles educativos, sobre todo de las mujeres, la gran migración rural-urbana, los procesos de urbanización y el aumento de la participación laboral femenina son los principales motores de dichos cambios (Bonilla & Rodríguez. 1992, Flórez 2000). Sin embargo, los canales de cómo estos cambios han afectado la TMI y el impacto preciso de cada uno de estos factores aún no se ha medido con precisión en el país, trabajo que busca satisfacer este artículo.

Esta tendencia a la baja en la TMI se puede observar en la figura 1. Sin embargo, Colombia aún se encuentra lejos de los niveles alcanzados por países

desarrollados. Alrededor de 2007, según las cifras oficiales de cada país, la TMI de países desarrollados varía entre 2.4 de Islandia, 2.8 Japón y 6.5 de Estados Unidos; mientras que Colombia alcanzó 19 por cada mil nacidos vivos. En particular, para el año de análisis, 1993, en Colombia la TMI oficial fue de 20 por cada mil niños nacidos, mientras que las de países desarrollados se sitúan alrededor de 5 por cada mil niños nacidos vivos.

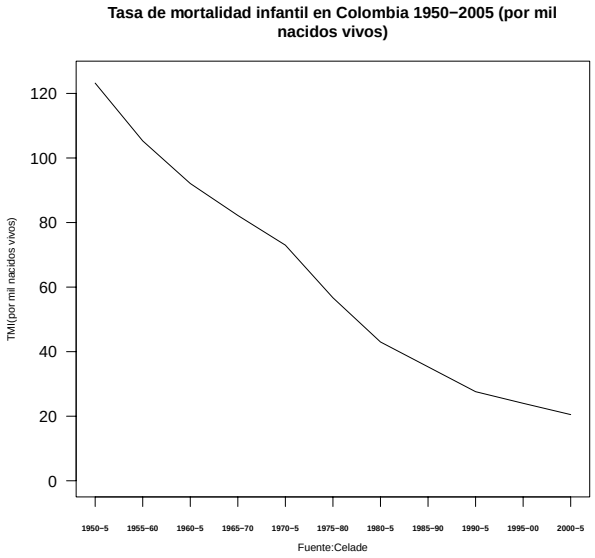

FIGURA 1: Tasa de mortalidad infantil en Colombia (por cada mil nacidos vivos), 1950-2005.

Dentro del contexto latinoamericano, Colombia se sitúa en los países con mortalidad media-baja (CELADE 1995) junto con Panamá, Argentina, Venezuela y Uruguay cuyo promedio es de 26 por mil nacidos vivos y siendo superados sólo por el grupo de países de baja mortalidad (Chile, Costa Rica y Cuba con TMI promedio de 14 por mil nacidos vivos). Dado que este artículo utilizará la información del último censo de población colombiano disponible al público, 1993, a partir de este punto las cifras harán referencia a dicho período con fines comparativos.

En el ámbito teórico, la mortalidad infantil se define como la probabilidad o riesgo de morir en el primer año de vida y se explica por factores endógenos al niño (deterioro biológico y genético) y a los exógenos a él y su familia, como condiciones sociales, económicas y ambientales. Desde el punto de vista analítico, el primer modelo que estudia los determinantes de la salud es el documento seminal de Grossman (1972). Su modelo establece una función de demanda por salud ("buena salud"), donde un individuo hereda un acervo inicial de salud, que se deprecia en el tiempo y que puede ser mejorado haciendo inversiones en salud, para producir finalmente una cantidad total de tiempo saludable que la persona dedica no sólo a sus inversiones en salud (p. ej. deporte), sino también a otras actividades lucrativas como trabajar. Esta idea, claramente, aplica mejor a adultos que a los niños y en

particular en el caso de la mortalidad infantil en el tiempo que invierten las madres al cuidado de sus hijos (Miller & Urdinola 2010).

Siguiendo el espíritu de este documento, Schultz (1984) propone un modelo analítico donde la supervivencia del niño depende de las dotaciones biológicas y de salud del niño y de los insumos en salud escogidos por la familia sujetos a los recursos de la misma. Estos insumos están determinados por la situación económica de la familia y las restricciones de la comunidad como la disponibilidad de servicios médicos, salarios y precios regionales y las condiciones ambientales. En general, sin tener en cuenta los factores biológicos, la mortalidad infantil es explicada por factores socioeconómicos como: las condiciones propias de la ocupación de aquellas personas que sean económicamente activas al interior del hogar y, sobre todo, del jefe del hogar, niveles de educación y en especial el de la madre, nutrición del niño y de la madre, niveles de fecundidad, condiciones y calidad de la vivienda, asistencia médica a la madre gestante y al niño después del nacimiento, niveles de ingreso del hogar, costumbres sociales, hábitos higiénicos y de preparación de alimentos, utilización adecuada de los programas y centros de asistencia de salud pública y privada.

Paralelamente, Mosley & Chen (1984) proponen su modelo de "determinantes próximos" a la mortalidad infantil, que hacen referencia a las variables que afectan directamente la determinación técnica de la salud del niño y a través de ellas todos los determinantes socioeconómicos deben operar. Por ejemplo, la educación no actúa directamente sobre la mortalidad infantil, pero si puede afectar una o más variables próximas o el nivel de ingreso de la familia. Así, a mayores ingresos hay mayor poder adquisitivo, mejor calidad y cantidad de las dietas consumidas por el niño y la madre gestante, mejor calidad de la vivienda, mayor capacidad de compra de bienes para higiene y vestuario y acceso a servicios de salud de buena calidad y mayor tecnología. Estos autores clasifican los determinantes socioeconómicos de acuerdo con las siguientes categorías de variables: variables a nivel individual, variables a nivel del hogar, variables a nivel comunitario y variables del sistema de salud.

Mosley & Chen (1984) definen cinco determinantes próximos: 1) factores maternales como edad de la madre, paridez, intervalo intergenésico, alimentación materna exclusiva, niveles de fecundidad y fertilidad; 2) factores ambientales que se relacionan con la generación de enfermedades infectocontagiosas, y se clasifican en aire, comida, agua, higiene personal y aseo del hogar; 3) deficiencias nutricionales, tanto de la madre como del niño; 4) accidentalidad y 5) control personal de enfermedades que incluye medidas preventivas y curativas en el niño.

Estas teorías se dan al tiempo que los primeros estudios empíricos de medición de "salud" en la infancia. Sobresale el trabajo de Rosenzweig & Schultz (1983) que estimó la función de producción de salud infantil para Estados Unidos entre 1967 y 1969. La salud infantil fue medida como peso al nacer o la duración del período gestacional. Mientras que las variables independientes incluyeron variables citadas como relevantes por la literatura médica[1], información local para los precios de

[1]Los autores incluyeron las siguientes variables: atención médica prenatal, número de meses que la madre trabajó durante el embarazo, número de meses antes que la madre realizara su

insumos y bienes, infraestructura en salud, gastos públicos y condiciones del mercado laboral[2]. El principal aporte de este trabajo fue la incorporación en el modelo de salud de los problemas de endogeneidad y autoselección de salud. Sin embargo, los autores reconocen que hay grandes debilidades por la ausencia de mediciones en diferentes variables que pueden generar confusión y poca consistencia en los parámetros, ya que varían de acuerdo con el modelo estadístico escogido.

2.1. Mediciones para el caso colombiano

En Colombia, los primeros estimativos de mortalidad infantil se realizaron hacia finales de los años 60 con registros vitales (defunciones y nacimientos), cuyas estimaciones resultaban deficientes, pues los datos no otorgaban confiabilidad ni eficiencia en la estimación (Zlotnik 1982). La Tabla 1 sintetiza las diferentes estimaciones indirectas llevadas a cabo en el país por los diferentes estudios, que a continuación se describen.

En Rosas & Rueda (1977) realizan la primera estimación indirecta utilizando la información del Censo de 1973. Estos autores aplican el método de Brass (ver siguiente sección) y obtienen el riesgo de morir en los 2 primeros años de vida para el período 1968-69, utilizando una muestra del 4 % de los hogares del censo. De esta estimación, obtienen una tasa de 88 por cada mil nacidos vivos para el total nacional, pero que los mismos autores consideran como una subestimación por deficiencias en los datos, en especial para la región Atlántica del país. Al generar las mismas estimaciones por zonas geográficas, nivel de educación de la mujer y ubicación rural-urbana, encuentran que tienen mayor probabilidad TMI quienes viven en zonas rurales y los hijos de las mujeres con menos escolaridad.

Somoza (1980) obtiene la TMI a partir de la Encuesta Mundial de Fertilidad (WFS-World Fertility Survey) de 1976. Este estudio estima dicha probabilidad por diferentes métodos y obtiene TMI decrecientes en el tiempo, con tasas mayores para los niños que para las niñas. Concluye que la muestra de la WFS es lo suficientemente buena como para obtener resultados eficientes y veraces.

Ochoa, Ordoñez & Richardson (1982) estiman la mortalidad infantil usando el módulo de Fecundidad de la Encuesta Nacional de Hogares de 1978 y 1980 y el Censo de 1973. Estiman la TMI indirectamente para los años de 1966, 1971, 1976 y 1981, que corresponden a tasas de 81, 74, 67 y 61 por cada mil nacidos vivos, respectivamente. Al estudiar la agrupación por departamentos y regiones, los resultados son coherentes con los de Rosas & Rueda (1977), así como los resultados sobre la educación materna. Este estudio concluye que las regiones con menor tasa de mortalidad infantil para 1981 fueron Bogotá, las regiones Oriental y Atlántica,

primera visita médica durante el embarazo, número de cigarrillos fumados por la madre durante el embarazo, paridez y edad de la madre al momento del parto.

[2]Las variables son: residencia metropolitana, raza, nivel educativo de los padres, ingreso del esposo, número de camas por hospital per capital, gastos públicos en salud per capital, número de hospitales y departamentos de salud con servicios de planificación familiar per capital, número de doctores médicos y ginecoobstetras per capital, tasa de desempleo para las mujeres entre 15-59 años, tasa total de desempleo, proporción de empleados en el sector público, gobierno e industria manufacturera, costo de los cigarrillos por paquete (incluyendo impuestos), impuesto a las ventas de los cigarrillos, precio por cuarto de leche, tamaño en población de las áreas metropolitanas.

con 45, 54 y 57 por mil nacidos vivos correspondientemente y la zona con mayor mortalidad fue la del Pacífico, por cada con 89 por mil nacidos.

Tabla 1: Resumen de las estimaciones indirectas de la tasa de mortalidad infantil. Total Colombia y según región, sexo y educación materna.

Autor	Fuente	Tasa de Mortalidad Infantil (por cada mil niños nacidos vivos)	Educación materna	Zona urbana	Sexo= hombres
Rosas & Rueda (1977)	Censo 1973	TMI*(67-68) = 88	−	−	
Ochoa, Ordóñez & Richardson (1982)	ENH 1978 ENH 1980 Censo 1973	TMI (66) = 81 TMI (71) = 74 TMI (76) = 67 TMI (81) = 61	−	−	
Bayona & Pabón (1982)	ENH 1978 ENH 1980 Censo 1973	TMI (50) = 135 TMI (82) = 57	−	−	
CCRP (1997)	Certificados de defunciones del DANE. Información censal de nacimientos y defunciones	TMI (79) = 48.26 TMI (76) = 30.94 TMI (81) = 24.42		−	+
Profamilia (1995) y Profamilia (2000)	ENDS-1986 ENDS-1990 ENDS-1995 ENDS-2000	TMI (72) = 62 TMI (76) = 47 TMI (77) = 48 TMI (82) = 37 TMI (87) = 34 TMI (92) = 28 TMI (90-95) = 27 TMI (95-00) = 21	−	−	+
Flórez & Mendez (1997)	Censo 1993	TMI (89-90) = 41.2 TMI (93) = 39.8			
Medina & Martínez (1999)	Censo 1993	TMI (85) = 44.24 TMI (90) = 32.56 TMI (94) = 34.15			

*La probabilidad de muerte se calcula entre cero y dos años de edad.
"−" representa una relación negativa en la medición de la TMI y la característica especial cuestión. Por ejemplo, a mayor nivel de educación de la madre, los autores encuentran una menor TMI y un "+" significa una relación positiva. ENH-Encuesta Nacional de Hogares.

Bayona & Pabón (1982) observaron una tendencia descendente en la TMI colombiana con grandes caídas al principio de período, que se desaceleraba hacia el final. Los resultados los presentan en rangos que varían desde una estimación máxima con valores de 146 en 1950 a 85 por cada mil en 1982, a una estimación mínima que oscila entre 135 en 1950 y 57 defunciones por cada mil nacidos vivos en 1982.

Un estudio con un enfoque diferente es el desarrollado por el Centro Corporación Regional de la Población-CCRP (1997), que evidencia las deficiencias del sistema de estadísticas vitales con el que cuenta el país. A partir de 1979, el DANE

comienza a hacer grandes esfuerzos en la recolección de las estadísticas de mortalidad apoyado por la UNICEF; sin embargo, seguía siendo la Registraduría la entidad encargada de mantener las cifras de nacimientos, lo que genera grandes inconsistencias en la medición directa de la mortalidad infantil, pues se corrige el numerador (defunciones), mas no el denominador (los nacimientos). Con fines ilustrativos, este estudio utiliza la información de los certificados de defunciones del DANE[3] y la información censal para corregir los nacimientos y estructura de edad, para medir directa e indirectamente la TMI. Obtienen así tasas para 1979, 1985 y 1990 de 48.26, 30.94 y 24.42 por cada mil nacidos vivos, respectivamente. Sin embargo, los resultados obtenidos para 1985 y 1990 parecen demasiado bajos frente a la tendencia y a los niveles observados hasta ese momento. Este trabajo diferencia las estimaciones de mortalidad infantil por región, departamento, zona y sexo y estima las principales enfermedades que causan la mortalidad infantil; desde este enfoque encontró los resultados teóricamente esperados. La TMI femenina siempre estuvo por debajo de la masculina en todos los departamentos y la tasa rural siempre fue mayor a la urbana. De igual manera, para los tres años estudiados, el departamento con menor TMI fue San Andrés, seguido de Bogotá, y los de mayores tasas fueron los antiguos territorios nacionales.

Profamilia ha utilizado la información de las Encuestas Nacionales de Demografía y Salud[4], que se llevan a cabo en el país y siguen la metodología propuesta por DHS-Macro internacional para la estimación indirecta de la TMI. Estas metodologías se han generado para poder tener una estimación comparable para los 75 países que han aplicado estas encuestas, que si bien permiten hacer comparaciones internacionales, pueden no ser las más convenientes para el país. La tabla 1 muestra una subestimación para los últimos años, cuando se comparan con los resultados de otras metodologías. Sin embargo, las estimaciones por género, zona y educación de la madre coinciden con los resultados de los demás estudios; es decir, mayores tasas para los infantes hombres, los hijos de madres menos educadas y los nacidos en zonas rurales.

Flórez & Méndez (1997) hacen una recopilación de las estimaciones hechas entre 1970 y 1992 de diferentes fuentes, y generan su propia estimación de la TMI para 1989-90 y 93 teniendo como fuente el censo de 1993. Los resultados muestran que la TMI llegó a 41.2 en 1989-90, mientras que en 1993 alcanza a 39.8 por cada mil niños nacidos vivos.

Finalmente, Medina & Martínez (1999) se concentran en evaluar la calidad de las cifras de defunciones infantiles (de 0 a 1 año) en el país, generando estimaciones indirectas usando el método Brass-Trussell. Además de la estimación de la TMI en 1985, 1990 y 1994 correspondientes a 44.24, 32.56 y 34.15, encuentran las mismas conclusiones de los estudios previos. La TMI es mayor para hombres que para mujeres, para los nacidos en zonas rurales y el departamento con mayor TMI es Chocó, mientras que las tasas más bajas se presentan en Bogotá y Valle del Cauca.

[3]Registros de defunciones en archivos especiales para los años: 79, 85, 89, 90, 91.

[4]Hasta el momento se han realizado cinco encuestas: 1986, 1990, 1995, 2000 y 2005. Se presentan los resultados hasta 2000, que incluyen las estimaciones para el período de referencia.

De estos dos últimos sobresalen tres resultados. Primero, la mejor fuente para la generación de estimaciones indirectas de la TMI es el censo de 1993, por cubrir el total de la población y por la consistencia que presenta en diferentes estimaciones. Segundo, paradójicamente, la calidad de los certificados de defunción hechos por el DANE parece decaer con el tiempo, hasta las fechas evaluadas por estos trabajos (1997). En particular por los cambios en la captura de las defunciones y nacimientos en el país (PAHO 1999, Urdinola 2004). Tercero, a pesar de las fallas intrínsecas de las cifras de mortalidad, la información de defunciones es de una calidad aceptable para las zonas urbanas y genera razones de sexo de credibilidad, es decir, dentro de los rangos demográficos observados en países con buenos registros vitales.

En general, esta revisión de literatura muestra un consenso entre los estudiosos de que siempre es preferible obtener medidas indirectas de la mortalidad que corrigen los graves problemas de subregistro que existen en el país. Asimismo, se puede esperar siempre encontrar mayores TMI para hombres que para mujeres, para los nacidos en zonas rurales y para los hijos de mujeres con menores niveles de educación; mayores tasas de mortalidad infantil en los llamados antiguos territorios nacionales y en el Chocó y menores en Bogotá. Finalmente, la fuente más confiable para generar estas estimaciones indirectas de mortalidad infantil resultan ser los censos nacionales, utilizando el total y no una muestra del mismo, que incluyen las preguntas necesarias para hacer esta estimación indirecta para el total de la población y dentro de los censos, el de 1993 presenta la mayor consistencia. Teniendo esto en cuenta, el presente trabajo generará las estimaciones indirectas a partir del censo de 1993.

2.2. Determinantes socioeconómicos de la mortalidad infantil en Colombia

Como ya se mencionó, no existe un estudio sistemático de los determinantes de la mortalidad infantil en Colombia. Las mejores aproximaciones fueron hechas por los estudios anteriormente descritos que, si bien no siguen un modelo estadístico para su medición, dan algunas luces.

Ochoa et al. (1982) muestran que la variable con mayores diferenciales en la mortalidad infantil es la educación de la madre y el grado de urbanización, seguido de los niveles de ingreso. De hecho, la TMI estimada varía entre 58.3 y 64.8 por cada mil entre los hogares de mayores y menores ingresos. Somoza (1980) muestra que la TMI se incrementa con la edad de la madre al momento de la encuesta, en parte porque existe la tendencia de las madres a omitir el número de hijos nacidos vivos, con un patrón de tasas relativamente alta para mujeres menores de 20 años, baja para aquellas entre 25 y 35 años y con tendencia a subir para mujeres mayores. Adicionalmente, que los hijos de mujeres con mayor educación dentro de una misma cohorte tienen mayor probabilidad de sobrevivir.

García (1986) estudia los diferenciales de la TMI por niveles de ingresos para el caso de Medellín durante la década del 70. Con la información de Censo de 1973

y la ENH de 1981 observó que el censo subestima la TMI[5] y que la calidad de la información para una población tan pequeña, comparada al total nacional, genera una menor calidad de la TMI. A pesar de esto, encuentra grandes diferenciales entre clases sociales, para ambos años, y que a mayor fecundidad de la madre mayor es el riesgo de muerte en los infantes. Sin embargo, este último hallazgo se hace endógeno, pues coincide con que son las mujeres de más bajo nivel social quienes tienen mayor número de hijos.

Flórez & Hogan (1990) estudian las zonas rurales del área cundiboyacense conectando las características demográficas y sociales y el estatus de la mujer con la mortalidad infantil en la zona. Los datos son tomados de un estudio longitudinal rural que cubre a los hogares cubiertos por el plan de Desarrollo Rural Integrado entre octubre y noviembre de 1986. Para la estimación dividen a la población de mujeres en 2 cohortes: 25-31 y 40-49 años. Este estudio calculó a través de un modelo logit de máxima verosimilitud con variable dependiente la TMI, y muestra un efecto negativo entre la TMI y el trabajo remunerado de las madres en trabajos agrícolas o del hogar y un efecto negativo pero muy pequeño de la lactancia al recién nacido y la educación secundaria. No se encontró alguna diferencia estadísticamente significativa en las probabilidades de supervivencia entre los hijos de mujeres con educación primaria y sin ninguna educación[6].

En síntesis, estos trabajos dan singular importancia al nivel de educación de la mujer y la zona, siendo menos estudiadas otras características sociales como los niveles de ingresos y la ocupación de los padres. Básicamente, falta ahondar en los determinantes socioeconómicos de la mortalidad infantil a nivel nacional, de acuerdo con las características de la familia y la vivienda.

2.3. Mediciones en el resto de Latinoamérica

Brass & Macrae (1985) estiman indirectamente la TMI siguiendo el método del "Hijo Previo" en Bolivia, Honduras, Argentina y República Dominicana. Esta práctica de medición indirecta consiste en preguntar a las mujeres que asisten a consulta durante el embarazo si su hijo anterior aún vive. Entonces, en una población con un intervalo intergenésico (diferencia de edad entre hijos consecutivos) medio cercano a los 30 meses, la división del número de madres con hijo previo fallecido por el número de madres con hijo previo vivo, proporcionaría una estimación de la probabilidad de morir entre el nacimiento y una edad x, que los autores recomiendan sea de 2 años. Sin embargo, también encontró que la población entrevistada tenía educación y edad menor al promedio de la población total, a excepción de un caso boliviano, que contaba con una sobrerepresentación de mujeres universitarias, quizás porque ellas son las que tienen mayor información y acceso a los servicios médicos hospitalarios. Esto llevaba a unos sesgos en la estimación indirecta, que implicaban tasas de mortalidad infantil muchísimo más altas que las de estadísticas vitales en los casos de sobrerepresentación de mujeres más

[5]Pues el censo no contabiliza las comunas 8 y 9 como urbanas y así se tuvo que hacer la estimación.

[6]Cabe anotar que a diferencia de los demás trabajos, el nivel de educación que aquí se toma es en el momento del nacimiento del niño y no en el momento en que se realiza la encuesta.

jóvenes y en el caso boliviano sucedió lo contrario. La TMI estimada era inferior a la reportada por las estadísticas vitales, por tener el sesgo de mujeres altamente educadas, quienes son las principales usuarias del sistema de hospitales.

Taucher (1988) investiga la relación entre fecundidad y mortalidad infantil para cinco países en largos períodos que culminan en los años 70: Costa Rica 1955-75, Chile 1972-78, México 1955-75, Paraguay 1958-78 y Perú 1956-76. Este estudio encuentra que el aumento en la mortalidad infantil tiene una relación ambigua en el número de hijos por mujer, mientras que la fecundidad tiene una relación negativa con la supervivencia infantil. Para todos los países se estimaron los diferenciales de mortalidad infantil por paridad (orden de nacimiento), edad y nivel de instrucción de la madre y en los países que los datos lo permitían se incluyó la longitud del intervalo intergenésico previo. Se observó, para todos los países, que la mortalidad infantil aumenta con la paridad, a excepción de México donde el primer hijo resultó más propenso a morir frente al resto, y que los hijos de madres muy jóvenes o de mayores edades tienen mayor probabilidad de morir antes del primer año. Cuando se incluyó el intervalo intergenésico, éste fue el principal determinante de la mortalidad con una relación negativa. En cuanto a la educación de la madre, se encontró que los países con mayor fecundidad son los mismos con menores niveles de educación femenina: Perú y México. A su vez, se comprobó que los diferenciales en educación son más notorios en países de baja mortalidad que en los de alta.

Naciones Unidas aplica en Costa Rica, Honduras y Paraguay el modelo de regresión planteado por Guzmán (1990), que mide la influencia de ciertos factores sociales sobre la mortalidad infantil. En cada país la selección de estos factores depende de la información en disponible[7]. En general, concluye que la mortalidad es menor en los hijos de padres en grupos ocupacionales de mayor rango y en ocupaciones manufactureras, para los hijos de asalariados, seguramente porque tienen completo acceso a la seguridad social. Asimismo, la mortalidad es menor para los residentes urbanos frente a los rurales y una vez más el determinante más importante es la educación de la madre. Sin embargo, a mayor participación materna en el mercado laboral menor es la salud del niño y la educación paterna no resultó significativa al controlar por otras variables en el análisis. Finalmente, que la vivienda no sea moderna, las condiciones sanitarias deficientes y la falta de electricidad aumentan la probabilidad de muerte en los infantes.

Por último, Castañeda (1994) usa datos agregados en una serie de tiempo entre 1975 y 1982 para Chile y encontró que el menor número de hijos, el aumento del consumo de leche en madres gestantes, el incremento en las consultas médicas y la mayor cobertura urbana de alcantarillado público reducen la TMI. Mientras que a mayor leche consumida por los menores, mayor es la mortalidad infantil, resultado que quizás se explica por sesgos de simultaneidad.

En resumen, la experiencia latinoamericana muestra que la TMI está negativamente relacionada con la educación materna e ingresos del padre y positivamente

[7]En Costa Rica se tomó el grupo socioocupacional del padre: agrario y no agrario; educación materna y paterna; lugar de residencia; en Honduras: las mismas que para Costa Rica y adicionó servicio de agua y servicio sanitario; y en Paraguay: fue igual a Honduras sin lugar de residencia y adicionó sistema de eliminación de basuras y calidad de la vivienda.

con los niveles de fecundidad e inexistencia de los servicios de salud y públicos. Hay que tener en cuenta que dentro de estos estudios de determinantes socioeconómicos no se encuentra ningún país con la estructura demográfica de mortalidad infantil semejante a la colombiana (grupo de países con TMI media-baja), por lo que este trabajo, en este sentido, constituye también una innovación para los países de la región.

3. Estimaciones indirectas de mortalidad y modelo estadístico

3.1. Mediciones indirecta de la TMI

La falta de estadísticas vitales confiables, sobre todo en los países en desarrollo, incitaron a demógrafos como William Brass a la creación de técnicas indirectas de medición de la mortalidad. La recolección de las Encuestas Mundiales de Fecundidad, desde los años 70, y la inclusión de las preguntas necesarias para su medición en censos de población han hecho posible estas mediciones en la mayoría de los países con falencias en sus registros vitales.

Este artículo aplica el método desarrollado por Brass (1964)[8], modificado posteriormente por Trusell (1975), especificando el modelo oeste de las tablas de Coale & Demeny (1966), dado que es el que mejor explica el comportamiento y la tendencia de los patrones de mortalidad en Colombia, describiendo el patrón más general de mortalidad.

El método de Brass mide la proporción de niños muertos clasificados por cohortes de edad de las mujeres con al menos un nacimiento vivo y éstas son ponderadas por el número de hijos nacidos tenidos vivos. De allí se estima D_i, que es la proporción de niños muertos con respecto a los vivos en grupos de edad de las mujeres en edad fértil ($i : 1 = 15 - 19, 2 = 20 - 24, \ldots$). Brass convirtió los valores de D_i en estimaciones de la probabilidad de morir entre el nacimiento y la edad x, en nuestro caso igual a 1, denotado por $q(x)$ y lo expresa de la forma específica:

$$q(x) = k_i D_i \tag{1}$$

donde k_i ajusta los factores de no-mortalidad, determinando el valor de D_i. Brass (1975) encontró que la proporción de niños muertos, D_i, y la medida de mortalidad en tablas de vida, $q(x)$, está principalmente influenciado por los patrones etarios de la fecundidad. Este determina la distribución de los niños en un grupo de mujeres de acuerdo con la exposición del riesgo de morir. Entonces, desarrolló un grupo de multiplicadores que convierte los valores observados de D_i en estimaciones de $q(x)$, con multiplicadores que son seleccionados según la razón del número promedio de niños nacidos reportados por las mujeres de los dos primeros grupos de edad, $P1/P2$,[9] que es un buen indicador de las condiciones de fecundidad en

[8]Explicado en: Naciones Unidas. Manual X: Indirect techniques for demographic estimation. 1983.

[9]P1 = mujeres de edades 15-19 y P2 = mujeres de edades 20-24.

los años más jóvenes. Finalmente Brass estimó k_i usando un polinomio de tercer grado de forma fija, pero con asignación de la edad variable. Y para representar la fecundidad obtuvo un sistema Logit generado por el estándar general de mortalidad y así obtuvo el elemento de mortalidad y una tasa de crecimiento del 2 % por año, con una distribución estable de las mujeres.

El supuesto esencial de este método es que el riesgo de muerte de un niño es una función que depende sólo de la edad del niño y no influyen otros factores como la edad materna o el orden de nacimiento del niño. Sin embargo, en la práctica se observa que los hijos de mujeres menores de 20 años y mayores de 35 años son más propensos a morir. Es por esto que para evitar sobre estimaciones en la TMI el rango de edad recomendado, y utilizado en este artículo, es el de mujeres de 20 a 34 años, así como para cada uno de los grupos quinquenales dentro de este rango.

Desde que Brass publicó su trabajo han surgido diferentes variantes a su método de estimación. La que se estimará en este artículo es la propuesta por Trusell (1975) que calcula los multiplicadores, k_i, diferentes al método original de Brass. En lugar de estimar un Logit, se obtienen de una regresión lineal por mínimos cuadrados ordinarios para articular la ecuación (1) a los datos generados de las tablas de vida de Coale & Demeny (1966) y al comportamiento de la fecundidad observada desarrollado por Coale & Demeny (1966) y Trusell (1975). El supuesto principal en esta estimación es que la mortalidad infantil y la fecundidad permanecen constantes en el pasado reciente. Este supuesto concuerda con la realidad demográfica colombiana y no afecta el análisis a realizar pues se está estimando la TMI para un año específico. Los estimativos de mortalidad indirecta se realizaron con el software especializado en fecundidad y mortalidad infantil "AFEMO 2" y el modelo de regresión se corrió en SAS.

3.2. Modelo estadístico

El modelo a utilizar expresa un indicador de mortalidad infantil como una función lineal del conjunto de variables socioeconómicas escogidas, representativas de la población a estudiar de acuerdo con la información obtenida del censo, que sigue la ecuación:

$$MI = C + \sum_{k=1}^{K} \sum_{j=1}^{J_{k-1}} b_{jk} X_{jk} + \varepsilon \tag{2}$$

donde:

MI: Indicador de mortalidad.
C: Constante de regresión.
b_{jk}: Coeficiente de la regresión de categoría j de la variable k.
X_{jk}: Variable independiente que representa la categoría j de la variable k.
J_k: Número total de categorías de la variable k.
K: Número total de variables.
ε: Término del error.

MI es una variable continua con media aproximada de 1, mientras que todas las variables independientes son categóricas y el modelo se estima por el método de

Mínimos Cuadrados Ordinarios (MCO). En este modelo, cada mujer está ponderada por el número de niños que haya tenido, por tanto es el niño y no la mujer la unidad del análisis. La variable dependiente representa la mortalidad de los hijos de cada mujer con relación al nivel nacional de mortalidad, estandarizado por la duración de exposición al riesgo de muerte, lo que es el principio de estimación del método de Brass, expuesto anteriormente, y que dentro del modelo explica el exceso relativo de riesgo de muerte en el primer año de vida de los hijos de una mujer particular con respecto a la probabilidad esperada para las madres de su misma edad en la población total. Su análisis es sencillo, por ejemplo, si la única variable explicativa fuera la ocupación materna, entonces sería el riesgo de muerte de los hijos de mujeres a cada subgrupo ocupacional con respecto al total nacional.

Se tomarán los grupos quinquenales de las mujeres: $i : 1 = 20 - 24, 2 = 25 - 29$ y $3 = 30 - 34$; pues existe evidencia que al tomar los grupos más jóvenes de la población se asegura que la estimación de la mortalidad pertenece a períodos muy cercanos al momento del censo y además minimiza los errores que frecuentemente se cometen por incluir las mujeres de mayor edad, que tienden a ocultar o distorsionar la información de fecundidad. De igual manera, lo más recomendable es excluir las mujeres menores de 20 años por la tendencia marcada de mayor TMI.

Para cada grupo i, el indicador de mortalidad se puede expresar como:

$$MI_i(a) = \frac{PD_i^o(a)}{PD^e(a)} \tag{3}$$

donde:

$PD_i^o(a)$: Es el número de hijos fallecidos por el total de hijos nacidos vivos.

$PD^e(a)$: Es la proporción esperada de hijos fallecidos para una mujer de edad a si tiene el riesgo de morir promedio del país.

Para todas las mujeres, MI tiene un promedio cercano a 1, si toma un valor por encima de la unidad, entonces la cantidad de niños que murieron fue mayor a la esperada y si es inferior a 1 ocurre lo contrario.

Entonces, la proporción esperada de niños muertos se obtiene aplicando el método inverso de Brass que estima la probabilidad de morir de las proporciones promedio:

$$PD^e(a) = \frac{q_s(x)}{k_i} \tag{4}$$

donde:

$q_s(x)$: Es la probabilidad estándar de morir desde que nace hasta la edad x, estimado con la ecuación (1).

k_i: Es el factor multiplicador que convierte el porcentaje de niños fallecidos en la probabilidad de muerte tomado del método de Brass (1975) con la modificación de Trusell (1975).

4. Resultados

El análisis se basa en la totalidad de los hogares registrados en el Censo Nacional de Población de Colombia en 1993[10]. El censo por definición cubre el 100 % de la población colombiana y la calidad del censo ha sido considerada como buena por diferentes especialistas por reducir errores y problemas en la medición (Brass 1996). Sin embargo, la experiencia ha demostrado la omisión de información sobre nacimientos y defunciones de niños, para las mujeres de mayor edad y una tendencia de mayor TMI en las madres adolescentes por razones biológicas, más que sociales, razón por la que se ha excluido estos grupos extremos de edad.

Las variables explicativas son dicotómicas, construidas para representar las diferentes categorías j de todas las k-variables. Cada variable k es expresada por un conjunto de $J-1$ variables dicotómicas que supone el valor de 1 si la mujer pertenece a la categoría y 0 si no. Las categorías con el menor nivel esperado de mortalidad infantil son las seleccionadas como categorías de referencia. Si la variable tiene un signo positivo, quiere decir que el riesgo de muerte en el niño se incrementa con respecto a la categoría de referencia y lo contrario sucederá si el signo es negativo.

Las variables incluidas para este artículo se pueden clasificar en 1) Variables de la vivienda: que incluyen la tenencia de sanitario, servicio de recolección de basuras, electricidad, acueducto, tipo de vivienda (igual a 1 si es casa o apartamento), material predominante de las paredes (con 4 variables dicotómicas: 1 si es tapia o bahareque, 2 si es madera burda o guadua caña, 3 si es zinc, tela, cartón o no tiene paredes, y la categoría de referencia que es si es bloque o ladrillo), material predominante de los pisos (igual a 1 si es de cemento y 0 para tierra o arena, madera burda u otro), distribución no hacinada (se define como número normal de cuartos el equivalente a la mitad del número de personas que habitan la vivienda más uno). 2) Variables individuales: nativo, nivel de educación materna (4 variables dicotómicas: 1 si no tiene ningún grado de educación o de preescolar, 2 si cursó primaria, 3 si alcanzó secundaria y la categoría de referencia que es si realizó estudios universitarios y de postgrado), nivel de educación paterna (se clasificó de manera idéntica que para la madre y además se incluyó una categoría adicional en la que se incorporan hogares sin información del padre), ocupación paterna (5 variables dummies: trabajador asalariado, trabajador independiente, no trabaja, sin compañero y finalmente la categoría de referencia trabajador familiar). 3) Variables geográficas: zona urbana, región geográfica (6 variables dicotómicas: 1 si es de la región Atlántica, 2 si pertenece a la región Oriental, 3 si vive en la región Central, 4 si es de la Pacífica, 5 si habita en Territorios Nacionales, y 6 la categoría de referencia si vive en Bogotá o en el Valle).

[10]Se utiliza la información del formulario de viviendas y hogares exclusivamente (formulario 1) por ser el que contiene las principales características socioeconómicas del hogar y la vivienda, así como las preguntas de fecundidad y mortalidad necesarias para el cálculo de la TMI indirecta.

4.1. Estimaciones de la tasa de mortalidad infantil y sus diferenciales

Los resultados fueron los teóricamente esperados y para todos los casos resultan coherentes con los estimativos recientes realizados por otros autores y además consistentes con la estructura de mortalidad infantil colombiana. La tasa nacional se estimó en 42.2 por cada mil niños nacidos vivos. El departamento con menor TMI es Atlántico, con 25.9, y el de mayor, Chocó con 79.1, con una tasa muy elevada frente a los demás departamentos del país si se considera que la tasa anterior es la de Caquetá de 60.3 por cada mil niños nacidos vivos. Además, la TMI para la mayoría de los departamentos (13) se concentra entre 40 y 50 por mil niños nacidos vivos.

La mayor TMI femenina a nivel nacional (ver Apéndice) fue la de Chocó: 71.1 y la menor la de Atlántico con 24.6. Mientras que las masculinas oscilaron entre 86.9 de Chocó y 25.9 de Atlántico. La TMI femenina siempre se encontró por debajo de la masculina, menos para San Andrés, donde fueron de 54.4 para las niñas y 46.5 por mil niños nacidos vivos, lo que puede deberse a la composición poblacional de alta inmigración en la isla o a fallas en los reportes de mortalidad de las madres.

Así mismo, a nivel rural las tasas femeninas estuvieron entre 72.3 de Chocó y 30.1 de Atlántico, y las masculinas entre 107.4 de Putumayo y 32.2 de Atlántico. A nivel urbano, las mayores tasas fueron las de Chocó: 69.5 en la femenina y 82.9 en la masculina; mientras que las menores fueron para Atlántico: 24.4 femenina y 26.5 masculina. Pero para las zonas urbanas las excepciones fueron San Andrés, y Casanare, con 41.4 para la TMI femenina, frente a 38 por mil niños nacidos vivos en la masculina y tasas casi iguales para los departamentos de Córdoba, Cundinamarca y Sucre, condiciones que se explican mejor por la proporción de niños nacidos vivos y sobrevivientes entre sexos, en lugar que por alguna condición social especial en estos departamentos.

Para el caso rural, se encontró que la mayor tasa fue la de Putumayo con 107.4, precedida por Chocó 81.1, y Caquetá 70.1. Los demás departamentos se encuentran entre el 61.8 de Nariño y el 32.2 de Atlántico; mientras que para el caso urbano, la mayor tasa se encuentra en 76.3 del Chocó muy lejos de los demás departamentos ubicados entre el 25.4 de Atlántico y 57.7 de Vichada. De todas formas, para todos los departamentos se encontró que la TMI rural es mayor; fue la urbana menos para Córdoba, donde la rural es de 40.8 y la urbana de 45.3 por cada mil niños nacidos vivos, y Quindío, donde la rural es de 38.2 y la urbana de 43.6 por cada mil niños nacidos vivos, y en Caldas casi se igualan las tasas en 39 por mil. Vale resaltar dos casos donde la TMI rural está muy por encima a la urbana, estos son Putumayo con una donde es más que el doble: 107.4 y 48 por cada mil niños nacidos vivos, respectivamente, y Caquetá, donde la rural es de 70.1, mientras que la urbana es de 46.7; poniendo en claro los problemas de la población rural colombiana; sobre todo para estos dos departamentos, pues es bien sabido que las áreas rurales cuentan con menor distribución y calidad de los servicios públicos básicos, así como los de salud, menores ingresos familiares y bajos niveles de educación, todos ellos factores determinantes en la supervivencia infantil. Asimismo, un departamento con graves dificultades, frente a los demás departamentos, es el Chocó. Pues no importa cuál

sea la medición de la TMI (rural, urbano, femenina, masculina o ambos sexos), siempre tiene una de las tasas más elevadas, lo que demuestra las bajas condiciones de este departamento en términos de desarrollo.

Ahora, al comparar las tasas femeninas urbanas y rurales se halló que la TMI rural siempre estuvo por debajo de la urbana menos en los casos de Arauca (36.1 frente a 48.3 por mil niños nacidos vivos, respectivamente); Caldas (37.8 y 33.9 correspondientemente) y Córdoba (38.9 y 45 por mil). Un caso excepcional vuelve a ser Putumayo, con una tasa rural de 198, mientras que la urbana sólo fue de 45. De igual forma, en el caso de las TMI masculinas la rural fue siempre inferior menos para Arauca, Córdoba, Quindío y San Andrés, y unas diferencias considerables (por encima de 14 puntos) para los departamentos de Caquetá, Casanare, Norte de Santander y, obviamente, Putumayo.

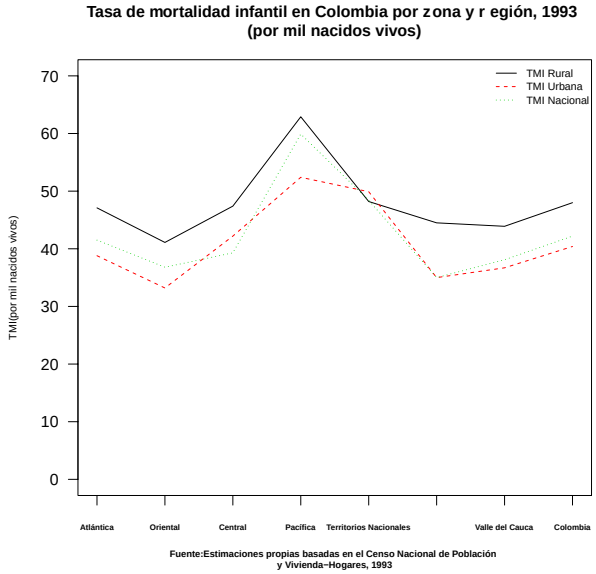

FIGURA 2: TMI por región geográfica y zona de residencia, 1993.

1=Región Atlántica = Atlántico, Bolivar, Cesar, Córdoba, Guajira, Magdalena, San Andrés y Sucre.
2=Región Oriental = Boyacá, Cundimarca, Norte Santander, Santander y Meta.
3=Región Central = Antioquia, Caldas, Caquetá, Huila, Quindío, Risaralda y Tolima.
4=Región Pacífica = Cauca, Chocó y Nariño.
5=Territorios Nacionales = Amazonas, Arauca, Casanare, Guaviare, Putumayo y Vichada.

Al agrupar los departamentos por regiones geográficas (figura 1), se encontró que la región con mayor TMI fue, como se esperaba la región Pacífica, 59.9 por cada mil y la de menor fue el Distrito Capital con 35 por mil niños nacidos vivos, seguida por la región Oriental y el Valle del Cauca: 36.8 y 38.1, respectivamente. Si miramos por zona, vemos que la mayor TMI rural y urbana fue la región Pacífica: 62.9 y 52.4, respectivamente, y que las menores fueron las de la región oriental: 41.1 en el área rural y 33.2 en la urbana. Esto nos lleva a concluir la zona del

pacífico, los antiguos territorios nacionales y, sobre todo, las áreas rurales son las que necesitan con premura, mejorar las condiciones socioeconómicas de sus habitantes y encontrar soluciones prontas al problema en estudio.

Al realizar estimaciones de mortalidad infantil de acuerdo con características socioeconómicas del hogar (figura 3), se encontró para todos los rangos de edad y para el total nacional que los niveles más altos se presentan en los hogares donde la mujer no tiene compañero y donde no tuvo ningún grado de educación.

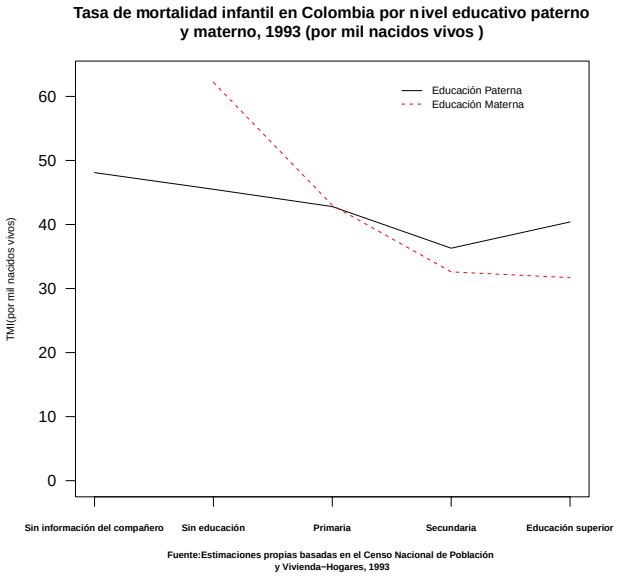

FIGURA 3: TMI por educación de la madre y del padre, 1993.

Las mediciones por otras características del hogar, que no han sido medidas en Colombia, muestran que para el total nacional todos los grupos de edad, las TMI más altas corresponden a las de hogares de mujeres sin compañero, sin ninguna ecuación y sin un adecuado servicio de sanitario (48.1, 62.2, y 48 por cada mil niños nacidos vivos, respectivamente). Además, siempre se encontró que a medida que se avanza en la educación materna, se reduce el riesgo de muerte del niño. En contraste, las mediciones de la TMI por educación paterna, muestra una menor TMI para padres con bachillerato, seguida por estudios superiores, primaria, sin educación y sin compañero. En este orden de ideas, la educación paterna puede ser una característica importante sobre la mortalidad infantil, pero que no se refleja directamente, sino a través de otros canales como la ocupación y, por consiguiente, los ingresos percibidos por el compañero o incluso la educación materna.

4.2. Resultados del modelo de regresión: determinantes socioeconómico de la mortalidad infantil en Colombia

Esta sección muestra los resultados del modelo de regresión[11]. Los resultados son muy similares para los tres grupos de edad y en el agregado, razón por la que sólo se presentan los resultados de mujeres entre 25 y 29 años (ver tablas en Apéndice para los demás grupos de edad). Como se muestra en la tabla 2, la variable más importante y significativa resultó ser, tal como se esperaba, la educación materna. Siempre el mayor coeficiente fue el de las mujeres sin ningún grado de educación con respecto a aquellas mujeres que alcanzan niveles universitarios o mayores 0.6915; seguidos de los coeficientes de educación primaria de 0.2489 y, por último, aquellas que alcanzan niveles de secundaria 0.3285. Esto quiere decir que a medida que se incrementa el grado de educación, el riesgo de morir disminuye, independientemente del rango de edad al que pertenezca la mujer; y el hecho de que en el agregado no sea significativo que la mujer alcance niveles de secundaria, implica que el mayor efecto de la educación se da entre las mujeres que no alcanzan ningún nivel de educación. Lo anterior, combinado con las estimaciones de TMI de acuerdo con la educación materna, confirma el hecho de que esta característica es la pieza clave dentro de los determinantes socioeconómicos de la mortalidad infantil.

En cuanto a la educación paterna, a pesar de que no pudo incluirse en la regresión final por generar problemas de multicolinealidad con la ocupación paterna, en los ejercicios de regresión previos al modelo definitivo se encontró que los coeficientes son siempre inferiores a los de educación materna, pero significativamente diferentes a cero. Es posible, entonces, que los efectos culturales e ideológicos estén bien representados en la educación materna, mientras que los de ingresos se relacionan con la educación paterna y los canales a través de los que se refleja no son directos.

Con respecto a la ocupación paterna, si el compañero es trabajador independiente, se incrementa riesgo de muerte relativo en 0.0496 y que no exista un compañero, lo hace en 0.0409. Aunque teóricamente esta variable se asocia a mayores ingresos en el hogar, este hecho no se puede asegurar con los datos a la mano. Bien puede ser que los trabajadores independientes y familiares sean predominantemente las personas que se dedican a labores propias de la economía informal, y por tanto, no perciban mayores ingresos o tengan acceso a los servicios de seguridad social.

Los resultados sobre la ocupación paterna pueden explicarse por dos posibles factores. Primero, las limitaciones existentes de información sobre ingresos como tal, éstos no están perfectamente representadas en grupos homogéneos y no se reflejan directamente a través de las categorías ocupacionales del padre. Segundo, manejar información de corte transversal no permite tener las características precisas del hogar en el momento en que el niño tenía un año de vida o menos, sino la información en el momento del censo.

[11] El método seguido para la estimación fue el de introducción o enter, en el que todas las variables del bloque se añaden como un grupo a la ecuación. Es decir, en un bloque subsiguiente se añadirán las variables para dicho bloque como un grupo al modelo final del bloque precedente.

Tabla 2: Coeficientes de regresión para las variables estudiadas, de acuerdo con la edad de la madre, 1993.

Variable	Mujeres entre 20 y 24 años	Mujeres entre 25 y 29 años	Mujeres entre 30 y 34 años	Total
1.VARIABLES DEL HOGAR				
Con servicio sanitario	−.107869*	−.139561*	−.168418*	−.141009*
Con Recolección de basuras	−.090218*	−.122449*	−.115794*	−.110780*
Con servicio de electricidad	−.082027*	−.069795*	−.097012*	−.082253*
Con servicio de acueducto	−.085137*	−.011809	−.028712*	−.035421*
Casa o apartamento	−.107037*	−.042265	−.067984*	−.060184*
Pisos en cemento	.031282*	.034186*	.062809*	.040897*
Material de las paredes (a)				
Tapia o bahareque	−.005738	.038918*	.028091*	.022107*
Madera burda o guadua cana	.194136*	.236716*	.247366*	.230479*
Zinc, tela, cartón o sin paredes	.156556*	.302824*	.160865*	.216974*
Sin hacinamiento	−.038793*	−.064079*	−.060508*	−.055262*
2. VARIABLES INDIVIDUALES				
No inmigrantes	−.033369*	−.053519*	−.045277*	−.045721*
Nivel de educación materna (b)				
Sin educación	.538216*	.691593*	.656799*	.655831*
Primaria	.119754*	248947*	.252178*	.228989*
Secundaria	−.105921*	.032855*	.043330*	.011051
Ocupación paterna (c)				
Sin compañero	.071606*	.040914*	.027068	.041638*
No trabaja	.028583	.013262	.031534*	.019684*
Trabajador asalariado	.037506	.022000	.036220*	.030264*
Trabajador independiente	.048268*	.049648*	.050130*	.048340*
3.VARIABLES GEOGRÁFICAS (d)				
Residencia urbana	−.012618	.003847	−.010175	−.003351
Región Atlántica	−.187216*	−.115596*	−.121862*	−.126347*
Región Oriental	−.184248*	−.085066*	−.125220*	−.108452*
Región Central	−.206224*	−.121150*	−.137467*	−.134073*
Región Pacífica	−.312569*	−.251208*	−.296461*	−.270705*
Antiguos Territorios Nacionales	−.247710*	−.155302*	−.195229*	-.181753*
4. ESTADÍSTICOS DE LA REGRESIÓN				
Número de casos	626.114	1.046.748	1.247.708	2.920.570
Constante	1.080080*	.840323*	.962125*	.911169*
R-cuadrado	.00420	.00582	.00786	.00578
R-cuadrado ajustado	.00415	.00580	.00784	.00577
F	90.06821	213.31006	313.34778	595.87781
Significancia de la prueba F	.0000	.0000	.0000	.0000

* Variable significativamente diferente de cero al 95 % de confianza.
(a) Categoría de referencia: Bloque o ladrillo
(b) Categoría de referencia: Educación superior y más
(c) Categoría de referencia: Trabajos familiares
(d) Categoría de referencia: Bogotá y Valle

El hecho de que la mujer no sea inmigrante, reduce el riesgo de muerte de los niños en un 0.0535; esto puede ser porque las condiciones en que viven estas mujeres no son las óptimas, quizás porque han sido desplazadas por la violencia o porque al cambiar su lugar de residencia, abandonan los posibles alternativas que

tenían en servicios de salud e higiene. En general, habría que profundizar sobre este grupo de la población y sus características, y su relación con el problema.

De otro lado, el lugar, las características de la vivienda y los servicios básicos ayudan a definir un poco mejor el estatus socioeconómico de la familia. De hecho, se encontró en todos los casos que contar con mejores servicios sanitarios, de electricidad, sistema de recolección de basuras, buenos materiales de construcción, en condiciones que no sean de hacinamiento y, caunto más amplia sea la vivienda, menor es el riesgo de muerte para el niño. Esto se debe no sólo al mejor nivel de vida económico del hogar, sino también a mejores condiciones de higiene (haciendo referencia sólo a los servicios públicos). En el modelo, dentro de los servicios públicos básicos y de higiene, los servicios sanitarios son los más importantes, ya que contar con este servicio disminuye el riesgo de muerte en 0.1395, seguido de la recolección de basuras (0.1224) y el de electricidad (0.0697).

Sobresalen los casos de materiales de las paredes hechas en zinc, tela, cartón o sin paredes, que incrementa el riesgo en 0.3028, y cuando el material es madera burda, guadua o caña, se aumenta el riesgo en 0.2367. Estos coeficientes están por encima de todos los servicios públicos y de higiene. Adicionalmente, disminuir las condiciones de hacinamiento reduce el riesgo de muerte en 0.064, así como vivir en casa o apartamento frente a vivir en cuarto u otro tipo de vivienda reduce el riesgo de muerte en 0.0422. Finalmente, en todos los casos es significativo menos en el segundo rango de edad contar con el servicio de acueducto, tal vez debeido a la estructura de datos[12].

Sin embargo, se encontró que el lugar de residencia (rural-urbana) no fue significativamente diferente de cero y para aislar mejor el efecto de este hecho, se corre el modelo para cada caso específico (ver siguiente subsección). De igual manera, los signos de las variables de región geográfica y de material de los pisos no fueron los esperados lo que puede ser aclarado por problemas de colinealidad entre variables.

En resumen, las variables que hacen referencia a la calidad de la vivienda se relacionan con ciertas enfermedades, en especial las infectocontagiosas, comprobándose que los mayores problemas a combatir son: vivir con deficiencias en los servicios públicos y sanitarios y en condiciones de pobreza reflejados en los malos materiales de construcción de la vivienda y condiciones de hacinamiento y los bajos niveles de ingresos del hogar interactuando con otras variables que afectan la mortalidad infantil determinan los grupos de mayor riesgo de muerte para los niños. No acceder a servicios básicos en la vivienda, sumados a los bajos niveles de educación paterna y sobretodo los bajos niveles de la educación materna influyen negativemente en la supervivencia infantil.

De hecho, contar con buenos materiales de construcción de las viviendas, reflejadas básicamente en los materiales del piso, o con servicios de letrinas, bajamar o simplemente no contar con servicios sanitarios higiénicos, es riesgoso, aún más que no contar con distribución de agua potable, lo que puede ser un punto de referencia a los programas de salud e higiene en el país.

[12]El 80 % de los hogares de este rango cuentan con el servicio de acueducto.

4.3. La mortalidad infantil por zona

La tabla 3 muestra gran similitud entre los determinantes de la TMI para ambas zonas. Sin embargo, en el caso rural, las variables que resultaron sin relevancia estadística fueron el material de los pisos y todas las categorías del material de las paredes, a excepción de madera burda que incrementa el riesgo de muerte en 0.203, con respecto a materiales como bloque o ladrillo. Del mismo modo, vivir en casa o apartamento en lugar de cuarto u otro es bastante significativo en el área rural, ya que reduce el riesgo de muerte en 0.1467, mientras que en zonas urbanas lo reduce en 0.035. Esto se explica porque no existen grandes diferencias entre los materiales de construcción de la vivienda para la mayoría de los hogares del área rural, así como la calidad de la vivienda. En las áreas urbanas, todas las categorías del material de las paredes son relevantes y sobresale que los materiales sean madera burda o guadua caña y zinc, tela, cartón o sin paredes, incrementando el riesgo de muerte en el niño en 0.0134 y 0.0347, respectivamente. Así mismo, que los pisos sean en cemento frente a otros materiales como tierra, arena, madera burda u otro, incrementan el riesgo en 0.05, siendo ésta la variable de menor influencia en el caso urbano.

De otro lado, la variable con mayor impacto en ambos casos fue la educación materna y dentro de ésta el mayor efecto se da para las mujeres que no alcanzan ningún nivel de educación, incrementando el riesgo de muerte del niño en 0.68 en el caso rural y en 0.61 en el caso urbano, frente a 0.21 de las mujeres que alcanzan primaria en ambos casos (teniendo como categoría de referencia la educación superior y más). De manera que tiene mayor impacto esta variable en las áreas rurales que en las urbanas.

Igualmente, en el caso rural, que el hombre se desempeñe en trabajos diferentes a los familiares incrementa el peligro de muerte, a excepción del caso en que no trabaja, con efectos de 0.074 si no tiene compañero, 0.066 si es trabajador asalariado y 0.048 si es trabajador independiente, mientras que en el caso urbano esta variable no tiene significancia estadística dentro del modelo. Entonces, la categoría ocupacional del padre tiene mayor efecto en las áreas rurales que en las urbanas, en donde la relación entre ingresos y la ocupación paterna puede estar distorsionada por cuestiones ya mencionadas de la estructura de información. Así mismo, el riesgo es mayor para los niños de áreas rurales cuyas madres no cuenten con un compañero, demostrando lo explicado sobre el papel de la ocupación masculina en relación al nivel de ingresos de los hogares rurales e igualmente que las concepciones culturales se encuentran más arraigadas en tales sectores del país que conciben al hombre como principal generador de ingresos del hogar.

De otro lado, contar con servicios de recolección de basuras, servicios sanitarios y electricidad en el área rural reduce el riesgo en 0.948, 0.1585 y 0.0921, respectivamente. En las áreas urbanas, los coeficientes homólogos reducen el riesgo de muerte en 0.0941, 0.1362 y 0.0873, respectivamente. El efecto de estas variables casi igual en ambos casos y confirmándose que un adecuado servicio sanitario es el servicio público más importante en la determinación de la mortalidad infantil, sin importar el lugar de residencia. Mientras que contar con servicio de acueducto sólo es relevante para el caso rural, reduciendo el riesgo de muerte en 0.05. El impacto

de esta variable es más notorio en las áreas rurales, lo que se puede explicar porque en sectores urbanos existen mayor facilidad y acceso para obtener agua potable para la cocción, así como un servicio de acueducto más eficiente que en el caso rural.

Tabla 3: Coeficientes de regresión para las variables estudiadas, de acuerdo con la residencia,1993.

Variable	Residencia rural	Residencia urbana
1.VARIABLES DEL HOGAR		
Con servicio sanitario	−.158595*	−.136258*
Con recolección de basuras	−.094830*	−.094141*
Con servicio de electricidad	−.092112*	−.087377*
Con servicio de acueducto	−.050505*	−.018592
Casa o apartamento	−.146756*	−.035634*
Pisos en cemento	−.020924	.050960*
Material de las paredes (a)		
Tapia o bahareque	.016601	.024735*
Madera burda o guadua cana	.203542*	.234574*
Zinc, tela , cartón o sin paredes	.078752	.265333*
Sin hacinamiento	.008221	−.074150*
2. VARIABLES INDIVIDUALES		
No inmigrantes	−.060502*	−.042105*
Nivel de educación materna (b)		
Sin educación	.680063*	.610019*
Primaria	.218633*	.216589*
Ocupación paterna (c)		
Sin compañero	.074497*	.003474
No trabaja	.018302	.019031
Trabajador asalariado	.066109*	−.013016
Trabajador independiente	.048555*	.017131
3.VARIABLES GEOGRÁFICAS (d)		
Región Atlántica	−.228575*	−.089305*
Región Oriental	−.062835*	−.128415*
Región Central	−.065741*	−.156515*
Región Pacífica	−.245112*	−.289469*
Antiguos Territorios Nacionales	−.305601*	−.132426*
4. ESTADÍSTICOS DE LA REGRESIÓN		
Número de casos	678.984	1.797.097
Constante	1.018269*	.906570*
R-cuadrado	.00507	.00403
R-cuadrado ajustado	.00504	.00402
F	156.31828	328.44988
Significancia para la prueba F	.0000	.0000

* Variable significativamente diferente de cero al 95 % de confianza.
a) categoría de referencia: bloque o ladrillo
b) categoría de referencia: educación secundaria y más
c) categoría de referencia: trabajos familiares
d) categoría de referencia: Bogotá y Valle

Finalmente, el hecho de que las mujeres no sean inmigrantes pesa más para el caso rural (0.06) que para el urbano (0.042), pues bien puede ser que al migrar del

área rural a la urbana se mejoran varias de las condiciones socioeconómicas que influyen en el proceso salud-enfermedad del niño.

5. Conclusiones y recomendaciones

Los niveles de mortalidad infantil en Colombia no son bajos frente a otros países desarrollados y algunos de la región latinoamericana. Aún se requieren esfuerzos en las áreas de salud, servicios médicos y mejorar las condiciones socioeconómicas de la población para reducir estos niveles.

Este estudio muestra novedosas mediciones indirectas de la mortalidad infantil por diferentes características socioeconómicas del hogar, no antes medidas en el país. Dentro de los determinantes socioeconómicos de la TMI se encontró a la educación materna como la principal variable. Esto coincide con los resultados de diferentes trabajos en el área, lo que no excluye la educación paterna como una variable significativa.

Identificar los grupos más vulnerables de la población frente a este problema ayuda a focalizar los esfuerzos que intenten solucionar este problema. Este artículo concluye que los hogares más vulnerables son los de áreas rurales, donde las mujeres no tengan compañero o cuya ocupación representen bajos ingresos, que vivan en malas condiciones físicas de la vivienda y sin educación. Las limitaciones en cuanto a la información no permiten esclarecer diferencias significativas entre las clases sociales en el país; sin embargo, otras variables relacionadas al poder adquisitivo de los hogares como los materiales de la vivienda, la amplitud de la misma, el hacinamiento, el acceso a los servicios públicos básicos y el acceso a servicio sanitario confirman que los hogares en peores condiciones sufren mayores riesgos de mortalidad infantil.

Los resultados del modelo demuestran que aumentar el acceso a servicios públicos como electricidad, adecuada recolección de basuras y agua potable son esfuerzos indispensables si se quiere reducir el riesgo de muerte en los infantes. Sobre todo un adecuado servicio sanitario resulta clave. Por tanto, es claro que si se quiere reducir los niveles de mortalidad infantil deben desarrollarse programas sanidad básica sumado a las actividades directas dentro del sector salud.

Otras variables resultaron determinantes como el hecho que la mujer sea inmigrante. Este resultado, sin embargo abre más interrogantes y, por ende, se propone un estudio particular para esta fracción de la población que esclarezca si es que estas personas mantienen menores condiciones socioeconómicas como escasez de recursos económicos, baja información sobre los servicios de salud y la falta de seguridad social frente al resto de la población, o si es efecto del fenómeno de desplazamiento interno, o si simplemente corresponde a un problema de autoselección, donde estas mujeres ya tenían características preexistentes que se mantienen aunque migren.

Asimismo, los resultados de la educación plantean una solución más compleja. Por un lado, la educación materna resultó ser el principal determinante de la mortalidad infantil, pero aumentar la cobertura de la misma se relaciona a condiciones

estructurales y de largo plazo. Teniendo esto en mente, lo ideal sería aumentar los niveles educativos en todas las instancias, pero dadas las limitaciones de recursos la focalización debería hacerse en los niveles de primaria y, sobre todo, en áreas rurales. Otra alternativa son las campañas educativas que toman mucho menos tiempo y recursos, y quizás tengan efectos similares. Estas campañas, sin embargo, deben dirigirse a los cuidados básicos de la madre durante el período de gestación y la alimentación y cuidados especiales de los niños en sus primeros meses de vida.

Estos esfuerzos pueden ir acompañados de la implementación de hospitales materno-infantiles en las áreas rurales y de mayor pobreza, dado que la inversión necesaria para la implementación de esta clase de hospitales es menor que la de aquellos que intentan cubrir toda clase de necesidades y puedes ser más efectivo sobre este problema.

[Recibido: enero de 2008 — Aceptado: enero de 2011]

Referencias

Bayona, A. & Pabón, A. (1982), La mortalidad en colombia 1970-1982, *in* M. de Salud, ed., 'Estudio Nacional de Salud', Bogotá, Colombia.

Bonilla, E. & Rodríguez. (1992), *Fuera del Cerco: Mujeres, Estructura y Cambio Social en Colombia*, ACDI - Agencia Canadiense de Desarrollo Internacional, Bogotá, Colombia.

Brass, W. (1964), Uses of Census or Survey Data for the Estimation of Vital Rates, Seminario Africano de Estadísticas Vitales, Addis Ababa, Etiopía.

Brass, W. (1975), *Methods for Estimating Fertility and Mortality from Limited and Defective Data*, Carolina Population Center, Laboratories for Population Statistics, North Carolina, United States.

Brass, W. (1996), 'Demographic data analysis in less developed countries: 1946-1996', *Population Studies* **50**(3), 451–467.
*http://dx.doi.org/10.1080/0032472031000149566

Brass, W. & Macrae, S. (1985), Childhood Mortality Estimated From Reports On Previous Births Given by Mothers at The Time of a Maternity I: Preceding Birth Technique, *in* W. Brass, ed., 'Advances in Methods for Estimating Fertility and Mortality From Limited and Defective Data', Centre For Population Studies, London, United Kingdom.

Castañeda, T. (1994), Contexto socioeconómico y causa del descenso de la mortalidad infantil en chile, Documento de trabajo 28, Centro de Estudios Públicos, Santiago de Chile, Chile.

CCRP (1997), Geografía de la mortalidad infantil. obtención y análisis de las tasas de mortalidad infantil, Informe técnico, Corporación Centro Regional de Población, Bogotá, Colombia.

CELADE (1995), *Mortalidad en la Niñez, una Base de Datos Actualizada en 1995. América Latina*, CELADE y UNICEF, Santiago de Chile.

Coale, A. & Demeny, P. (1966), *Regional Model Life Tables and Stable Populations*, Princenton University Press, New Jersey.

Flórez, C. E. (2000), *Las Transformaciones Sociodemográficas en Colombia durante el Siglo XX*, Banco de La República, Bogotá, Colombia.

Flórez, C. E. & Hogan, D. (1990), 'Women's status and infant mortality in rural colombia', *Biología Social* **37**, 188–203.

Flórez, C. E. & Méndez, R. (1997), *La Cobertura de las Defunciones en 1993*, Reporte Entregado al Ministerio de Salud, Bogotá, Colombia.

García, C. (1986), *Mortalidad Infantil y Clases Sociales: el Caso de Medellín en la Década del 70*, Universidad Pontificia Bolivariana, Medellín, Colombia.

Grossman, M. (1972), 'On the concept of health capital and the demand for health', *The Journal of Political Economy* **80**, 223–255.

Guzmán, J. M. (1990), Metodología, *in* N. Unidas, ed., 'CELADE. Factores sociales de riesgo de muerte en la infancia', Santiago de Chile, Chile.

Medina, M. & Martínez, C. (1999), *Geografía de la Mortalidad Infantil en Colombia, 1985-1994*, Departamento Administrativo Nacional de Estadística (DANE), Bogotá, Colombia. Estudios Censales #12.

Miller, G. & Urdinola, B. P. (2010), 'Cyclicality, mortality, and the value of time: The case of coffee price fluctuations and child survival in Colombia', *Journal of Political Economy* **118**(1), 113–155.

Mosley, H. & Chen, L. (1984), 'An analitical framework for the study of child survival in developing countries', *Population and Development* **10**. Supp 84.

Ochoa, L. H., Ordoñez, M. & Richardson, P. (1982), Resumen de resultados 1963-1983, *in* M. de Salud, ed., 'La mortalidad en Colombia 1970-1982', Bogotá, Colombia. Estudio Nacional de Salud. V6.

Pabón, A. (1993), *La Mortalidad en Colombia, 1953-1991*, Instituto Nacional de Salud, Bogotá, Colombia.

PAHO (1999), *La Salud en las Americas*, Vol. 2, 1998 edn, Pan American Health Organization (PAHO), Washington, States United.

Profamilia (1995), *Encuesta Nacional de Demografía y Salud. Resultados*, Profamilia, Bogotá, Colombia.

Profamilia (2000), *Encuesta Nacional de Demografía y Salud. Resultados*, Profamilia, Bogotá, Colombia.

Rosas, H. & Rueda, J. O. (1977), *La Mortalidad en los Primeros Años de Vida en Países de América Latina. Colombia 1968-1969*, Celade, San José, Colombia.

Rosenzweig, M. & Schultz, T. P. (1983), 'Estimation a household production function: Heterogeneity, the demand for health inputs, and their effects on birth weight', *Journal of Political Economy* **91**, 723–746. Issue 5.

Schultz, P. (1984), 'Studying the impact of household economic and community variables on child mortality', *Population and Development Review* (10), 215–235.

Somoza, J. (1980), Illustrative Analysis. Infant and Child Mortality in Colombia, Wfs Scientific Reports 10, United Kingdom.

Taucher, E. (1988), Efecto del Descenso de la Fecundidad en la Mortalidad Infantil, *in* 'Estudios sobre Mortalidad y Salud Infantil', number 57, Ottawa, Canada.

Trusell, J. (1975), 'A re-estimation of the multiplying factors for the brass technique for determining childhood survivorship rates', *Population Studies* **XXIX**(1), 97–108.

Urdinola, B. P. (2004), Could Political Violence Affect Infant Mortality? The Colombian Case, PhD thesis, University of California, Berkeley, States United.

Zlotnik, H. (1982), *Levels and Recent Trends in Fertility and Mortality in Colombia*, National Academy Press, Washington, D.C., States United.

Apéndice

Tabla 4: Tasa de mortalidad infantil (por cada mil nacidos vivos), por sexo y departamento. Colombia 1993.

Departamento	TMI Femenina	TMI Masculino	TMI Nacional
Amazonas	49.5	60.5	54.9
Antioquia	32.9	36.4	34.8
Arauca	43.9	52.1	48.1
Atlántico	24.6	27.0	25.9
Bogotá	33.3	36.4	35
Bolívar	43.6	48.5	46.2
Boyacá	35.8	41.1	38.5
Caldas	36.3	42.2	39.4
Caquetá	56.8	63.7	60.3
Casanare	44.6	45.7	45.4
Cauca	39.3	48.9	44.2
Cesar	49.3	55.4	52.4
Chocó	71.1	86.9	79.1
Córdoba	41.7	43.7	42.8
Cundinamarca	30.1	31.8	31.0
Guajira	44.9	57.0	51.0
Guaviare	47.0	54.9	42.5
Huila	41.2	46.1	43.8
Magdalena	43.0	47.3	45.3
Meta	42.6	48.1	45.5
Nariño	52.8	60.9	57.0
Norte de Santander	35.2	40.4	37.9
Putumayo	43.1	55.7	49.4
Quindío	39.6	45.3	42.5
Risaralda	40.6	44.9	42.8
San Andrés	54.4	46.5	50.6
Santander	34.4	38.2	36.5
Sucre	35.0	35.9	35.7
Tolima	42.0	43.6	43.0
Valle	36.5	39.4	38.1
Vichada	44.2	51.4	57.7
Total Colombia	39.8	44.3	42.2

Tabla 5: Tasa de mortalidad infantil (por cada mil nacidos vivos), por zona y departamento. Colombia 1993.

Departamento	TMI Rural	TMI Urbana	TMI Nacional
Amazonas	-	54.9	54.9
Antioquia	44.3	30.0	34.8
Arauca	43.4	50.9	48.1
Atlántico	32.2	25.4	25.9
Bogotá	44.5	35.0	35.0
Bolívar	53.1	42.1	46.2
Boyacá	43.1	31.1	38.5
Caldas	39.1	39.6	39.4
Caquetá	70.1	46.7	60.3
Casanare	50.4	39.9	45.4
Cauca	55.1	44.2	44.2
Cesar	54.2	51.3	52.4
Chocó	81.1	76.3	79.1
Córdoba	40.8	45.3	42.8
Cundinamarca	33.0	28.8	31.0
Guajira	52.5	50.7	51
Guaviare	-	42.5	42.5
Huila	51.6	43.7	43.8
Magdalena	47.5	44.1	45.3
Meta	48.4	43.5	45.5
Nariño	61.8	49.6	57.0
Norte de Santander	45.8	34.0	37.9
Putumayo	107.4	48.0	49.4
Quindío	38.2	43.6	42.5
Risaralda	48.5	40.6	42.8
San Andrés	58.2	47.8	50.6
Santander	43.0	32.3	36.5
Sucre	41.9	32.2	35.7
Tolima	46.6	40.1	43.0
Valle	43.9	36.7	38.1
Vichada	-	57.7	57.7
Total Colombia	48.0	40.4	42.2

Tabla 6: Tasa de mortalidad infantil (por cada mil nacidos vivos), por sexo y departamento para zonas urbanas. Colombia 1993.

Departamento	TMI Femenina	TMI Masculina	TMI Nacional urbana
Amazonas	50.0	60.0	54.9
Antioquia	29.1	30.8	30.0
Arauca	48.3	53.4	50.9
Atlántico	24.2	26.5	25.4
Bogotá	33.3	36.4	35.0
Bolívar	40.1	44.0	42.1
Boyacá	29.4	32.7	31.1
Caldas	37.8	41.3	39.6
Caquetá	44.2	49.0	46.7
Casanare	41.4	38.0	39.9
Cauca	39.3	48.9	44.2
Cesar	48.8	53.6	51.3
Chocó	69.5	82.9	76.3
Córdoba	45.0	45.2	45.3
Cundinamarca	28.8	28.6	28.8
Guajira	44.7	56.5	50.7
Guaviare	35.3	49.5	42.5
Huila	41.2	46.1	43.7
Magdalena	42.5	45.3	44.1
Meta	41.8	45.0	43.5
Nariño	46.4	52.7	49.6
Norte de Santander	32.4	35.3	34.0
Putumayo	41.5	54.4	48.0
Quindío	40.7	46.4	43.6
Risaralda	38.9	42.1	40.6
San Andrés	48.5	46.7	47.8
Santander	31.8	32.6	32.3
Sucre	32.2	32.1	32.2
Tolima	39.3	40.8	40.1
Valle	35.7	37.4	36.7
Vichada	56.0	59.2	57.7
Total Colombia	38.6	41.9	40.4

Tabla 7: Tasa de mortalidad infantil (por cada mil nacidos vivos), por sexo y departamento para zonas rurales. Colombia 1993.

Departamento	TMI Femenina	TMI Masculina	TMI Nacional rural
Antioquia	40.7	47.7	44.3
Arauca	36.1	50.2	43.4
Atlántico	30.1	34.2	32.2
Bogotá	41.4	47.5	44.5
Bolivar	49.8	56.1	53.1
Boyacá	39.8	46.1	43.1
Caldas	33.9	43.9	39.1
Caquetá	66.0	74.0	70.1
Casanare	47.9	52.7	50.4
Cauca	49.5	60.5	55.1
Cesar	50.1	58.1	54.2
Chocó	72.3	89.5	81.1
Córdoba	38.9	42.6	40.8
Cundinamarca	31.3	34.6	33.0
Guajira	45.7	59.1	52.5
Huila	47.8	55.3	51.6
Magdalena	43.9	50.7	47.5
Meta	43.8	52.6	48.4
Nariño	57.1	66.2	61.8
Norte de Santander	41.1	50.2	45.8
Putumayo	98.0	118.9	107.4
Quindío	35.2	41.0	38.2
Risaralda	45.1	51.6	48.5
San Andrés	69.2	45.9	58.2
Santander	38.6	47.1	43.0
Sucre	40.4	43.2	41.9
Tolima	45.6	47.3	46.6
Valle	40.2	47.4	43.9
Total Colombia	44.2	51.4	48.0

Tabla 8: Tasa de mortalidad infantil (por cada mil nacidos vivos), por características socioeconómicas. Colombia 1993.

Educación paterna	Femenina
Sin información del compañero	48.1
Sin educación	45.5
Primaria	42.8
Secundaria	36.3
Educación superior	40.4
Educación materna	
Sin educación	62.2
Primaria	43.0
Secundaria	32.6
Educación superior	31.7
Servicio Sanitario	
Sin Servicio	48.0
Con Servicio	34.1

Tabla 9: Tasa de mortalidad infantil (por cada mil nacidos vivos), por región Ggeográfica. Colombia 1993.

Región	TMI Rural	TMI Urbana	TMI Femenina	TMI Masculina	TMI Nacional
Atlántica	47.1	38.8	39.4	43.3	41.5
Oriental	41.1	33.2	34.6	38.7	36.8
Central	47.4	42.2	37.2	41.2	39.3
Pacífica	62.9	52.4	54.6	64.8	59.9
Territorios Nacionales	48.2	49.9	44.5	52.0	48.4
Distrito Capital	44.5	35.0	33.3	36.4	35.0
Valle del Cauca	43.9	36.7	36.5	39.4	38.1
Total Colombia	48.0	40.4	39.8	44.3	42.2

Atlántica= Atlántico, Bolívar, Cesar, Córdoba, Guajira, Magdalena, San Andrés y Sucre

Oriental= Boyacá, Cundinamarca, N. de Santander, Santander y Meta

Central= Antioquia, Caldas, Caquetá, Huila, Quindío, Risaralda y Tolima

Pacífica= Cauca, Chocó y Nariño

Territorios Nacionales= Amazonas, Arauca, Casanare, Guaviare, Putumayo y Vichada

Tabla 10: Determinantes socioeconómicos de la mortalidad infantil. Regresión lineal simple. Variable dependiente: MI (Indicador de mortalidad infantil).

	Total			
	Mujeres de 20-24 años	Mujeres de 25-29 años	Mujeres de 30-34 años	Mujeres de 20-34 años
1. Variables del hogar				
Con servicio sanitario	−0.107869**	−0.139561**	−0.168418**	−0.141009***
	(0.017241)	(0. 012493)	(0. 011314)	(0.007463)
Con recolección de basuras	−0.090218**	−0.122449**	−0.115794**	−0.11078
	(0.018427)	(0. 012794)	(0. 011372)	(0.007656)
Con electricidad	−0.082027**	−0.069795**	−0.097012**	−0.082253***
	(0.019236)	(0. 014371)	(0. 013398)	(0.0086)
Con acueducto	−0.085137**	−0.011809**	−0.028712**	−0.035421***
	(0.018418)	(0. 013591)	(0. 012375)	(0.008117)
Casa o apartamento	−0.107037**	−0.042265**	−0.067984**	−0.060184***
	(0.022898)	(0.016936)	(0. 01628)	(0.01024)
Pisos en cemento	0.031282*	0.034186***	0. 062809***	0.040897***
	(0.012474)	(0. 008463)	(0. 007424)	(0.005069)
Material de las paredes				
Tapia o bahareque	−0.005738	0.038918**	0. 028091**	0.022107***
	(0.01757)	(0. 012345)	(0. 011074)	(0.007378)
Madera burda o guadua caña	0.194136**	0.236716**	0. 247366**	0.230479***
	(0.02241)	(0. 016384)	(0. 015054)	(0.009744)
Zinc, tela, cartón o sin paredes	0.156556**	0.302824**	0. 160865**	0.216974***
	(0.066772)	(0. 049925)	(0. 047059)	(0.030185)
Sin hacinamiento	−0.038793**	−0.064079**	−0.060508**	−0.055262***
	(0.013004)	(0. 008834)	(0. 007787)	(0.005289)
2. Variables individuales				
Nativo	−0.033369***	−0.053519***	−0.045277***	−0.045721***
	(0.01208)	(0. 008082)	(0. 006978)	(0.004827)
Educación materna				
Sin educación	0.538216***	0.691593***	0. 650799***	0.655831***
	(0.037902)	(0. 023903)	(0. 019343)	(0.013987)
Primaria	0.119754***	0.248947***	0. 252178**	0.228989***
	(0.028248)	(0. 015568)	(0. 012214)*	(0.00922)
Secundaria	−0.105921**	0.032855**	0. 043330**	0.011051
	(0.027307)	(0. 014466)	(0. 011192)	(0.008612)
Ocupación paterna				
Sin compañero	0.071606***	0. 040914***	0.027068***	0.041638***
	(0.022165)	(0. 015968)	(0.014078)	(0.009513)
No trabaja	0.028583***	0. 013262***	0.031534***	0.019684**
	(0.01909)	(0. 015483)	(0. 013282)	(0.009238)
Trabajador asalariado	0.037506***	0.02200***	0.03622***	0.030264***
	(0.020884)	(0. 014999)	(0.013388)	(0.008993)
Trabajador independiente	0.048268***	0. 049648***	0.05013***	0.04834***
	(0.021016)	(0. 015194)	(0.013523)	(0.009087)
3. Variables geográficas				
Residencia urbana	−0.012618	0. 003847	−0.010175	−0.003351
	(0.019344)	(0. 013757)	(0. 012427)	(0.00822)
Región Atlántica	−0.187216***	−0.115596***	−0.121862***	−0.126347***
	(0.022352)	(0. 013079)	(0. 011128)	(0.007805)
Región Oriental	−0.184248***	−0.085066***	−0.12522***	−0.108452***
	(0.022594)	−0.12522***	−0.108452***	

Tabla 10: Determinantes Socioeconómicos de la mortalidad infantil. Regresión lineal simple. Variable dependiente: MI (Indicador de mortalidad infantil). (continuación)

	Total			
	Mujeres de 20-24 años	Mujeres de 25-29 años	Mujeres de 30-34 años	Mujeres de 20-34 años
Región Central	−0.206224***	−0.12115**	−0.13746***	−0.134073***
	(0.020837)	(0. 011522)	(0. 00969)	(0.006869)
Región Pacífica	−0.312569***	−0.251208***	−2.96461***	−0.270705***
	(0.026852)	(0. 017883)	(0. 015680)	(0.010652)
Territorios Nacionales	−0.24771***	−0.155302***	−0.195229***	−0.181753***
	(0.056529)	(0. 034858)	(0. 031585)	(0.020745)
Estadísticos de la regresión				
Número de casos	626114	1046748	1247708	
Constante	1.08008**	0. 840323***	0. 962125***	2.920570***
R-cuadrado	0.0042	0. 00582	0. 00786	0.911169
R-cuadrado ajustado	0.00415	0. 0058	0. 00784	0.00578
F	90.06821	213.31006	313.34778	0.00577
Significancia prueba F	0	0	0	595.87781

*** Variables significativas al 99 %

Tabla 11: Determinantes socioeconómicos de la mortalidad infantil. Regresión lineal simple por zona. Variable dependiente: MI (Indicador de mortalidad infantil).

	Zona Rural	Zona Urbana
1. Variables del hogar		
Con Servicio sanitario	−0.158595***	−0.136258***
	(0.012402)	(0.00964)
Con Recolección de basuras	−0.09483***	−0.094141***
	(0.023588)	(0.008405)
Con electricidad	−0.092112***	−0.087377***
	(0.012098)	(0.013681)
Con acueducto	−0.050505***	−0,018592*
	(0.011274)	(0,01261)
Casa o apartamento	−0.146756***	−0.035634***
	(-0.026097)	(0.011051)
Pisos en cemento	−0.020924***	0.05096***
	(0.010854)	(0.005775)
Material de las paredes		
Tapia o bahareque	0,016601**	0.024735***
	(0,012104)	(0.009886)
Madera burda o guadua caña	0.203542***	0.234574***
	(0.015736)	(0,013429)
Zinc, tela, cartón o sin paredes	0,078752*	0.265333***
	(0,060906)	(0.03473)
Sin hacinamiento	0,008221	−0.07415***
	(0,011095)	(0.005991)
2. Variables individuales		
Nativo	−0.060502***	−0.042105***
	(0.010304)	(0,005452)
Educación materna		
Sin educación	0.680063***	0.610019***
	(0.018555)	(0.017259)

Tabla 11: Determinantes socioeconómicos de la mortalidad infantil. Regresión lineal simple por zona. Variable dependiente: MI (Indicador de mortalidad infantil). (continuación)

	Zona Rural	Zona Urbana
Primaria	0.218633***	0.216589***
	(0.012245)	(0.006124)
Ocupación paterna		
Sin compañero	0.074497***	0,003474
	(0.019146)	(0,011422)
No trabaja	0,018302	0.019031***
	(0,027102)	(0.009664)
Trabajador asalariado	0.066109***	−0,013016
	(0.016212)	(0,011094)
Trabajador independiente	0.048555***	0,017131*
	(0.015652)	(0,011427)
3. Variables geográficas		
Región Atlántica	−0.228575***	−0.089305***
	(0.02335)	(0.008421)
Región Oriental	−0.062835***	−0.128415***
	(0.021914)	(0.00852)
Región Central	−0.065741***	−0.156515***
	(0.021283)	(0.007339)
Región Pacífica	−0.245112***	−0.289469***
	(0.024022)	(0.013901)
Territorios Nacionales	−0.305601***	−0.132426***
	(0.015733)	(0.023497)
Estadísticos de la regresión		
Número de casos	678.984	1.797.097
Constante	1018269***	0.90657***
	(0.039225)	(0.022445)
R-cuadrado	0,00507	0,00403
R-cuadrado ajustado	0,00504	0,00402
F	156,31828	328,44988
Significancia prueba F	0	0

*** Variables significativas al 99 %

Revista Colombiana de Estadística
Junio 2011, volumen 34, no. 1, pp. 73 a 94

Testing Linearity against a Univariate TAR Specification in Time Series with Missing Data

Sobre una prueba de linealidad en presencia de datos faltantes contra la alternativa de no linealidad especificada por un modelo TAR

Fabio H. Nieto[1,a], Milena Hoyos[2,b]

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

[2]Facultad de Economía, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

Nowadays, procedures for testing the null hypothesis of linearity of a (univariate or multivariate) stochastic process are well known, almost all of them based on the assumption that their paths (i.e. observed time series) are complete. This paper describes an approach for testing this null hypothesis in the presence of missing data, using an extension of one of the test statistics used in the literature. The alternative hypothesis is that the univariate stochastic process of interest follows a threshold autoregressive (TAR) model. It is found that if the missing-data percentage is low, the null distribution of the proposed test statistic is maintained; while if it is high, it is not. A threshold value for the missing-data percentage is detected, which can be utilized in practice.

***Key words*:** Linearity test, Missing data, Nonlinear time series, Threshold autoregressive model.

Resumen

Las pruebas estadísticas que se conocen actualmente para examinar la hipótesis nula de linealidad de un proceso estocástico (univariado o multivariado) están basadas, casi todas, en el supuesto de que las series temporales observadas son completas. En este trabajo, se presenta un nuevo procedimiento para examinar esta hipótesis nula, en presencia de datos faltantes, el cual es una extensión de un método muy citado en la literatura. La hipótesis alternativa especifica que el proceso estocástico de interés obedece a un modelo autoregresivo de umbrales (TAR). Se encuentra que si el porcentaje de observaciones faltantes es bajo, la distribución nula de la estadística de prueba se mantiene; en otro caso no. El estudio arroja un valor umbral para este porcentaje, el cual puede ser usado en la práctica.

***Palabras clave*:** datos faltantes, modelos autoregresivos de umbrales, prueba de linealidad, series de tiempo no linales.

[a]Profesor titular. E-mail: fhnietos@unal.edu.co
[b]Profesora auxiliar. E-mail: nmhoyosg@unal.edu.co

1. Introduction

Nieto (2005) developed a procedure for modeling a univariate threshold-autoregressive processes (TAR) in the presence of missing data. The approach was based on the assumption that one knows a priori that the dynamic relationship between the two stochastic processes is nonlinear. This model can be seen as a particular case of Tsay's (1998) multivariate threshold model, where a test for the null hypothesis of linearity was considered. An important contribution of Nieto's (2005) paper is the development of a smoother for estimating the missing data in the two time series involved.

There are several methods for testing the null hypothesis of linearity in a univariate or multivariate stochastic process. However, almost all of these methods have been developed on the basis that the time series are complete or equally spaced. Sometimes, this is not the case and one is faced with the problem of performing those tests in the presence of partial or missing observations. Tong & Yeung (1991*a*, 1991*b*), Brockwell (1994) and Tsai & Chan (2000) have worked on this topic, but only Tong & Yeung (1991*a*) have considered *discrete* time series, while the other authors have addressed the problem under the continuous-time context. Specifically, Tong & Yeung (1991*b*) have studied the case of partially observed time series, where the main characteristic is that, by nature, the observations are not equally spaced, as happens with financial variables that are not observed in the weekends or holidays. The underlying model in that paper was a univariate self-exciting threshold (SETAR) model. In this paper, we will consider the case where the missing data appear because, for different reasons, the values of a variable were not recorded although they actually occurred. Of course, this situation causes unequally-spaced time series.

Unfortunately, Tong & Yeung's (1991*a*) procedure has a drawback, in the sense that the state space model they used for basing their *adapted* tests is not appropriate, as we will show in Section 3 below. Then, their arranged-autoregression ideas cannot be extended to the case of TAR models via state space forms. Instead, in this paper, we extend Tsay's (1998) test statistic and look for its null distribution under three scenarios: (1) complete data, (2) low missing-data percentage, and (3) medium and high missing-data percentage. Our goal is to find a threshold value for the missing-data rate up to which our extended test statistic maintains its null distribution.

The idea behind our work is the following: under the null hypothesis of linearity, one can estimate the missing data in the time series, using a linear-model based procedure as that of Gómez & Maravall (1994). Now, if the so-called input time series is nonlinear, we use a simplification of Nieto's (2005) smoother; if not, we also use the same linear-model based procedure. Then, one completes the time series with the estimated values and computes the test statistic. At the bottom line, we will find the distribution of the proposed test statistic under the null hypothesis of linearity taking into account the uncertainty of the missing data estimates. This work is done by means of Monte Carlo simulations.

The paper is organized as follows. In Section 2, we present the basic TAR model and its simplification under the null hypothesis. Section 3 describes Tsay's (1998) nonlinearity test, the extended test statistic and its null distribution for complete time series. In Section 4, we include a theoretical example that shows the drawback of Tong & Yeung's (1991*a*) procedure and analyze the effect that the missing-data-estimates uncertainty has on the null distribution of the proposed test statistic. Section 5 presents a real-data application and Section 6 concludes.

2. Specification of the TAR Model

Let $\{X_t\}$ and $\{Z_t\}$ be stochastic processes related by the equation (TAR model)

$$X_t \quad = \; a_0^{(j)} + \sum_{i=1}^{k_j} a_i^{(j)} X_{t-i} + h^{(j)} \varepsilon_t \;,\; r_{j-1} < Z_t \leq r_j \tag{1}$$

where $j = 1, \ldots, l-1$ indicate the presence of l regimes in the process $\{X_t\}$, which are determined by the threshold values r_0,r_1,...,r_{l-1}, and r_l of process $\{Z_t\}$, with $r_0 = -\infty$ and $r_l = \infty$. Here, $a_i^{(j)}$ and $h^{(j)}$; $j = 1, \ldots, l$; $i = 0, 1, \ldots, k_j$; are real numbers and $\{\varepsilon_t\}$ is a Gaussian zero-mean white noise process with variance 1. Additionally, the nonnegative integer numbers $k_1, \ldots, k_l$ denote, respectively, the autoregressive orders of $\{X_t\}$ in each regime. We shall use the symbol TAR$(l; k_1, \ldots, k_l)$ to denote this model and call l,r_1,...,r_{l-1}, k_1,...,k_{l-1} and k_l the model structural parameters.

These models were introduced by Tong (1978) and Tong & Lim (1980), specifically, in the case where the threshold variable is the lagged variable X_{t-d}, where d is some positive integer. In this case, the model is known as the self-exciting TAR (SETAR) model and, at present, there is a lot of literature about the topic of analyzing these models, under the frequent assumption that we know the number l of regimes and the autoregressive orders $k_1, \ldots, k_l$.

We also assume that $\{Z_t\}$ is exogenous in the sense that there is no feedback of $\{X_t\}$ towards it and that $\{Z_t\}$ is a homogeneous pth order Markov chain with initial distribution $F_0(z, \boldsymbol{\theta}_z)$ and kernel distribution $F_p(z_t \mid z_{t-1}, \ldots, z_{t-p}, \boldsymbol{\theta}_z)$, where $\boldsymbol{\theta}_z$ is a parameter vector in an appropriate numerical space. Furthermore, we assume that these distributions have densities in the Lebesgue-measure sense. Let $f_0(z, \boldsymbol{\theta}_z)$ and $f_p(z_t \mid z_{t-1}, \ldots, z_{t-p}, \boldsymbol{\theta}_z)$ be, respectively, the initial and kernel density functions of the distributions above. In what follows, we assume that the p-dimensional Markov chain $\{Z_t\}$ has an invariant or stationary distribution $f_p(z, \boldsymbol{\theta}_z)$.

Nieto's (2005) algorithms are based strongly in the regime-switching state-space form of the TAR model, given by the following: let $k = \max\{k_1, \ldots, k_l\}$, $\alpha_t = (X_t, X_{t-1}, \ldots, X_{t-k+1})'$, $\omega_t = (\varepsilon_t, 0, \ldots, 0)'$, and $\{J_t\}$ be a sequence of indicator variables such that $J_t = j$ if and only if $Z_t \in B_j$ for some j, $j = 1, \ldots, l$. Now, let $\boldsymbol{H} = (1, 0, \ldots, 0)'$ and for $j = 1, \ldots, l$, let $\boldsymbol{C}_j = (a_0^{(j)}, 0, \ldots, 0)'$,

$$\boldsymbol{A}_j = \left(\begin{array}{cccc|c} a_1^{(j)} & a_2^{(j)} & \cdots & a_{k-1}^{(j)} & a_k^{(j)} \\ & \boldsymbol{I}_{k-1} & & & \boldsymbol{0} \end{array} \right)$$

where $a_i^{(j)} = 0$ for $i > k_j$ and $\boldsymbol{I}_{k-1}$ denotes the identity matrix of order $k-1$, and

$$\boldsymbol{R}_j = \begin{pmatrix} h^{(j)} & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Then, the state space form for the TAR$(l; k_1, \ldots, k_l)$ model becomes

$$X_t = \boldsymbol{H}\boldsymbol{\alpha}_t \tag{2}$$

as the observation equation, and

$$\boldsymbol{\alpha}_t = \boldsymbol{C}_{J_t} + \boldsymbol{A}_{J_t}\boldsymbol{\alpha}_{t-1} + \boldsymbol{R}_{J_t}\omega_t \tag{3}$$

as the system or state equation, where it is understood that $\boldsymbol{C}_{J_t} = \boldsymbol{C}_j$ if at time t, $J_t = j$. The same remark holds for the values of the matrices $\boldsymbol{A}_{J_t}$ and $\boldsymbol{R}_{J_t}$. This kind of *nonlinear* state space models, where appart from the observation and system equations there is an underlying *indicator process* that defines the structure of these equations and the probability distributions of the error terms, have been studied in the literature by Shumway & Stoffer (1991), Carter & Kohn (1994, 1996) and Kim & Nelson (1999), among others.

The situation of interest we shall consider is that there are missing observations in the two time series, in such a way that the observed data are located at the unequally-spaced time points $t_1, \ldots, t_N$, with $1 \leq t_1 \leq \cdots \leq t_N \leq T$, for $\{X_t\}$, and at $s_1, \ldots, s_M$, $1 \leq s_1 \leq \cdots \leq s_M \leq T$, for $\{Z_t\}$, where T is the sample size. Nieto (2005) solved the problem of estimating both the model parameters, including the structural parameters, and the missing observations on the basis that $l > 1$. In particular, for estimating missing values in the observed time series of process $\{Z_t\}$, he found that the posterior densities for the variables Z are given by

$$p(\mathbf{z}_T \mid \boldsymbol{\alpha}, \mathbf{x}) \propto \prod_{j=T-p+1}^{T} p(\alpha_j \mid \mathbf{z}_T, \alpha_{j-1}) f_p(\mathbf{z}_T) \tag{4}$$

and

$$p(z_t \mid \mathbf{z}_{t+p}, \boldsymbol{\alpha}_t, \mathbf{x}_t) \propto p(\alpha_t \mid \mathbf{z}_{t+p-1}, \alpha_{t-1}) f_p(z_{t+p} \mid \mathbf{z}_{t+p-1}) f_p(\mathbf{z}_{t+p-1}) \tag{5}$$

for $t = T-p, \ldots, 1$, where, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)$, $\boldsymbol{\alpha}_t = (\alpha_1, \ldots, \alpha_t)$, $\mathbf{x}_t = (x_1, \ldots, x_t)$, $\mathbf{x} = (x_1, \ldots, x_T)$, and, in general, $\mathbf{z}_t = (z_{t-p+1}, \ldots, z_t)$. Then, for estimating the missing data at the time points $s_1, \ldots, s_M$, one obtains draws from their corresponding posterior densities given by expressions (4) and (5) via MCMC procedures.

Now, for estimating the missing data in the time series of process $\{X_t\}$, one has to take into account that (see Nieto's (2005) paper)

$$p(\boldsymbol{\alpha} \mid \mathbf{z}, \mathbf{x}) = p(\boldsymbol{\alpha}_T \mid \mathbf{z}, \mathbf{x}) \prod_{t=1}^{T-1} p(\boldsymbol{\alpha}_t \mid \boldsymbol{\alpha}_{t+1}, \mathbf{z}, \mathbf{x}_t) \tag{6}$$

where $\mathbf{z} = (z_1, \ldots, z_T)$. Since the first component of $\boldsymbol{\alpha}_t$ is X_t, one obtains draws from the posterior density $p(\boldsymbol{\alpha} \mid \mathbf{z}, \mathbf{x})$, then marginalizes it at the time points $t_1, \ldots, t_N$ and picks the first component up. Under the assumption of Gaussianity for the process $\{\varepsilon_t\}$, each factor in (6) is the density of a multivariate normal distribution (see Carter & Kohn's (1994) paper for details)

If $l = 1$, $\{X_t\}$ reduces to a linear AR(k_1) model and there is no influence of $\{Z_t\}$ onto $\{X_t\}$, in the sense of the dynamic causality explained by the TAR model (1). That is to say, for any value of the variable Z, the variable X has the same dynamic autoregressive answer, with parameters $a_i = a_i^{(1)}$ for all i; $i = 0, 1, \ldots, k_1$. Likewise, one has that the white-noise-process weight is $h = h^{(1)}$. Importantly, the process $\{Z_t\}$ can be either linear or nonlinear.

3. A Nonlinearity Test for Complete Multivariate Stochastic Processes

Let $\{\mathbf{X}_t = (X_{1t}, \ldots, X_{kt})'\}$, $\{\mathbf{Y}_t = (Y_{1t}, \ldots, Y_{vt})'\}$ and $\{Z_t\}$ be stochastic processes, where the first two are multivariate and the last one is univariate. Tsay (1998) proposed the following multivariate threshold model for $\{\mathbf{X}_t\}$ with threshold process $\{Z_t\}$ and delay $d > 0$:

$$\mathbf{X}_t = \mathbf{a}_0^{(j)} + \sum_{i=1}^{p} \mathbf{a}_i^{(j)} \mathbf{X}_{t-i} + \sum_{i=1}^{q} \mathbf{b}_i^{(j)} \mathbf{Y}_{t-i} + \boldsymbol{\varepsilon}_t^{(j)} \tag{7}$$

if Z_{t-d} belongs to the real interval $B_j = (r_{j-1}, r_j]$ for some j; $j = 1, \ldots, l$; where $-\infty = r_0 < r_1 < \cdots < r_{l-1} < r_l = \infty$, $\mathbf{a}_0^{(j)}$ are constant vectors, $\mathbf{a}_i^{(j)}$ and $\mathbf{b}_i^{(j)}$ are constant matrices, and p and q are nonnegative integers. The innovations satisfy $\boldsymbol{\varepsilon}_t^{(j)} = \Sigma_j^{1/2} \mathbf{u}_t$, where $\Sigma_j^{1/2}$ is a symmetric positive definite matrix, $j = 1, \ldots, l$, and $\{\mathbf{u}_t\}$ is a zero-mean vector white noise process with covariance matrix $\mathbf{I}$, the identity matrix. The threshold process $\{Z_t\}$ is assumed to be stationary and have a continuous distribution. Notice the presence of the process $\{\mathbf{Y}_t\}$ in the autoregressive equation for $\{\mathbf{X}_t\}$. This is to explain for exogenous variables.

Model (1) can be seen as a particular case of model (7) if one puts $k = 1$, no exogenous variables, and $d = 0$ (although this value is not strictly covered by Tsay's (1998) model, in a mathematical sense). However, in model (1), one can have different autoregressive orders $k_1, \ldots, k_l$ in each regime and the threshold process is specified to be an invariant Markov chain, a more general concept than that of a stationary process. This point is important under the null hypothesis, as noted in the previous section, because, at present, the only method for estimating missing data in process $\{Z_t\}$, when it is nonlinear, is Nieto's (2005) approach.

Now, we describe Tsay's (1998) test. Consider the null hypothesis that $\{\mathbf{X}_t\}$ is linear, i.e. $l = 1$, versus the alternative hypothesis that it follows the multivariate threshold model given in (7). Using the arranged regression scheme, one has the following: given observations $\mathbf{x}_t$, $\mathbf{y}_t$, and z_t, $t = 1, 2, \ldots, n$, the goal is to detect the threshold nonlinearity of $\{\mathbf{X}_t\}$, assuming that p, q, and d are known.

Let $h = \max\{p, q, d\}$, $\mathbf{W}_t = (1, \mathbf{x}'_{t-1}, \ldots, \mathbf{x}'_{t-p}, \mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-q})$ (a $(pk + qv + 1)$–dimensional vector) and $\mathbf{\Phi}$ an unknown matrix. If the null hypothesis holds, then the model collapses to

$$\mathbf{X}'_t = \mathbf{W}'_t \mathbf{\Phi} + \boldsymbol{\varepsilon}'_t \tag{8}$$

as explained by Tsay (1998), for $t = h + 1, \ldots, n$, independent of the values of variable Z. Let $S = \{z_{h+1-d}, \ldots, z_{n-d}\}$ be the set of values of Z_{t-d}. Consider the order statistics of S and denote its ith smallest element by $z_{(i)}$. Then, the arranged regression based on the increasing order of the threshold variable Z_{t-d} is

$$\mathbf{X}'_{t(i)+d} = \mathbf{W}'_{t(i)+d} \mathbf{\Phi} + \boldsymbol{\varepsilon}'_{t(i)+d} \tag{9}$$

for $i = 1, \ldots, n - h$.

Now, let $\widehat{\mathbf{\Phi}}_m$ be the least square estimator of $\mathbf{\Phi}$ of equation (9) based on the first m observations, that is, those associated with the m smallest values of S. Let

$$\widehat{\mathbf{e}}_{t(m+1)+d} = \mathbf{X}_{t(m+1)+d} - \widehat{\mathbf{\Phi}}'_m \ \mathbf{W}_{t(m+1)+d} \tag{10}$$

and

$$\widehat{\boldsymbol{\eta}}_{t(m+1)+d} = \widehat{\mathbf{e}}_{t(m+1)+d} / [1 + \mathbf{W}'_{t(m+1)+d} \mathbf{V}_m \mathbf{W}_{t(m+1)+d}]^{1/2} \tag{11}$$

where $\mathbf{V}_m = \left[\sum_{i=1}^{m} \mathbf{W}_{t(i)+d} \mathbf{W}'_{t(i)+d}\right]^{-1}$, be the predictive residual and the standardized predictive residual of regression (9). These quantities can be obtained by the recursive least square algorithm. Next, consider the regression

$$\widehat{\boldsymbol{\eta}}'_{t(l)+d} = \mathbf{W}'_{t(l)+d} \mathbf{\Psi} + \boldsymbol{\epsilon}'_{t(l)+d} \tag{12}$$

for $l = m_0 + 1, \ldots, n - h$, where m_0 denotes the starting point of the recursive least squares estimation. The problem of interest is then to test the hypothesis $H_0 : \boldsymbol{\psi} = \mathbf{0}$ versus the alternative $H_a : \boldsymbol{\psi} \neq \mathbf{0}$. Tsay (1998) proposed the test statistic

$$C(d) = [n - h - m_0 - (kp + vq + 1)][\ln|\mathbf{S}_0| - \ln|\mathbf{S}_1|] \tag{13}$$

where the argument d signifies that the test depends strongly on the delayed threshold variable Z_{t-d}, $|\boldsymbol{A}|$ denotes the determinant of the matrix $\boldsymbol{A}$,

$$\mathbf{S}_0 = \frac{1}{n - h - m_0} \sum_{l=m_0+1}^{n-h} \widehat{\boldsymbol{\eta}}_{t(l)+d} \widehat{\boldsymbol{\eta}}'_{t(l)+d}$$

and

$$\mathbf{S}_1 = \frac{1}{n - h - m_0} \sum_{l=m_0+1}^{n-h} \widehat{\boldsymbol{\epsilon}}_{t(l)+d} \widehat{\boldsymbol{\epsilon}}'_{t(l)+d}$$

where $\widehat{\boldsymbol{\epsilon}}_{t(l)+d}$ is the least squares residual of regression (12). Under the null hypothesis that $\mathbf{X}_t$ is linear, Tsay (1998) showed that $C(d)$ is asymptotically a chi-squared

random variable with $k(pk + qv + 1)$ degrees of freedom. This paper shows that this test statistic has good optimal properties, which are reflected in having a greater power function than other statistical tests for the same null hypothesis. As a by-product, the statistic $C(d)$ can be used for choosing adequate threshold variables, when one has several candidate variables. The idea is to select that variable for which its corresponding value of $C(d)$ is the largest. Furthermore, if $k = 1$, i.e. $\{X_t\}$ is univariate, and there are not exogenous variable, i.e. $q = 0$, then the corresponding chi-squared distribution has $p + 1$ degrees of freedom.

Now, we consider the test statistic $C(d)$ above and set $d = 0$, then the previous regressions can still be conducted and thus the statistic $C(0)$ is computed. To find its null distribution for complete time series, we proceed via simulation and obtained tha,t even for small sample sizes, it is practically a chi-squared distribution with $p + 1$ degrees of freedom. Table 1 presents the results for a Monte Carlo simulation experiment, where the autoregressive linear model under the null is $X_t = 2 + 0.5X_{t-1} + \varepsilon_t$, where $\{\varepsilon_t\}$ is a Gaussian zero-mean white noise process with variance 1. The sample size was $n = 150$ and we run 5000 replicates. In the body of the table appear the quantiles of the $\chi^2(2)$ distribution and of the empirical distribution of $C(0)$. The p-value for the Kolmogorov-Smirnov test statistic was 0.36, approximately, which signals a no rejection of the null hypothesis of equal distributions. Additionally, we used a sample size $n = 10000$ and another AR(1) models with coefficients -0.5 and 1 (nonstationary process), and found analogous results. These are presented in Tables 4 and 5 of the Appendix. Furthermore, we considered AR(2) and AR(3) models with coefficients that make the processes stationary and nonstationary and, in the first case, we considered scenarios where the roots are either real numbers or some of them complex numbers. In this way, we take into account different characteristics in the time and frequency domain. We can provide these results upon request. The overall conclusion was the same, i.e. the distribution of $C(0)$ is practically a $\chi^2(p + 1)$ distribution, $p = 1, 2, 3$. We feel that the maximum value 3 for the AR order is enough in this simulation study because of model parsimony and that the exercise can be extended to seasonal AR processes obtaining the same global result.

Table 1: Comparison of empirical quantiles of the null distribution of $C(0)$ with those of the χ^2, in the case of an AR model with parameter 0.5.

	Quantiles								
Distribution	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
$\chi^2(2)$	0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21
$C(0)$	0.02	0.05	0.10	0.22	1.42	4.59	5.97	7.31	9.19

4. The Null Distribution of the Proposed Test Statistic in the Case of Missing Data

4.1. A Tray to Use State-Space-Model Based Approaches

In the SETAR-model univariate context, Tong & Yeung (1991*a*) presented a state-space-model based procedure for implementing known tests of the null hypothesis $H : l = 1$, in the presence of partial data. Following the arranged-regression philosophy, they argued that under the null hypothesis the arranged regression can be cast in state space form. Setting $t(i)$ in place of t everywhere in equations 2 and 3, one would obtain

$$\boldsymbol{X}_{t(i)} = \boldsymbol{H}\boldsymbol{\alpha}_{t(i)} \tag{14}$$

as the observation equation, and

$$\boldsymbol{\alpha}_{t(i)} = \boldsymbol{C}_{J_{t(i)}} + \boldsymbol{A}_{J_{t(i)}}\boldsymbol{\alpha}_{t(i)-1} + \boldsymbol{R}_{J_{t(i)}}\omega_{t(i)} \tag{15}$$

as the system or state equation, where it is remarkably noted that $J_{t(i)} = 1$ for all $i = 1, \ldots, n-h$. Hence, the system equation becomes

$$\boldsymbol{\alpha}_{t(i)} = \boldsymbol{C}_1 + \boldsymbol{A_1}\boldsymbol{\alpha}_{t(i)-1} + \boldsymbol{R}_1\omega_{t(i)} \tag{16}$$

Equations (14) and (16) define Tong & Yeung's (1991*b*) state space model and they will be referred as an arranged state space model.

Apparently, the usual statistical assumptions and properties of state space models continue to be valid (see Harvey's (1989) book, for example), and thus the Kalman filter, its associated smoothing algorithms and the Nieto's (2005) approach might still be used. However, this is not possible. Indeed, (i) an important argument in deducting the Kalman filter and then the well-known smoothing algorithms (see, among others, Harvey (1989), Catlin (1989) and Brockwell & Davis (1991)) is that the time points at which observations are made need to be in a monotone order although not necessarily equally spaced. In the present scheme, it can happen that $i < j$ and even $t(i) > t(j)$. (ii) The so-called predictive residuals are not orthogonal among them and orthogonal to lagged variables of the output process $\{X_t\}$ neither. Hence, the probabilistic behaviour of the so-called *adapted* test statistics of Tong & Yeung (1991*a*), which is necessary for implemeting their tests, is not necessarily guaranteed. The following example illustrates these facts.

An AR(1) model will be considered for process $\{X_t\}$ given by $X_t = a_1 X_{t-1} + h\varepsilon_t$ (as happens under the null hypothesis), where a_1 and h are real numbers with $h > 0$, and $\{\varepsilon_t\}$ is a zero-mean white noise process for which $\mathrm{E}(X_s\varepsilon_t) = 0$ for $s < t$. Then, trivially, the state-space-model elements are $\boldsymbol{\alpha}_t = (X_t)$, $\boldsymbol{\omega}_t = (\varepsilon_t)$, $\boldsymbol{H} = 1$, $\boldsymbol{C}_1 = 0$, $\boldsymbol{A}_1 = a_1$, and $\boldsymbol{R}_1 = h$. We assume that the sample size is $n = 1000$, and that $t(1) = 200$, where there is a missing data, $t(2) = 27$, and $t(3) = 379$. After some simple calculations, using the algorithms presented by Tong & Yeung (1991*a*) for computing the state-space-model based predictive errors $\boldsymbol{\eta}$'s, one founds that $\eta_{27} = X_{27} - a_1^2 X_{199}$ and $\eta_{379} = X_{379} - a_1^2 X_{27}$, which are clearly

correlated. Consequently, we leave Tong & Yeung's (1991*a*) state-space-model based approach and consider the alternative of using directly the test described in Subsection 3.1.

4.2. The Null Distribution of $C(0)$

Now, the idea is to assess the influence that the uncertainty in the missing data estimates has on the distribution of $C(0)$. Let $\widehat{C}(0)$ be the statistic that is obtained when we use missing data estimates to compute $C(0)$. We proceed via Monte Carlo simulation as in the case of complete data, maintaining the same AR(p) models for process $\{X_t\}$, that is with $p = 1, 2, 3$ and different values for the autoregressive parameters, and considering different sample sizes. The new element is to consider several rates of missing observations, going from low to high percentages, and to detect a threshold rate up to which the χ^2 null distribution is preserved. In this paper, we consider values in the set $\{0\%, 10\%, 20\%, \ldots, 80\%\}$.

The design of the simulation experiment was the following: we fix a stationary AR(1) model for process $\{Z_t\}$; in this way, $\{Z_t\}$ is a Markov chain of order 1 with invariant distribution. The chosen model is $Z_t = 0.25Z_{t-1} + \alpha_t$, where $\{\alpha_t\}$ is a Gaussian zero-mean white noise process with variance 1.5^2. Then,

(1) we draw time series for each stochastic process $\{X_t\}$ and $\{Z_t\}$, say $\{x_t\}$ and $\{z_t\}$, in an independent way.

(2) We select randomly two sets of time points in the set $\{1, \ldots, T\}$. The first one of size $T - N$ for $\{x_t\}$ and the second one of size $T - M$ for $\{z_t\}$. These time points are fixed. Then, we discard the observations in the time series $\{x_t\}$, located at the first set of time points, and those for $\{z_t\}$ that correspond to the second set.

(3) Using Gómez & Maravall (1994) procedures, specifically their fixed-point smoother algorithm, we estimated the missing observations in the time series $\{x_t\}$. Since $\{Z_t\}$ is linear, the same procedure is used to estimate the missing data in its simulated time series.

(4) Compute $\widehat{C}(0)$ with these "completed" time series.

***Note* 1.** Thinking in practice, if process $\{Z_t\}$ is not linear, we can use the smoother given by equations (4) and (5) for estimating its missing data, with the following modification: since there is no influence of $\{Z_t\}$ onto $\{X_t\}$, the posterior densities presented in equations (4) and (5) are reduced, respectively, to

$$p(\mathbf{z}_T \mid \boldsymbol{\alpha}, \mathbf{x}) \propto f_p(\mathbf{z}_T) \tag{17}$$

for $t = T - p + 1, \ldots, T$, and

$$p(z_t \mid \mathbf{z}_{t+p}, \boldsymbol{\alpha}_t, \mathbf{x}_t) \propto f_p(z_{t+p} \mid \mathbf{z}_{t+p-1}) f_p(\mathbf{z}_{t+p-1}) \tag{18}$$

for $t = T-p, \ldots, 1$. In this way, drawings for $\mathbf{Z}_T = (Z_{T-p+1}, \ldots, Z_T)$ are obtained directly from the invariant distribution of the Markov chain $\{Z_t\}$, and for $t =$

$T-p,\ldots,1$, drawings for Z_t are obtained from the distribution given by the product of the kernel density with the invariant density.

The above procedure is repeated I times, $I \geq 1$, to obtain a sample of size I for the statistic $\widehat{C}(0)$, maintaining fixed the missing-data percentages $(T-N)/T$ and $(T-M)/T$ through all the iterations. With this sample for $\widehat{C}(0)$, we obtained its empirical cumulative distribution function and then compare it with that of the χ^2_{p+1} distribution.

The results of the simulation experiment, with the AR(1) model with parameter 0.5 for $\{X_t\}$, are presented in Table 2, using samples of size $n = 150$ and $I = 5000$, a fixed value that we will use in all the remaining simulations. We can see the following important facts: (i) when there are not missing observations in the time series $\{x_t\}$, for any percentage of missing data in $\{z_t\}$, the empirical quantiles are practically equal to the $\chi^2(2)$ distribution. This reflects that under H_0, the process $\{Z_t\}$ does not influence $\{X_t\}$. (ii) Fixing the missing-data percentage of $\{x_t\}$ and varying that of $\{z_t\}$, the corresponding empirical quantiles are very similar, reflecting once more again the observation made in (i). (iii) When the rate of missing data in the time series $\{x_t\}$ gets larger, the discrepancy between empirical and theoretical quantiles gets larger, too, independent of the missing-data percentage in the time series $\{z_t\}$. This fact suggests that the null distribution of $\widehat{C}(0)$ departs from the $\chi^2(2)$ distribution.

Table 2: Comparison of empirical quantiles of the null distribution of $\widehat{C}(0)$ with those of the $\chi^2(2)$ for $n = 150$ and $\phi = 0.5$.

% of missing data		Quantiles								
x-data	z-data	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
10	0	0.02	0.04	0.10	0.21	1.37	4.58	5.92	7.43	9.31
20	0	0.02	0.05	0.09	0.20	1.34	4.48	5.80	7.06	8.94
30	0	0.02	0.05	0.10	0.22	1.51	5.12	6.67	8.04	10.36
0	10	0.02	0.05	0.10	0.21	1.40	4.62	5.98	7.33	8.80
10	10	0.02	0.05	0.09	0.21	1.35	4.50	5.88	7.28	8.72
20	10	0.02	0.04	0.09	0.19	1.33	4.37	5.65	6.87	8.59
30	10	0.02	0.05	0.12	0.22	1.51	5.10	6.48	8.22	10.23
0	20	0.02	0.05	0.10	0.21	1.41	4.63	6.27	7.73	9.57
10	20	0.02	0.05	0.10	0.20	1.39	4.74	6.17	7.56	9.40
20	20	0.02	0.05	0.11	0.22	1.29	4.25	5.61	6.91	8.84
30	20	0.02	0.05	0.11	0.22	1.32	4.48	5.75	7.09	9.30
0	30	0.02	0.05	0.10	0.22	1.37	4.64	6.06	7.51	9.22
10	30	0.02	0.05	0.10	0.21	1.40	4.70	6.15	7.57	9.88
20	30	0.02	0.04	0.09	0.20	1.30	4.24	5.61	6.94	8.80
30	30	0.02	0.04	0.09	0.18	1.29	4.21	5.43	6.71	8.65
$\chi^2(2)$		0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21

In Table 3 we present the Kolmogorov-Smirnov statistics for the same missing-data rates considered in Table 2 and, as we can see, when the percentage of missing data in $\{x_t\}$ is greater or equal than 20%, approximately, the null hypothesis of

equal distributions is rejected in almost every case. Consequently, we always halt the simulations at the rate 30%.

TABLE 3: p-values of the Kolmogorov-Smirnov statistic for the distributions in Table 2.

% of missing data		
x-data	z-data	p-value*
10	0	0.14
20	0	0.28
30	0	0.00
0	10	0.80
10	10	0.26
20	10	0.01
30	10	0.00
0	20	0.76
10	20	0.85
20	20	0.00
30	20	0.02
0	30	0.86
10	30	0.90
20	30	0.00
30	30	0.00

* Rounded to two decimal digits.

We repeated the simulation exercise for this AR model with sample sizes $n = 1000, 10000$ and obtained very similar results to the last ones. Also, we used the other AR(1) models proposed in Section 3 with the same sample sizes $n = 150$ and $n = 10000$, and we find analogous results, which are in Tables 6-17 in the Appendix. In this last case, we omitted the sample size $n = 1000$ because we do not observe important differences with respect to the sample size $n = 150$. Furthermore, we consider the same AR(2) and AR(3) models of Section 3 with the same sample sizes and, once more again, we obtained similar results, which can be provided upon request.

As a global conclusion, when the missing data percentage in the time series $\{x_t\}$ is less than 20%, approximately, we can say that the null distribution of $\widehat{C}(0)$ is still a $\chi^2(p+1)$ distribution. For rates of missing data greater than this approximate threshold, the influence of the missing-data estimates uncertainty on the distribution of $\widehat{C}(0)$ is so relevant that its empirical distribution departs from the χ^2 distribution. It is important to remark here that one could extend this simulation study via the selection of percentage values between 10% and 20%, to find a more precise bound for the percentage of missing values in $\{x_t\}$ up to which the χ^2 distribution continues to be valid as the null distribution of $\widehat{C}(0)$.

5. An Empirical Example

Nieto (2005) considers an application with actual data. The time series considered were daily rainfall (in mm), as the threshold variable, and a daily river flow (in m^3/s), as the response variable, in a certain Colombian geographical region. The data set corresponds to the sample period from January 1, 1992, up to November 30, 2000 (3256 data), and it was assembled by IDEAM, the official Colombian agency for hydrological and meteorological studies. In Figure 1, one can see the two time series, where is clear the dynamical relationship between the two variables. Additionally, one can see certain stable path in both variables and bursts of large values, specifically in the river flow.

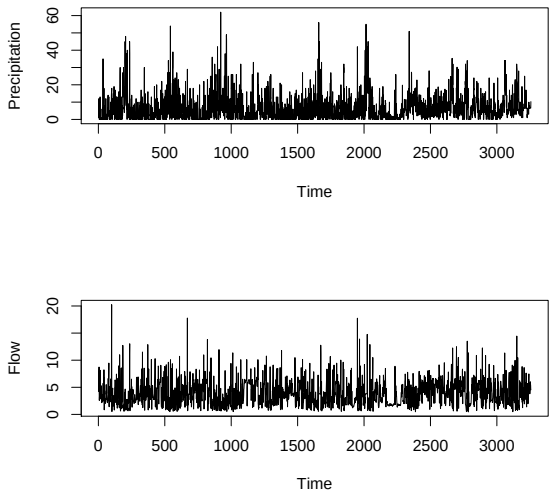

FIGURE 1: Time series for the real data example: (a) Precipitation (b) Flow.

Let P_t and X_t be respectively the rainfall and the river flow at day t. Because of the universal convention for measuring these two variables, he set up $Z_t = P_{t-1}$. That is, the precipitation was lagged one period back for relating it to the river flow. The flow time series was adjusted with two transformations: (1) square root of the data and (2) an adjustment for conditional heteroscedasticity via an ARCH(1) model. From now on, the flow data to be analyzed will be the transformed ones and we denote them as $\{\widetilde{x}_t\}$. The two time series have missing data, 52 in $\{z_t\}$ and 32 in $\{x_t\}$. In percentage terms, these are 1.6% for time series $\{z_t\}$ and 1% for $\{\widetilde{x}_t\}$, which are lesser than 20%.

Under the null hypothesis, $\{X_t\}$ is a linear AR process. Using Akaike's information criterion, we found that $p = 2$ is a reasonable autoregressive order for this process. In this way, we are in the conditions studied in Section 4 and, consequently, we can use the $\chi^2(3)$ distribution for running the statistical test for

the null hypothesis. We note firstly that the process $\{Z_t\}$ has two regimes with $r_1 = 6.0$ mm and that it was modeled as a 1st order Markov chain with approximate intial and transition kernel distributions given by the mixtures

$$f_0(z) = 0.26h_n(z) + 0.74g(z)$$

and

$$f_1(z_t \mid z_{t-1}) = 0.87h_n(z_t) + 0.13g(z_t \mid z_{t-1})$$

respectively, where

$$h_n(z) = \begin{cases} 0, & \text{if } -\infty < z < -1/n, \\ (n\pi/2)\cos(n\pi z + \pi/2), & \text{if } -1/n \leq z \leq 0 \\ 0, & \text{if } z > 0 \end{cases}$$

$g(z)$ denotes the truncated density of a $N(3.24, 7.76^2)$ at the point $z = 0$, $g(z_t \mid z_{t-1})$ is the truncated density of a $N(z_{t-1}, 7.76^2)$ at the same point $z = 0$, $P(Z_t = 0) = 0.26$, and $P(Z_t = 0 \mid Z_{t-1} \in B_1) = 0.87$. For more details on the modeling of process $\{Z_t\}$, the reader can see Gómez & Maravall (1994) paper. Then, we used the procedure described in Subsection 3.3 for estimating the missing data in the two time series, specifically, we used the TSW software (Caporello & Maravall 2003) for estimating the missing data in $\{\widetilde{x}_t\}$ and the smoother given in equations (17) and (18) for estimating the missing observations in $\{z_t\}$. Next, I complete the time series with the estimated data and obtained that $\widehat{C}(0) = 52.57$ with p-value equal to 0. These results signal the strong threshold nonlinearity of $\{X_t\}$, which is explained by $\{Z_t\}$.

6. Conclusions

In this paper, we have shown the feasibility of a well-known statistical procedure for testing the null hypothesis of linearity against the alternative of TAR nonlinearity, when there are missing data in a bivariate time series. The statistical test is an extension of Tsay's (1998) statistic. The extension consists in allowing the number zero to be the delay parameter of the threshold variable. Then, to compute values of this statistic we use estimates of the missing data as observed values. Strictly speaking, this statistic is other than the extended one. Via Monte Carlo simulations, we found that the extended statistic also follows a χ^2 distribution with complete data and that the modified statistic also has this distribution if the proportion of missing data in the output time series is less or equal than 20%, approximately. We feel that if the missing-data percentage is larger than this value, we should use additional variables that help to explain the dynamical behavior of the output one, via for example regression models, to get more observed data and to do a frequentist statistical test. Another alternative might be to use a Bayesian testing approach to compensate the large uncertainty produced by the high percentage of missing data. This route would need to consider appropriate prior distributions. This is a challenging problem for future research.

In the lines of the above recommendation, another interesting problem for future research would be to consider test statistics other than Tsay's (1998) one, as for example Hansen's (1996) test. And then to do a comparison about the size and power of the different tests, under scenarios of missing-data proportions where the known null distributions are preserved.

As a by-product of this study, we have also shown that state-space-model based approaches, which aim to take into account the arranged autoregression philosophy, are not adequate. This means that the appealing idea of detecting change points, via arranged regressions, should be used directly for designing inferential procedures in the context of TAR models.

7. Acknowledgments

We gratefully acknowledge Professor Ruey S. Tsay at the University of Chicago for his disposal to discuss with us the topic and for his valuable advising on the development of the research. We also thank an anonymous referee for useful comments and suggestions on a previous version of the paper, which help to improve substantially the initial manuscript.

[Recibido: enero de 2008 — Aceptado: febrero de 2011]

References

Brockwell, P. J. (1994), 'On continuous-time threshold ARMA processes', *Journal of Statistical Planning and Inference* **39**, 291–303.

Brockwell, P. J. & Davis, R. A. (1991), *Time Series: Theory and Methods*, Springer-Verlag, New York.

Caporello, G. & Maravall, A. (2003), *Software TSW*, Banco de España, Madrid.

Carter, C. K. & Kohn, R. (1994), 'On Gibbs sampling for state space models', *Biometrika* **81**, 541–553.

Carter, C. K. & Kohn, R. (1996), 'Markov chain Monte Carlo in conditionally gaussian state space models', *Biometrika* **83**, 589–601.

Catlin, D. (1989), *Estimation, Control, and the Discrete Kalman Filter*, Springer-Verlag, New York.

Gómez, V. & Maravall, A. (1994), 'Estimation, prediction, and interpolation for nonstationary series with the Kalman filter', *Journal of the American Statistical Association* **89**, 611–624.

Hansen, B. E. (1996), 'Inference when a nuisance parameter is not identified under the null hypothesis', *Econometrica* **64**, 413–460.

Harvey, A. C. (1989), *Forecasting, Structural Time Series, and the Kalman filter*, Cambridge University Press, Cambridge.

Kim, C. & Nelson, C. R. (1999), *State Space Models with Regime Switching*, The MIT Press, Cambridge.

Nieto, F. H. (2005), 'Modeling bivariate threshold autoregressive processes in the presence of missing data', *Communications in Statistics - Theory and Methods* **34**(4), 905–930.

Shumway, R. H. & Stoffer, D. S. (1991), 'Dynamic linear models with switching', *Journal of the American Statistical Association* **86**, 411–430.

Tong, H. (1978), On a threshold model in pattern recognition and signal processing, *in* C. H. Chen, ed., 'Pattern recognition and signal processing', Sijhoff & Noordhoff, Amsterdam.

Tong, H. & Lim, K. S. (1980), 'Threshold autoregression, limit cycles, and cyclical data', *Journal of the Royal Statistical Society, Series B* **42**, 245–292.

Tong, H. & Yeung, I. (1991*a*), 'On tests for self-exciting threshold autoregressive non-linearity in partially observed time series', *Applied Statistics* **40**, 43–62.

Tong, H. & Yeung, I. (1991*b*), 'Threshold autoregressive modeling in continuous time', *Statistica Sinica* **1**, 411–430.

Tsai, H. & Chan, K. S. (2000), 'Testing for nonlinearity with partially observed time series', *Biometrika* **87**, 805–821.

Tsay, R. S. (1998), 'Testing and modeling multivariate threshold models', *Journal of the American Statistical Association* **93**, 1188–1202.

Appendix

TABLE 4: Comparison of empirical quantiles of the null distribution of $C(0)$ with those of a $\chi^2(2)$, in the case of an AR(1) process with parameter -0.5 and sample size $n = 150$.

	Quantiles								
Distribution	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
$\chi^2(2)$	0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21
$C(0)$	0.02	0.05	0.11	0.21	1.40	4.62	6.06	7.36	9.15

The Kolmogorov-Smirnov statistical test yields a p-value of 0.67, which implies a no rejection of the null hypothesis of equal distributions. With a sample size $n = 10000$ the p-value of this statistical test is 0.38, which leads to the same decision.

Table 5: Empirical quantiles of the null distribution of $C(0)$ and those of the $\chi^2(2)$, in the case of an AR(1) process with parameter 1.0 and sample size $n = 150$.

	Quantiles								
Distribution	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
$\chi^2(2)$	0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21
$C(0)$	0.03	0.06	0.12	0.23	1.40	4.62	6.05	7.42	9.35

Table 6: Comparison of empirical quantiles of the null distribution of $\widehat{C}(0)$ for the AR(1) model with parameter 0.5 and sample size 1000.

% of missing data		Quantiles								
x-data	z-data	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
10	0	0.02	0.05	0.11	0.22	1.37	4.42	5.83	7.16	9.08
20	0	0.02	0.06	0.10	0.22	1.42	4.63	6.19	7.64	9.54
30	0	0.02	0.05	0.11	0.22	1.46	4.76	6.17	7.59	9.95
0	10	0.02	0.06	0.10	0.21	1.37	4.56	6.06	7.54	9.17
10	10	0.02	0.05	0.11	0.21	1.38	4.39	5.73	7.36	9.23
20	10	0.01	0.05	0.10	0.22	1.43	4.86	6.21	7.67	9.88
30	10	0.02	0.04	0.09	0.20	1.44	4.76	6.20	7.77	9.63
0	20	0.02	0.05	0.10	0.22	1.39	4.63	6.01	7.31	8.93
10	20	0.02	0.05	0.11	0.23	1.40	4.48	5.77	7.26	9.01
20	20	0.02	0.05	0.10	0.21	1.42	4.60	6.00	7.28	9.29
30	20	0.02	0.05	0.11	0.21	1.45	4.77	6.16	7.34	9.78
0	30	0.03	0.06	0.12	0.22	1.43	4.69	6.18	7.53	9.44
10	30	0.02	0.06	0.11	0.23	1.43	4.77	6.15	7.41	9.47
20	30	0.02	0.05	0.10	0.23	1.44	4.82	6.16	7.46	9.67
30	30	0.02	0.05	0.11	0.24	1.52	4.91	6.34	7.96	9.92
χ^2		0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21

The Kolmogorov-Smirnov statistical test yields a p-value of 0.74, which signals a no rejection of the null hypothesis of equal distributions. With a sample size $n = 10000$ the p-value of this statistical test is 0.12, which signals the same decision.

TABLE 7: p-values of the Kolmogorov-Smirnov statistics in the AR(1) model with coefficient 0.5 and sample size 1000.

% of missing data		p-value*
x-data	z-data	
10	0	0.56
20	0	0.31
30	0	0.00
0	10	0.78
10	10	0.27
20	10	0.11
30	10	0.11
0	20	0.98
10	20	0.23
20	20	0.41
30	20	0.09
0	30	0.07
10	30	0.01
20	30	0.11
30	30	0.00

* Rounded to two decimal digits.

TABLE 8: Comparison of empirical quantiles of the null distribution of $\widehat{C}(0)$ for the AR(1) model with parameter 0.5 and sample size 10000.

% of missing data		Quantiles								
x-data	z-data	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
10	0	0.02	0.05	0.10	0.20	1.37	4.68	6.05	7.50	9.15
20	0	0.02	0.05	0.10	0.21	1.40	4.76	6.34	7.67	9.29
30	0	0.02	0.05	0.11	0.22	1.48	4.81	6.24	7.82	9.93
0	10	0.02	0.05	0.11	0.20	1.35	4.54	5.87	7.09	8.65
10	10	0.02	0.05	0.10	0.21	1.35	4.59	5.98	7.28	8.68
20	10	0.02	0.05	0.10	0.20	1.42	4.82	6.23	7.63	9.33
30	10	0.02	0.05	0.11	0.22	1.46	4.77	6.12	7.59	9.91
0	20	0.02	0.05	0.10	0.20	1.35	4.56	5.87	7.25	8.89
10	20	0.02	0.05	0.11	0.21	1.40	4.62	5.85	7.33	9.33
20	20	0.02	0.05	0.10	0.20	1.39	4.82	6.28	7.56	9.57
30	20	0.02	0.06	0.12	0.22	1.43	4.73	6.13	7.52	9.75
0	30	0.02	0.05	0.10	0.20	1.36	4.45	5.94	7.13	8.75
10	30	0.02	0.05	0.11	0.22	1.38	4.57	6.05	7.24	8.75
20	30	0.02	0.05	0.10	0.22	1.38	4.77	6.07	7.46	9.17
30	30	0.02	0.05	0.10	0.22	1.44	4.71	6.12	7.51	9.37
χ^2		0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21

TABLE 9: p-values of the Kolmogorov-Smirnov statistics in the AR(1) model with coefficient 0.5 and sample size 10000.

% of missing data		p-value*
x-data	z-data	
10	0	0.27
20	0	0.23
30	0	0.00
0	10	0.35
10	10	0.43
20	10	0.12
30	10	0.02
0	20	0.15
10	20	0.74
20	20	0.22
30	20	0.30
0	30	0.66
10	30	0.80
20	30	0.44
30	30	0.02

* Rounded to two decimal digits.

TABLE 10: Comparison of empirical quantiles of the null distribution of $\widehat{C}(0)$ for the AR(1) model with parameter -0.5 and sample size 150.

% of missing data		Quantiles								
x-data	z-data	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
10	0	0.02	0.05	0.11	0.21	1.39	4.65	6.02	7.15	9.11
20	0	0.02	0.06	0.12	0.22	1.35	4.48	5.96	7.26	8.70
30	0	0.02	0.05	0.11	0.22	1.51	5.14	6.72	8.16	10.36
0	10	0.02	0.04	0.09	0.20	1.40	4.57	5.92	7.17	9.50
10	10	0.02	0.05	0.10	0.21	1.35	4.56	5.92	7.15	8.95
20	10	0.02	0.05	0.11	0.21	1.29	4.23	5.74	7.12	8.40
30	10	0.02	0.05	0.11	0.24	1.49	5.06	6.48	8.03	9.65
0	20	0.02	0.05	0.10	0.21	1.41	4.67	6.13	7.62	9.76
10	20	0.02	0.04	0.10	0.21	1.40	4.65	6.07	7.79	9.65
20	20	0.02	0.04	0.08	0.19	1.25	4.22	5.44	6.64	8.62
30	20	0.02	0.05	0.10	0.22	1.42	4.68	6.33	7.70	9.74
0	30	0.02	0.05	0.11	0.21	1.38	4.52	6.02	7.45	9.25
10	30	0.02	0.05	0.10	0.21	1.43	4.85	6.33	7.73	9.50
20	30	0.02	0.05	0.09	0.19	1.21	4.16	5.36	6.58	8.59
30	30	0.02	0.05	0.10	0.19	1.28	4.44	5.88	7.14	8.71
	χ^2	0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21

TABLE 11: p-values of the Kolmogorov-Smirnov statistics in the AR(1) model with coefficient -0.5 and sample size 150.

% of missing data		p-value*
x-data	z-data	
10	0	0.91
20	0	0.22
30	0	0.00
0	10	0.77
10	10	0.56
20	10	0.00
30	10	0.00
0	20	0.53
10	20	0.97
20	20	0.00
30	20	0.08
0	30	0.79
10	30	0.09
20	30	0.00
30	30	0.00

* Rounded to two decimal digits.

TABLE 12: Comparison of empirical quantiles of the null distribution of $\widehat{C}(0)$ for the AR(1) model with parameter -0.5 and sample size 10000.

% of missing data		Quantiles								
x-data	z-data	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
10	0	0.02	0.04	0.09	0.20	1.38	4.66	5.95	7.20	8.98
20	0	0.02	0.05	0.10	0.22	1.45	4.77	6.20	7.52	9.58
30	0	0.02	0.06	0.12	0.23	1.46	4.82	6.20	7.88	9.93
0	10	0.02	0.05	0.10	0.20	1.37	4.52	5.93	7.11	8.92
10	10	0.02	0.04	0.10	0.21	1.36	4.62	5.97	7.30	9.26
20	10	0.02	0.05	0.12	0.23	1.47	4.82	6.20	7.60	9.75
30	10	0.02	0.05	0.10	0.21	1.44	4.74	6.28	7.68	9.47
0	20	0.02	0.04	0.10	0.20	1.33	4.47	5.76	7.28	8.77
10	20	0.01	0.04	0.10	0.20	1.37	4.45	5.91	7.57	9.40
20	20	0.02	0.06	0.10	0.21	1.43	4.64	6.06	7.44	9.34
30	20	0.02	0.05	0.11	0.23	1.40	4.76	6.16	7.63	9.48
0	30	0.02	0.04	0.10	0.20	1.35	4.47	5.86	7.10	8.61
10	30	0.02	0.04	0.09	0.20	1.39	4.49	5.98	7.43	8.90
20	30	0.03	0.06	0.11	0.21	1.42	4.81	6.25	7.66	9.40
30	30	0.02	0.05	0.11	0.23	1.44	4.72	6.05	7.46	9.43
χ^2		0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21

TABLE 13: p-values of the Kolmogorov-Smirnov statistics in the AR(1) model with coefficient -0.5 and sample size 10000.

% of missing data		p-value*
x-data	z-data	
10	0	0.96
20	0	0.02
30	0	0.00
0	10	0.60
10	10	0.78
20	10	0.01
30	10	0.01
0	20	0.14
10	20	0.79
20	20	0.06
30	20	0.15
0	30	0.48
10	30	0.50
20	30	0.04
30	30	0.08

* Rounded to two decimal digits.

TABLE 14: Comparison of empirical quantiles of the null distribution of $\widehat{C}(0)$ for the AR(1) model with parameter 1.0 and sample size 150.

% of missing data		Quantiles								
x-data	z-data	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
10	0	0.02	0.05	0.10	0.20	1.38	4.55	5.93	7.22	9.09
20	0	0.02	0.06	0.12	0.24	1.40	4.81	6.18	7.58	9.24
30	0	0.02	0.06	0.11	0.23	1.61	5.34	6.94	8.42	10.30
0	10	0.03	0.06	0.11	0.22	1.40	4.60	6.08	7.52	9.12
10	10	0.02	0.05	0.10	0.21	1.39	4.52	5.90	7.33	8.99
20	10	0.02	0.04	0.10	0.22	1.40	4.62	6.09	7.43	9.45
30	10	0.02	0.07	0.13	0.25	1.59	5.24	6.74	8.30	10.01
0	20	0.02	0.05	0.11	0.22	1.41	4.68	6.03	7.40	9.01
10	20	0.02	0.05	0.11	0.23	1.42	4.67	6.13	7.48	9.28
20	20	0.02	0.05	0.11	0.23	1.49	4.77	6.04	7.41	9.47
30	20	0.02	0.05	0.11	0.22	1.45	4.69	6.11	7.68	9.34
0	30	0.02	0.05	0.11	0.22	1.41	4.45	5.77	7.34	9.27
10	30	0.02	0.05	0.10	0.21	1.42	4.46	5.75	7.16	8.96
20	30	0.02	0.05	0.10	0.23	1.49	4.76	6.17	7.64	9.87
30	30	0.02	0.05	0.10	0.20	1.39	4.49	5.67	7.02	8.80
χ^2		0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21

TABLE 15: p-values of the Kolmogorov-Smirnov statistics in the AR(1) model with coefficient 1.0 and sample size 150.

% of missing data		p-value*
x-data	z-data	
10	0	0.83
20	0	0.31
30	0	0.00
0	10	0.25
10	10	0.92
20	10	0.48
30	10	0.00
0	20	0.28
10	20	0.09
20	20	0.00
30	20	0.01
0	30	0.48
10	30	0.34
20	30	0.00
30	30	0.05

* Rounded to two decimal digits.

TABLE 16: Comparison of empirical quantiles of the null distribution of $\widehat{C}(0)$ for the AR(1) model with parameter 1.0 and sample size 10000.

% of missing data		Quantiles								
x-data	z-data	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
10	0	0.01	0.04	0.09	0.16	1.14	3.79	5.03	6.20	7.78
20	0	0.02	0.05	0.11	0.22	1.48	4.88	6.39	7.90	10.21
30	0	0.02	0.06	0.12	0.23	1.42	4.79	6.26	8.14	10.38
0	10	0.02	0.06	0.11	0.22	1.32	4.69	6.22	7.58	9.36
10	10	0.01	0.04	0.10	0.19	1.21	3.96	5.14	6.42	7.84
20	10	0.03	0.06	0.12	0.23	1.46	4.91	6.49	7.88	10.04
30	10	0.02	0.05	0.10	0.22	1.44	4.87	6.37	8.07	10.08
0	20	0.03	0.05	0.10	0.21	1.37	4.72	6.37	7.78	9.51
10	20	0.03	0.05	0.10	0.22	1.30	4.03	5.29	6.78	8.54
20	20	0.02	0.06	0.12	0.23	1.51	5.03	6.56	7.96	9.77
30	20	0.02	0.05	0.10	0.22	1.48	4.96	6.53	8.04	10.23
0	30	0.02	0.04	0.10	0.20	1.37	4.60	5.90	7.32	9.31
10	30	0.02	0.03	0.08	0.16	1.03	3.63	4.79	5.94	7.53
20	30	0.02	0.05	0.11	0.23	1.48	4.86	6.32	7.64	9.82
30	30	0.02	0.05	0.10	0.22	1.47	5.09	6.54	8.31	11.03
χ^2		0.02	0.05	0.10	0.21	1.39	4.60	5.99	7.38	9.21

TABLE 17: p-values of the Kolmogorov-Smirnov statistics in the AR(1) model with coefficient 1.0 and sample size 10000.

% of missing data		p-value*
x-data	z-data	
10	0	0.12
20	0	0.01
30	0	0.08
0	10	0.09
10	10	0.00
20	10	0.03
30	10	0.00
0	20	0.72
10	20	0.00
20	20	0.00
30	20	0.00
0	30	0.99
10	30	0.00
20	30	0.00
30	30	0.01

* Rounded to two decimal digits.

Revista Colombiana de Estadística
Junio 2011, volumen 34, no. 1, pp. 95 a 109

Análisis bayesiano para la distribución lognormal generalizada aplicada a modelos de falla con censura

Bayesian Analysis for the Generalized Lognormal Distribution Applied to Failure Time Analysis

Freddy Hernández[1,a], Olga Cecilia Usuga[1,2,b]

[1]Departamento de Estadística, Instituto de Matemática y Estadística, Universidad de São Paulo, São Paulo, Brasil

[2]Ingeniería Industrial, Facultad de Ingenierías, Universidad de Antioquia, Medellín, Colombia

Resumen

Existen varias versiones de la distribución lognormal en la literatura estadística y una de ellas está basada en la transformación exponencial de la distribución normal generalizada (NG). En el presente artículo se presenta el análisis Bayesiano para la distribución lognormal generalizada (logNG) considerando distribuciones a priori de Jeffreys independientes para los parámetros; así como el procedimiento para implementar el muestreador de Gibbs que permite obtener las distribuciones a posteriori de los parámetros. Los resultados obtenidos son usados para analizar modelos de tiempo de falla con datos no censurados y censurados a derecha Tipo I. El procedimiento propuesto es ilustrado usando una base de datos real relacionada con tiempos de falla de computadores.

***Palabras clave*:** análisis de tiempo de falla, censura a derecha, distribución lognormal generalizada, inferencia bayesiana, muestreador de Gibbs.

Abstract

There are several versions of the lognormal distribution in the statistical literature, one is based in the exponential transformation of generalized normal distribution (GN). This paper presents the Bayesian analysis for the generalized lognormal distribution (logGN) considering independent non-informative Jeffreys distributions for the parameters as well as the procedure for implementing the Gibbs sampler to obtain the posterior distributions of parameters. The results are used to analyze failure time models with right-censored and uncensored data. The proposed method is illustrated using actual failure time data of computers.

***Key words*:** Bayesian inference, Failure time analysis, Gibbs sampling, Lognormal distribution, Right censoring.

[a]Estudiante de doctorado. E-mail: fhernanb@ime.usp.br
[b]Profesora asistente. E-mail: ousuga@udea.edu.co

1. Introducción

En estudios de confiabilidad la distribución exponencial tiene un papel fundamental desde el punto de vista conceptual y práctico; sin embargo, algunas veces esta distribución no proporciona ajustes apropiados para modelar los datos obtenidos de un experimento, esto mismo sucede con otras distribuciones como la Weibull y Gamma; por lo tanto, una buena opción consiste en analizar los datos usando la distribución lognormal (Chen 1995). Se han obtenido buenos ajustes usando la distribución lognormal para el caso de conjuntos de datos observados y datos experimentales (Aitchison & Brown 1957) para modelar fallas en pruebas de vida (Chen & Papadopoulos 1997) y ha sido usada especificamente en el campo de la electrónica para analizar tiempos de vida de mecanismos de conducción eléctrica (Howard & Dodson 1961) y en tiempos de vida de transistores de germanio (Adam 1962).

Desde el punto de vista bayesiano para problemas de tiempos de falla la distribución lognormal de tres parámetros propuesta por Lawless (1982) fue analizada por Upadhyay & Peshwani (2001) para el caso en el que existen observaciones censuradas y no censuradas; este análisis fue realizado usando el muestreador de Gibbs, el cual es una herramienta importante para obtener la distribución a posteriori de los parámetros y que exige diseño, implementación y validación de algoritmos apropiados (Barrera & Correa 2008). Otra versión de la distribución lognormal generalizada con tres parámetros basada en la transformación exponencial de la distribución normal generalizada propuesta por Nadarajah (2005) fue estudiada por Martín & Pérez (2009) usando el muestreador de Gibbs pero sólo para el caso de observaciones sin censura.

El objetivo de este artículo consiste en extender la propuesta de análisis bayesiano presentada por Martín & Pérez (2009) usando el enfoque de Upadhyay & Peshwani (2001) para la distribución logNG propuesta por Nadarajah (2005). Se consideraron dos situaciones: la primera, cuando se tienen observaciones sin censura, y la segunda, cuando se tiene censura a derecha Tipo I. El análisis bayesiano propuesto está basado en el muestreador de Gibbs y se presenta una aplicación para una base de datos relacionada a los tiempos para que se presente la primera falla en un conjunto de computadores nuevos.

En la segunda sección de este artículo se presenta la versión de la distribuición lognormal generalizada con tres parámetros a estudiar y el procedimiento para generar observaciones aleatorias de esta distribución. En la tercera sección se presenta el análisis bayesiano y los algoritmos para la implementación del muestreador de Gibbs considerando el enfoque no informativo en los casos con y sin observaciones censuradas. En la cuarta sección se presenta la aplicación del procedimento propuesto en el artículo para analizar una base de datos real.

2. Distribución lognormal generalizada

En la literatura estadística se pueden encontrar varias versiones de la distribución lognormal propuestas por Lawless (1982), Chen (1995) y otra versión pre-

sentada en esta sección que es obtenida através de la transformación exponencial de una variable que sigue una distribución normal generalizada propuesta por Nadarajah (2005).

Una variable aleatoria X sigue distribución lognormal generalizada con tres parámetros si la transformación $Y = \log X$ sigue una distribución normal generalizada. La función de densidad para una variable logNG con parámetros μ, σ y s es dada por

$$f(x \mid \mu, \sigma, s) = \frac{s}{2x\sigma\Gamma(1/s)} \exp\left(-\left|\frac{\log x - \mu}{\sigma}\right|^s\right) \tag{1}$$

donde $x > 0$, $-\infty < \mu < +\infty$, $\sigma > 0$, $s \geq 1$ y $\Gamma(\cdot)$ corresponde a la función Gamma.

Otras distribuciones como la lognormal y la log-Laplace se obtienen a partir de la expresión (1), tomando $s = 2$ y cambiando σ por $\sqrt{2}\sigma$ se obtiene la distribución lognormal y cuando $s = 1$ se obtiene la distribución log-Laplace. Martín & Pérez (2009) afirman que las densidades de la logNG con $s \in (1, 2) \cup (2, 3)$ son apropiadas para el ajuste de datos en los cuales la lognormal no genera ajustes satisfactorios. Una característica importante de la familia logNG es que todas las densidades están concentradas a la izquierda (véase figura 1), y cuando X tiende infinito, la densidad disminuye lo cual es apropiado para modelar variables asociadas a tiempos de vida.

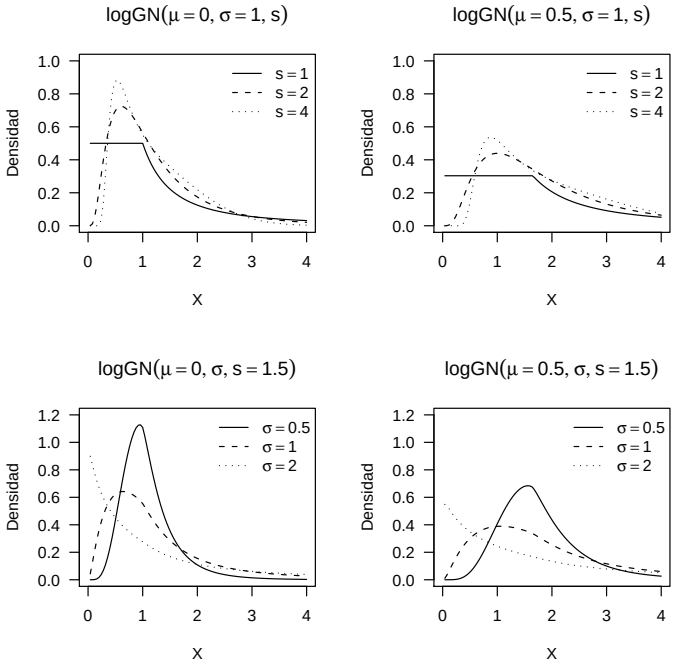

FIGURA 1: Densidades para la familia logNG

La siguiente proposición fue presentada por Martín & Pérez (2009) y muestra la relación entre la distribución Gamma y la logNG y, por tanto, un método para generar variables aleatorias de una distribución logNG.

Proposición 1. *Sean U y X dos variables aleatorias tales que $U \sim Gamma(\alpha = 1 + 1/s, \gamma = 1)$ y $f(x \mid u) = \mathcal{I}\left[\exp(\mu - \sigma u^{1/s}) < x < \exp(\mu + \sigma u^{1/s})\right] / (2x\sigma u^{1/s})$ entonces $X \sim \log NG(\mu, \sigma, s)$.*

donde $\mathcal{I}[\cdot]$ corresponde a la función indicadora.

El proceso propuesto por Martín & Pérez (2009) basado en la proposición anterior para generar observaciones de una distribución logNG(μ, σ, s) consta de los siguientes tres pasos.

1. Generar $U \sim Gamma(\alpha = 1 + 1/s, \gamma = 1)$
2. Generar $V \sim Unif(0, 1)$
3. Hacer $X = \exp(\sigma U^{1/s} V + \mu)$

La función de sobrevivencia definida como $S(x) = P(X > x)$ corresponde a la probabilidad de que un individuo sobreviva más allá del tiempo x y en el contexto de fallas de equipos $S(x)$ es llamada función de confiabilidad (Klein & Moeschberger 2003); para la distribución logNG, la función de sobrevivencia está dada por:

$$S(x) = \begin{cases} 1 - \dfrac{\Gamma\left(\frac{1}{s}, \left(\dfrac{\mu - \log(x)}{\sigma}\right)^s\right)}{2\Gamma(1/s)} & \text{si } 0 < x \leq \exp(\mu) \\ \dfrac{\Gamma\left(\frac{1}{s}, \left(\dfrac{\log(x) - \mu}{\sigma}\right)^s\right)}{2\Gamma(1/s)} & \text{si } x > \exp(\mu) \end{cases}$$

donde $\Gamma(\cdot, \cdot)$ denota la función Gamma incompleta.

La función de riesgo definida como $r(x) = f(x)/S(x)$ se conoce en sobrevivencia como tasa condicional de falla (Klein & Moeschberger 2003) y para la distribución logNG es dada por:

$$r(x) = \begin{cases} \dfrac{s \exp\left(-\left(\dfrac{\mu - \log(x)}{\sigma}\right)^s\right)}{x\sigma\left(2\Gamma(1/s) - \Gamma\left(\frac{1}{s}, \left(\dfrac{\mu - \log(x)}{\sigma}\right)^s\right)\right)} & \text{si } 0 < x \leq \exp(\mu) \\ \dfrac{s \exp\left(-\left(\dfrac{\log(x) - \mu}{\sigma}\right)^s\right)}{x\sigma\Gamma\left(\frac{1}{s}, \left(\dfrac{\log(x) - \mu}{\sigma}\right)^s\right)} & \text{si } x > \exp(\mu) \end{cases}$$

En la figura 2 se presentan las funciones de riesgo para dos densidades particulares de la distribución logNG. Para ver propiedades de la función de sobrevivencia y riesgo, véase Gupta & Lvin (2005).

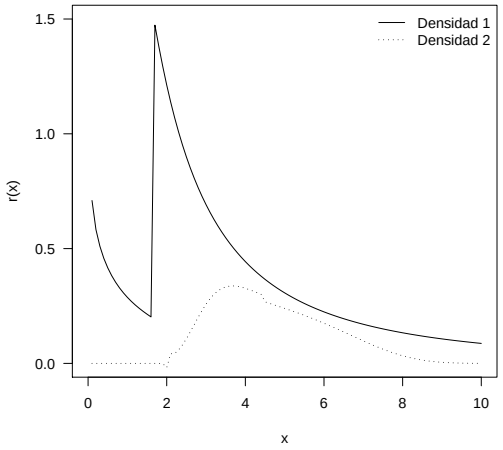

FIGURA 2: Función de riesgo. Densidad 1 ($\mu = 0.5$, $\sigma = 2$, $s = 1.5$) y Densidad 2 ($\mu = 1.5$, $\sigma = 0.5$, $s = 3$)

3. Análisis bayesiano

Según Robert (2001), el principal propósito de la teoría estadística es que a partir de un modelo estadístico basado en las observaciones recolectadas de un fenómeno aleatorio se puedan obtener inferencias sobre la distribución de probabilidad asociada al fenómeno estudiado. El modelo estadístico bayesiano está compuesto de dos elementos, el primer elemento corresponde al modelo estadístico paramétrico dado por $f(x \mid \theta)$ donde x corresponde a la información obtenida de los datos y θ al parámetro de la distribución asociada al fenómeno; el segundo elemento $\pi(\theta)$ corresponde a la distribución a priori para el parámetro. Estos dos elementos combinados dan lugar al modelo bayesiano $\pi(\theta \mid x)$ dado por $\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$. La novedad de la aproximación bayesiana es que pone las causas (observaciones) y los efectos (parámetros) en el mismo nivel considerando para ambos distribuciones de probabilidad (Robert 2001).

Berger (1985) asegura que el enfoque bayesiano ofrece la posibilidad de incluir en el modelo la opinión de especialistas por medio de la distribución a priori en el proceso de inferencia. Diversos trabajos relacionados con la distribución lognormal tradicional han usado a priori de Jeffreys, Padgett & Johnson (1983), Upadhyay & Peshwani (2001, 2003, 2008). Martín & Pérez (2009) propusieron la utilización de la distribución a priori no informativa de Jeffreys para la distribución logNG. Portela & Gómez-Villegas (2004) sugieren usar distribuciones a priori independientes para cada uno de los parámetros en las distribuciones de la familia de distribuciones Exponencial Potencia a la cual pertenece la logNG. En este trabajo se consideran distribuciones a priori de Jeffreys independientes para los parámetros, este tipo de

distribuciones no informativas son útiles cuando las opiniones de los especialistas o conocimientos previos difieren (Gelman, Stern & Rubin 2004).

3.1. Enfoque sin censura

En virtud de la proposición 1 es posible escribir la distribución para logNG como la distribución marginal para x de $f(x \mid \mu, \sigma, s, u)f(u \mid s)$, donde

$$f(x \mid \mu, \sigma, s, u) = \frac{1}{2x\sigma u^{1/s}} \mathcal{I}\left[\exp(\mu - \sigma u^{1/s}) < x < \exp(\mu + \sigma u^{1/s})\right] \tag{2}$$

$$f(u \mid s) = Gamma(1 + 1/s, 1) \tag{3}$$

Así la función de verosimilitud para los parámetros μ, σ, s y $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)'$ dada una muestra aleatoria $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$ es dada por

$$\begin{aligned} L(\mu, \sigma, s, \boldsymbol{u} \mid \boldsymbol{x}) &= \prod_{i=1}^{n} f(x_i \mid \mu, \sigma, s, u_i) f(u_i \mid s) \\ &= \frac{s^n}{(2\sigma)^n \Gamma^n(1/s)} \prod_{i=1}^{n} \frac{e^{-u_i}}{x_i} \mathcal{I}\left[\exp(\mu - \sigma u_i^{1/s}) < x_i < \exp(\mu + \sigma u_i^{1/s})\right] \end{aligned} \tag{4}$$

Considerando el siguiente conjunto de distribuciones a priori de Jeffreys independientes para los parámetros

$$\pi(\mu) \propto 1, \quad \pi(\sigma) \propto \frac{1}{\sigma}, \quad \pi(s) \propto \frac{1}{s}$$

La distribución a posteriori obtenida es dada por

$$\pi(\mu, \sigma, s, \boldsymbol{u} \mid \boldsymbol{x}) \propto \frac{s^{n-1}}{\sigma^{n+1} \Gamma^n(1/s)} \prod_{i=1}^{n} \frac{e^{-u_i}}{x_i} \mathcal{I}\left[\exp(\mu - \sigma u_i^{1/s}) < x_i < \exp(\mu + \sigma u_i^{1/s})\right] \tag{5}$$

En el caso de modelos multiparamétricos la distribución a posteriori conjunta no siempre presenta una forma conocida y, por lo, tanto es difícil obtener muestras aleatorias; sin embargo, es posible que a partir de las distribuciones a posteriori marginales de cada uno de los parámetros se puedan obtener muestras aleatorias con mayor facilidad. En estos casos, una aproximación a la distribución a posteriori conjunta puede realizarse usando el muestreador de Gibbs, consiste en un algoritmo iterativo para construir una secuencia dependiente de valores para los parámetros que convergen a los parámetros de la distribución a posteriori conjunta estudiada (Hoff 2009). El conjunto de distribuciones a posteriori marginal para cada uno de los parámetros de la distribución a posteriori conjunta de la expresión (5) es dada

por:

$$\pi(\mu \mid \sigma, s, \boldsymbol{u}, \boldsymbol{x}) \propto 1, \quad \max_i \left\{\log(x_i) - \sigma u_i^{1/s}\right\} < \mu < \min_i \left\{\log(x_i) + \sigma u_i^{1/s}\right\} \quad (6)$$

$$\pi(\sigma \mid \mu, s, \boldsymbol{u}, \boldsymbol{x}) \propto \frac{1}{\sigma^{n+1}}, \quad \sigma > \max_i \left\{ \frac{|\mu - \log(x_i)|}{u_i^{1/s}} \right\} \quad (7)$$

$$\pi(s \mid \sigma, \mu, \boldsymbol{u}, \boldsymbol{x}) \propto \frac{s^{n-1}}{\Gamma^n(1/s)}, \quad s < \frac{\log(u_i)}{\log\left|\frac{\log(x_i)-\mu}{\sigma}\right|} \quad (8)$$

$$\pi(u_i \mid \sigma, s, \mu, \boldsymbol{x}) \propto e^{-u_i}, \quad u_i > \left(\frac{|\log(x_i) - \mu|}{\sigma}\right)^s, \quad i = 1, 2, \ldots, n \quad (9)$$

Para generar observaciones aleatorias de las distribuciones a posteriori anteriores es importante identificar los núcleos característicos, la densidad (6) corresponde a la distribución uniforme, en la densidad (7) se tiene el núcleo de la distribución Pareto y en la densidad (9) se tiene el núcleo de la exponencial, mientras que para la densidad (8) es necesario usar el método de rechazo para obtener observaciones aleatorias (véase Gamerman & Lopes 2006).

3.2. Enfoque con censura

Supongamos que la muestra aleatoria $\boldsymbol{x}$ obtenida tiene observaciones con censura a derecha Tipo I, es decir, existen r observaciones $\boldsymbol{x}_1 = (x_1, x_2, \ldots, x_r)'$ de las n en las cuales fueron observados tiempos de falla, mientras que $\boldsymbol{x}_2 = (x_{r+1}, x_{r+2}, \ldots, x_n)'$ corresponde a las $n - r$ observaciones con censura, donde $\boldsymbol{x} = \boldsymbol{x}_1 \cup \mathbf{x}_2$. Así la función de verosimilitud dada en (4) ahora se define como

$$L(\mu, \sigma, s, \boldsymbol{u} \mid \boldsymbol{x}) = \prod_{i=1}^{r} f(x_i \mid \mu, \sigma, s, u_i) \prod_{j=r+1}^{n} \int_{x_j}^{\infty} f(x_j \mid \mu, \sigma, s, u_j) dx_j \quad (10)$$

Considerando las mismas distribuciones a priori de la sección anterior, la distribución a posteriori es ahora dada por

$$\pi(\mu, \sigma, s, \boldsymbol{u} \mid \boldsymbol{x}_1, \boldsymbol{x}_2) \propto P_1 P_2 \quad (11)$$

donde

$$P_1 = \frac{s^{r-1}}{\sigma^{r+1}\Gamma^r(1/s)} \prod_{i=1}^{r} \frac{e^{-u_i}}{x_i} \mathcal{I}\left[\exp(\mu - \sigma u_i^{1/s}) < x_i < \exp(\mu + \sigma u_i^{1/s})\right]$$

$$P_2 = \prod_{j=r+1}^{n} \int_{x_j}^{\infty} f(x_j \mid \mu, \sigma, s, u_j) dx_j$$

El cálculo de la distribución a posteriori dada en (11) es complicado por causa de P_2 debido a las observaciones censuradas; una manera de tratar este problema, el cual es el objetivo del presente trabajo, consiste en considerar $\boldsymbol{x}_2$, el conjunto

de observaciones con censura, como desconocidos e incluirlo en el muestreador de Gibbs (Upadhyay & Peshwani 2001). Las distribuciones a posteriori para los parámetros de la logNG y para las observaciones con censura son entonces dadas por

$$\pi(\mu, \sigma, s, \boldsymbol{u} \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \pi(\mu, \sigma, s, \boldsymbol{u} \mid \boldsymbol{x}) \tag{12}$$

$$\pi(\boldsymbol{x}_2 \mid \mu, \sigma, s, \boldsymbol{u}, \boldsymbol{x}_1) = \pi(\boldsymbol{x}_2 \mid \mu, \sigma, s, \boldsymbol{u}) \tag{13}$$

El enfoque propuesto tiene dos etapas: la primera consiste en incluir el conjunto de observaciones con censura como conocido (suponiendo valores iguales a la censura) en el muestreador de Gibbs para calcular la distribución a posteriori de los parámetros de la logNG usando la expresión (12), la cual corresponde a la distribución a posteriori de los parámetros de la logNG, sin considerar censura de la expresión (5). La segunda etapa consiste en que una vez que son actualizados los valores de los parámetros de la distribución logNG se generan nuevos valores para las observaciones censuradas usando la expresión (13) que se reduce a generar $n - r$ observaciones independientes x de una distribución logNG truncada de tal forma que $x > x_j$ para $j = r + 1, \ldots, n$.

3.3. Sistemas de generación

A continuación se presenta el procedimiento usado para la aplicación del muestreador de Gibbs al problema estudiado.

3.3.1. Sin censura

1. Generar valores iniciales para μ, σ, s y u_i.
2. Generar una observación para μ según la distribución uniforme dada en (6).
3. Generar una observación para σ según la distribución Pareto dada en (7).
4. Generar una observación para s según la distribución dada en (8) usando el método de muestreo por rechazo implementado en el paquete Runuran de R Development Core Team (2010).
5. Generar u_i según la distribución exponencial dada en (9).

Una vez actualizados los valores de los parámetros μ, σ, s y u_i repetir los pasos 2 al 5 y continuar el proceso.

3.3.2. Con censura

Cuando la base de datos contiene observaciones censuradas, el procedimiento para la aplicación del muestreador de Gibbs es en esencia similar a los pasos de la subsección anterior, pero con algunos cambios. En la primera iteración es necesario considerar temporalmente las r observaciones censuradas como no censuradas, y

el valor de la observación corresponde al valor de la censura, esto se hace así para generar valores iniciales de los parámetros usando toda la información disponible. Después se aplican los pasos 2 al 5 más un sexto paso adicional que consiste en generar r observaciones provenientes de una distribución logNG truncada en los valores de censura. Después volver al paso 2 y continuar el proceso iterativo.

4. Aplicación

La aplicación se realizó usando la base de datos presentada por Barrera & Correa (2008) obtenida del departamento de apoyo técnico de la Universidad Nacional de Colombia, la cual contiene información sobre el tiempo para la presencia de la primera falla en un conjunto de 72 computadores nuevos con iguales características, comprados todos en la misma fecha y observados hasta transcurridos 66 meses.

En la tabla 1 se presentan los datos. Se observa que de los 72 computadores solo 17 fallaron antes de terminar el horizonte de observación mientras que los 55 datos restantes denotados por 66^+ indican que estos computadores no presentaron fallas durante el horizonte del estudio.

Tabla 1: Tiempos para la primera falla de 72 computadores (meses).

14.07	17.80	19.43	21.33	24.60	28.97	29.63	33.73	37.60	37.67	40.87	52.40
53.97	60.57	64.27	65.43	65.43	66^+	66^+	66^+	66^+	66^+	66^+	66^+
66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+
66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+
66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+	66^+

Se realizaron dos análisis a la base de datos. En el primero se consideró la distribución logNG para los tiempos de falla y luego aplicando los pasos de la sección 3.3.1 con diez mil iteraciones fueron obtenidas las distribuciones marginal a posteriori de los 3 parámetros. En el segundo análisis se consideró la distribución lognormal y nuevamente aplicando los pasos de la sección 3.3.2, pero teniendo en cuenta que la distribución lognormal es un caso particular de la logNG cuando $s = 2$ y $\sigma = \sqrt{2}\sigma$, así fueron nuevamente obtenidas las distribuciones marginal a posteriori.

En la figura 3 se presentan los resultados obtenidos del muestreador de Gibbs, en la parte izquierda se tienen las distribuciones marginales para el caso de la distribución logNG, mientras que en la parte derecha están las distribuciones marginales con la distribución lognormal; además, se incluyeron las regiones de mayor densidad (High Density Regions (HDR)) a un nivel de confianza del 90 % las cuales se calcularon con el paquete *hdrcde* de R Development Core Team (2010). Se observa que las distribuciones y las regiones de mayor densidad para μ son similares en ambos casos, se observa también que la distribución para S está concentrada cerca del valor de 2 y la región de mayor densidad incluye este valor de 2.

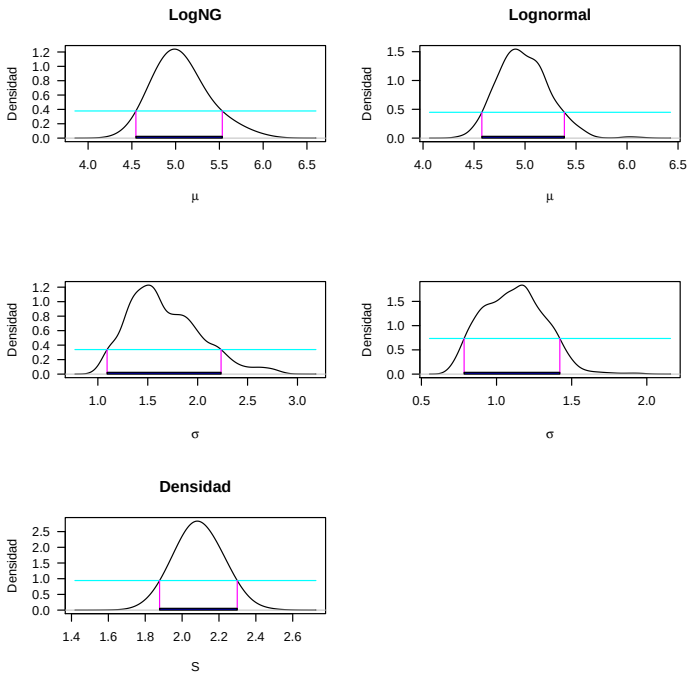

Figura 3: Distribución marginal a posteriori para los parámetros con HDR de 90 %.

En la tabla 2 se presentan las modas para las distribuciones marginales de los parámetros de la figura 3 en los casos considerados. Se observa que las modas de μ y σ son cercanas para ambos casos y la moda para S está cercana al valor de 2.

Tabla 2: Moda para las distribuciones marginales

	μ	σ	S
logNG	4.99	1.51	2.08
log	4.91	1.17	-

En la figura 4 se presentan dos distribuciones predictivas, una de ellas corresponde al caso de la distribución logNG y la otra a la distribución lognormal como posibles distribuciones para el tiempo de falla de los computadores. En la tabla 3 se presentan algunos cuantiles de interés para las distribuciones predictivas de la figura 4. De la figura se observa que existe una gran similitud entre las dos distribuciones, ya que las curvas son bastante cercanas entre sí; además, los percentiles son cercanos, especialmente el percentil 25.

Se usó el criterio de información de desvianza (CID) propuesto por Spiegelhalter, Best, Carlin & van der Linde (2002), el cual es un criterio de comparación de modelos bayesianos, el modelo con el menor CID es elegido como el modelo que mejor predice el conjunto de datos. Al calcular el CID para la aplicación se

encontró que $\text{CID}_{logNG} = 916$ y $\text{CID}_{lognormal} = 941$. Esto indica que el modelo más apropiado corresponde al modelo logNG.

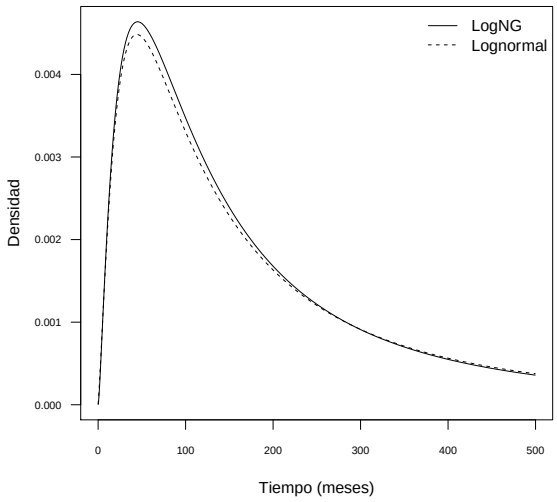

FIGURA 4: Distribución predictiva.

TABLA 3: Percentiles para las distribuciones predictivas.

Percentil	25	50	75
logNG	69.9	150.4	338.0
log	68.0	142.0	306.3

5. Conclusiones

En el artículo se ha propuesto la metodología bayesiana para usar la distribución lognormal generalizada con tres parámetros con el objetivo de estudiar modelos de falla con censura a derecha. Se consideraron a prioris de Jeffreys independientes para los parámetros y se implementó el muestreador de Gibbs para obtener las distribuciones a posteriori de los parámetros. El procedimiento se ilustró usando una base de datos real correspondiente a los tiempos de falla de computadores, teniendo la base de datos observaciones con censura a derecha. Se consideraron las distribuciones lognormal y lognormal generalizada como posibles distribuciones para los tiempos de falla.

Como posible trabajo futuro se pueden generar nuevos procedimientos considerando otras distribuciones a priori para los parámetros, así como otro tipo de distribuciones para los tiempos de falla. La comparación del desempeño de

los procedimientos propuestos se podría realizar por medio del escore logarítmico propuesto por Bernardo (1979) mediante simulaciones.

Agradecimientos

Agradecemos a los revisores anónimos por sus comentarios que permitieron mejorar el presente trabajo.

[Recibido: agosto de 2010 — Aceptado: febrero de 2011]

Referencias

Adam, J. (1962), 'Failure time distribution estimation', *Semiconductor Reliability* **2**, 41–52.

Aitchison, J. & Brown, J. (1957), *The Lognormal Distribution*, Cambridge University Press, United Kingdom.

Barrera, C. & Correa, J. (2008), 'Distribución predictiva bayesiana para modelos de pruebas de vida vía MCMC', *Revista Colombiana de Estadística* **31**(2), 145–155.

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer, New York.

Bernardo, J. (1979), 'Expected information as expected utility', *Annals of Statistics* **7**(3), 686–690.

Chen, G. (1995), 'Generalized log-normal distributions with reliability application', *Computational Statistics and Data Analysis* **19**(3), 309–319.

Chen, K. & Papadopoulos, A. (1997), 'Shortest Bayes credibility intervals for the lognormal failure model', *Microelectron Reliability* **37**(12), 1859–1863.

Gamerman, D. & Lopes, H. (2006), *Markov Chain Monte Carlo Stocastic Simulation for Bayesian Inference*, Chapman and Hall/CRC, Boca Raton.

Gelman, A., Stern, J. & Rubin, H. (2004), *Bayesian Data Analysis*, Chapman & Hall-CRC.

Gupta, R. & Lvin, S. (2005), 'Reliability functions of generalized log-normal model', *Mathematical and Computer Modelling* **42**, 939–946.

Hoff, P. (2009), *A First Course in Bayesian Statistical Methods*, Springer, New York.

Howard, B. & Dodson, G. (1961), 'High stress aging to failure of semiconductor devices', *Proc. Seventh National Symposium on Reliability and Quality Control* pp. 201–207.

Klein, J. & Moeschberger, M. (2003), *Survival Analysis: Techniques for Censored and Truncated Data*, Springer-Verlag, New York.

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, Wiley, New York.

Martín, J. & Pérez, C. (2009), 'Bayesian analysis of a generalized lognormal distribution', *Computational Statistics and Data Analysis* **53**, 1377–1387.

Nadarajah, S. (2005), 'A generalized normal distribution', *Journal of Applied Statistics* **37**(7), 685–694.

Padgett, W. & Johnson, M. (1983), 'Lower bounds on reliability in the lognormal distribution', *The Canadian Journal of Statistics - La Reveu Canadienne de Statistique* **11**(2), 137–147.

Portela, J. & Gómez-Villegas, M. (2004), 'Implementation of a robust Bayesian method', *Journal of Statistical Computation and Simulation* **74**(4), 235–248.

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Robert, C. (2001), *The Bayesian Choice*, second edn, Springer, New York.

Spiegelhalter, D., Best, N., Carlin, B. & van der Linde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of The Royal Statistical Society* **64**, 583–639.

Upadhyay, S. & Peshwani, M. (2001), 'Full posterior analysis of three parameter Lognormal distribution using Gibbs sampler', *Journal of Statistical Computation and Simulation* **71**(3), 215–230.

Upadhyay, S. & Peshwani, M. (2003), 'Choice between Weibull and lognormal models: A simulation based Bayesian study', *Communications in Statistics - Theory and Methods* **32**(2), 381–405.

Upadhyay, S. & Peshwani, M. (2008), 'Posterior analysis of lognormal regressions models using the Gibbs sampler', *Statistical Papers* **49**, 59–85.

Apéndice

A continuación se presenta el código usado en R Development Core Team (2010) para la implementación del muestreador de Gibbs.

```
#!---------------------------- LIBRERIAS REQUERIDAS ----------------------------
require(hbmem)
require(Runuran)
require(VGAM)
require(hdrcde)
#!-------------------------   FUNCIONES   ----------------------------------
#! fdp para la lognormal generalizada con los parametros mu sigma e s
flog=function(x,mu,sigma,s) s*exp(-abs((log(x)-mu)/sigma)^s)/(2*x*sigma*gamma(1/s))
#!-------  Generador de observaciones aleatoria logGN truncadas ----------------
rlogGN = function(n,MU,SIGMA,S,minimo=0,center=50) {
flog  = function(x,mu,sigma,s) s*exp(-abs((log(x)-mu)/sigma)^s)/(2*x*sigma*gamma(1/s))
logflog = function(x,mu,sigma,s) log(s/(2*x*sigma*gamma(1/s)))+-abs((log(x)-mu)/sigma)^s
a = function(x) flog(x,mu=MU,sigma=SIGMA,s=S)      # f(x)
b = function(x) logflog(x,mu=MU,sigma=SIGMA,s=S)   # log f(x)
generador = pinv.new(pdf=a,lb=minimo, ub=Inf, islog=FALSE, center=center)
ur(generador,n) }
#!--------------------------------------------------------------------------------
rexp.trunc=function(n=1,rate,mini)  {
x = urexp(n=n, rate=1/rate, lb=mini, ub=Inf)
x }
#!---------- Funciones para muestrear de las a posteriori, caso logNG ----------
rpost.mu=function(sigma,s,u,x) {       # u y x son vectores
l.inf = max(log(x)-sigma*u^(1/s))
l.sup = min(log(x)+sigma*u^(1/s))
if (l.inf<=l.sup) {
                   rmu=runif(n=1,min=l.inf,max=l.sup)
                  }
else rmu=runif(n=1,min=l.sup,max=l.inf)
rmu  }
#!--------------------------------------------------------------------------------
rpost.sigma=function(mu,s,u,x) {               # u y x son vectores
n=length(x)
minimo = max(abs(mu-log(x))/u^(1/s))
rpareto(n=1, shape=n, location=minimo) }
#!--------------------------------------------------------------------------------
rpost.u = function(x,mu,sigma,s) {
mini = (abs(log(x)-mu)/sigma)^s
n=length(x)
muestra=NULL
for (i in 1:n) muestra[i]=urexp(n=1,rate=1,lb=mini[i],ub=Inf)
muestra  }
#!--------------------------------------------------------------------------------
rpost.s = function(sigma,mu,u,x) {   # u y x son vectores
N = length(x)
limite = log(u)/log(abs(mu-log(x))/sigma)  # calcula los limites para s
s.1 = which(limite>=1)           # obtiene los índices donde el limite es >=1
mini =  min(c(1,s.1))
maxi =  max(s.1)
pdf.s = function(x,n=N) { x^(n-1)/(gamma(x))^n }
gen.s = tdr.new(pdf=pdf.s,n=N,lb=1,ub=10)
ur(gen.s,n=1) }
#!---------- Funciones para muestrear de las a posteriori, caso lognormal ------
rpost.mu.l=function(sigma,u,x) {       # u y x son vectores
l.inf = max(log(x)-sigma*(2*u)^(1/2))
l.sup = min(log(x)+sigma*(2*u)^(1/2))
if (l.inf<=l.sup) {
                   rmu=runif(n=1,min=l.inf,max=l.sup)
                  }
else rmu=runif(n=1,min=l.sup,max=l.inf)
rmu }
#!--------------------------------------------------------------------------------
rpost.sigma.l=function(mu,u,x) {               # u y x son vectores
n=length(x)
```

```
minimo = max(abs(mu-log(x))/(2*u)^(1/2))
rpareto(n=1, shape=n, location=minimo)  }
#!-------------------------------------------------------------------------------
rpost.u.l = function(x,mu,sigma) {
mini = (abs(log(x)-mu)/(sqrt(2)*sigma))^2
n=length(x)
muestra=NULL
for (i in 1:n) muestra[i]=urexp(n=1,rate=1,lb=mini[i],ub=Inf)
muestra  }
#!-------------------------------------------------------------------------------
#!------------ Análisis de los datos con censura considerando logNG  -----------
#!-------------------------------------------------------------------------------
dsc=c(14.07, 17.80, 19.43, 21.33, 24.60, 28.97, 29.63, 33.73, 37.60, 37.67,
       40.87, 52.40, 53.97, 60.57, 64.27, 65.43, 65.43)
dcc=rep(x=66,times=55) ; datos = c(dsc,dcc)
#! -----------------------------------------------------------------------------
#! Definiendo los valores iniciales de los parametros
n=length(dsc) ; n.censu=length(dcc)
mu0=3.5 ; sigma0=1 ; s0=20 ; x=datos
Mu=Sigma=S=NULL ; Mu[1]=mu0 ; Sigma[1]=sigma0 ; S[1]=s0
#!-------------------------------------------------------------------------------
for (i in 2:10000) {
x[18:72]  =rlogGN(n=n.censu,MU=Mu[i-1],SIGMA=Sigma[i-1],S=S[i-1],minimo=66,center=80)
u.temporal=rpost.u(x=x,mu=Mu[i-1],sigma=Sigma[i-1],s=S[i-1])
Mu[i]     =rpost.mu(sigma=Sigma[i-1],s=S[i-1],u=u.temporal,x=x)
Sigma[i]  =rpost.sigma(mu=Mu[i],s=S[i-1],u=u.temporal,x=x)
S[i]      =rpost.s(sigma=Sigma[i],mu=Mu[i],u=u.temporal,x=x) }
#!-------------------------------------------------------------------------------
parametros_logNG = data.frame( Mu=Mu[-(1:50)] , Sigma=Sigma[-(1:50)] , S=S[-(1:50)] )
#!-------------------------------------------------------------------------------
#!------------ Análisis de los datos con censura considerando logN  ------------
#!-------------------------------------------------------------------------------
#! Definiendo los valores iniciales de los parametros
Mu=Sigma=NULL ; Mu[1]=mu0 ; Sigma[1]=sigma0
#!-------------------------------------------------------------------------------
for (i in 2:10000) {
x[18:72]  =urlnorm(n=n.censu , meanlog=Mu[i-1] , sdlog=Sigma[i-1] , lb=66 , ub=Inf)
u.temporal=rpost.u.l(x=x,mu=Mu[i-1],sigma=Sigma[i-1])
Mu[i]     =rpost.mu.l(sigma=Sigma[i-1],u=u.temporal,x=x)
Sigma[i]  =rpost.sigma.l(mu=Mu[i],u=u.temporal,x=x)   }
#!-------------------------------------------------------------------------------
parametros_logN = data.frame( Mu=Mu[-(1:50)] , Sigma=Sigma[-(1:50)]  )
#!-------------------------------------------------------------------------------
Titulo = c('Mu','Sigma','S') ; par(mfrow=c(2,3))
for (i in 1:3) hdr.den(parametros_logNG[,i],xlab=Titulo[i],main='logNG',prob=90)
for (i in 1:2) hdr.den(parametros_logN[,i] ,xlab=Titulo[i],main='logN' ,prob=90)
#!-------------------------------------------------------------------------------
```

Revista Colombiana de Estadística
Junio 2011, volumen 34, no. 1, pp. 111 a 131

A Bayesian Analysis in the Presence of Covariates for Multivariate Survival Data: An example of Application

Análisis bayesiano en presencia de covariables para datos de sobrevivencia multivariados: un ejemplo de aplicación

Carlos Aparecido Santos[1,a], Jorge Alberto Achcar[2,b]

[1]Departamento de Estatística, Centro de Ciências Exatas, UEM - Universidade Estadual de Maringá, Maringá-PR, Brasil

[2]Departamento de Medicina Social, FMRP - Faculdade de Medicina de Ribeirão Preto, USP - Universidade de São Paulo, Ribeirão Preto-SP, Brasil

Abstract

In this paper, we introduce a Bayesian analysis for survival multivariate data in the presence of a covariate vector and censored observations. Different "frailties" or latent variables are considered to capture the correlation among the survival times for the same individual. We assume Weibull or generalized Gamma distributions considering right censored lifetime data. We develop the Bayesian analysis using Markov Chain Monte Carlo (MCMC) methods.

***Key words*:** Bayesian methods, Bivariate distribution, MCMC methods, Survival distribution, Weibull distribution.

Resumen

En este artículo, se introduce un análisis bayesiano para datos multivariados de sobrevivencia en presencia de un vector de covariables y observaciones censuradas. Diferentes "fragilidades" o variables latentes son consideradas para capturar la correlación entre los tiempos de sobrevivencia para un mismo individuo. Asumimos distribuciones Weibull o Gamma generalizadas considerando datos de tiempo de vida a derecha. Desarrollamos el análisis bayesiano usando métodos Markov Chain Monte Carlo (MCMC).

***Palabras clave*:** distribución bivariada, distribución de sobrevivencia, distribución Weibull, métodos bayesianos, métodos MCMC.

[a]Adjoint professor. E-mail: casantos@uem.br
[b]Professor. E-mail: jorge.achcar@pq.cnpq.br

1. Introduction

Different parametric regression models are introduced in the literature to analyse lifetime data in the presence of censored data (see for example, Lawless 1982). A popular semi-parametric regression model to analyse survival data was introduced by Cox (1972) assuming proportional hazards (see also, Cox & Oakes 1984). In these models, the survival times are independent, that is, the individuals are not related to each other.

In many practical situations, especially in medical studies, to have dependent survival times is common, when the individuals are related to each other (same family, repeated measurements in the same individual or two or more measuremenst in the same patient).

As an example, we could consider a survival data set introduced by McGilchrist & Aisbett (1991) related to kidney infection where the recurrence of infection of 38 kidney patients, using portable dialysis machines, is recorded. Infections may occur at the location of insertion of the catheter. The time recorded, called infection time, is either the survival time (in days) of the patient until an infection occurred and the catheter had to be removed, or the censored time, where the catheter was removed by others reasons. The catheter is reinserted after some time and the second infection time is again observed or censored (data set in Table 1).

Different survival multivariate models are introduced in the literature to analyse dependent lifetime data in the presence of a covariate vector and censored observations.

To capture the correlation among two or more survival times, we could consider the introduction of "frailties" or latent variables (see for example, Clayton & Cuzick (1985), Oakes (1986, 1989) and Shih & Louis (1992)), assuming proportional hazard models.

Clayton (1991) uses a Levy process (Kalbfleisch 1978) as a nonparametric Bayesian model for the baseline hazard, applied to continuous data, that is, data with no ties.

In this paper, we assume parametric regression models for dependent survival data in the presence of censored observations considering the special Weibull distribution, a popular lifetime model and the Generalized Gamma distribution, a supermodel that generalizes some common models used for lifetime data as the Weibull, the Gamma, and the log-normal distributions.

Different "frailties" are assumed to model the dependent structure of the data, under the Bayesian paradigm.

For a Bayesian analysis of the proposed models, we use MCMC (Markov Chain Monte Carlo) methods to obtain posterior summaries of interest (see for example, Gelfand & Smith (1990) and Chib & Greenberg (1995)).

The paper is organized as follows: in Section 2, we introduce a Weibull regression model for multivariate survival data; in Section 3, we introduce a Bayesian analysis; in Section 4, we consider the use of a generalized Gamma distribution for

multivariate survival data; in Section 5 we present an analysis for the recurrence times introduced in Table 1.

TABLE 1: Recurrence times of infections in 38 kidney patients.

Patient	First time	Second time	Censoring first time	Censoring second time	Sex
1	8	16	1	1	1
2	23	13	1	0	2
3	22	28	1	1	1
4	447	318	1	1	2
5	30	12	1	1	1
6	24	245	1	1	2
7	7	9	1	1	1
8	511	30	1	1	2
9	53	196	1	1	2
10	15	154	1	1	1
11	7	333	1	1	2
12	141	8	1	0	2
13	96	38	1	1	2
14	149	70	0	0	2
15	536	25	1	0	2
16	17	4	1	0	1
17	185	117	1	1	2
18	292	114	1	1	2
19	22	159	0	0	2
20	15	108	1	0	2
21	152	562	1	1	1
22	402	24	1	0	2
23	13	66	1	1	2
24	39	46	1	0	2
25	12	40	1	1	1
26	113	201	0	1	2
27	132	156	1	1	2
28	34	30	1	1	2
29	2	25	1	1	1
30	130	26	1	1	2
31	27	58	1	1	2
32	5	43	0	1	2
33	152	30	1	1	2
34	190	5	1	0	2
35	119	8	1	1	2
36	54	16	0	0	2
37	6	78	0	1	2
38	63	8	1	0	1

(Censoring (0); infection ocurrence (1); male (1); female (2))

2. A Weibull Regression Model for Multivariate Survival Data

Let T_{ji} be a random variable denoting the survival time of the i^{th} individual ($i = 1, 2, \ldots, n$) in the j^{th} repeated measurement for the same individual ($j = 1, 2, \ldots, k$) with a Weibull (1951) distribution with density,

$$f(t_{ji} \mid \nu_j, \lambda_j(i)) = \nu_j \lambda_j(i) t_{ji}^{\nu_j - 1} \exp\{-\lambda_j(i) t_{ji}^{\nu_j}\} \tag{1}$$

where $t_{ji} > 0$; $\nu_j > 0$ is the shape parameter and $\lambda_j(i)$ is the scale parameter.

To capture the correlation among the repeated measures $T_{1i}, T_{2i}, \ldots, T_{ki}$ for the same individual, we introduce a "frailty" or latent variable W_i, $i = 1, 2, \ldots, n$ with a normal distribution, that is,

$$W_i \overset{\text{iid}}{\sim} N(0, \sigma_w^2) \tag{2}$$

In the presence of a covariate vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ and the latent variable W_i, we assume the regression model in (1), given by

$$\lambda_j(i) = \exp\{w_i + \boldsymbol{\beta}_j' \mathbf{x}_i\} \tag{3}$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jp})$ is the vector of regression parameters, $j = 1, 2, \ldots, k$.

The hazard function is given by

$$h_j(t_{ji} \mid \mathbf{x}_i, w_i) = \nu_j t_{ji}^{\nu_j - 1} \exp\{w_i + \boldsymbol{\beta}_j' \mathbf{x}_i\} \tag{4}$$

The survival function for a given t_{ji} is

$$S(t_{ji} \mid \mathbf{x}_i, w_i) = \exp\{-t_{ji}^{\nu_j} e^{w_i + \boldsymbol{\beta}_j' \mathbf{x}_i}\} \tag{5}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$.

Let us denote the model defined by (1)-(5) as "model 1".

From (4), we observe that we can have constant, decreasing or increasing hazards, assuming, respectively, $\nu_j = 1$, $\nu_j < 1$ or $\nu_j > 1$.

The conditional mean and variance for T_{ji} given $\mathbf{x}_i$ and w_i, are given, respectively, by

$$E(T_{ji} \mid \mathbf{x}_i, w_i) = \frac{\Gamma(1 + 1/\nu_j)}{\exp\left\{\frac{1}{\nu_j}\left(w_i + \boldsymbol{\beta}_j' \mathbf{x}_i\right)\right\}}$$

and

$$\text{Var}(T_{ji} \mid \mathbf{x}_i, w_i) = \frac{1}{\exp\left\{\frac{2}{\nu_j}\left(w_i + \boldsymbol{\beta}_j' \mathbf{x}_i\right)\right\}} \left\{\Gamma\left(1 + \frac{2}{\nu_j}\right) - \Gamma^2\left(1 + \frac{1}{\nu_j}\right)\right\}$$

for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$.

The unconditional mean for T_{ji} is obtained from the result, $E(T_{ji} \mid \mathbf{x}_i) = E[E(T_{ji} \mid \mathbf{x}_i, w_i)]$, that is,

$$E(T_{ji} \mid \mathbf{x}_i) = \frac{\Gamma(1 + 1/\nu_j)}{\exp\left(\frac{\boldsymbol{\beta}_j' \mathbf{x}_i}{\nu_j}\right)} E\left\{e^{-W_i/\nu_j}\right\}$$

Observe that, since $W_i \sim N(0, \sigma_w^2)$, we have

$$g(W_i) = e^{-W_i/\nu_j} \overset{\text{a}}{\sim} N\left\{g(0); [g'(0)]^2 \sigma_w^2\right\}$$

("delta method"), that is,

$$e^{-W_i/\nu_j} \overset{\text{a}}{\sim} N\left[1; \frac{\sigma_w^2}{\nu_j^2}\right]$$

Thus, the unconditional mean for T_{ji} given $\mathbf{x}_i$ is,

$$E(T_{ji} \mid \mathbf{x}_i) = \frac{\Gamma(1+1/\nu_j)}{\exp\left(\frac{\boldsymbol{\beta}_j'\mathbf{x}_i}{\nu_j}\right)} \tag{6}$$

for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$.

The unconditional variance for T_{ji} is obtained from $\text{Var}(T_{ji} \mid \mathbf{x}_i) = \text{Var}\{E(T_{ji} \mid \mathbf{x}_i, \mathbf{w}_i)\} + E\{\text{Var}(T_{ji} \mid \mathbf{x}_i, w_i)\}$, that is,

$$\begin{aligned}\text{Var}(T_{ji} \mid \mathbf{x}_i) = {} & \frac{\Gamma^2(1+1/\nu_j)}{\exp\left(\frac{2\beta_j'\mathbf{x_i}}{\nu_j}\right)}\text{Var}(e^{-W_i/\nu_j}) \\ & + \frac{\left[\Gamma(1+2/\nu_j) - \Gamma^2(1+1/\nu_j)\right]}{\exp\left(\frac{2\boldsymbol{\beta}_j'\mathbf{x}_i}{\nu_j}\right)}E(e^{-2W_i/\nu_j})\end{aligned}$$

Also using the "delta method", we observe that $g(W_i) = e^{-2W_i/\nu_j} \overset{\text{a}}{\sim} N\left[1; \frac{4\sigma_w^2}{\nu_j^2}\right]$, that is,

$$\begin{aligned}\text{Var}(T_{ji} \mid \mathbf{x}_i) = {} & \frac{\sigma_w^2\Gamma^2(1+1/\nu_j)}{\nu_j^2\exp\left(\frac{2\boldsymbol{\beta}_j'\mathbf{x}_i}{\nu_j}\right)} \\ & + \frac{1}{\exp\left(\frac{2\boldsymbol{\beta}_j'\mathbf{x}_i}{\nu_j}\right)}\left\{\Gamma(1+2/\nu_j) - \Gamma^2(1+1/\nu_j)\right\}\end{aligned} \tag{7}$$

Observe that not considering the presence of the "frailty" W_i, the variance for T_{ji}, given, $\mathbf{x}_i$ is

$$\text{Var}(T_{ji} \mid \mathbf{x}_i) = \frac{\Gamma(1+2/\nu_j) - \Gamma^2(1+1/\nu_j)}{\exp\left(\frac{2\boldsymbol{\beta}_j'\mathbf{x}_i}{\nu_j}\right)} \tag{8}$$

From (7) and (8), we observe that the extra-Weibull variability in the presence of the "frailty" W_i with normal distribution (2) is given by

$$\frac{\sigma_w^2\Gamma^2(1+1/\nu_j)}{\nu_j^2\exp\left(\frac{2\boldsymbol{\beta}_j'\mathbf{x}_i}{\nu_j}\right)}$$

for $j = 1, 2, \ldots, k$; $i = 1, 2, \ldots, n$.

A different model could be considered replacing (3) by

$$\lambda_j(i) = w_i e^{\boldsymbol{\beta}'\mathbf{x}_i} \tag{9}$$

with

$$W_i \overset{\text{iid}}{\sim} \text{Gamma}(\phi^{-1}, \phi^{-1}) \tag{10}$$

From (9), observe that $E(W_i) = 1$ and $\text{Var}(W_i) = 0$.

Let us assume the model defined by (1) and (9) as "model 2".

From "model 2", the conditional mean and variance for T_{ji} given $\mathbf{x}_i$ and w_i, are given, respectively, by,

$$E(T_{ji} \mid \mathbf{x}_i, w_i) = \frac{\Gamma(1 + 1/\nu_j)}{w_i^{1/\nu_j} e^{\boldsymbol{\beta}_j' \mathbf{x}_i/\nu_j}} \tag{11}$$

and

$$\text{Var}(T_{ji} \mid \mathbf{x}_i, w_i) = \frac{\Gamma(1 + 2/\nu_j) - \Gamma^2(1 + 1/\nu_j)}{w_i^{2/\nu_j} e^{2\boldsymbol{\beta}_j' \mathbf{x}_i/\nu_j}}$$

for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$.

Following the same arguments used in the determination of the unconditional mean and variance for T_{ji} assuming "model 1", and observing that the "frailty" W_i has a Gamma (ϕ^{-1}, ϕ^{-1}) distribution, the unconditional mean for T_{ji} assuming "model 2" is (see section 6) given by

$$E(T_{ji} \mid \mathbf{x}_i) = \frac{\Gamma(1 + 1/\nu_j)(\phi^{-1})^{1/\nu_j} \Gamma(\phi^{-1} - \nu_j^{-1})}{\exp\{\boldsymbol{\beta}_j' \mathbf{x}_i/\nu_j\} \Gamma(\phi^{-1})}$$

for $\phi^{-1} > \nu_j^{-1}$, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$.

The unconditional variance for T_{ji} (see Section 7) is given by,

$$\text{Var}(T_{ji} \mid \mathbf{x}_i) = \frac{(\phi^{-1})^{2/\nu_j}}{\exp(2\boldsymbol{\beta}_j' \mathbf{x}_i/\nu_j)} \times \left\{ \frac{\Gamma(1 + 2/\nu_j)\Gamma(\phi^{-1} - 2/\nu_j)}{\Gamma(\phi^{-1})} - \left[\frac{\Gamma(1 + 1/\nu_j)\Gamma(\phi^{-1} - 1/\nu_j)}{\Gamma(\phi^{-1})} \right]^2 \right\}$$

for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$.

Generalization of "model 1" and "model 2" could be considered assuming that the covariate vector $\mathbf{x}_i$ also affect the shape parameter ν_j, that is, assuming the regression model $\nu_j(i) = \exp\{\boldsymbol{\alpha}_j' \mathbf{x}_i\}$, where $\boldsymbol{\alpha}_\mathbf{j} = (\alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jp})$ is another vector of regression parameters, $j = 1, 2, \ldots, k$. Let us denote these models as "model 3" and "model 4", respectively.

3. A Bayesian Analysis

Assuming lifetime in the presence of censored observations and a covariate vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$, let us define an indicator variable for censoring or not censoring observations, by

$$\delta_{ji} = \begin{cases} 1 & \text{for observed lifetime} \\ 0 & \text{for censored observation} \end{cases} \tag{12}$$

Assuming "model 1" defined by (1), (2) and (3), the likelihood function is given by

$$f(\mathbf{t} \mid \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\nu}, \mathbf{w}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left[f(t_{ji} \mid \mathbf{x}_i, w_i)\right]^{\delta_{ji}} \left[S(t_{ji} \mid \mathbf{x}_i, w_i)\right]^{1-\delta_{ji}}$$

where $S(t_{ji} \mid \mathbf{x}_i, w_i)$ is the survival function defined by (5), $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)$, $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jp})$, $j = 1, 2, \ldots, k$; $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_k)$, $\mathbf{w} = (w_1, w_2, \ldots, w_n)$, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, $i = 1, 2, \ldots, n$.

That is,

$$f(\mathbf{t} \mid \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\nu}, \mathbf{w}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \nu_j^{\delta_{ji}} t_{ji}^{\delta_{ji}(\nu_j - 1)} \exp\left\{\delta_{ji}[w_i + \boldsymbol{\beta}_j' \mathbf{x}_i]\right\} \exp\{-t_{ji}^{\nu_j} e^{w_i + \boldsymbol{\beta}_j' \mathbf{x}_i}\} \tag{13}$$

Assuming "model 2", the likelihood function is (from (12) and (9)) given by

$$f(\mathbf{t} \mid \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\nu}, \mathbf{w}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \nu_j^{\delta_{ji}} t_{ji}^{\delta_{ji}(\nu_j - 1)} w_i^{\delta_{ji}} e^{\delta_{ji} \boldsymbol{\beta}' \mathbf{x}_i} \exp\{-t_{ji}^{\nu_j} w_i e^{\boldsymbol{\beta}' \mathbf{x}_i}\}$$

For a hierarchical Bayesian analysis of "model 1", we assume in the first stage, the following prior distributions for the parameters:

$$\begin{aligned} \nu_j &\sim \text{Gamma}(a_j, b_j) \\ \beta_{jl} &\sim \text{N}(0; c_{jl}^2) \end{aligned} \tag{14}$$

where $j = 1, 2, \ldots, k$; $l = 1, 2, \ldots, p$; a_j, b_j, c_{jl} are known hyperparameters and $\text{Gamma}(a, b)$ denotes a gamma distribution with mean a/b and variance a/b^2.

In the second stage of the hierarchical Bayesian analysis, we assume a gamma prior distribution for σ_w^2, that is,

$$\sigma_w^2 \sim \text{Gamma}(d, e) \tag{15}$$

where d and e are known hyperparameters.

We further assume independence among the parameters.

Combining (2), (13), (14) and (15), we get the joint posterior distribution for $\mathbf{w}, \boldsymbol{\nu}, \boldsymbol{\beta}$ and σ_w^2, given by

$$\begin{aligned} \pi(\boldsymbol{\nu}, \boldsymbol{\beta}, \mathbf{w}, \sigma_w^2 \mid \mathbf{x}, \mathbf{t}) \propto & \left\{\prod_{i=1}^{n} \exp\left(-\frac{w_i^2}{2\sigma_w^2}\right)\right\} \left\{\prod_{j=1}^{k} \prod_{l=1}^{p} \exp\left(-\frac{\beta_{jl}^2}{2c_j^2}\right)\right\} \\ & \times (\sigma_w^2)^{d-1} \exp(-e\sigma_w^2) \left(\prod_{j=1}^{k} \nu_j^{a_j - 1} e^{-b_j \nu_j}\right) \\ & \times \prod_{i=1}^{n} \prod_{j=1}^{k} \nu_j^{\delta_{ji}} t_{ji}^{\delta_{ji}(\nu_j - 1)} \exp\{\delta_{ji}(w_i + \boldsymbol{\beta}_j' \mathbf{x}_i)\} \\ & \times \exp\{-t_{ji}^{\nu_j} e^{w_i + \boldsymbol{\beta}_j' \mathbf{x}_i}\} \end{aligned} \tag{16}$$

To get the posterior summaries of interest, we simulate samples of the joint posterior distribution (16) using MCMC methods as the popular Gibbs sampling algorithm (see for example, Gelfand & Smith 1990) or the Metropolis-Hastings algorithm (see for example, Chib & Greenberg 1995).

A great simplification in the simulation of the samples for the joint posterior distribution is given by the WinBugs software (Spiegelhalter, Thomas, Best & Lunn 2003), which requires only the specification of the joint distribution for the data and the prior distributions for the parameters.

Assuming "model 2", we consider the same priors (14) for ν_j and β_{jl}, and an uniform prior distribution for ϕ, that is,

$$\phi \sim \mathrm{U}(0, f) \tag{17}$$

where $U(a, b)$ denotes an uniform distribution in the interval (a, b) and f is a known hyperparameter.

The joint posterior distribution for $\mathbf{w}, \boldsymbol{\nu}, \boldsymbol{\beta}$ and ϕ is given by

$$\begin{aligned}\pi(\boldsymbol{\beta}, \boldsymbol{\beta}, \mathbf{w}, \phi \mid \mathbf{x}, \mathbf{t}) \propto & \left\{\prod_{i=1}^{n} w_i^{\phi^{-1}-1} e^{-\phi^{-1} w_i}\right\} \phi^{f-1} e^{-g\phi} \\ & \times \left\{\prod_{j=1}^{k}\prod_{l=1}^{p} \exp\left(-\frac{\beta_{jl}^2}{2c_j^2}\right)\right\} \left\{\prod_{j=1}^{k} \nu_j^{a_j-1} e^{-b_j\nu_j}\right\} \\ & \times \prod_{i=1}^{n}\prod_{j=1}^{k} \nu_j^{\delta_{ji}} t_{ji}^{\delta_{ji}(\nu_j-1)} w_i^{\delta_{ji}} \exp\{\delta_{ji}\boldsymbol{\beta}_j'\mathbf{x}_i\} \exp\{-t_{ji}^{\nu_j} w_i e^{\boldsymbol{\beta}_j'\mathbf{x}_i}\}\end{aligned}$$

4. Use of a Generalized Gamma Distribution for Multivariate Survival Data

In this section, we assume that the lifetime T_{ji} has a generalized gamma distribution with density

$$f(t_{ji} \mid \nu_j, \mu_j(i), \theta_j) = \frac{\theta_j}{\Gamma(\nu_j)} [\mu_j(i)]^{\theta_j\nu_j} t_{ji}^{\theta_j\nu_j - 1} \exp\left\{-[\mu_j(i) t_{ji}]^{\theta_j}\right\} \tag{18}$$

where $t_{ji} > 0$, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$; $\theta_j > 0$; $\nu_j > 0$ and $\mu_j(i) > 0$.

The generalized gamma distribution is a fairly flexible family of distributions that includes as special cases the exponential ($\theta_j = \nu_j = 1$), Weibull ($\nu_j = 1$) and gamma ($\theta_j = 1$) distributions. The log-normal distribution also arises as a limiting form of (18), that is, the generalized gamma model includes as special cases all of the most commonly used lifetime distributions. This makes it useful for discriminating among these other models.

The survival function for a given value of T_{ji} is given by

$$S(t_{ji} \mid \nu_j, \mu_j(i), \theta_j) = P(T_{ji} > t_{ji}) = \frac{\theta_j}{\Gamma(\nu_j)} [\mu_j(i)]^{\theta_j\nu_j} \int_{t_{ji}}^{\infty} z^{\theta_j\nu_j - 1} e^{-[\mu_j(i) z]^{\theta_j}} dz$$

To capture the correlation among the repeated measures $T_{1i}, T_{2i}, \ldots, T_{ki}$ for the same individual, we introduce "frailties" W_i, $i = 1, 2, \ldots, n$. In the presence of a covariate vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$, we assume the regression models

$$\mu_j(i) = \exp\{w_i + \boldsymbol{\beta}_j' \mathbf{x}_i\}$$

where W_i has a normal distribution (2) $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$, denoted as "model 5", or,

$$\mu_j(i) = w_i e^{\boldsymbol{\beta}_j' \mathbf{x}_i}$$

where W_i has a gamma distribution (10) denoted as "model 6".

Assuming "model 5", we consider the following prior distributions in a first stage of a hierarchical Bayesian analysis:

$$\begin{aligned} \nu_j &\sim \text{Gamma}(a_j, b_j); \\ \theta_j &\sim \text{Gamma}(c_j, d_j); \\ \beta_{jl} &\sim \text{N}(0, e_{jl}^2); \end{aligned} \tag{19}$$

where $j = 1, 2, \ldots, k$; $l = 1, 2, \ldots, p$; a_j, b_j, c_j, d_j and e_{jl} are known hyperparameters. In a second stage of the hierarchical Bayesian analysis, let us assume a gamma prior (15) for σ_w^2.

Assuming "model 6", we consider the same priors (19) for ν_j, θ_j and β_{jl}, and a gamma prior (17) for ϕ.

To develop a Bayesian analysis for the generalized gamma distribution of multivariate survival data in the presence of covariates and censored observations, we need informative prior distributions to get convergence for the Gibbs sampling algorithm. Observe that using the generalized gamma distribution usually we have great difficulties to get classical inferences of interest (see for example, Stacy & Mihram (1965), Parr & Webster (1965) and Hager & Bain (1970)).

Samples of the joint posterior distribution for the parameters of "model 3" or "model 4" are obtained using MCMC methods.

5. Model Selection

Different model selection methods could be used to choose the most adequate model to analyse multivariate survival data in the presence of covariates and censored observartions. As a special situation, we could use the generalized gamma distribution (see Section 4). In this way, if credible intervals for the parameters ν_j, $j = 1, 2, \ldots, k$ include the value one, this is an indication that the use of Weibull distribution in the presence of "frailties" gives good fit for the survival data.

We also could consider the Deviance Information Criterion (DIC), which is a criterion specifically useful for selection models under the Bayesian approach where samples of the posterior distribution for the parameters of the model are obtained using MCMC methods.

The deviance is defined by

$$D(\theta) = -2\log L(\theta) + c$$

where θ is a vector of unknown parameters of the model, $L(\theta)$ is the likelihood function of the model and c is a constant that does not need to be known when the comparison between models is made.

The DIC criterion defined by Spiegelhalter, Best, Carlin & Van der Linde (2002) is given by,

$$DIC = D(\widehat{\theta}) + 2n_D$$

where $D(\widehat{\theta})$ is the deviance evaluated at the posterior mean $\widehat{\theta} = E(\theta \mid data)$ and n_D is the effective number of parameters of the model given by $n_D = \overline{D} - D(\widehat{\theta})$, where $\overline{D} = E(D(\theta) \mid data)$ is the posterior deviance measuring the quality of the data fit for the model. Smaller values of DIC indicates better models. Note that these values could be negative.

6. Some Results About Gamma Distribution

Let W_i be a random variable with a Gamma(a, b) distribution, with density

$$f(w_i \mid a, b) = \frac{b^a}{\Gamma(a)} w_i^{a-1} e^{-bw_i} \tag{20}$$

where $w_i > 0$, $a > 0$, $b > 0$, $i = 1, 2, \ldots, n$.

From (20) we observe that

$$\int_0^\infty w_i^{a-1} e^{-bw_i} dw_i = \frac{\Gamma(a)}{b^a} \tag{21}$$

Also observe that

$$E(w_i^{-k}) = \int_0^\infty w_i^{-k} \frac{b^a}{\Gamma(a)} w_i^{a-1} e^{-bw_i} dw_i = \frac{b^a}{\Gamma(a)} \int_0^\infty w_i^{(a-k)-1} e^{-bw_i} dw_i$$

From (21), we have:

$$E(w_i^{-k}) = \frac{b^k \Gamma(a-k)}{\Gamma(a)}$$

for $a > k$.

Assuming $a = b = \phi^{-1}$, we have:

i) With $k = 1/\nu_j$,

$$E(w_i^{-1/\nu_j}) = \frac{(\phi^{-1})^{\nu_j^{-1}} \Gamma(\phi^{-1} - \nu_j^{-1})}{\Gamma(\phi^{-1})} \tag{22}$$

for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$;

ii) With $k = 2/\nu_j$,

$$E(w_i^{-2/\nu_j}) = \frac{(\phi^{-1})^{2/\nu_j}\Gamma(\phi^{-1} - 2/\nu_j)}{\Gamma(\phi^{-1})} \tag{23}$$

for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$.

From (10), we have:

$$E(T_{ji} \mid \mathbf{x}_i) = E[E(T_{ji} \mid \mathbf{x}_i, w_i)] = \frac{\Gamma(1 + 1/\nu_j)}{\exp\{\boldsymbol{\beta}_j'\mathbf{x}_i/\nu_j\}} E(W_i^{-1/\nu_j})$$

Thus, from (22), we find the unconditional mean for T_{ji}, given by

$$E(T_{ji} \mid \mathbf{x}_i) = \frac{\Gamma(1 + 1/\nu_j)(\phi^{-1})^{1/\nu_j}\Gamma(\phi^{-1} - 1/\nu_j)}{\Gamma(\phi^{-1})\exp\{\boldsymbol{\beta}_j'\mathbf{x}_i/\nu_j\}} \tag{24}$$

From (10) and (11) and using the result $\mathrm{Var}(T_{ji} \mid \mathbf{x}_i) = E\{\mathrm{Var}(T_{ji} \mid \mathbf{x}_i, w_i)\} + \mathrm{Var}\{E(T_{ji} \mid \mathbf{x}_i, w_i)\}$, we have:

$$\mathrm{Var}(T_{ji} \mid \mathbf{x}_i) = \frac{[\Gamma(1 + 2/\nu_j) - \Gamma^2(1 + 1/\nu_j)]}{\exp\{2\boldsymbol{\beta}_j'\mathbf{x}_i\}} E(W_i^{-2/\nu_j}) + \tag{25}$$

$$+ \frac{\Gamma^2(1 + 1/\nu_j)}{\exp\{2\boldsymbol{\beta}_j'\mathbf{x}_i/\nu_j\}} \mathrm{Var}(W_i^{-1/\nu_j}) \tag{26}$$

Observe that $\mathrm{Var}(W_i^{-1/\nu_j}) = E(W_i^{-2/\nu_j}) - [E(W_i^{-1/\nu_j})]^2$, that is, from (22) and (23),

$$\mathrm{Var}(W_i^{-1/\nu_j}) = \frac{(\phi^{-1})^{2/\nu_j}\Gamma(\phi^{-1} - 2/\nu_j)}{\Gamma(\phi^{-1})} - \frac{(\phi^{-1})^{2/\nu_j}\Gamma^2(\phi^{-1} - 1/\nu_j)}{\Gamma^2(\phi^{-1})}$$

That is, from (23) and (24), we find the unconditional variance for T_{ji} given by

$$\mathrm{Var}(T_{ji} \mid \mathbf{x}_i) = \frac{(\phi^{-1})^{2/\nu_j}}{\exp(2\boldsymbol{\beta}_j'\mathbf{x}_i/\nu_j)} \times$$
$$\times \left\{ \frac{\Gamma(1 + 2/\nu_j)\Gamma(\phi^{-1} - 2/\nu_j)}{\Gamma(\phi^{-1})} - \left[\frac{\Gamma(1 + 1/\nu_j)\Gamma(\phi^{-1} - 1/\nu_j)}{\Gamma(\phi^{-1})} \right]^2 \right\}$$

7. Analysis of the Recurrence Times of Infections for Kidney Patients

To analyse the recurrence times of infections (see Table 1), let us assume a Weibull regression model ("model 1") in the presence of a "frailty" W_i with a normal distribution (2).

In this case, we have only a covariate x_i (sex; $x_i = 1$ for male; $x_i = 0$ for female) and $k = 2$ recurrence times.

From (3), we have the regression model

$$\lambda_j(i) = \exp\{\beta_{1j} + \beta_{2j}x_i + w_i\}$$

$i = 1, 2, \ldots, 38$; $j = 1, 2$.

For a Bayesian analysis of "model 1", let us assume the prior distributions (14) and (15) with $a_1 = b_1 = a_2 = b_2 = 1$; $c_{11} = c_{21} = c_{12} = c_{22} = 10$ and $d = e = 0.1$.

Using the WinBugs software (Spiegelhalter et al. 2003), we discarded the first 5000 simulated Gibbs samples ("burn-in-sample") to eliminate the effect of the initial values for the parameters of the model. Choosing every 20^{th} simulated Gibbs sample, we obtained a final sample of size 2000 to get the posterior summaries of interest (see Table 2). Convergence of the Gibbs sampling algorithm was monitored using existing methods as time series plots for the simulated samples and Gelman & Rubin (1992) indexes. This simulation procedure also was employed for the other models considered in this section.

In Table 2, we also have the Monte Carlo estimate for the posterior mean of the median survival time in each recurrence time. Observe that the median survival time not including the covariate x_i is given by $\text{Med}_j = [(\log 2)e^{-\beta_{1j}}]^{1/\nu_j}$, $j = 1, 2$.

Table 2: Posterior summaries ("model 1").

Parameter	Mean	S.D.	95% Credible Interval
β_{21}	1.9820	0.5694	(0.8885; 3.1550)
β_{22}	0.7594	0.5570	(−0.3279; 1.8510)
β_{11}	−5.5490	0.8571	(−7.3400; −3.9880)
β_{12}	−5.6110	0.8805	(−7.4620; −3.9940)
med 1	120.40	31.780	(69.350; 194.50)
med 2	106.60	29.520	(62.720; 173.10)
ν_1	1.0880	0.1610	(0.8012; 1.4270)
ν_2	1.1310	0.1742	(0.8113; 1.5100)
$1/\sigma_w^2$	3.1350	3.1060	(0.7188; 11.270)

In Figure 1, we have the time series plots for the simulated Gibbs samples under "model 1". From these plots, we observe convergence of the algorithm in all cases.

Assuming "model 2", that is, defined by the Weibull density (1) where $\lambda_j(i)$ is given by (9), we have

$$\lambda_j(i) = w_i \exp\{\beta_{1j} + \beta_{2j}x_i\}$$

$i = 1, 2, \ldots, 38$; $j = 1, 2$. Let us assume the prior distributions (14) and (17) with $a_1 = b_1 = a_2 = b_2 = 1$; $c_{11} = c_{21} = c_{12} = c_{22} = 10$ and $f = 5$.

Following the same simulation steps considered in the generation of samples for the joint posterior distribution of the parameters of "model 1", we have, in Table 3, the posterior summaries of interest assuming the final Gibbs sample of size 2000.

In Figure 2, we have plots for the simulated Gibbs samples under "model 2".

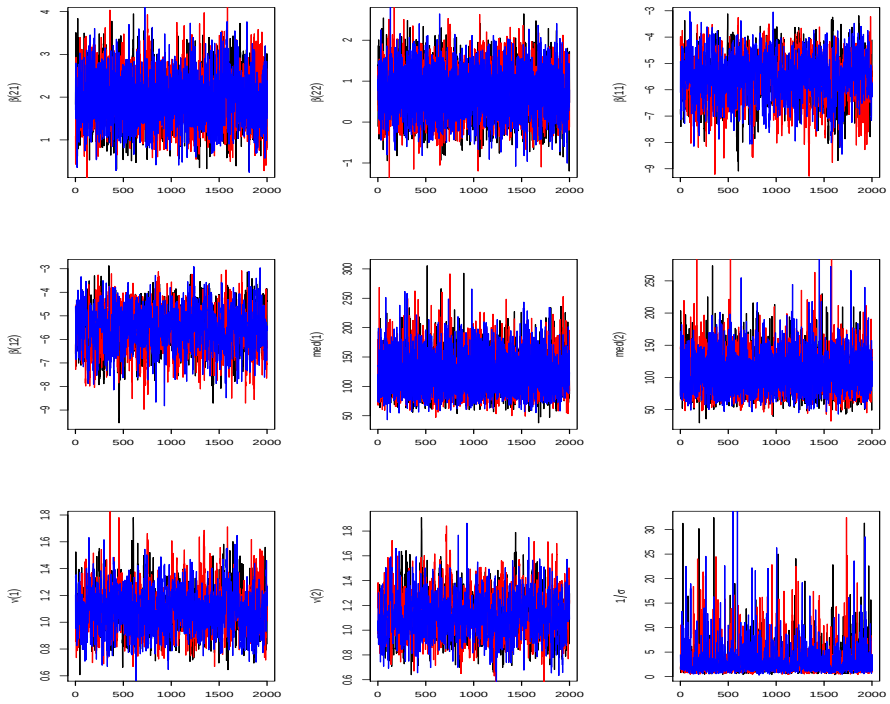

FIGURE 1: Simulated Gibbs samples ("model 1").

TABLE 3: Posterior summaries ("model 2").

Parameter	Mean	S.D.	95% Credible Interval
β_{21}	2.2370	0.6192	$(1.0770; 3.5080)$
β_{22}	1.0180	0.6234	$(-0.1979; 2.2800)$
β_{11}	-5.6340	0.8506	$(-7.3790; -4.1350)$
β_{12}	-5.6690	0.8794	$(-7.5240; -4.0420)$
med 1	98.590	26.260	$(54.860; 157.90)$
med 2	87.340	23.140	$(50.320; 141.90)$
ν_1	1.1560	0.1719	$(0.8508; 1.5130)$
ν_2	1.1950	0.1863	$(0.8604; 1.5900)$
$1/\phi$	2.4670	2.6360	$(0.7685; 7.7670)$

From the results of Tables 2 and 3, we observe similar results considering "model 1" and "model 2". We observe that in both models, we have a significative effect of sex for the first recurrence time, since zero is not included in the 95% credible interval for β_{21}; in the same way, we observe that sex does not have a significative effect in the second recurrence time, since zero is included in the 95% credible interval for β_{22}.

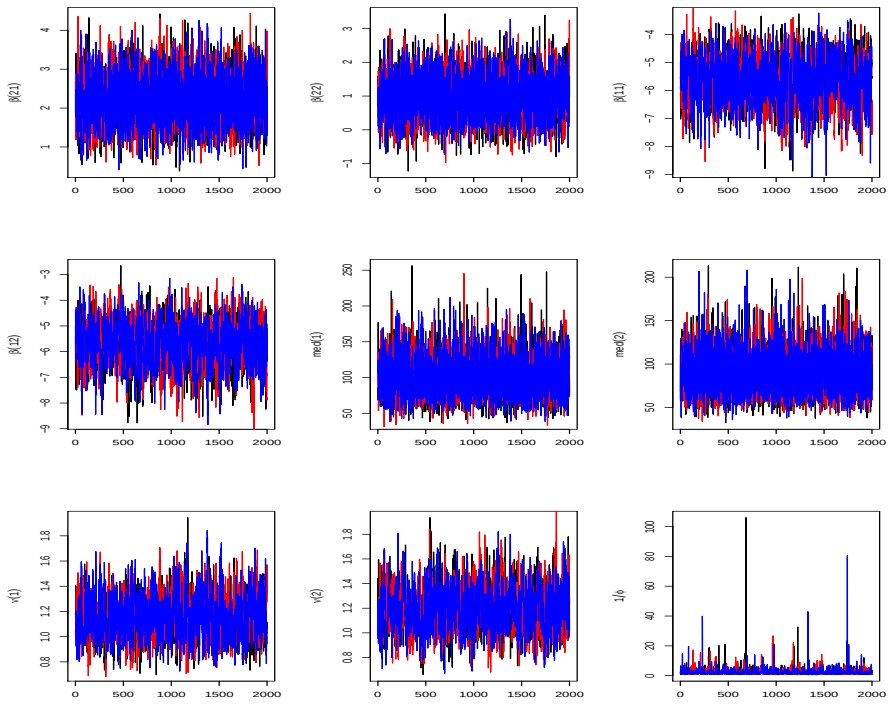

FIGURE 2: Simulated Gibbs samples ("model 2").

A Monte Carlo estimate for DIC (see Section 5), based on the 2000 simulated Gibbs samples considering "model 1", is given by DIC = 667.07. Considering "model 2", we have DIC = 662.86. That is, since we have a small decreasing in the value of DIC assuming "model 2", we could conclude that "model 2" is better fitted by the recurrence times of infection for kidney patients. To point out that other discrimination methods also could be used to decide by the best model is important.

A further modification could be assumed for "model 1" and "model 2", introducing the effect of covariate sex (x_i) in the shape parameter ν_j, $j = 1, 2$.

In this way, we assume for "model 1" and "model 2" the regression model for the shape parameter given by

$$\nu_j(i) = \exp\{\alpha_{1j} + \alpha_{2j} x_i\}$$

$i = 1, 2, \ldots, 38$; $j = 1, 2$.

Let us denote these models as "model 3" and "model 4".

For "model 3" and "model 4", we assume informative normal prior distributions for β_{1j} and β_{2j} considering means close to the obtained posterior means for β_{1j} and β_{2j} assuming "model 1" and "model 2", respectively. We also assume normal priors for α_{1j} and α_{2j}, $j = 1, 2$, considering small variances.

In Table 4, we have the posterior summaries obtained from 2000 simulated Gibbs samples for the joint posterior distributions of interest.

Table 4: Posterior summaries ("model 3" and "model 4").

Model	Parameter	Mean	S.D.	95% Credible Interval
"model 3"	β_{21}	2.0050	0.2932	$(1.4410; 2.5880)$
DIC = 660.87	β_{22}	1.9470	0.3007	$(1.3610; 2.5370)$
	β_{11}	-5.9650	0.2970	$(-6.5460; -5.3840)$
	β_{12}	-6.0650	0.2937	$(-6.6580; -5.4970)$
	α_{21}	0.0603	0.1186	$(-0.1818; 0.2850)$
	α_{22}	-0.1855	0.1180	$(-0.4252; 0.0356)$
	α_{11}	0.1395	0.0662	$(0.0064; 0.2670)$
	α_{12}	0.1899	0.0697	$(0.0452; 0.3159)$
	$1/\sigma_w^2$	1.7700	0.8131	$(0.6983; 3.8310)$
"model 4"	β_{21}	2.0170	0.3044	$(1.4210; 2.5960)$
DIC = 658.06	β_{22}	1.9510	0.3013	$(1.3610; 2.5440)$
	β_{11}	-5.9410	0.2938	$(-6.5280; -5.3700)$
	β_{12}	-6.0510	0.2921	$(-6.6460; -5.4680)$
	α_{21}	0.1142	0.1242	$(-0.1362; 0.3476)$
	α_{22}	-0.1263	0.1281	$(-0.3848; 0.1203)$
	α_{11}	0.1772	0.0696	$(0.0430; 0.3143)$
	α_{12}	0.2267	0.0691	$(0.0838; 0.3593)$
	$1/\phi$	2.1050	1.9220	$(0.7696; 5.5800)$

In Figures 3 and 4, we have plots for the simulated Gibbs samples considering "model 3" and "model 4", respectively.

From the results in Table 4, we observe that "model 3" and "model 4" give similar inferences. We observe that the covariate x_i (sex) does not have a significative effect on the shape parameter of the Weibull distribution for the recurrences times, since zero is included in the 95% credible intervals for α_{21} and α_{22} assuming both models. We also observe that "model 4" gives a smaller value for DIC (658.06) when compared to models 1, 2 and 3.

Another way, to check if the Weibull regression model is well fitted by the data, is to assume a generalized gamma distribution.

Considering "model 5" with a generalized gamma density (18) with regression model,

$$\mu_j(i) = \exp\{\beta_{1j} + \beta_{2j} x_i + w_i\}$$

where the "frailty" W_i has a normal distribution (2), let us assume the priors (19) and (15) for the parameters of the model with hyperparameter values $a_1 = a_2 = b_1 = b_2 = c_1 = c_2 = d_1 = d_2 = 1$ and normal distributions for β_{11}, β_{12}, β_{21} and β_{22} with variance equals to one.

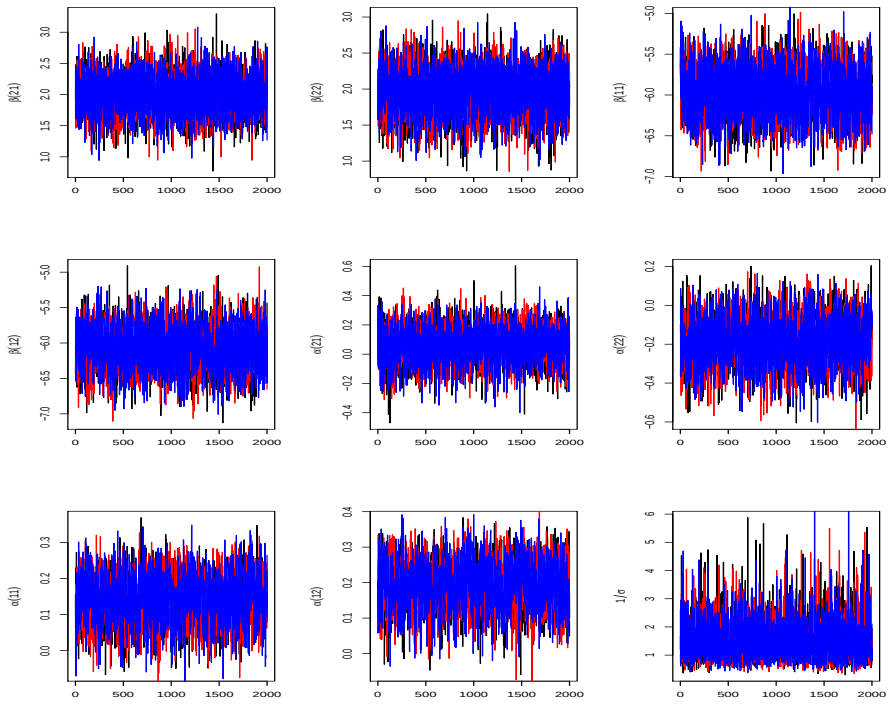

FIGURE 3: Simulated Gibbs samples ("model 3").

Assuming "model 6", with a generalized gamma density (18), and a regression model,

$$\mu_j(i) = w_i \exp\{\beta_{1j} + \beta_{2j} x_i\}$$

where the "frailty" W_i has a gamma distribution (10), let us assume the same prior distributions considered for "model 5", in the first stage of the hierarchical Bayesian analysis and a Gamma$(1, 1)$ prior for the parameter ϕ.

In Table 5, we have the posterior summaries of interest considering "model 5" and "model 6".

From the results of Table 5, we observe that assuming "model 5" or "model 6", the 95% credible intervals for ν_1 and ν_2 include the value one, that is, an indicator that the Weibull models in the presence of "frailties" give good fit for the multivariate survival data introduced in table 1.

8. Discussion and Concluding Remarks

Longitudinal survival data is common in many studies as in medicine or in engineering. Usually, we have repeated measures for the same patient or unit. In these studies, the presence of covariates and censoring data is common . The use of

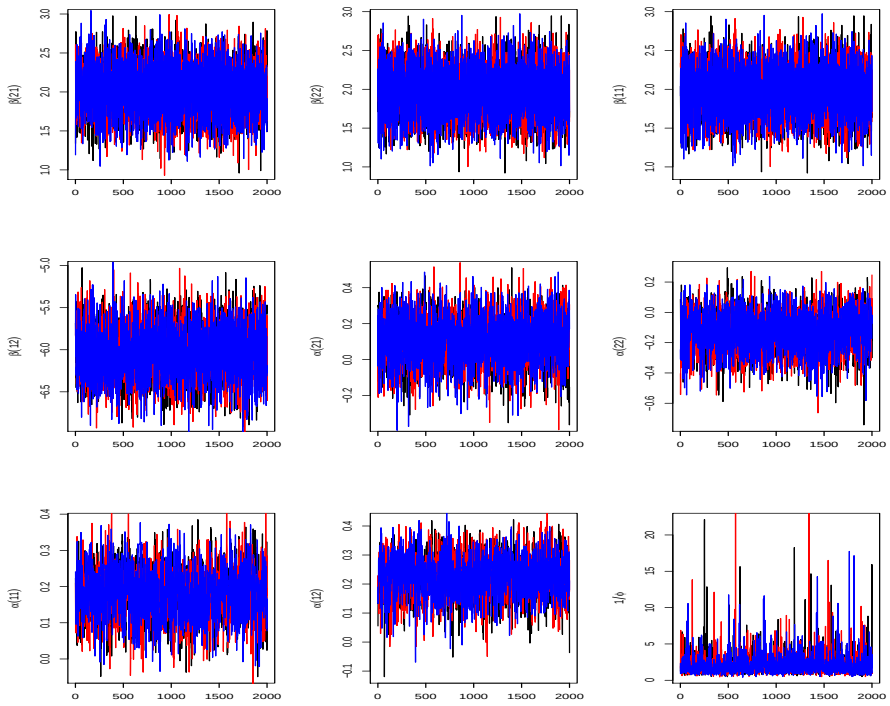

FIGURE 4: Simulated Gibbs samples ("model 4").

Bayesian hierarchical models with "frailties" or latent variables assuming different structures is a powerful way to get the inferences of interest.

Observe that considering independent survival times assuming Weibull distribution (1) and regression model (3) to analyse the survival data introduced in Table 1, we have the value of DIC given by 678.82 considering non-informative priors for the parameters of the model and the same Gibbs algorithm steps assumed for the other proposed models. That is, since DIC is larger assuming independent Weibull models, we have a great indication of the presence of a correlation structure for the survival data of Table 1.

In Table 6, we have the posterior summaries assuming independent Weibull models.

Since we have only a covariate x_i (sex; $x_i = 1$ for male and $x_i = 0$ for female), we can compare the obtained means and variances assuming independent Weibull distributions and "model 1" in the presence of a "frailty". Observe that for "model 1" we use the approximate formulas (6) and (7) for the unconditional means and variances for the survival times (see Table 7). In Table 7, we also have the sample means and sample variances for each combination sex versus response assuming only the uncensored data.

Table 5: Posterior summaries ("model 5" and "model 6").

Model	Parameter	Mean	S.D.	95% Credible Interval
"model 5"	β_{21}	1.8270	0.4312	$(0.9458; 2.6560)$
	β_{22}	0.8333	0.4353	$(-0.0450; 1.6790)$
	β_{11}	-5.1650	0.6007	$(-6.1090; -3.7340)$
	β_{12}	-4.7660	0.5579	$(-5.7620; -3.5470)$
	θ_1	1.4920	0.7905	$(0.6186; 3.7110)$
	θ_2	1.3930	0.6994	$(0.6453; 3.3830)$
	ν_1	0.9544	0.5290	$(0.2398; 2.2420)$
	ν_2	1.2580	0.5967	$(0.3553; 2.6830)$
	$1/\sigma_w^2$	1.9060	0.7860	$(0.8513; 3.9330)$
"model 6"	β_{21}	1.9010	0.4240	$(1.0380; 2.6750)$
	β_{22}	0.9467	0.4284	$(0.1055; 1.7830)$
	β_{11}	-4.8950	0.6403	$(-5.8470; -3.3000)$
	β_{12}	-4.6930	0.5428	$(-5.6470; -3.4300)$
	θ_1	1.3900	0.7163	$(0.6086; 3.3030)$
	θ_2	1.4630	0.6724	$(0.6761; 3.2120)$
	ν_1	1.0330	0.5737	$(0.2642; 2.4950)$
	ν_2	1.1450	0.5517	$(0.3419; 2.5530)$
	$1/\phi$	2.8710	2.1210	$(1.1220; 7.3010)$

Table 6: Posterior summaries (independent Weibull model).

Parameter	Mean	S.D.	95% Credible Interval
β_{21}	1.5540	0.4271	$(0.6873; 2.3750)$
β_{22}	0.2536	0.4351	$(-0.6191; 1.1000)$
β_{11}	-4.8710	0.7078	$(-6.3190; -3.5540)$
β_{12}	-4.8880	0.7298	$(-6.4030; -3.5710)$
med 1	123.80	31.640	$(71.730; 193.80)$
med 2	104.10	28.470	$(61.170; 171.00)$
ν_1	0.9388	0.1238	$(0.7112; 1.1960)$
ν_2	0.9792	0.1392	$(0.7279; 1.2720)$

Table 7: Means and variances ("model 1" and independent Weibull distributions).

	data without censoring		independent Weibull		"model 1"	
	sample mean	sample var	mean	var	unc mean	unc var
$(x = 1)$, resp 1	32.8	2052.09	34.36	1338.18	25.68	565.55
$(x = 0)$, resp 1	162.2	28358.6	186.95	39619.1	86.22	14307.6
$(x = 1)$, resp 2	105.8	36214.1	117.29	14512.9	69.76	3836.03
$(x = 0)$, resp 2	115.9	10609.0	148.36	23243.7	136.51	14658.7

(resp=response; unc=unconditional; male $(x = 1)$; female $(x = 0)$)
var = variance

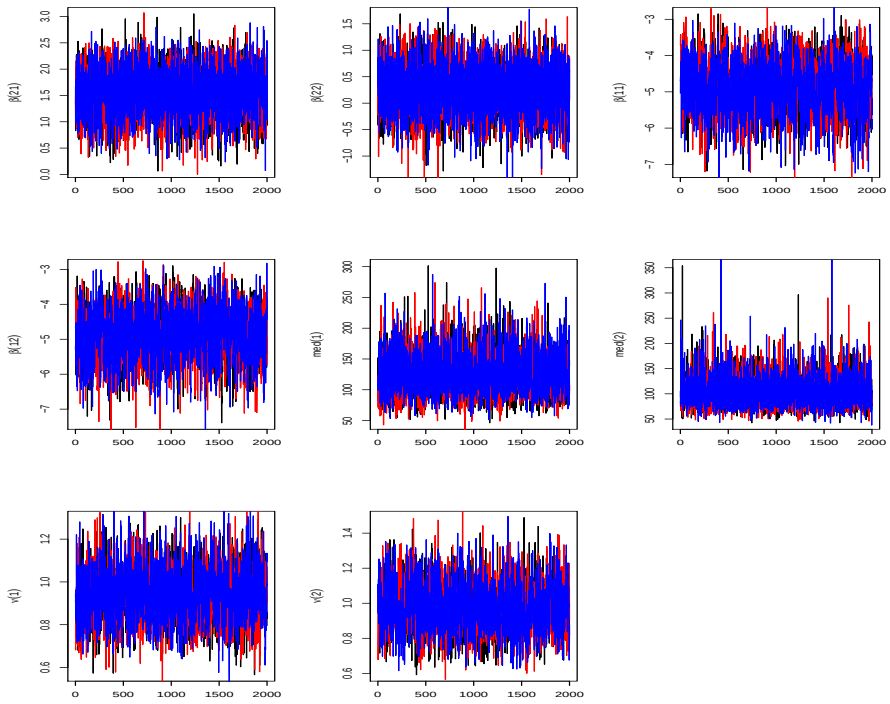

FIGURE 5: Simulated Gibbs samples independent Weibull model.

From the results of Table 7, we observe that the variances of the survival times have a great influence of the presence of the "frailty". Also to point out that these differences could be affected by the sample sizes for each class sex x response is important.

Acknowledgments

The authors would like to thank the editor and referees for their helpful comments.

[Recibido: julio de 2009 — Aceptado: enero de 2011]

References

Chib, S. & Greenberg, E. (1995), 'Understanding the Metropolis-Hastings algorithm', *The American Statistician* (49), 327–335.

Clayton, D. (1991), 'A Monte Carlo Method for Bayesian inference in frailty models', *Biometrics* (47), 467–485.

Clayton, D. & Cuzick, J. (1985), 'Multivariate generalizations of the proportional hazards model', *Journal of the Royal Statistical Society* **A**(148), 82–117.

Cox, D. R. (1972), 'Regression models and life tables', *Journal of the Royal Statistical Society* **B**(34), 187–220.

Cox, D. R. & Oakes, D. (1984), *Analysis of Survival Data*, Chapman & Hall, London.

Gelfand, A. E. & Smith, A. F. M. (1990), 'Sampling based approaches to calculating marginal densities', *Journal of the American Statistical Association* (85), 398–409.

Gelman, A. & Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences (with discussion)', *Statistical Science* **7**(4), 457–472.

Hager, H. W. & Bain, L. J. (1970), 'Inferential procedures for the generalized gamma distribution', *Journal of the American Statistical Association* (65), 1601–1609.

Kalbfleisch, J. D. (1978), 'Nonparametric Bayesian analysis of survival time data', *Journal of the Royal Statistical Society* **B**(40), 214–221.

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, John Wiley, New York.

McGilchrist, C. A. & Aisbett, C. W. (1991), 'Regression with frailty in survival analysis', *Biometrics* (47), 461–466.

Oakes, D. (1986), 'Semiparametric inference in a model for association in bivariate survival data', *Biometrika* **73**, 353–361.

Oakes, D. (1989), 'Bivariate survival models induced by frailties', *Journal of the American Statistical Association* **84**, 487–493.

Parr, V. B. & Webster, J. T. (1965), 'A method for discriminating between failure density functions used in reliability predictions', *Technometrics* (7), 1–10.

Shih, J. A. & Louis, T. A. (1992), Models and analysis for multivariate failure time data, Technical report, Division of Biostatistics, University of Minnesota.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van der Linde, A. (2002), 'Bayesian measures of model complexity and fit (with discussion)', *Journal of the Royal Statistical Society* **B**(64), 583–639.

Spiegelhalter, D. J., Thomas, A., Best, N. G. & Lunn, D. (2003), *WinBugs version 1.4 user manual*, Institute of Public Health and Department of Epidemiology & Public Health, London.
*http://www.mrc-bsu.com.ac.uk/bugs

Stacy, E. W. & Mihram, G. A. (1965), 'Parameter estimation for a generalized gamma distribution', *Technometrics* (7), 349–358.

Weibull, W. (1951), 'A statistical distribution function of wide applicability', *Journal of Applied Mechanics* **18**(3), 292–297.

Revista Colombiana de Estadística
Junio 2011, volumen 34, no. 1, pp. 133 a 146

Nonparametric Cutoff Point Estimation for Diagnostic Decisions with Weighted Errors

Estimación no paramétrica del punto de corte asociado a una decisión diagnóstica con errores ponderados

Pablo Martínez-Camblor[1,2,a]

[1]CAIBER, Oficina de Investigación Biosanitaria, Oviedo, Spain
[2]Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Oviedo, Spain

Abstract

The study of diagnostic tests is a hot topic which has direct applications in biomedical sciences. Despite of the relevance, in a diagnostic process, of the threshold (or cutoff point) employed on the decision taken by the physician, the study and comparison of the accuracy among different diagnostic criterions has been the main field of study. In this paper, the authors are interested in the study of the involved cutoff point estimation in diagnostic tests with weighted errors. With this goal, a nonparametric smoothed utility function estimator is considered. The bootstrap and the asymptotic distributions for the related M-estimator are derived. Finally, the obtained results are applied to study the Procalcitonin level which determines whether a child within the Pediatric Intensive Care Unit (UCIP) has a virical sepsis.

***Key words*:** Kernel density estimator, Sensitivity, Specificity, Threshold, Utility function.

Resumen

El estudio de tests diagnósticos es un tema candente con aplicaciones directas en las ciencias biomédicas. Aunque en la práctica, a la hora de tomar una decisión, los clínicos deben fijar un valor umbral (o punto de corte) a pesar de la relevancia que este valor tiene, el estudio y la comparación de la calidad entre diferentes criterios diagnósticos ha sido el principal campo de estudio. En este trabajo, los autores están interesados en el estudio de la estimación del punto de corte involucrado en un test diagnóstico con errores ponderados. Con este objetivo, se considera un estimador suavizado para una función de utilidad. Se estudian las distribuciones *bootstrap* y asintóticas del M-estimador resultante. Finalmente, los resultados obtenidos son aplicados al estudio de los niveles de Procalcitonina que determinan si un niño ingresado en la Unidad de Cuidados Intensivos Pediátricos (UCIP) tiene infección vírica.

***Palabras clave*:** especificadas, estimador núcleo para la densidad, función de utilidad, sensibilidad, umbral.

[a]Biostatistics and Associate professor. E-mail: pablomc@ficyt.es

1. Introduction

Diagnostic methods play an important role in the medical attention. The estimation and comparison of the accuracy among different methods are the focus of a wide variety of studies (see, for example, Zhou, Obuchowski & McClish (2002) and references therein). The main goal in a diagnostic test is to determine whether one individual is ill (positive). With this purpose, usually, some physiologic measure, T, is taken (as a marker) on a patient; the patient is classified as positive (with the illness) if this measure is upper (or lower) than a previously fixed threshold. This classification process has associated two possible mistakes –to classify a healthy individual in the positive group and to classify an unhealthy individual in the negative group. Of course, to determine the diagnostic test accuracy, these errors are basic. The proportion of positives which are correctly identified is known as *sensitivity* (S_E) and the proportion of negatives which are correctly identified is known as *specificity* (S_P).

The *Receiver Operating Characteristic* (ROC) curve (Green & Swets 1966) is a popular graphical method of displaying the discriminatory accuracy of a diagnostic test (based on a marker) for distinguishing between two populations. It is a plot of true-positive fraction (S_E) against the false-positive fraction ($1 - S_P$) over all possible threshold values of the considered marker. Although alternative indices have been discussed (see, for example, Lee & Hsiao (1998) or more recently Hand (2009)), the area under ROC curve (AUC) is, probably, the most commonly used index for diagnostic global accuracy. The ROC curve and the AUC index have been studied from different approaches (see Rodríguez-Álvarez, Tahoces, Cadarso-Suárez & Lado (2011) and Airola, Pahikkala, Waegeman, De Baets & Salakoski (2011) for some recent references). They have also been involved in the solution of different practical problems; for instance, recently, López-de Ulibarri, Cao, Cadarso-Suárez & Lado (2008) used a smooth estimation of the conditional ROC curve and the AUC on task discriminations and Martínez-Camblor & Yáñez-Juan (2009) developed a test to compare the equality of the diagnostic effectiveness of one measure with respect differents features based on the respective AUC values.

The Youden Index (Youden 1950) is also frequently used as accuracy measure. It is defined as $J = \max_{t\in\mathbb{R}}\{S_E(t)+S_P(t)-1\}$ and ranges between 0 and 1. Chin-Ying, Tian & Schisterman (2011) derived a procedure to build exact confidence interval estimations for the Youden index and its corresponding optimal cut-point. A vast study about the Youden Index and its associated cut-point estimations have been conducted by Fluss, Faraggi & Reiser (2005). They concluded that, in the estimation of the Youden Index the kernel is generally the best (among four considered estimators) unless the data can be well transformed to achieve normality whereas in estimation of the optimal threshold value results are more variable.

Most considered indices assume that the sensitivity and the specificity have the same relevance. However, to understand that there exist situations in which the impact of the two possible mistakes is quite different is easy. Taking into acount these differences and, for each $\lambda \in (0, 1)$, we introduce the following linear *utility* function (although in other context, it has been previously considered by

Krzanowski & Hand 2009)

$$U_\lambda(t) = \lambda S_E(t) + (1-\lambda) S_P(t) \tag{1}$$

Because the λ value (weight) determines the final impact of the sensitivity and specificity, its election is really important and, usually, depends on the costs of the different decisions and the prevalence of the illness which is being studied. Obviously, for each particular problem, its real value will be previously fixed by the specialist who must taking into count the different misclassification effects. Note that, if for $0 \leq \lambda \leq 1$ it is considered the optimum reachable utility, i.e.

$$J_\lambda = \max_{t \in \mathbb{R}} \{U_\lambda(t)\} \qquad \lambda \in (0,1) \tag{2}$$

then $J = 2(J_{1/2} - 1/2)$. Therefore, J_λ generalizes J when the mistakes in the classification process have different weights.

In this paper, smoothed estimators for the coefficient J_λ and its associated threshold are studied. In Section 2, the asymptotic and the bootstrap approximations for the cutoff point smoothed estimator are derived. Finally, in Section 3, we apply the proposed methods on the data set which motivated this research. On this data set, we study the procalcitonin (PCT) level which determines whether a child into the Pediatric Intensive Care Unit (UCIP) has a virical sepsis.

2. Nonparametric Cutoff Point Estimation

Let T be a continuous marker, we can assume (without loss of generality) that an individual is classified within group E (positives) if $T > t$ and within group $\overline{E}$ ($\overline{E}$ denotes the complementary set of E) if $T \leq t$. Let F_N and f_N be the distribution and the density functions, respectively, of N_T (T in the negative population; without the characteristic), and let F_P and f_P be the distribution and the density functions, respectively, of P_T (T in the positive population; with the characteristic), we have the equalities

$$S_E(t) = \mathcal{P}\{T > t \mid \mathrm{E}\} = 1 - F_P(t) \tag{3}$$

$$S_P(t) = \mathcal{P}\left\{T \leq t \mid \overline{\mathrm{E}}\right\} = F_N(t) \tag{4}$$

As usual, to estimate S_E and S_P we must estimate the distribution functions involved in the above definitions. Following the conclusions obtained by Fluss et al. (2005), we employ the kernel estimator and put the respective Smoothed Empirical Cummulative Distribution Functions (SECDF) instead of the theoretical ones to estimate the sensitivity and the specificity. Let $X = \{x_1, \ldots, x_n\}$ be a random sample from a continuous distribution F, the SECDF introduced by Nadaraya (1962) is defined as

$$\widetilde{F}_n(X,t) = \frac{1}{n} \sum_{i=1}^{n} \widetilde{K}\left(\frac{t - x_i}{h_n}\right)$$

where $\widetilde{K}$ is a kernel function, usually taken to be a continuous probability function, with continuous and symmetrical about zero first derivative and $\{h_n\}_{n\in\mathbb{N}}$ is a sequence of deterministic bandwidths. The properties of the kernel estimator and its related curves have been widely studied and there exists a vast literature about this topic (see, for example, Mugdadi & Ghebregiorgis (2005) or Liu & Yang (2008) and references therein). Under some regularity conditions over the theoretical distribution and the used kernel function (it is enough, although not necessary, that both functions have three bounded and continuous derivatives), the mean ($\mathbb{E}$) and the variance ($\mathbb{V}$) for the SECDF are

$$\mathbb{E}[\widetilde{F}_n(X,t)] = F(t) + (1/2)f'(t)h_n^2 + \mathcal{O}(h_n^3) \tag{5}$$

$$n\mathbb{V}[\widetilde{F}_n(X,t)] = F(t)(1-F(t)) - 2h_n f(t)\int v\widetilde{K}'(v)\widetilde{K}(v)dv + \mathcal{O}(h_n^2) \tag{6}$$

Let $X_P = \{x_{P_1},\ldots,x_{P_n}\}$ and $X_N = \{x_{N_1},\ldots,x_{N_m}\}$ be two random samples from the positive and the negative populations, respectively, the natural *smoothed estimators* for S_P and S_N are

$$\widetilde{S}_E(t) = 1 - \widetilde{F}_n(X_P,t) \tag{7}$$

$$\widetilde{S}_P(t) = \widetilde{F}_m(X_N,t) \tag{8}$$

In the same way, replacing the sensitivity and the specificity by the above estimators, it is obtained the smoothed estimator for the utility function defined in (1),

$$\widetilde{U}_\lambda(t) = \lambda\widetilde{S}_E(t) + (1-\lambda)\widetilde{S}_P(t) \qquad \lambda\in(0,1) \tag{9}$$

Finally, the estimator for the associated cutoff point which is one of focus of this research, is the M-statistic

$$\widetilde{\theta}_\lambda = \min\{\text{argmax}_{t\in\mathbb{R}}\{\widetilde{U}_\lambda(t)\}\} \qquad \lambda\in(0,1) \tag{10}$$

The following result proves the asymptotic normality for the statistic $\widetilde{\theta}_\lambda$ under quite general conditions on the theoretical underlying distribution and on the parameters involved in the estimator definition (kernel function and used bandwidth).

Theorem 1. *Let X_N and X_P be two independent random samples (both independient and identically distributed, iid) with size n and m, respectively. Let $\widetilde{F}_n(X_N,t)$ be and $\widetilde{F}_m(X_P,t)$ the respective Smoothed Empirical Cummulative Distribution Functions (SECDF). Under the following assumptions*

A_1. *The real distribution function have three bounded and continuous derivatives.*

A_2. *Used kernel, $\widetilde{K}$, is a symmetrical about zero function with three bounded and continuous derivatives and $\int x^2 d\widetilde{K}(x) = 1$.*

A_3. $\exists \lim_n \sqrt{nh_n/mh_m} = \lim_n \alpha_n = \alpha < \infty$.

A_4. $U''_\lambda(\theta_\lambda) \neq 0$.

then,

$$\sqrt{nh_n}\frac{\widetilde{\theta}_\lambda - \theta_\lambda}{V_\lambda} \xrightarrow{\mathcal{L}}_n \mathcal{N}(0,1) \tag{11}$$

with

$$V_\lambda^2 = \frac{R(K)\left(\lambda^2 f_P(\theta_\lambda) + (1-\lambda)^2\alpha^2 f_N(\theta_\lambda)\right)}{\left(\lambda f_P'(\theta_\lambda) + (1-\lambda)f_N'(\theta_\lambda)\right)^2} \tag{12}$$

where for each real function, g, $R(g) = \int g^2(x)\,dx$.

As usual, the variance of the statistic $\widetilde{\theta}_\lambda$ depends on several theoretical and unknown parameters, in particular, on the density functions (and its first derivative) in the positive and negative populations evaluated at the real optimal cutoff point, θ_λ. These theoretical (unknown) parameters are replaced by their *natural estimators* (the smoothed ones in the present study) to compute confidence intervals for $\widetilde{\theta}_\lambda$ (plug-in method) or for conducting inference on the parameter.

The Kernel density estimator, introduced by Rosenblatt (1956), is the most popular and commonly used density function estimator. Let $X = \{x_1, \ldots, x_n\}$ be a random sample (iid), it is defined as

$$\widetilde{f}_n(X,t) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{t-x_i}{h_n}\right) \tag{13}$$

where $K = \widetilde{K}' = (\partial\widetilde{K}(t)/\partial t)$ is a kernel function and $\{h_n\}_{n\in\mathbb{N}}$ is a sequence of deterministic bandwidths. In this setting, the *natural estimator* for the first density function derivative is

$$\widetilde{f}_n'(X,t) = \frac{1}{nh_n^2}\sum_{i=1}^{n} K'\left(\frac{t-x_i}{h_n}\right) \tag{14}$$

The *bandwidth* selection for the kernel estimators was a very hot topic in the 80s and early 90s (and it is still the focus of several recent papers). Their optimal convergence rates were widely studied. Cao (1990), looking for the bandwidth which minimizes the mean integrated square error (MISE), proved that the optimum convergence ratio for the SECDF is $\mathcal{O}(n^{-1/3})$, $\mathcal{O}(n^{-1/5})$ for kernel density function estimator and $\mathcal{O}(n^{-1/7})$ for its first derivative.

Silverman (1978) proved that if the real density function, f, is continuous, the used kernel is a variation bounded function and the bandwidth, h_n, is such that $nh_n \to_n \infty$ and $h_n \to_n 0$, the kernel density estimator, $\widetilde{f}_n$ converges uniformly almost surely to the real density function, i.e. $\sup_{t\in\mathbb{R}} |\widetilde{f}_n(X,t) - f(t)| \to 0$ a.s. (almost surely). This result allows deriving the following theorem

Theorem 2. *Under the assumptions in Theorem 1 and if it is also satisfied that*

A_5. $\widetilde{U}_\lambda''(\widetilde{\theta}_\lambda) \neq 0$.

A_6. *All the used bandwidth have the previously written optimal convergence rates (i.e. $\mathcal{O}(n^{-1/3})$ for SECDF; $\mathcal{O}(n^{-1/5})$ for density estimator and $\mathcal{O}(n^{-1/7})$ for its first derivative).*

then

$$\sqrt{nh_n}\ \frac{\widetilde{\theta}_\lambda - \theta_\lambda}{V_{n,\lambda}} \xrightarrow{\mathcal{L}}_n \mathcal{N}(0,1) \tag{15}$$

with

$$V^2_{n,\lambda} = \frac{R(K)\left(\lambda^2 \widetilde{f}_n(X_P, \widetilde{\theta}_\lambda) + (1-\lambda)^2\alpha^2 \widetilde{f}_m(X_N, \widetilde{\theta}_\lambda)\right)}{\left(\lambda \widetilde{f}'_n(X_P, \widetilde{\theta}_\lambda) + (1-\lambda)\widetilde{f}'_m(X_N, \widetilde{\theta}_\lambda)\right)^2} \tag{16}$$

The main disadvantage of the previous result lies on the variance denominator estimator. Kernel estimators depend on the bandwidth selection which must be made by the investigator. There exist several automatic methods with this goal but does not exist an optimal solution (in adittion, the optimal bandwidth usual changes with each particular problem: density estimation, inference, etc.) For some discussion about this topic see Martínez-Camblor & De Uña-Álvarez (2009). The involved parameters on the denominator of the variance can be close to zero and, hence, small changes in their estimations can produce big changes on the final result. Trying to avoid these problems, as usual, we propose to use a resampling plan. Because the studied marker, T, is continuous and the expressions of the studied estimators depend on local properties (derivability), the *Smoothed Bootstrap* procedure (Hall, DiCiccio & Romano 1989) seems the most appropriate. The proposed algorithm is:

$\boldsymbol{B_1}$. From positive (X_P) and negative (X_N) samples, and for a fixed, or a grid of λ values ($\lambda \in (0,1)$), compute the SECDF and estimate: *sensitivity, specificity* and *utility functions*. Also compute the optimal cutoff point (threshold), $\widetilde{\theta}_\lambda(X_P, X_N) = \widetilde{\theta}_\lambda$.

$\boldsymbol{B_2}$. Run B pairs of bootstrap samples (X^b_P, X^b_N for $1 \leq b \leq B$) with the same sample sizes than the original ones from the respective SECDFs. On each bootstrap sample, compute and estimate functions which appear in $\boldsymbol{B}_1$. Also obtain the values for $\widetilde{\theta}^b_\lambda = \widetilde{\theta}^b_\lambda(X^b_P, X^b_N)$ with $1 \leq b \leq B$.

$\boldsymbol{B_3}$. The distribution of $\widetilde{\theta}_\lambda$ (and the other involved statistics) is approximated by $\{\widetilde{\theta}^1_\lambda, \ldots, \widetilde{\theta}^B_\lambda\}$.

Since the differences among the different resampling methods to make confidence intervals are, generally, negligible, we used the simplest and, probably, the most often used one; the percentile method (Efron & Tibshirani 1993). This method assumes that for a unknown monotone increasing transformation for the studied parameter, $h(\theta_\lambda)$ (in the present case, $\lambda \in (0,1)$), it is hold that

$$h(\widetilde{\theta}_\lambda) - h(\theta_\lambda) \sim \mathcal{N}(0, \sigma^2_{h(\widetilde{\theta}_\lambda)})$$

From this approach, a simple approximation for a $(1-\alpha)$ confidence interval can be found as $(\widetilde{\theta}^{(\alpha/2)}_\lambda, \widetilde{\theta}^{(1-\alpha/2)}_\lambda)$, where $\widetilde{\theta}^b_\lambda$ ($b \in 1, \ldots, B$) is obtained from the algorithm above.

The main goal of this algorithm is to approximate the $\widetilde{\theta}_\lambda$ distribution but, analogously, it can also be used to approximate the distribution for the other involved parameters (sensitivity, specificity and utility functions).

Despite of the AUC is widely used to summarize the global classification accuracy of a diagnostic rule, it is fundamentally incoherent in terms of misclassification costs (Hand 2009). The J_λ index defined in (2) provides a oportunity to define a global index for the diagnostic test accuracy which takes into count the different misclassification cost. With this goal, for each measure μ, it is define, by

$$\text{AUJ} = \int_0^1 J_\lambda d\mu(\lambda) \tag{17}$$

Note that AUJ index ranges between 0 and $\mu([0,1])$. If the chosen weight, μ, is the traditional Lebesgue measure, the AUJ stands for the area under J_λ curve and it means that all possible values of the weights are considered to be equally plausible.

3. Real Data Analysis

Bacterial sepsis is an important cause of mortality and morbidity in critically ill child. A delayed diagnosis of this condition is associated with worse prognosis. However, early detection of bacterial sepsis is difficult because the first signs of this disease may be minimal or non specific. Moreover, critically ill children present signs of sepsis such as fever, tachycardia, hyperventilation, and leukocytosis even in the absence of infection.

The availability of a laboratory test to accurately and rapidly identify critically ill children with sepsis would be of great value to improve the outcome of these patients. Early detection of the absence of infection would decrease the number of children started on antibiotics, shorten the length of hospital stay, and lessen the potential for emergent of resistant bacteria.

Body response to bacterial sepsis involves the release of several mediators. Recently, PCT, one of these mediators, has been proposed as an earlier marker of bacterial sepsis in children. Moreover, PCT levels are related to the severity of infection, presenting higher levels among patients with more severe sepsis. However, there is still some debate about the best cutoff levels to differentiate a patient with sepsis from a patient without sepsis (Rey, Los Arcos, Concha, Medina, Prieto, Martínez-Camblor & Prieto 2007). To define PCT cutoff levels with the optimum sensitivity and specifity to diagnose a critically ill child with sepsis can be very useful. Our goal is to study the cutoff point for the procalcitonin levels which determine that a child has a sepsis. With this objective we used the previous results for different λ values. We used information from patients admitted to the Pediatric Intensive Care Unit at the Hospital Universitario Central de Asturias (HUCA) from August 2002 until September 2004.

The descriptive statistics showed in Table 1 suggests a strongly asymmetry in the distribution of the PCT levels in both positive and negative considerated

groups. We know (see, for example, Silverman 1986) that the performance of the smoothed estimators is better for symmetrical distributions. To improve the estimations, we make a logarithmic transformation on the PCT levels. Kernel density estimations for the logarithmic of the PCT levels and for the function $\max_{t\in\mathbb{R}}\{\widetilde{U}_\lambda(t)\}$ ($\lambda \in (0,1)$) are shown in Figure 1.

TABLE 1: Descriptive statistics (mean, standard deviation (SD), minimum (Min), percentiles 25 (P_{25}), 50 (P_{50}), 75 (P_{75}), maximum (Max) and sample size (N)) for the Procalcitonin levels in the different considerated groups.

	Mean	**SD**	**Min**	$\mathbf{P}_{25}$	$\mathbf{P}_{50}$	$\mathbf{P}_{75}$	**Max**	**N**
Positive Group	22.89	39.83	0.11	2.81	10.64	27.53	347.10	125
Negative Group	1.48	3.98	0.01	0.12	0.30	1.00	39.01	232
Totals	8.98	25.83	0.01	0.18	0.95	5.73	347.10	357

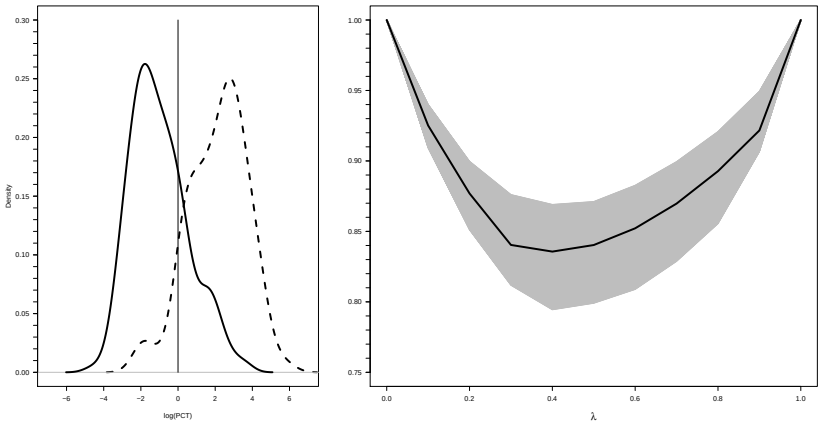

FIGURE 1: Kernel density estimations (left) for the logarithmic of the PCT levels in the positive (dotted line) and negative (continuous line) populations and the function $J_\lambda = \max_{t\in\mathbb{R}}\{\widetilde{U}_\lambda(t)\}$ with a 95% bootstrap confidence band (right).

Table 2 shows the obtained estimations for $\widetilde{\theta}_\lambda$, $\widetilde{U}_\lambda(\widetilde{\theta}_\lambda)$, $\widetilde{S}_E(\widetilde{\theta}_\lambda)$, $\widetilde{S}_P(\widetilde{\theta}_\lambda)$ and the square root for the asymptotic ($SD(\widetilde{\theta}_\lambda)$) and the bootstrap ($SD_B(\widetilde{\theta}_\lambda)$) (based on 10 000 Monte Carlo simulations) variance (SD) for $\widetilde{\theta}_\lambda$ for several λ values. For this data set, the asymptotic variance is, smaller than the bootstrap one. This fact suggests a slow speed for the asymptotic convergence. The value for the AUJ index when μ is the Lebesgue measure is 0.894 (really, the AUJ value represents the global utility of the particular diagnostic test when all the possible values of the weights are chosen to be equally plausible).

Figure 2 depicts 95% asymptotic and bootstrap confidence intervals (upper). In the lower plots, utility functions at the extremes of these confidence intervals are shown. The difference among the values is always quite small, which suggests robustness with respect to the chosen threshold.

TABLE 2: Values for $\widetilde{\theta}_\lambda$, $\widetilde{U}_\lambda(\widetilde{\theta}_\lambda)$, $\widetilde{S}_E(\widetilde{\theta}_\lambda)$, $\widetilde{S}_P(\widetilde{\theta}_\lambda)$, square root for asymptotic variance of $\widetilde{\theta}_\lambda$ ($SD(\widetilde{\theta}_\lambda)$) and bootstrap variance of $\widetilde{\theta}_\lambda$ ($SD_B(\widetilde{\theta}_\lambda)$) based on 10 000 Monte Carlo simulations for several λ values.

λ	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
$\widetilde{\theta}_\lambda$	12.67	9.02	2.34	1.65	1.25	1.03	0.85	0.69	0.52
$\widetilde{U}_\lambda(\widetilde{\theta}_\lambda)$	0.92	0.88	0.84	0.83	0.84	0.85	0.87	0.89	0.92
$\widetilde{S}_E(\widetilde{\theta}_\lambda)$	0.44	0.53	0.77	0.34	0.88	0.91	0.93	0.94	0.95
$\widetilde{S}_P(\widetilde{\theta}_\lambda)$	0.98	0.96	0.87	0.83	0.80	0.76	0.72	0.69	0.63
$SD(\widetilde{\theta}_\lambda)$	8.69	3.37	0.73	0.26	0.12	0.09	0.08	0.07	0.05
$SD_B(\widetilde{\theta}_\lambda)$	9.39	2.28	2.11	0.66	0.26	0.18	0.13	0.11	0.19

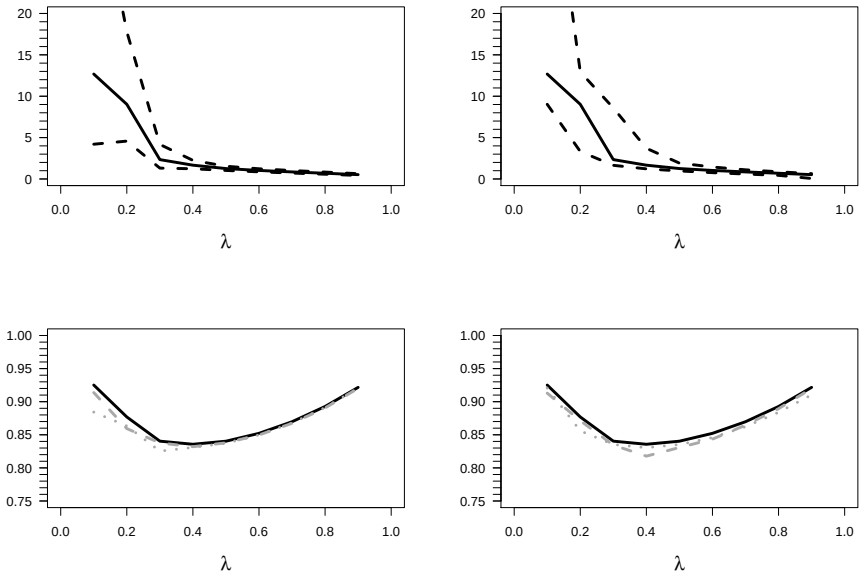

FIGURE 2: Upper, asymptotic (left) and bootstrap (right) 95% confidence intervals for the associated cutoff point estimation. Lower, utility function evaluated at the optimal estimated cutoff point (continuous lines) and at the extremes of the previous confidence intervals, asymptotic (left) and bootstrap (right) upper bound (grey dashed lines) and lower bound (grey dotted lines).

When sensitivity and specificity have the same relevance, both the AUC (0.913) and the Youden Index (0.680) suggest that the procalcitonin is a very good sepsis marker for the studied population. If different weights are assigned to S_E and S_P, in spite of the results are still quite goods, when we pay more attention to the sensitivity, the obtained utility is bigger. On the contrary the lower plots in the Figure 2 show that the final *gain* not changes when the cutoff points are within a reasonable interval.

4. Main Conclusions

Two important points of diagnostic medicine research, which are usually omitted, are the possible different impact of the two involved errors on the decision process and the effects on the final results in the variability of the associated cutoff point estimator. In this paper, we deal with the first problem introducing a linear *utility* function (obviously, most complex utility function could be considered with this goal) which allows study different weights for the sensitivity and the specificity. These weights must be previously chosen for the specialist which will take into count the different cost of the possible misclassification. The methods to cancer diagnostic tests are a special intersting field of application. There are continuous advance in this field with the aparition of new diagnostic markers (usually related with genes but which sensitivity and specificity are, generally, not large) and new (customized) drugs. The cost of the misclassification in this situation is usually different with great advantages for the early diagnostic.

We studied a nonparametric smoothed estimator for a linear utility function which allows to weight sensitivity and specificity and the corresponding associated cutoff point. We also derived its asymptotic distribution. In addition, the smoothed bootstrap procedure is considered. Because in the case of discrete markers all possible cutoff points could be studied and the researcher could chose among all the possibilities, we focus on continuos markers.

The obtained asymptotic variance for the threshold estimator is strongly depending on the first derivative of the density function. Because the convergence speed of the usual (kernel) estimator for this function is quite slow (see, for example, Silverman 1978), the use of the bootstrap approximation is advised when sample size is not large. To obtain adequate asymptotic conficence intervals, the required sample size depends on the variability and, in special, on the shape of the functions but, under simmetry, sizes around 100σ (σ^2 denotes the variance population) are advisables.

The effect that a little change on the used threshold produces on the final utility function is a specially interesting issue. In our analysis, this change seems to have a minor effect and the developed methods seem to be robust in this sense.

The AUC is a very widely used measure of performance for classification and diagnostic rules. It is mainly used in medicine and, recently, its use has been generalized to measure the accuracy in evaluating learning algorithms (see, for example, Huang & Ling (2005) and references therein). It has the appealing property of being objective, requiring no subjective input from the user but it is incoherent in terms of misclassification costs (Hand 2009). From the J_λ an coherent alternative (AUJ) to the AUC index (in cost terms) is also defined and studied.

Acknowledgements

The author is very grateful with Corsino Rey Galan and Marta Los Arcos Solas from the Hospital Universitario Central de Asturias (HUCA) for permission to use their data and for suggesting this research. The author is also grateful with the three anonymous referees whose suggestions and comments have really improved the paper.

[Recibido: junio de 2010 — Aceptado: diciembre de 2010]

References

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B. & Salakoski, T. (2011), 'An experimental comparison of cross-validation techniques for estimating the area under the ROC curve', *Computational Statistics & Data Analysis* **55**(4), 1828–1844.

Cao, R. (1990), Aplicaciones y nuevos resultados del Método Bootstrap en la estimación no paramétrica de curvas, PhD thesis, University of Santiago de Compostela, Santiago de Compostela, Spain.

Chin-Ying, L., Tian, L. & Schisterman, E. F. (2011), 'Exact confidence interval estimation for the Youden index and its corresponding optimal cut-point', *Computational Statistics & Data Analysis* . In Press, Corrected Proof. DOI: 10.1016/j.csda.2010.11.023.
*http://www.sciencedirect.com/science/article/B6V8V-51N223Y-1/2/9ac9b12dcddcf280e599a152375ca56c

Efron, B. & Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, London, United Kingdom.

Fluss, R., Faraggi, D. & Reiser, B. (2005), 'Estimation of the Youden index and its associated cutoff point', *Biometrical Journal* **47**(4), 458–472.

Green, D. M. & Swets, J. A. (1966), *Signal Detection Theory and Psychophysics*, Wiley, New York, United States.

Hall, P., DiCiccio, J. T. & Romano, J. P. (1989), 'On smoothing and the bootstrap', *Annals of Statistics* **17**, 692–702.

Hand, D. J. (2009), 'Measuring classifier performance: A coherent alternative to the area under the ROC curve', *Machine Learning - ML* **77**(1), 103–123.

Huang, J. & Ling, C. X. (2005), 'Using AUC and accuracy in evaluating learning algorithms', *IEEE Transactions on Knowledge and Data Engineering* **17**(3), 299–310.

Krzanowski, W. J. & Hand, D. J. (2009), *ROC Curves for Continuous Data*, Chapman and Hal, New York, United States.

Lee, W. C. & Hsiao, C. K. (1998), 'Alternative summary indices for the receiver operating characteristic curve', *Epidemiology* **7**, 605–611.

Liu, R. & Yang, L. (2008), 'Kernel estimation of multivariate cumulative distribution function', *Journal of Nonparametric Statistics* **20**(8), 661–667.

López-de Ulibarri, I., Cao, R., Cadarso-Suárez, C. & Lado, M. J. (2008), 'Nonparametric estimation of conditional ROC curves: Application to discrimination tasks in computarized detection of early breast cancer', *Computational Statistics & Data Analysis* **52**(5), 2623–2631.

Martínez-Camblor, P. & De Uña-Álvarez, J. (2009), Studying the bandwidth in k-sample smooth tests, Technical Report 09, Universidad de Vigo, Vigo, Spain.

Martínez-Camblor, P. & Yáñez-Juan, A. (2009), 'Testing the equality of diagnostic effectiveness of one measure with respect to k different features', *Journal of Applied Statistics* **36**(4), 359–367.

Mugdadi, A. R. & Ghebregiorgis, G. S. (2005), 'The kernel distribution estimator of functions of random variables', *Journal of Nonparametric Statistics* **17**(7), 807–818.

Nadaraya, E. A. (1962), 'Some new estimates for distribution functions', *Theory Probability Application* **9**, 497–500.

Rey, C., Los Arcos, M., Concha, A., Medina, A., Prieto, S., Martínez-Camblor, P. & Prieto, B. (2007), 'Procalcitonin and C-reactive protein as markers of systemic inflammatory response syndrome severity in critically ill children', *Intensive Care Medicine* **33**(3), 477–484.

Rodríguez-Álvarez, M. X., Tahoces, P. G., Cadarso-Suárez, C. & Lado, M. J. (2011), 'Comparative study of ROC regression techniques-applications for the computer-aided diagnostic system in breast cancer detection', *Computational Statistics & Data Analysis* **55**(1), 888–902.

Rosenblatt, M. (1956), 'Remarks on some nonparametric estimates of a density function', *Annals of Mathematical Statistics* **17**, 832–837.

Silverman, B. W. (1978), 'Weak and strong uniform consistency of the density estimation and its derivatives', *Annals of Statistics* **6**, 1177–1184.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, United States.

Youden, W. J. (1950), 'Index for rating diagnostic test', *Cancer* **3**, 32–35.

Zhou, X. H., Obuchowski, N. A. & McClish, D. K. (2002), *Statistical Methods in Diagnostic Medicine*, Wiley & Sons, New York, United States.

Appendix
Proof of the Results

Following, we deal with the proofs for the Theorems 1 and 2. Both demonstrations, quite similar, are based on the smoothed estimators and M-statistic properties and on the regularity conditions asked to the involved functions.

Proof. (Theorem 1) Conditions A_1 and A_2 guarantee the uniformly almost surely convergence for the kernel density estimator and its two first derivatives (Silverman 1978), therefore we can derive that $(\widetilde{U}_\lambda(\theta_\lambda) - U_\lambda(\theta_\lambda)) \to_P 0$.

$\widetilde{\theta}_\lambda = \text{argmax}\{\widetilde{U}_\lambda(t)\}$ and $\theta_\lambda = \text{argmax}\{U_\lambda(t)\}$, hence $\widetilde{U}'_\lambda(\widetilde{\theta}_\lambda) = 0 = U'_\lambda(\theta_\lambda)$. From the Theorem of the Mean Value, there exists ξ_λ between $\widetilde{\theta}_\lambda$ and θ_λ such that

$$\widetilde{U}'_\lambda(\theta_\lambda) - U'_\lambda(\theta_\lambda) = \widetilde{U}'_\lambda(\theta_\lambda) - \widetilde{U}'_\lambda(\widetilde{\theta}_\lambda) = \widetilde{U}''_\lambda(\xi_\lambda)(\theta_\lambda - \widetilde{\theta}_\lambda) \quad \lambda \in (0,1)$$

therefore $(\theta_\lambda - \widetilde{\theta}_\lambda) \to_P 0$.

Applying a three-term Taylor expansion on the first derivative of the utility function at point $\widetilde{\theta}_\lambda$, there exists η_λ between $\widetilde{\theta}_\lambda$ and θ_λ such that

$$\begin{aligned} 0 =& \widetilde{U}'_\lambda(\widetilde{\theta}_\lambda) = \widetilde{U}'_\lambda(\theta_\lambda + \widetilde{\theta}_\lambda - \theta_\lambda) \\ =& \widetilde{U}'_\lambda(\theta_\lambda) + \widetilde{U}''_\lambda(\theta_\lambda)(\widetilde{\theta}_\lambda - \theta_\lambda) + (1/2)\widetilde{U}'''_\lambda(\eta_\lambda)(\widetilde{\theta}_\lambda - \theta_\lambda)^2 \quad \lambda \in (0,1) \end{aligned}$$

then

$$\sqrt{nh_n}\,(\widetilde{\theta}_\lambda - \theta_\lambda) = -\frac{\sqrt{nh_n}\,\widetilde{U}'_\lambda(\theta_\lambda)}{\widetilde{U}''_\lambda(\theta_\lambda) + \frac{1}{2}\widetilde{U}'''_\lambda(\eta_\lambda)(\widetilde{\theta}_\lambda - \theta_\lambda)}$$

The A_2 assumption also implies that $\widetilde{U}'''_\lambda(t)$ is a bounded function $\forall t \in \mathbb{R}$, therefore $\widetilde{U}'''_\lambda(\eta_\lambda)(\widetilde{\theta}_\lambda - \theta_\lambda) \to_P 0$ for $\lambda \in (0,1)$. Kernel estimator convergence properties (cited at the beginning of the proof) imply $(\widetilde{U}''_\lambda(\theta_\lambda) - U''_\lambda(\theta_\lambda)) \to_P 0$, and then

$$\left(\sqrt{nh_n}(\widetilde{\theta}_\lambda - \theta_\lambda) + \frac{\sqrt{nh_n}U'_\lambda(\theta_\lambda)}{U''_\lambda(\theta_\lambda)}\right) \xrightarrow{P} 0$$

The Central Limit Theorem leads us to the convergence

$$\sqrt{nh_n}\,\widetilde{U}'_\lambda(\theta_\lambda) = \sqrt{nh_n}(\widetilde{U}'_\lambda(\theta_\lambda) - U'_\lambda(\theta_\lambda)) \xrightarrow{\mathcal{L}}_n \mathcal{N}(0, \sigma_\lambda)$$

with $\sigma^2_\lambda = R(K)\left(\lambda^2 f_P(\theta_\lambda) + (1-\lambda)^2\alpha^2 f_N(\theta_\lambda)\right)$.

The Slutski Lemma allows deducing, inmediately, that $\sqrt{nh_n}\,(\widetilde{\theta}_\lambda - \theta_\lambda)$ is asymptotically normal distributed with mean zero and variance

$$\frac{R(K)\left(\lambda^2 f_P(\theta_\lambda) + (1-\lambda)^2\alpha^2 f_N(\theta_\lambda)\right)}{\left(U''_\lambda(\theta_\lambda)\right)^2} = V^2_\lambda$$

□

Proof. To prove the Theorem 2 we only need to check that $(V^2_{n,\lambda} - V^2_\lambda) \to_P 0$. From the regularity assumptions (conditions A_1, A_2 and A_5) and the convergence rates of the used bandwidth (A_6), the already known kernel estimator convergence properties, we can write for $t, s \in \mathbb{R}$,

$$\begin{aligned}\widetilde{f}_n(X,t) =& f(s) + \mathcal{O}(t-s) + \mathcal{O}_P(n^{-2/5})\\ \widetilde{f}'_n(X,t) =& f'(s) + \mathcal{O}(t-s) + \mathcal{O}_P(n^{-2/7})\end{aligned}$$

Therefore

$$\begin{aligned}V^2_{n,\lambda} &= \frac{R(K)\left(\lambda^2\widetilde{f}_n(X_P,\widetilde{\theta}_\lambda) + (1-\lambda)^2\alpha_n^2\widetilde{f}_m(X_N,\widetilde{\theta}_\lambda)\right)}{\left(\lambda\widetilde{f}'_n(X_P,\widetilde{\theta}_\lambda) + (1-\lambda)\widetilde{f}'_m(X_N,\widetilde{\theta}_\lambda)\right)^2}\\ &= \frac{R(K)\left(\lambda^2\widetilde{f}_n(X_P,\widetilde{\theta}_\lambda) + (1-\lambda)^2\alpha_n^2\widetilde{f}_m(X_N,\widetilde{\theta}_\lambda)\right)}{\left(\lambda f'_P(\theta_\lambda) + (1-\lambda)f'_N(\theta_\lambda)\right)^2}\\ &\qquad + \mathcal{O}(\widetilde{\theta}_\lambda - \theta_\lambda) + \mathcal{O}_P(n^{-2/7}) + \mathcal{O}_P(m^{-2/7})\end{aligned}$$

therefore

$$\begin{aligned}(V^2_{n,\lambda} - V^2_\lambda) =& \\ &\frac{R(K)[\lambda^2(\widetilde{f}(X_P,\widetilde{\theta}_\lambda) - f_P(\theta_\lambda)) + (1-\lambda)^2\alpha_n^2(\widetilde{f}(X_N,\widetilde{\theta}_\lambda) - f_N(\theta_\lambda))]}{\left(\lambda f'_P(\theta_\lambda) + (1-\lambda)f'_N(\theta_\lambda)\right)^2}\\ &\qquad + \mathcal{O}(\widetilde{\theta}_\lambda - \theta_\lambda) + \mathcal{O}_P(n^{-2/7}) + \mathcal{O}_P(m^{-2/7})\\ &= \frac{R(K)[\lambda^2(\mathcal{O}(\widetilde{\theta}_\lambda - \theta_\lambda) + \mathcal{O}_P(n^{-2/5}))]}{\left(\lambda f'_P(\theta_\lambda) + (1-\lambda)f'_N(\theta_\lambda)\right)^2}\\ &\quad + \frac{R(K)[(1-\lambda)^2\alpha_n^2(\mathcal{O}(\widetilde{\theta}_\lambda - \theta_\lambda) + \mathcal{O}_P(m^{-2/5}))]}{\left(\lambda f'_P(\theta_\lambda) + (1-\lambda)f'_N(\theta_\lambda)\right)^2}\\ &\qquad + \mathcal{O}(\widetilde{\theta}_\lambda - \theta_\lambda) + \mathcal{O}_P(n^{-2/7}) + \mathcal{O}_P(m^{-2/7}) \to_P 0\end{aligned}$$ □

Revista Colombiana de Estadística
Junio 2011, volumen 34, no. 1, pp. 147 a 176

Nuevas cartas de control basadas en la distribución Birnbaum-Saunders y su implementación

New Control Charts Based on the Birnbaum-Saunders Distribution and their Implementation

Víctor Leiva[1,a], Gerson Soto[2,b], Enrique Cabrera[1,3,c], Guillermo Cabrera[4,d]

[1]Departamento de Estadística, Universidad de Valparaíso, Valparaíso, Chile
[2]Instituto Nacional de Estadísticas, Santiago, Chile
[3]Instituto de Estadística, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
[4]Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile

Resumen

El modelo Birnbaum-Saunders (BS) es una distribución de vida que tiene propiedades interesantes y aplicaciones en varias áreas. Esto la ha convertido en un foco de investigación importante en el último tiempo. Sin embargo, la suma de variables aleatorias independientes BS (BSsum) no sigue una distribución BS. A través de la distribución BSsum, se pueden monitorear los tiempos de vida de productos expuestos a fallas mediante una carta de control de calidad. Los procedimientos clásicos de cartas de control suponen normalidad en la distribución de los datos. No obstante, una de las características principales de los tiempos de vida es que éstos generalmente siguen distribuciones asimétricas. Por tanto, si se quiere monitorear estos tiempos, se deben considerar cartas de control para distribuciones asimétricas, como es el caso de la distribución BS. El monitoreo de los tiempos de vida se realiza generalmente mediante el tiempo acumulado o el tiempo promedio hasta la ocurrencia de cierto número de fallas. Entonces, usando la distribución BSsum, desarrollamos, implementamos y aplicamos una nueva metodología para cartas de control basada en la distribución BS.

***Palabras clave*:** lenguaje de computación R, métodos de verosimilitud, tiempos de vida.

[a]Profesor titular. E-mail: victor.leiva@uv.cl
[b]Ingeniero estadístico. E-mail: gerson.soto@ine.cl
[c]Profesor adjunto. E-mail: enrique.cabrera@uv.cl
[d]Profesor permanente. E-mail: guillermo.cabrera@ucv.cl

Abstract

The Birnbaum-Saunders (BS) model is a life distribution with interesting properties and applications in several fields. This has transformed the BS model in an important research focus in recent decades. However, the sum of BS (BSsum) independent random variables does not follow a BS distribution. By means of the BSsum distribution, we can monitor the lifetime of products subject to failures using a quality control chart. Classic procedures for control charts assume normality in the distribution of the data. Nevertheless, one of the main characteristics of the lifetimes is that them generally follow asymmetric distributions. Therefore, if we want to monitor these lifetimes, we must consider control charts for asymmetric distributions, such as it is the case of the BS distribution. The monitoring of the lifetimes is carried out generally by the accumulated lifetime or the lifetime average until than a number of failures occurs. Thus, by using the BSsum distribution, we develop, implement and apply a new methodology for control charts based on the BS distribution.

***Key words*:** Lifetime data, Likelihood methods, R computer language.

1. Introducción

Motivados por un problema de vibración en aviones comerciales, lo que produce fatiga de materiales, y basados en la Ley de Miner o de daño acumulativo, Birnbaum & Saunders (1969*b*) desarrollaron un modelo probabilístico asimétrico de dos parámetros que describe los tiempos de falla (o de vida) de especímenes de materiales expuestos a cargas cíclicas bajo estrés y que producen fatiga; ver Johnson, Kotz & Balakrishnan (1995, pp. 651-663) y Sanhueza, Leiva & Balakrishnan (2008). La distribución Birnbaum-Saunders (BS) está relacionada a la distribución normal y describe el tiempo de falla que transcurre hasta que cierta clase de daño acumulativo excede un umbral de resistencia máximo, provocando la falla del material. Desmond (1985) proporcionó una derivación más general de esta distribución basada en un modelo biológico y consolidó una justificación física para el uso de esta distribución relajando los supuestos hechos por Birnbaum & Saunders (1969*b*) indicando, por ejemplo, que la distribución BS puede también obtenerse desde modelos distintos al normal. Basados en este principio y usando argumentos estadísticos, Díaz-García & Leiva (2005) generalizaron la distribución BS obteniendo su densidad, algunas propiedades y varios casos particulares de la distribución Birnbaum-Saunders generalizada (BSG). Esta generalización permite obtener el modelo BS desde una clase de distribuciones simétricas en la recta de los números reales produciendo una gran flexibilidad y una estimación de parámetros robusta.

La distribución BS fue implementada por Leiva, Hernández & Riquelme (2006) en el software estadístico R (`http://www.R-project.org`) en un paquete llamado `bs`; ver R Development Core Team (2009). En este paquete, se pueden encontrar funciones para calcular probabilidades, estimar parámetros, generar números aleatorios y hacer estudios de bondad de ajuste y análisis de confiabilidad. Por otro lado, Barros, Paula & Leiva (2009) desarrollaron un paquete R llamado `gbs` para la

distribución BSG. Los tres generadores de números aleatorios BS y BSG existentes pueden revisarse en Leiva, Sanhueza, Sen & Paula (2008).

La distribución BS tiene características interesantes, pero no posee la propiedad reproductiva. Así, la suma de variables aleatorias (v.a.) BS (que llamaremos BS-sum) no sigue una distribución BS. Raaijmakers (1980, 1981) halló la distribución del tiempo de vida útil de un sistema "standby" cuyas componentes fallan independientemente, de acuerdo con una distribución BS. La vida útil de tal sistema corresponde a la suma de los tiempos de falla de cada componente de este sistema siguiendo una distribución BSsum. Raaijmakers (1980) se basó en la transformada de Laplace para hallar la distribución de la suma de v.a. independientes (convolución). Este resultado permite relacionar la función de densidad de probabilidades (f.d.p.) de la vida útil de cada componente a la f.d.p. de la vida útil de todo el sistema.

Debido a que algunos procesos industriales evidencian la presencia de observaciones con un comportamiento asimétrico, distribuciones que tengan este patrón son las adecuadas para analizar características de calidad de este tipo de procesos. Como se mencionó, una de estas distribuciones es el modelo BS, la que es principalmente útil para modelar tiempos de vida de productos expuestos a fallas. Este tipo de tiempos, debería monitorearse mediante cartas de control basadas en distribuciones asimétricas. A la fecha, existen pocos trabajos en esta dirección, menos aún basados en la distribución BS, salvo el estudio desarrollado por Lio & Park (2008) para monitorear percentiles de esta distribución.

Conocer la distribución BSsum es particularmente importante en estadística industrial, por ejemplo, para implementar cartas de control para el tiempo de fallas de productos siguiendo una distribución BS. Al conocer la distribución BSsum, se pueden obtener de forma exacta los límites de control inferior (LCI) y superior (LCS) de la carta de control. Aunque el modelo BS se ha usado ampliamente como distribución de vida en ingeniería, trabajos recientes han aplicado este modelo a otras áreas, permitiendo considerarlo como una distribución probabilística general más que restringirla solamente al modelamiento de datos de tiempos de vida. Así, aunque una metodología de cartas de control basada en la distribución BS puede ser más apropiada para tiempos de falla, esta metodología puede usarse para cualquier v.a. positiva. Para más detalles acerca de nuevas aplicaciones del modelo BS, ver Leiva, Sanhueza & Saunders (2009). Para una revisión de distribuciones de vida, ver Marshall & Olkin (2007) y Saunders (2007). Para aplicaciones de la distribución BS en áreas diferentes a la ingeniería, ver Leiva, Barros, Paula & Galea (2007), Podlaski (2008), Barros, Paula & Leiva (2008), Leiva, Barros, Paula & Sanhueza (2008), Leiva, Sanhueza & Angulo (2009), Bhatti (2010), Vilca, Sanhueza, Leiva & Christakos (2010) y Leiva, Vilca, Balakrishnan & Sanhueza (2010).

El supuesto básico para implementar cartas de control clásicas para la media de un proceso es que la v.a. a controlar debe seguir una distribución normal; ver Duncan (1996). Este supuesto no siempre se cumple; ver Schoonhoven & Does (2010). La falta de normalidad es particularmente frecuente cuando se estudian tiempos de vida de productos sujetos a fallas. Como la información que aportan

estos tiempos es crucial para la estabilidad de un proceso productivo, se deben considerar cartas de control para distribuciones asimétricas, las que, como se mencionó, no han sido ampliamente desarrolladas. Algunas cartas de este tipo fueron propuestas por Cheng & Xie (2000) y Surucu & Sazak (2009) para las distribuciones lognormal y Weibull, respectivamente; ver también Vargas & Montaño (2005). Como se mencionó, una carta de control para la distribución BS fue propuesta por Lio & Park (2008), quienes crearon una metodología para el monitoreo de los percentiles de la distribución BS. Cartas de control para la media y el tiempo acumulado de procesos de producción gobernados por una distribución BS no han sido propuestas.

El objetivo principal de este artículo es desarrollar una metolodogía de cartas de control para la media y el tiempo acumulado de fallas de un proceso basadas en la distribución BS. Específicamente, en este artículo: i) introducimos la distribución BSsum; ii) desarrollamos cartas de control basadas en esta distribución; iii) implementamos la distribución BSsum en el software R, así como cartas de control basadas en la distribución BS; iv) llevamos a cabo un estudio de simulación que detecta la sensibilidad de la metolodogía propuesta a salidas de control del proceso productivo, y v) aplicamos los resultados obtenidos a datos industriales reales. Cabe destacar que en este artículo discutimos también los aspectos computacionales de este trabajo. Específicamente, en dos apéndices localizados después de las conclusiones, discutimos una nueva versión del paquete `bs` llamada `bs` 2.0 y un paquete para la distribución BSsum llamado `bssum` que contiene las cartas de control BS. El paquete `bs` 2.0 es más completo que el `bs` 1.0, ya que incorpora, por ejemplo, funciones para datos censurados. Este tipo de datos no son analizados aquí, ya que una metodología como la desarrollada en este trabajo no puede aplicarse directamente cuando existen datos censurados. Esto nos propone un desafío para un futuro trabajo. El lector interesado en este tipo de métodos, puede revisar Steiner & Mackay (2000), Zhang & Chen (2004) y Vargas & Montaño (2005). El paquete `bs` 1.0 puede descargarse desde CRAN (`http://CRAN.R-project.org`), mientras que los paquetes `bs` 2.0 y `bssum` puede descargarse desde `http://staff.deuv.cl/leiva/archivos`.

El resto de este artículo está organizado como sigue. En la sección 2 proporcionamos algunos elementos preliminares de la distribución BS útiles para desarrollar la metodología propuesta. En la sección 3, introducimos la distribución BSsum. En la sección 4 discutimos una carta de control para percentiles de la distribución BS y desarrollamos una nueva metodología para cartas de control basadas en esta distribución. En esta sección, llevamos a cabo también un estudio de simulación y dos ejemplos con datos industriales reales. En la sección 5, bosquejamos algunas conclusiones de este trabajo. En la parte final, en dos apéndices analizamos algunas características de los paquetes `bs` 1.0, `bs` 2.0 y `bssum`.

2. Preliminares

En esta sección proporcionamos algunas propiedades y características de la distribución BS.

2.1. Distribución Birnbaum-Saunders

Una v.a. T con distribución BS tiene dos parámetros, uno de forma (α) y otro de escala (β), con β siendo además un parámetro de posición, pues corresponde a la mediana. Esto se denota por $T \sim \text{BS}(\alpha, \beta)$. Las variables aleatorias BS y normal estándar, denotadas respectivamente por T y Z, están relacionadas mediante

$$T = \beta \left[\frac{\alpha Z}{2} + \sqrt{\left\{ \frac{\alpha Z}{2} \right\}^2 + 1} \right]^2 \quad \text{y} \quad Z = \frac{1}{\alpha} \left[\sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}} \right] \tag{1}$$

Esto nos permite llegar a

$$W = \frac{1}{\alpha^2} \left[\frac{T}{\beta} + \frac{\beta}{T} - 2 \right] \sim \chi^2(1) \tag{2}$$

cuyo resultado es útil para bondad de ajuste y detectar datos atípicos mediante la distancia de Mahalanobis. Si $T \sim \text{BS}(\alpha, \beta)$, entonces las siguientes características se cumplen. La f.d.p. de T es

$$f_T(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2\alpha^2} \left[\tfrac{t}{\beta} + \tfrac{\beta}{t} - 2 \right] \right) \frac{t^{-3/2}[t+\beta]}{2\alpha\sqrt{\beta}}, \quad t > 0,\ \alpha > 0,\ \beta > 0 \tag{3}$$

La función de distribución acumulativa (f.d.a.) de T es

$$F_T(t) = \mathbb{P}(T \leq t) = \Phi\left(\tfrac{1}{\alpha} \left[\sqrt{\tfrac{t}{\beta}} - \sqrt{\tfrac{\beta}{t}} \right] \right) \quad t > 0 \tag{4}$$

donde $\Phi(\cdot)$ es la f.d.a. de $Z \sim \text{N}(0,1)$. La función cuantil (f.q.) de T (cuantil q-ésimo) es

$$t(q) = F_T^{-1}(q) = \frac{\beta}{4} \left[\alpha z(q) + \sqrt{\{\alpha z(q)\}^2 + 4} \right]^2, \quad 0 < q < 1 \tag{5}$$

donde $F_T^{-1}(\cdot)$ es la función inversa de $F_T(\cdot)$ y $z(q)$ es el cuantil q-ésimo de $Z \sim \text{N}(0,1)$. Por tanto, desde (5), $t(0.5) = \beta$ y así β es la mediana del modelo BS, tal como fue mencionado. Algunas propiedades de $T \sim \text{BS}(\alpha, \beta)$ son: (i) $cT \sim \text{BS}(\alpha, c\beta)$, con $c > 0$, y (ii) $1/T \sim \text{BS}(\alpha, 1/\beta)$. Estas propiedades indican que la distribución BS pertenece a las familias de transformaciones de escala y cerradas bajo recíproco. El momento k-ésimo de T está dado por

$$\text{E}\left[T^k\right] = \beta^k \sum_{j=0}^{k} \binom{2k}{2j} \sum_{i=0}^{j} \binom{j}{i} \frac{(2k-2j+2i)!}{2^{k-j+i}(k-j+k)!} \left[\frac{\alpha}{2} \right]^{2[k-j+i]}, \quad k = 1, 2 \ldots \tag{6}$$

La media, la varianza y los coeficientes de variación (CV), sesgo (CS) y curtosis (CC) de T están dados respectivamente por

$$\text{E}[T] = \frac{\beta}{2}\left[2 + \alpha^2\right], \quad \text{V}[T] = \frac{\beta^2}{4}\left[5\alpha^4 + 4\alpha^2\right], \quad \text{CV}[T] = \frac{\sqrt{5\alpha^4 + 4\alpha^2}}{\alpha^2 + 2} \tag{7}$$

$$\mathrm{CS}[T] = \frac{44\alpha^3 + 24\alpha}{[5\alpha^2+4]^{3/2}} \quad \text{y} \quad \mathrm{CC}[T] = 3 + \frac{558\alpha^4 + 240\alpha^2}{[5\alpha^2+4]^2} \tag{8}$$

Entonces, desde (8), $\mathrm{CS}[T] \to 0$ y $\mathrm{CC}[T] \to 3$, cuando $\alpha \to 0$, es decir, cuando α es pequeño, el sesgo y la curtosis de la distribución BS se acercan al sesgo y la curtosis de la distribución normal degenerando en el parámetro β. Los CV, CS y CC son invariantes bajo escala, es decir, estos coeficientes son funciones independientes del parámetro de escala β. Por otra parte, si T tiene una distribución BS con parámetros α y β y ya que $1/T$ tiene también una distribución BS con los correspondientes parámetros α y $1/\beta$, respectivamente, entonces se tiene que

$$\mathrm{E}\left[\frac{1}{T}\right] = \frac{1}{2\beta}\left[\alpha^2+2\right] \quad \text{y} \quad \mathrm{V}\left[\frac{1}{T}\right] = \frac{1}{4\beta^2}\left[5\alpha^4+4\alpha^2\right] \tag{9}$$

La figura 1 muestra el comportamiento de la f.d.p. de la distribución BS para algunos valores del parámetro de forma α. Note que a medida que α decrece, la f.d.p. es aproximadamente simétrica. Gráficos para diferentes valores de β no se han considerado, porque al ser éste un parámetro de escala y de posición, la forma de la f.d.p. no cambia al variar este parámetro. Así, sin pérdida de generalidad, hemos considerado $\beta = 1$ en esta figura.

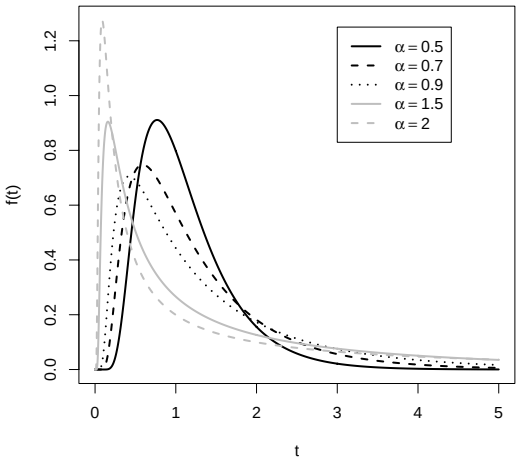

FIGURA 1: f.d.p. de $T \sim \mathrm{BS}(\alpha, \beta = 1.0)$.

2.2. Métodos de estimación para la distribución BS

Se han planteado varios estimadores para los parámetros de forma α y de escala β de la distribución BS. A continuación se presentan tres métodos para estimar estos parámetros. Estos métodos son el de verosimilitud máxima (VM), uno gráfico que permite la estimación de α y β por mínimos cuadrados y otro de momentos modificado; ver Birnbaum & Saunders (1969*a*), Pérez & Correa (2008), Chang & Tang (1994) y Ng, Kundu & Balakrishnan (2003). Para cada uno de estos

métodos, considere que $T_1, \ldots, T_n$ es una muestra aleatoria de tamaño n, donde $T_i \sim \text{BS}(\alpha, \beta)$, para $i = 1, \ldots, n$.

2.2.1. Método de verosimilitud máxima

La función de log-verosimilitud para α y β está dada por

$$\ell(\alpha, \beta) \propto \frac{n}{\alpha^2} - n\log(\alpha) - \frac{n}{2}\log(\beta) + \sum_{i=1}^{n}\left\{\log(t_i + \beta) - \frac{1}{2\alpha^2}\left[\frac{t_i}{\beta} + \frac{\beta}{t_i}\right]\right\} \quad (10)$$

Derivando (10) con respecto a los parámetros α y β, se obtiene, respectivamente,

$$\frac{\partial \ell(\alpha, \beta)}{\partial \alpha} = -\frac{2\,n}{\alpha^3} - \frac{n}{\alpha} + \sum_{i=1}^{n}\left\{\frac{1}{\alpha^3}\left[\frac{t_i}{\beta} + \frac{\beta}{t_i}\right]\right\} \quad \text{y} \quad (11)$$

$$\frac{\partial \ell(\alpha, \beta)}{\partial \beta} = -\frac{n}{2\beta} + \sum_{i=1}^{n}\left\{\frac{1}{t_i + \beta}\right\} + \frac{1}{2\alpha^2}\sum_{i=1}^{n}\left\{\frac{t_i}{\beta^2} - \frac{1}{t_i}\right\} \quad (12)$$

Al igualar ecuaciones de log-verosimilitud dadas en (11) y (12) a cero, se obtiene

$$\widehat{\alpha}^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{t_i}{\widehat{\beta}} + \frac{\widehat{\beta}}{t_i}\right\} - 2 \quad \text{y} \quad (13)$$

$$\widehat{\alpha}^2 = \frac{2\widehat{\alpha}^2\widehat{\beta}}{n}\sum_{i=1}^{n}\left\{\frac{1}{t_i + \widehat{\beta}}\right\} + \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{t_i}{\widehat{\beta}} - \frac{\widehat{\beta}}{t_i}\right\} \quad (14)$$

Considere la media aritmética y la media armónica de un conjunto de números positivos, digamos $t_1, \ldots, t_n$, respectivamente, dadas por

$$s = \frac{1}{n}\sum_{i=1}^{n} t_i \quad \text{y} \quad r = \frac{1}{\left[\frac{1}{n}\sum_{i=1}^{n} 1/t_i\right]} \quad (15)$$

Considere además una función media armónica $K(x) = n/\sum_{i=1}^{n}\{1/(x+t_i)\}$, para $x > 0$. Sustituyendo s y r y $K(x)$ en (13) y (14), se obtiene

$$\widehat{\alpha}^2 = \frac{s}{\widehat{\beta}} - \frac{\widehat{\beta}}{r} + \frac{2\widehat{\alpha}^2\widehat{\beta}}{K(\widehat{\beta})} \quad \text{y} \quad (16)$$

$$\frac{\widehat{\beta}}{r} = 1 + \frac{\widehat{\alpha}^2\widehat{\beta}}{K(\widehat{\beta})} \quad (17)$$

Si se sustituye (16) en (17), se obtiene $\widehat{\beta}K(\widehat{\beta})/r = K(\widehat{\beta}) + \widehat{\beta}[s/\widehat{\beta} + \widehat{\beta}/r - 2]$, de modo que $\widehat{\beta}^2 - \widehat{\beta}[2r - K(\widehat{\beta})] + r[s + K(\widehat{\beta})] = 0$. Ahora, tomando $g(x) = x^2 - x[2r - K(x)] + r[s + K(x)]$, se tiene que el estimador de VM de β, digamos $\widehat{\beta}$, es la solución de $g(x) = 0$, la cual es única ya que $0 < x < \infty$ y $s > \widehat{\beta} > r$; ver

Birnbaum & Saunders (1969*a*). Así, el estimador de VM de α, digamos $\widehat{\alpha}$, queda expresado en términos de $\widehat{\beta}$ mediante

$$\widehat{\alpha} = \left[\frac{s}{\widehat{\beta}} + \frac{\widehat{\beta}}{r} - 2\right]^{1/2} \tag{18}$$

mientras que $\widehat{\beta}$ debe obtenerse usando un método numérico iterativo por lo que se puede usar la mediana muestral como valor de partida o la estimación media-media para β propuesta por Birnbaum & Saunders (1969*a*) y dada por $\overline{\overline{\beta}} = [sr]^{1/2}$.

2.2.2. Método de mínimos cuadrados

Desde (4) se tiene que

$$t = \beta + \alpha\sqrt{\beta}\sqrt{t}\,\Phi^{-1}(F_T(t)) \tag{19}$$

Note que (19) no es una función lineal en t, lo que es esencial para gráficos de probabilidad de bondad de ajuste. Se puede eludir este problema reemplazando $p = \sqrt{t}\,\Phi^{-1}(F_T(t))$ en (19), dando como resultado la forma lineal $y = a + bx$, donde la ordenada es $y = t$, el intercepto $a = \beta$, la abscisa es $x = p$ y la pendiente es $b = \alpha\sqrt{\beta}$, es decir, $t = \beta + \alpha\sqrt{\beta}\,p$. Luego, se debe graficar t_i versus p_i, para $i = 1, \ldots, n$, donde $p_i = \sqrt{t_i}\,\Phi^{-1}(F(t_i))$. Existen varias opciones para obtener $F_n(t_i)$. Aquí se utiliza el rango medio, es decir, $F_n(t_i) = [i - 0.3]/[n + 0.4]$. Así, este método puede usarse como una herramienta de bondad de ajuste para detectar si un conjunto de datos podría o no provenir de una distribución BS. En particular, si el gráfico de t_i versus p_i, para $i = 1, \ldots, n$, sigue aproximadamente una línea recta, entonces esto es una indicación de que la muestra aleatoria podría provenir de una distribución BS. Chang & Tang (1994) obtuvieron estimaciones del intercepto a y de la pendiente b mediante el método de cuadrados mínimos, lo que es denotado por $\widetilde{a}$ y $\widetilde{b}$, respectivamente. Ellos indicaron que una vez obtenidas estas estimaciones, uno puede determinar los parámetros de la distribución BS mediante las expresiones

$$\widetilde{\beta} = \widetilde{a} \quad \text{y} \quad \widetilde{\alpha} = \frac{\widetilde{b}}{\sqrt{\widetilde{a}}}$$

2.2.3. Método de momentos modificados

En el caso de distribuciones de dos parámetros, sus estimadores de momentos se obtienen igualando los primeros dos momentos poblacionales a sus momentos muestrales correspondientes. Por tanto, los estimadores de momentos de α y β de la distribución BS se pueden obtener como las soluciones de α y β de estas ecuaciones de momentos. Si el CV muestral es mayor que $\sqrt{5}$, entonces el estimador de momentos de α no existe. Por el contrario, si el CV es menor que $\sqrt{5}$, el estimador de momentos sí existe, pero el estimador de β no es único. En lugar de utilizar expresiones para los momentos de la distribución BS dadas en (7), Ng et al. (2003) propusieron utilizar las expresiones del lado izquierdo de las ecuaciones (7)

y (9) e igualarlas a sus momentos muestrales correspondientes. En este caso, se tienen las ecuaciones de momentos

$$s = \beta\left[1 + \frac{1}{2}\alpha^2\right] \quad \text{y} \quad \frac{1}{r} = \frac{1}{\beta}\left[1 + \frac{1}{2}\alpha^2\right] \tag{20}$$

Resolviendo las ecuaciones dadas en (20) para α y β, se obtienen las estimaciones de momentos modificados para α y β denotadas por $\overline{\alpha}$ y $\overline{\beta}$ como

$$\overline{\alpha} = \left[2\left\{\left(\frac{s}{r}\right)^{1/2} - 1\right\}\right]^{1/2} \quad \text{y} \quad \overline{\beta} = [s\,r]^{1/2} \tag{21}$$

2.3. El modelo BS como una distribución de vida

Una de las principales herramientas del análisis de confiabilidad es la función tasa de riesgo (f.r.) o tasa de fallas. En la ciencia actuarial, por ejemplo, la f.r. es la probabilidad por año que una persona de una edad determinada muera en el instante siguiente, expresada como una tasa de mortalidad por año. Una propiedad de la f.r. es que hace posible caracterizar el comportamiento de las distribuciones de vida. Este tipo de distribuciones describen los tiempos de vida de unidades hasta la ocurrencia de una falla, pudiendo estas unidades ser, por ejemplo, sistemas, componentes, órganos, personas o productos; para más detalles acerca de distribuciones de vida, ver Marshall & Olkin (2007) y Saunders (2007). Por ejemplo, modelos de probabilidad con densidades con formas similares en algunos casos tienen f.r. distintas. Éste es el caso de las distribuciones BS, Gaussiana inversa (GI), lognormal, gamma y Weibull; ver Balakrishnan, Leiva & López (2007). Para revisar algunas relaciones entre las distribuciones BS y GI, ver Balakrishnan, Leiva, Sanhueza & Cabrera (2009). Por tanto, es muy importante tener en cuenta la f.r. para seleccionar o descartar distribuciones, incluso cuando no se estén analizando datos de tiempos de vida. Por ejemplo, si la f.r. es creciente, la probabilidad de sobrevivencia disminuye con el tiempo. Este tipo de f.r. puede considerarse para modelar tiempos de vida cuando el desgaste o envejecimiento está presente. Sin embargo, una f.r. creciente puede ser inadecuada cuando se modela mortalidad humana, ya que, en este caso, el riesgo está disminuyendo, se estabiliza durante un período y luego éste es cada vez mayor, de manera que la forma que se debe suponer para la f.r. es aquella conocida como gráfico de bañera (en inglés *bathtub*). Otro tipo de f.r. tiene forma de bañera invertida, la que aparece con frecuencia cuando se analizan datos de fatiga y de contaminación ambiental. Una f.r. decreciente puede ser adecuada para modelar la sobrevivencia después de una cirugía exitosa, cuando hay un riesgo inicial alto causado, por ejemplo, por una infección o hemorragia, pero este riesgo comienza luego a disminuir a medida que el paciente se va recuperando. Por último, existen casos de f.r. constante, la que se observa, por ejemplo, cuando se estudia la duración de chips de computadores, los que no envejecen con el tiempo.

Distribuciones de vida que modelan los diferentes tipos de riesgo mencionados arriba pueden escogerse mediante la tasa de falla. Entonces, establecer la forma de la f.r. que un conjunto de datos tiene es de vital importancia en análisis de

confiabilidad para determinar la distribución que estos datos pueden tener. La confiabilidad o sobrevivencia se define como la probabilidad de que una entidad sobreviva o funcione adecuadamente en un período de tiempo t; ver Lawless (1982). Así, si T es el tiempo de vida de una unidad, entonces la función de confiabilidad (o sobrevivencia), lo cual abreviamos como f.c., al tiempo t está dada por $R_T(t) = \mathbb{P}(T \geq t) = 1 - F_T(t) = \int_t^\infty f_T(s)\, ds$, para $t > 0$, donde $f_T(\cdot)$ y $F_T(\cdot)$ son la f.d.p. y la f.d.a. de T, respectivamente. La f.r. es un indicador muy importante en el análisis de tiempos de vida y se define como la tasa instantánea de fallas en el tiempo $t + \triangle t$, dado que la unidad ha sobrevivido a t. Entonces, la f.r. de una v.a. T es

$$h_T(t) = \lim_{\triangle t \to 0} \frac{\mathbb{P}(t < T < \Delta t | T > t)}{\Delta t} = \frac{f_T(t)}{R_T(t)} = -\frac{d \log(R_T(t))}{dt}, \quad t > 0 \quad (22)$$

donde $0 < R_T(t) < 1$. La f.d.p. puede estimarse fácilmente en forma empírica a través de un histograma. Sin embargo, una estimación como ésa para la f.r. no es fácil. Una herramienta que permite determinar la forma de la f.r. es el gráfico TTT (en inglés, *total time on test*). Específicamente, si T es una v.a. positiva, sus funciones TTT y TTT escalada son $H_T^{-1}(u) = \int_0^{F_T^{-1}(u)} R_T(y)\, dy$ y $W_T(u) = H_T^{-1}(u)/H_T^{-1}(1)$, respectivamente, para $0 < u < 1$. La función TTT escalada $W_T(\cdot)$ puede aproximarse en forma empírica mediante

$$W_n\,(k/n) = \frac{\sum_{i=1}^{k} T_{(i)} + [n-k]\, T_{(k)}}{\sum_{i=1}^{n} T_{(i)}}, \quad k = 1, \ldots, n \quad (23)$$

donde $T_{(i)}$ es el estadístico de orden i-ésimo. Esto permite construir el gráfico TTT empírico mediante los puntos $\big[k/n, W_n(k/n)\big]$. Así, por medio de este gráfico se puede detectar el tipo de f.r. que los datos tienen. En la figura 8 (izquierda) es posible observar diferentes formas teóricas para el gráfico TTT, las cuales están asociadas con la f.r. respectiva. Si este gráfico arroja una curva cóncava, la f.r. es creciente. Por el contrario, si esta curva es convexa, la f.r. es decreciente. Ahora bien, si el gráfico TTT produce una curva que es primero cóncava y luego convexa, la f.r. tiene forma de bañera invertida. Si esta curva es convexa y luego concava, la f.r. tiene forma de bañera. Por último, si el gráfico TTT produce una línea recta, la f.r. es constante, como es el caso de la distribución exponencial. Así, el gráfico TTT nos puede dar una indicación del tipo de distribución asociada con los datos; ver la figura 2.

El modelo BS ha sido usado ampliamente como distribución de vida debido a la justificación física que lo originó; ver Birnbaum & Saunders (1969*b*) y Johnson et al. (1995, pp. 651-652). Si $T \sim \mathrm{BS}(\alpha, \beta)$, entonces la f.r. de T es

$$h_T(t) = \frac{\phi\big([1/\alpha][\sqrt{t/\beta} - \sqrt{\beta/t}]\big)\, t^{-3/2}[t+\beta]}{\Phi\big([1/\alpha][\sqrt{t/\beta} - \sqrt{\beta/t}]\big)\, 2\,\alpha\,\sqrt{\beta}} \quad t > 0 \quad (24)$$

donde $\phi(\cdot)$ es la f.d.p. N(0,1). La figura 3 (izquierda) muestra que la f.c. BS decrece a medida que α aumenta. La figura 3 (derecha) indica que la f.r. BS tiende a ser creciente a medida que α decrece.

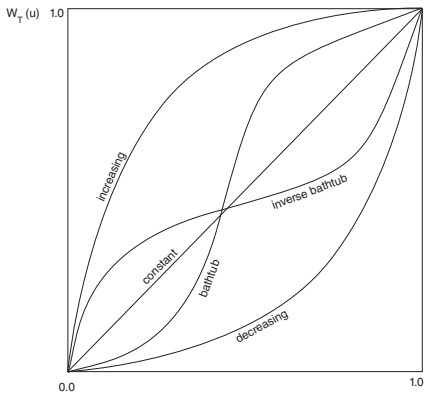

FIGURA 2: Curvas TTT para distribuciones con la f.r. indicada.

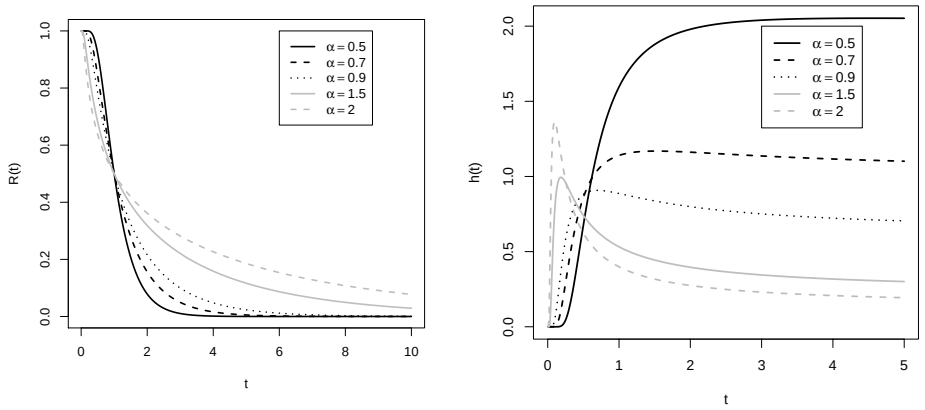

FIGURA 3: f.c. (izquierda) y f.r. (derecha) BS para diferentes valores de α.

3. Distribución BSsum

En esta sección introducimos algunas propiedades y características de la distribución BSsum.

3.1. Vida útil de un sistema de k componentes independientes

Un sistema se muestra consistente cuando la primera unidad falla y la segunda unidad, sin demora, se hace cargo de la operación. Este proceso continúa hasta que todas las k unidades que componen el sistema han fallado. Esto se conoce como sistema "standby". Considere T_i como la vida útil de la unidad i-ésima en el sistema, para $i = 1, \ldots, k$, y f como la f.d.p. para todos los T_i, que se suponen independientes. Considere además que Y_k es la vida útil del sistema de k unidades

independientes y f_k es la f.d.p. de Y_k, para $k \in \mathbb{N}$. Claramente, se puede concluir que $Y_k = \sum_{i=1}^{k} T_i$. Es posible establecer una relación entre las funciones f y f_k por medio de sus transformadas de Laplace. Específicamente, para cualquier función g definida entre $(0, \infty)$, digamos Lg, cuando existe, es la transformada de Laplace definida como $Lg(s) = \int_0^\infty \exp(-s\,u)\,g(u)\,du$, para $s > 0$. La relación existente entre las funciones f y f_k es $Lf_k = [Lf]^k$. Por tanto, para obtener la distribución BSsum se debe calcular la transformada de Laplace de la f.d.p. BS, elevarla a la potencia k (donde k es el número de términos de la suma) y después calcular la transformada de Laplace inversa de esta potencia.

3.2. Distribución BSsum

Considere $f_{\alpha,\beta}$ como la f.d.p. BS con parámetros α y β definida en (3). Sin pérdida de generalidad, se puede tomar el parámetro de escala $\beta = 1$ y aplicar la transformada de Laplace a esta f.d.p. Así, la transformada de Laplace de $f_{\alpha,1}$ es

$$Lf_{\alpha,1}(s) = \frac{\exp\left(\frac{1}{\alpha^2}\right)}{2\alpha\sqrt{2\pi}} \left[\sqrt{\frac{\pi}{u}} + \sqrt{\frac{\pi}{a}}\right] \exp(-2\sqrt{ua}) \tag{25}$$

donde $u = s + a$ y $a = 1/2\alpha^2$. Se define la función $q(\cdot)$ como

$$q(s) = \frac{1}{2}\left[1 + \frac{1}{\sqrt{s}}\right] \exp\left(\frac{[1-\sqrt{s}]}{\alpha^2}\right) \tag{26}$$

Después de algunos cálculos se desprende que $Lf_{\alpha,1}(s) = r[1+2\alpha^2 s]$. Este resultado se debe elevar a la cantidad de términos que tiene la suma. Para esto se define $Q(s) = q(s)^k$ y $g = k/\alpha^2$. Entonces, se obtiene

$$H(s) = \frac{\exp(g)}{2^k}\left[1 + \frac{1}{\sqrt{s}}\right]^k \exp(-g\sqrt{s}) = \frac{\exp(g)}{2^k}\sum_{i=0}^{k}\binom{k}{i}\frac{1}{s^{i/2}}\exp(-g\sqrt{s}) \tag{27}$$

Luego, considere $\mu_i(t)$ con

$$L\mu_i(s) = \frac{1}{s^{i/2}}\exp(-\sqrt{s}) \tag{28}$$

y una función $z(t)$, tal que $Lz(s) = H(s)$. Sustituyendo las ecuaciones (27) y(28) en $Lz(s)$, se tiene que

$$Lz(s) = \frac{\exp(g)}{2^k}\sum_{i=0}^{k}\binom{k}{i} g^i L\mu_i(g^2 s) \tag{29}$$

lo cual implica

$$z(t) = \frac{\exp(g)}{2^k}\sum_{i=0}^{k}\binom{k}{i} g^{i-2}\mu_i\left(\frac{t}{g^2}\right)$$

Así, después del uso de algunos teoremas de la transformada de Laplace, la f.d.p. BSsum es

$$f_k(t) = \frac{a}{2^k} \exp(g - at) \sum_{i=0}^{k} \binom{k}{i} g^{i-2} \mu_i \left(\tfrac{a\,t}{g^2} \right) \tag{30}$$

donde $a = 1/2\alpha^2$ y $g = k/\alpha^2$. La f.d.a. de Y_k, que es la suma de v.a. BS, está dada por $F_k(t) = \mathbb{P}(Y_k \leq t) = \int_0^t f_k(u)\, du$, para $t > 0$, donde $f_k(t)$ está dada en (30), esto es,

$$F_k(t) = \frac{1}{2^k} \exp(g - at) \sum_{i=2}^{k} \left[l_i g^{i-2} \mu_i \left(\tfrac{a\,t}{g^2} \right) \right] + \Phi(\alpha^{-1} \varphi_k(t)), \quad t > 0 \tag{31}$$

donde $l_i = l_{i+2} - \binom{k}{i}$, con $l_{k+2} = l_{k+1} = 0$ y $\varphi_m(t) \approx \sqrt{t} - m/\sqrt{t}$, para $m = 1, 2, \ldots,$ y $i = 2, \ldots, k$.

3.3. Análisis de forma

La figura 4 (izquierda) muestra gráficas para la f.d.p. de la suma de cinco v.a. BS para diferentes valores del parámetro de forma α. Se ve que a medida que el parámetro de forma α crece, la forma de la distribución es más sesgada hacia la derecha. La figura 4 (derecha) muestra gráficas de la f.d.a. BSsum con parámetro de forma $\alpha = 1.5$ y escala $\beta = 1$. Se ve que a medida que se suman más v.a. la forma de la f.d.p. tiende a ser más simétrica.

4. Cartas de control para la distribución BS

En esta sección introducimos dos cartas de control basadas en la distribución BS. La primera de ellas es una carta de control bootstrap para percentiles BS disponible en la literatura y propuesta por Lio & Park (2008). La segunda es una nueva carta propuesta en este artículo.

4.1. Carta de control bootstrap para percentiles BS

Como se mencionó, la carta de control $\overline{T}$ clásica de Shewhart supone que los datos observados en el proceso provienen de una distribución normal. No obstante, cuando la distribución es desconocida o no normal, la distribución muestral del estimador del parámetro de interés para controlar el proceso puede no estar disponible en teoría y así los límites de control no pueden obtenerse. El método bootstrap se puede utilizar para construir los límites para el seguimiento de un determinado percentil de la distribución BS. Si bien el método funciona para cualquier percentil, la preocupación por parte de los ingenieros está especialmente en los percentiles inferiores, ya que un cambio a la baja en un percentil inferior de la distribución puede ser algo más grave que en otro tipo de percentil.

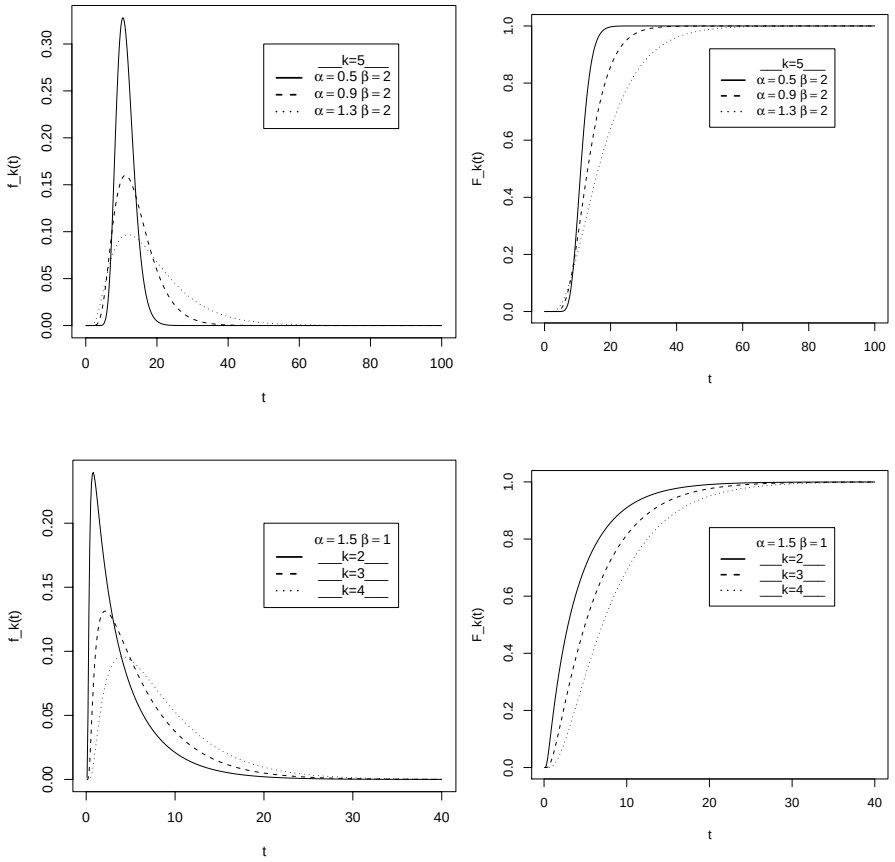

FIGURA 4: f.d.p. (izquierda) y f.d.a. (derecha) de la distribución BSsum para diferentes valores de α y k.

4.1.1. Construcción de la carta de control bootstrap

Para construir una carta de control para percentiles BS asumiendo que el proceso es estable y está bajo control se debe seguir un algoritmo que contiene los siguientes pasos de acuerdo a lo mencionado en Lio & Park (2008):

(B1) Obtener k muestras aleatorias independientes de tamaño n_j, para $j = 1, \ldots, k$, asumiendo que provienen de una distribución BS con parámetros α y β desconocidos, donde las observaciones de la muestra j-ésima se denotan como t_{ij}, para $i = 1, \ldots, n_j$.

(B2) Encontrar los estimaciones de VM de α y β con la muestra combinada de tamaño $N = \sum_{j=1}^{k} n_j$.

(B3) Generar observaciones $t_1^*, t_2^*, \ldots, t_m^*$ mediante el método bootstrap de tamaño m desde una distribución BS y utilizar las estimaciones obtenidas en el

paso (B2). Aquí, m es el tamaño de muestra, el cual se usa en el futuro como tamaño de los subgrupos.

(B4) Encontrar las estimaciones de VM de α y β basadas en la muestra generada en el paso (B3) y denotarlas como $\widehat{\alpha}^*$ y $\widehat{\beta}^*$.

(B5) Calcular las observaciones bootstrap $\widehat{t}_p^* = [\widehat{\beta}^*/4]\big[\widehat{\alpha}^* z(p) + [\widehat{\alpha}^{*^2} z(p)^2 + 4]^{1/2}\big]^2$, donde p representa el percentil que se quiere monitorear y $z(p)$ es el cuantil p-ésimo de la distribución N(0, 1), para la muestra bootstrap del paso (B3) y con las estimaciones obtenidas en el paso (B4).

(B6) Repetir los pasos (B3), (B4) y (B5) una gran cantidad de veces, obteniendo B (por ejemplo, $B = 10000$) muestras bootstrap de $\widehat{t}_p$, denotada por $\widehat{t}_{p1}^*, \ldots, \widehat{t}_{pB}^*$.

(B7) Hallar los percentiles $(\gamma/2) \times 100\,\%$ y $(1 - \gamma/2) \times 100\,\%$ usando la muestra bootstrap obtenida en paso (B6). Estos percentiles son los LCI y LCS, respectivamente.

Ejemplo 1. Para ilustrar el procedimiento de la carta de control BS bootstrap, considere el monitoreo de la resistencia a la ruptura de un material de aluminio. Para ello se han simulado datos de un proceso productivo de hojas de aluminio 6061-T6. Así, se simularon veinte grupos de tamaño cinco en forma independiente desde un proceso bajo control con distribución BS de parámetros $\alpha = 0.280$ y $\beta = 1.358$. Estos veinte subgrupos se presentan en la tabla 1. A partir de estos datos simulados se obtienen los límites de control bootstrap para el monitoreo del primer percentil de la distribución BS, realizando $B = 10000$ veces los pasos (B2)-(B7). Así, el LCI y el LCS de la carta de control para monitorear el primer percentil de la distribución BS con razón de falsa alarma $\delta = 0.003$ son LCI= 0.920 y LCS = 1.874, respectivamente; ver el libro de Duncan (1996) para más detalles del concepto de falsa alarma. Se supone ahora que el proceso se ha convertido en uno fuera de control. Entonces, se simularon veinte subgrupos de tamaño cinco correspondientes a un modelo BS cuyo parámetro de forma ha cambiado a $\alpha = 0.878$. Estos datos se presentan también en la tabla 1. La figura 5 muestra la carta de control generada mediante el método bootstrap para el monitoreo de la resistencia a la ruptura a través del primer percentil de la distribución BS. Los límites se obtienen a partir de la muestra bajo control. Se grafican los 20 subgrupos que se asumen como fuera de control. Esto es para evidenciar si la carta puede detectar rápidamente una señal fuera de control o no. El límite central (LC) también está incluido en la carta de control. Esto es para mostrar que la distribución muestral del primer percentil de la distribución BS es asimétrica. Se ve claramente que el proceso inmediatamente da señales de estar fuera de control, ya que sólo seis puntos de los veinte subgrupos están dentro de los límites y sólo uno de éstos se encuentra por sobre el límite central.

4.2. Cartas de control $\overline{T}$ para la distribución BS

Como se mencionó, una gran cantidad de datos generados desde procesos productivos siguen distribuciones asimétricas, tal como ocurre con los tiempos de vida. Cartas de control para estas distribuciones son de gran importancia para el

control adecuado de los procesos en cuestión. En la subsección 4.1.1, se mostró la única carta de control que se ha desarrollado para la distribución BS. Los pocos desarrollos de cartas de control para esta distribución se deben básicamente a que esta distribución no posee la propiedad reproductiva. Por tanto, cartas de control para el monitoreo del tiempo acumulado o promedio hasta la ocurrencia de m fallas para esta distribución no han sido implementadas, ya que éstas necesitan de la distribución BSsum.

Tabla 1: datos simulados de tiempos de falla por fractura debido a estrés de piezas de aluminio 6061-T6.

Subgrupo	Datos de un proceso bajo control					Datos de un proceso fuera de control				
1	1.140	1.430	1.076	2.113	1.489	0.280	8.532	0.627	0.767	0.414
2	1.080	1.556	1.669	1.595	1.247	0.519	1.290	1.689	2.051	1.592
3	2.066	1.514	1.142	0.738	1.857	2.344	0.717	1.384	0.807	4.217
4	1.341	1.352	1.767	1.708	1.603	0.416	3.147	0.565	1.083	0.753
5	1.754	1.689	1.387	0.784	1.615	1.511	1.688	0.369	2.564	0.536
6	1.337	1.300	0.902	1.188	1.526	0.219	0.772	1.173	0.807	1.649
7	1.981	1.320	1.513	1.338	0.926	1.436	3.991	0.957	1.344	1.907
8	1.209	1.216	1.336	1.845	1.680	1.726	0.771	5.854	1.631	3.047
9	1.297	1.265	1.650	1.586	1.121	0.988	0.854	7.217	3.632	1.221
10	1.115	1.504	1.683	1.316	1.736	4.562	1.186	4.448	1.575	0.595
11	1.518	1.145	1.494	0.992	2.022	4.474	0.436	2.346	0.444	0.568
12	2.346	1.226	1.015	1.592	1.308	0.546	1.198	2.190	3.836	2.028
13	2.625	1.343	1.646	1.369	1.104	0.435	1.126	5.999	1.145	1.155
14	1.432	0.824	2.040	1.417	2.470	1.779	0.703	1.221	3.087	0.524
15	1.551	1.114	1.610	1.047	0.958	1.946	0.532	2.614	1.399	5.211
16	1.473	1.200	1.358	1.387	1.152	2.201	0.624	4.550	0.886	1.387
17	1.159	1.308	1.885	0.890	1.603	0.565	3.595	1.409	0.303	0.767
18	1.490	1.826	1.247	1.506	1.463	1.531	1.391	1.639	3.103	2.940
19	1.167	1.900	1.876	1.651	2.108	0.532	0.774	0.817	0.537	5.010
20	1.587	0.952	1.157	0.966	1.190	1.436	1.117	1.044	1.038	0.816

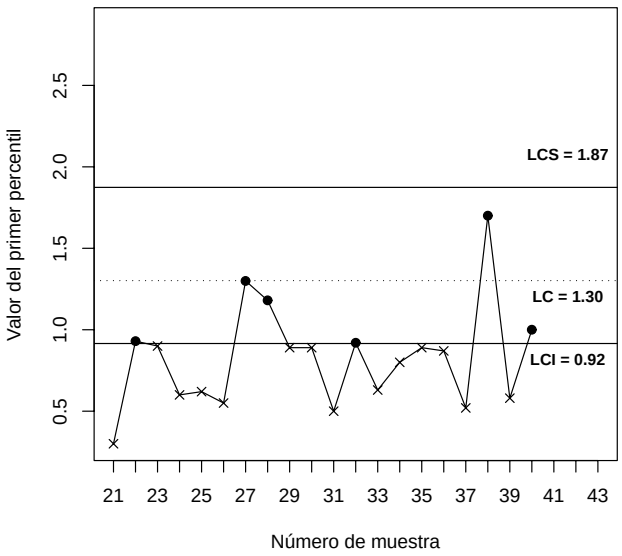

Figura 5: Carta de control *bootstrap* BS.

4.2.1. Construcción de la carta de control $\overline{T}$

Siguiendo la propuesta de Surucu & Sazak (2009), implementamos cartas de control para monitorear la ocurrencia de fallas de un proceso que genera datos provenientes de una distribución BS. La ventaja de la metodología desarrollada aquí frente a la carta de control creada (Surucu & Sazak 2009), donde se tuvo que aproximar la distribución de la suma de v.a. Weibull con tres parámetros a la distribución normal y así aproximar los límites de control, es que al tener implementada la distribución BSsum en forma exacta, los límites de control para la distribución BS también lo son. Lo primero que se debe decidir cuando se quiere monitorear el tiempo de ocurrencia de fallas en un proceso productivo es el número de fallas que se van a monitorear. Este número se denota por m. Una vez tomada esta decisión, para construir una carta de control para la media $\overline{T}$ basada en la distribución BS, asumiendo que el proceso es estable y está bajo control, proponemos un algoritmo con los siguientes pasos:

(C1) Registrar los tiempos de fallas, denotados por T_i, para $i = 1, 2, \ldots, k$, los cuales son v.a. i.i.d. con distribución BS. Estos tiempos deben tomarse desde un proceso que esté bajo control, ya que son ocupados para determinar los límites de la carta de control.

(C2) Calcular el tiempo acumulado hasta la ocurrencia de la falla m-ésima, generando n muestras de tamaño m, el cual se obtiene mediante la v.a. $Y_i = \sum_{j=m[i-1]+1}^{i\,m} X_j$, para $i = 1, 2, \ldots, n$.

(C3) Estimar los parámetros de la distribución de la suma de m variables aleatorias BS para la muestra obtenida en el paso (C2). Esto se puede realizar mediante la función `estbssum()` del paquete `bssum`; ver detalles en Apéndice 2. Estas estimaciones se denotan por $\widehat{\alpha}_{\texttt{bssum}}$ y $\widehat{\beta}_{\texttt{bssum}}$.

(C4) Calcular el LCI y el LCS de la carta de control, los que están determinados por los percentiles $(\gamma/2) \times 100\,\%$ y $(1 - \gamma/2) \times 100\,\%$ de la distribución BSsum, respectivamente. Estos percentiles se pueden obtener con la función `qbssum()` del paquete `bssum`. Se debe mencionar que en esta carta de control no tiene mucho sentido incorporar el LC de control, ya que éste es más apropiado para datos que presentan símetría. Sin embargo, con el objetivo de detectar si existe aleatoriedad en los datos, se ha construido igualmente un LC basado en la mediana.

(C5) Finalmente, determinados los límites de control de la carta, se pueden graficar los tiempos acumulados de fallas, para así corroborar que estos tiempos de fallas están bajo control.

Ahora, estamos en condiciones de monitorear cualquier conjunto de datos que siga una distribución BS para los que se desee saber su estado de control o fuera de control en relación al tiempo acumulado hasta la ocurrencia de m fallas. Si los tiempos graficados están dentro de los límites de control y no presentan un comportamiento anómalo, se juzga que el proceso está bajo control. Por el contrario, si los tiempos están por debajo del LCI, se concluye que las fallas del proceso están

ocurriendo con demasiada frecuencia y se juzga que el proceso está fuera de control. Ahora bien, si los tiempos están por sobre el LCS, se concluye que las fallas están ocurriendo lentamente, lo cual es un buen indicador del proceso productivo.

4.3. Implementación

Aquí ilustramos mediante un ejemplo la metodología de construcción de cartas de control desarrolladas para la distribución BS. Específicamente, presentamos una carta de control para el monitoreo de la ocurrencia de fallas de un proceso que genera datos provenientes desde una distribución BS. Esta carta de control está implementada en el paquete `bssum` e incorpora todos los pasos de la metodología de construcción en una única función.

Ejemplo 2. Para ilustrar el procedimiento de construcción de la carta de control para la ocurrencia de fallas de un proceso que genera datos provenientes desde una distribución BS, se usan los datos presentados en Cheng & Xie (2000) empleados allí para ilustrar cartas de control para la distribución lognormal. Estos datos se muestran en la tabla 2, representan los tiempos de falla de válvulas en un proceso industrial y se encuentran implementados en el paquete `bssum` con el nombre de `valve`.

Tabla 2: Tiempos de falla (en horas) de la válvula i-ésima en la muestra j-ésima indicada.

j	t_{ij}					j	t_{ij}				
1	4.55	4.99	3.62	3.52	3.77	16	2.53	4.16	3.78	3.77	1.72
2	1.93	3.95	4.10	4.16	1.61	17	3.41	3.10	6.02	1.09	2.92
3	2.22	1.73	5.10	4.52	4.06	18	2.85	4.46	3.17	2.50	3.91
4	2.71	2.45	4.60	2.09	1.90	19	3.16	3.70	2.61	2.65	3.42
5	2.91	5.68	4.33	3.51	3.24	20	2.54	4.77	1.63	2.64	3.59
6	2.20	5.66	3.71	3.35	1.61	21	3.61	2.13	5.08	2.01	1.92
7	2.82	5.22	3.75	3.50	3.31	22	3.16	4.20	2.32	2.44	1.62
8	2.76	4.40	3.13	1.55	3.70	23	2.96	6.09	3.78	2.29	4.16
9	4.98	4.05	4.00	7.20	3.18	24	2.47	3.49	3.38	4.45	2.61
10	4.88	2.71	3.51	3.15	4.81	25	3.55	3.35	3.18	4.75	8.72
11	4.50	1.95	3.41	2.87	1.90	26	1.35	2.50	2.51	4.20	3.50
12	3.07	4.02	4.17	4.33	4.06	27	2.30	2.26	2.22	1.60	9.70
13	2.39	2.91	3.09	3.15	2.52	28	3.71	3.06	1.53	2.45	6.40
14	2.92	4.25	3.02	2.26	5.72	29	9.48	1.72	4.20	3.37	5.58
15	2.56	4.38	1.24	2.62	1.92	30	1.90	2.56	4.28	3.18	1.94

Para implementar cartas de control para la distribución BS para el proceso de producción de válvulas, se debe comprobar que los datos de la tabla 2 provienen de una distribución BS. Para esto se utiliza la función `ksbs()` de los paquetes `bs` 1.0 y `bs` 2.0; ver detalles en Apéndice 1. A continuación se presentan los resultados de la

aplicación de la instrucción `data(valve)` en el paquete `bssum` y de la instrucción `ksbs(valve, graph = FALSE, alternative = two.sided)` en el paquete `bs`:

```
One-sample Kolmogorov-Smirnov test
Data: valve
D = 0.1038
p-value = 0.8702
Alternative hypothesis: two-sided
```

Para apoyar la decisión obtenida mediante la prueba de Kolmogorov-Smirnov (KS) que indica que prácticamente no hay envidencia (valor-p = 0.8702) como para indicar que los datos no siguen una distribución BS, producimos la figura 6, la que respalda gráficamente lo indicado por la prueba KS.

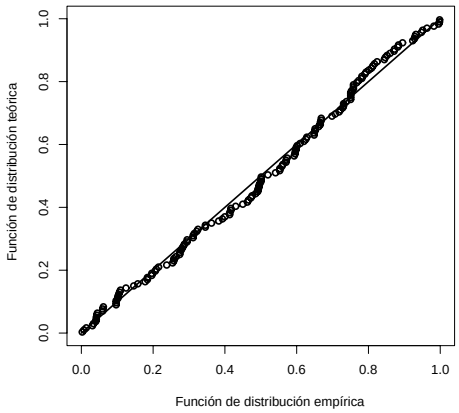

FIGURA 6: PP para los tiempos de fallas de las válvulas.

Con esta buena relación entre la distribución BS y los datos `valve`, podemos implementar una carta de control BS que permite monitorear la ocurrencia de fallas del proceso productivo de válvulas. El monitoreo de la producción de válvulas se realiza hasta que se acumulan cinco fallas. Una vez determinado esto, se construye la variable Y usando el paso (C2). En la tabla 3 se muestran estos tiempos.

TABLA 3: Tiempo acumulado hasta la quinta falla, Y_i, para $i = 1, 2, \ldots, 30$.

1	2	3	4	5	6	7	8	9	10
20.45	15.75	17.63	13.75	19.67	16.53	18.60	15.54	23.41	19.06
11	12	13	14	15	16	17	18	19	20
14.63	19.65	14.06	18.17	12.72	15.96	16.54	16.59	15.54	15.17
21	22	23	24	25	26	27	28	29	30
14.75	13.74	19.28	16.40	23.55	14.06	18.08	17.15	24.35	13.86

A continuación, para la muestra generada de tiempos acumulados Y, se estiman los parámetros de la distribución de la suma de m v.a. independientes BS. Esto se realiza con la función `estbssum()` del paquete `bssum`. La instrucción

`estbssum(5, valve)` arroja las estimaciones de α y β de la suma de m v.a. BS independientes, que son $\widehat{\alpha} = 0.358$ y $\widehat{\beta} = 3.306$. Una vez estimados los parámetros, se deben calcular los límites de la carta de control que monitorea la ocurrencia de fallas en un proceso productivo. Como se menciona en el paso (C4), los límites están determinados por los percentiles $(\gamma/2) \times 100\,\%$ y $(1-\gamma/2) \times 100\,\%$ de la distribución BSsum, respectivamente. A continuación se muestra la instrucción que calcula estos percentiles en el paquete `bssum`, que en definitiva son respectivamente el LCI y el LCS de la carta BS:

```
> qbssum(q = 0.0027/2, k = 5, alpha = 0.358, beta = 3.306,
+        lower.tail = TRUE)
[1] 10.720
> qbssum(q = 1 - 0.0027/2, k = 5, alpha = 0.358, beta = 3.306,
+        lower.tail = TRUE)
[1] 28.024
```

El conjunto de datos de la tabla 3 se usa para construir los límites de control de la carta, los que son graficados para comprobar que el proceso está bajo control en relación a los tiempos de fallas. Sin embargo, se necesita un conjunto de datos simulados con un cambio en el proceso con la idea de ver si la carta de control es capaz de detectar dicho cambio en los tiempos de fallas. Para esto, se simularon cincuenta (50) observaciones que representan el tiempo de falla de cincuenta (50) válvulas desde un proceso fuera de control. Estos datos simulados están cargados en el paquete `bssum` con el nombre de `valve1`. A continuación se muestran los tiempos acumulados hasta la ocurrencia de cinco (5) fallas en el proceso productivo, a partir de estos datos simulados. Estos datos se generan usando paso (C2) y se muestran en la tabla 4.

Tabla 4: Tiempo acumulado hasta la quinta falla, y_i, para $i = 1, \ldots, 10$.

1	2	3	4	5	6	7	8	9	10
11.35	13.13	13.77	11.85	11.98	11.56	12.29	12.27	9.99	12.46

Con la idea de construir la carta de control para la distribución BS, se implementó en el paquete `bssum` la instrucción `chartControlT()`, la que incorpora los pasos (C1)-(C5). Se debe tener en cuenta la cantidad de fallas que son monitoreadas, los tiempos de fallas que se utilizan para obtener los límites de control (`valve)`, los datos a monitorear (`valve 1`) y la razón de falsa alarma. La instrucción `chartControlT(5, valve, valve1, 0.0027)` genera la carta de control graficada en la figura 7 (izquierda). Las primeras 30 muestras corresponden a los datos con los cuales se obtuvieron los límites de control. Las últimas 10 muestras corresponden a los datos simulados con un cambio en el proceso. Se ve claramente que la carta de control para el monitoreo de tiempos de falla detecta rápidamente el cambio en el proceso.

A medida que el valor del parámetro α disminuye, la forma de la distribución BS tiende a ser simétrica. En estos casos sería de gran ayuda tener también implementada una carta de control basada en la distribución BS para $\overline{T}$. Así,

primero se debe verificar que los tiempos que se quieren monitorear se distribuyan BS y que tengan un comportamiento tendiente a la simetría. En el paquete `bssum`, se implementó también una carta de control basada en la distribución BS para $\overline{T}$. Su uso es similar al de la función `chartControlT()`. La instrucción `chartControlMean(5, valve, valve1, 0.0027)` genera la carta de control de la figura 7 (derecha).

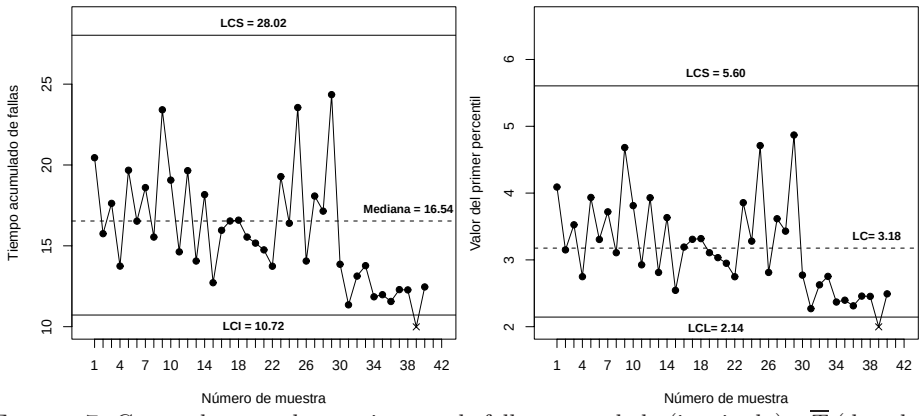

FIGURA 7: Cartas de control para tiempos de fallas acumulado (izquierda) y $\overline{T}$ (derecha) de la distribución BS.

5. Conclusiones

Debido a que muchos procesos industriales de productos expuestos a fallas presentan datos con un comportamiento asimétrico, cartas de control basadas en la Birnbaum-Saunders deberían estar disponibles. A la fecha existe poca literatura disponible en esta dirección, salvo el estudio desarrollado por Lio & Park (2008) para monitorear percentiles de esta distribución. En el presente trabajo, hemos desarrollado una nueva metodología para cartas de control basadas en la distribución Birnbaum-Saunders que permite monitorear la media de procesos de producción de unidades expuestas a fallas y el tiempo acumulado de estas fallas. Ya que últimamente el modelo Birnbaum-Saunders se ha considerado más ampliamente como una distribución de probabilidades y no sólo como una distribución de vida, la metodología de cartas de control desarrollada en este trabajo puede ser también aplicada a otros tipos de procesos y de fenómenos. Para producir esta metodología, introdujimos e implementamos la distribución de la suma de variables aleatorias independientes Birnbaum-Saunders. Esta suma no sigue una distribución Birnbaum-Saunders, lo que fue probado en un estudio de confiabilidad conducido por Raaijmakers (1980) en donde se obtuvo la distribución de vida de un sistema compuesto por unidades en standby. Cada una de estas unidades tiene tiempos de vida independientes que siguen una distribución Birnbaum-Saunders. Por tanto, la distribución de la vida útil del sistema corresponde a la distribución de la suma de variables aleatorias independientes Birnbaum-Saunders. La imple-

mentación de esta nueva distribución se desarrolló en el software R en el paquete `bssum`. Este paquete contiene funciones de probabilidad, un generador de números aleatorios y un método de estimación de parámetros para la distribución de la suma de variables aleatorias independientes Birnbaum-Saunders. Además, agregamos en este paquete dos funciones gráficas que generan las cartas de control para datos modelados mediante la distribución Birnbaum-Saunders. Estas cartas permiten monitorear el tiempo acumulado o el tiempo promedio hasta la ocurrencia de cierta cantidad de fallas. El paquete `bssum` es un primer acercamiento a la implementación de esta distribución en un paquete computacional estadístico. Éste puede usarse como base para una implementación más completa de esta distribución tal como ocurre con los paquetes `bs` y `gbs`. Otro desafío sobre el que los autores están trabajando es una metodología como la desarrollada en este artículo para datos censurados.

Agradecimientos

Los autores agradecen a los editores Dr. Leonardo Trujillo y Dra. Piedad Urdinola y a dos referees anónimos por sus valiosos comentarios que permitieron mejorar la versión preliminar de este artículo. Este estudio contó con el apoyo del proyecto FONDECYT 1080326 del gobierno de Chile.

[Recibido: octubre de 2010 — Aceptado: febrero de 2011]

Referencias

Balakrishnan, N., Leiva, V. & López, J. (2007), 'Acceptance sampling plans from truncated life tests based on the generalized Birnbaum-Saunders distribution', *Communications in Statistics: Simulation and Computation* **36**, 643–656.

Balakrishnan, N., Leiva, V., Sanhueza, A. & Cabrera, E. (2009), 'Mixture inverse Gaussian distribution and its transformations, moments and applications', *Statistics* **43**, 91–104.

Barros, M., Paula, G. A. & Leiva, V. (2008), 'A new class of survival regression models with heavy-tailed errors: robustness and diagnostics', *Lifetime Data Analysis* **14**, 316–332.

Barros, M., Paula, G. A. & Leiva, V. (2009), 'An R implementation for generalized Birnbaum-Saunders distributions', *Computational Statistics and Data Analysis* **53**, 1511–1528.

Bhatti, C. R. (2010), 'The Birnbaum-Saunders autoregressive conditional duration model', *Mathematics and Computers in Simulation* **80**, 2062–2078.

Birnbaum, Z. W. & Saunders, S. C. (1969*a*), 'Estimation for a family of life distributions with applications to fatigue', *Journal of Applied Probability* **6**, 328–347.

Birnbaum, Z. W. & Saunders, S. C. (1969*b*), 'A new family of life distributions', *Journal of Applied Probability* **6**, 319–327.

Chang, D. S. & Tang, L. C. (1994), 'Graphical analysis for Birbaum-Saunders distribution', *Microelectronics and Reliability* **34**, 17–22.

Cheng, S. W. & Xie, H. (2000), 'Control charts for lognormal data', *Tamkang Journal of Science and Engineering* **3**, 131–137.

Desmond, A. (1985), 'Stochastic models of failure in random environments', *Canadian Journal of Statistics* **13**, 171–183.

Díaz-García, J. A. & Leiva, V. (2005), 'A new family of life distributions based on elliptically contoured distributions', *Journal of Statistical Planning and Inference* **128**, 445–457.

Duncan, A. (1996), *Control de Calidad y Estadística Industrial*, Alfaomega Grupo Editor, México.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995), *Continuous Univariate Distributions*, Vol. 2, second edn, John Wiley & Sons, New York.

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York.

Leiva, V., Barros, M., Paula, G. A. & Galea, M. (2007), 'Influence diagnostics in log-Birnbaum-Saunders regression models with censored data', *Computational Statistics and Data Analysis* **51**, 5694–5707.

Leiva, V., Barros, M., Paula, G. & Sanhueza, D. (2008), 'Generalized Birnbaum-Saunders distributions applied to air pollutant concentration', *Environmetrics* **19**, 235–249.

Leiva, V., Hernández, H. & Riquelme, M. (2006), 'A new package for the birnbaum-saunders distribution', *R Journal* **6**, 35–40.
*http://www.R-project.org/doc/Rnews/Rnews_2006-4.pdf

Leiva, V., Sanhueza, A. & Angulo, J. M. (2009), 'A length-biased version of the Birnbaum-Saunders distribution with application in water quality', *Stochastic Environmental Research and Risk Assessment* **23**, 299–307.

Leiva, V., Sanhueza, A. & Saunders, S. C. (2009), New developments and applications on life distributions under cumulative damage, Technical Report 4, CIMFAV.
*http://www.cimfav.cl/reports.html#2009

Leiva, V., Sanhueza, A., Sen, P. K. & Paula, G. A. (2008), 'Random number generators for the generalized Birnbaum-Saunders distribution', *Journal of Statistical Computation and Simulation* **78**, 1105–1118.

Leiva, V., Vilca, F., Balakrishnan, N. & Sanhueza, A. (2010), 'A skewed sinh-normal distribution and its properties and application to air pollution', *Communications in Statistics: Theory and Methods* **39**, 426–443.

Lio, Y. L. & Park, C. (2008), 'A bootstrap control chart for Birnbaum-Saunders percentiles', *Quality and Reliability Engineering International* **24**, 85–600.

Marshall, A. W. & Olkin, I. (2007), *Life Distributions*, Springer, New York.

Ng, H. K. T., Kundu, D. & Balakrishnan, N. (2003), 'Modified moment estimation for the two-parameter Birnbaum-Saunders distribution', *Computational Statistics and Data Analysis* **43**, 283–298.

Pérez, R. A. & Correa, J. C. (2008), 'Intervalos de confianza vía verosimilitud relativa de los parámetros de la distribución Birnbaum-Saunders', *Revista Colombiana de Estadística* **37**, 645–670.

Podlaski, R. (2008), 'Characterization of diameter distribution data in near-natural forests using the Birnbaum-Saunders distribution', *Canadian Journal of Forest Research* **18**, 518–526.

R Development Core Team (2009), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org

Raaijmakers, F. J. M. (1980), 'The lifetime of a standby system of units having the Birnbaum-Saunders distribution', *Journal of Applied Probability* **17**, 490–497.

Raaijmakers, F. J. M. (1981), 'Reliability of standby system for units with the Birnbaum-Saunders distribution', *IEEE Transactions on Reliability* **30**, 198–199.

Sanhueza, A., Leiva, V. & Balakrishnan, N. (2008), 'The generalized Birnbaum-Saunders distribution and its theory, methodology and application', *Communications in Statistics: Theory and Methods* **37**, 645–670.

Saunders, S. C. (2007), *Reliability Life Testing and Prediction of Service Lives*, Springer, New York.

Schoonhoven, M. & Does, R. J. M. M. (2010), 'The control chart under non-normality', *Quality and Reliability Engineering International* **26**, 167–176.

Steiner, H. S. & Mackay, J. R. (2000), 'Monitoring processes with highly censored data', *Journal of Quality Technology* **32**, 199–208.

Surucu, B. & Sazak, H. (2009), 'Monitoring reliability for a three-parameter Weibull distribution', *Reliability Engineenring and System Safety* **94**, 503–508.

Vargas, J. & Montaño, T. (2005), 'Carta de control CEV $\overline{X}$ para distribuciones Weibull con datos censurados', *Revista Colombiana de Estadística* **28**, 125–139.

Vilca, F., Sanhueza, A., Leiva, V. & Christakos, G. (2010), 'An extended Birnbaum-Saunders model and its application in the study of environmental quality in Santiago, Chile', *Stochastic Environmental Research and Risk Assessment* **24**, 771–782.

Zhang, L. & Chen, G. (2004), 'EWMA charts for monitoring the mean of censored Weibull lifetimes', *Journal of Quality Technology* **36**, 321–328.

Apéndice A.
Implementación de la distribución BS

Aquí se discuten dos implementaciones para la distribución BS en el software R. La primera de éstas se encuentra en el paquete `bs` 1.0 elaborado por Leiva et al. (2006), el que incorpora funciones de probabilidad, de análisis de confiabilidad, de estimación de parámetros y de bondad de ajuste. El segundo es el paquete `bs` 2.0, una implementación elaborada recientemente que es una versión más completa del paquete `bs` 1.0. A continuación se detallan algunas de las funciones principales de ambos paquetes.

Funciones básicas

Para calcular la f.d.p., la f.d.a. y la f.q. de la distribución BS con parámetros α y β se usan las funciones `dbs()`, `pbs()` y `qbs()`, respectivamente. En la tabla 5 se encuentran las instrucciones que ilustran estas funciones.

Tabla 5: Funciones de probabilidad básicas BS.

Función	Instrucción	Resultado
f.d.p.	`dbs(3.0, alpha = 0.5, beta = 1.0, log = FALSE)`	0.021
f.d.a.	`pbs(1.0, alpha = 0.5, beta = 1.0, log = FALSE)`	0.500
f.q.	`qbs(0.5, alpha = 0.5, beta = 1.0, log = FALSE)`	1.000

Para generar números aleatorios desde el modelo BS se puede usar la f.q. dada en (5), la que requiere de un generador de números desde la distribución normal estándar. El paquete `bs` 1.0 tiene incorporado tres métodos de generación de números aleatorios BS y una función que selecciona automáticamente el método más apropiado (entre los tres existentes). El paquete `bs` 2.0 tiene disponible sólo uno de los tres generadores de la versión 1.0, mediante la función `rbs()`, el que es más eficiente y está basado en la f.q. BS. La siguiente instrucción ilustra este comando:

```
> rbs(n = 6, alpha = 0.5, beta = 2.5)
[1] 1.876529 2.785264 2.388317 2.898078 1.903702 1.366595
```

Para un análisis de confiabilidad basado en la distribución BS se ha implementado, entre otras funciones, la f.c. y la f.r. Las funciones mencionadas se muestran en la tabla 6.

Tabla 6: Funciones de confiabilidad BS.

Función	Instrucción	Resultado
f.r.	`frbs(3.0, alpha = 0.5, beta = 1.0)`	2.040
f.c.	`rfbs(1.0, alpha = 0.5, beta = 1.0)`	0.500

El paquete `bs` 2.0 tiene implementado el gráfico TTT mediante la función `TTT()`. Adicionalmente, esta función permite agregar en el gráfico una línea recta que representa a la distribución exponencial, tal como ocurre con los gráficos de probabilidad, usando la instrucción `expLine = TRUE`. Un ejemplo del uso de esta función para el conjunto de datos `psi31`, los cuales quedan disponible cuando se aplica la instrucción `data(psi31)`, es `TTT(psi31, expLine = TRUE)`, lo cual arroja el gráfico de la figura 8.

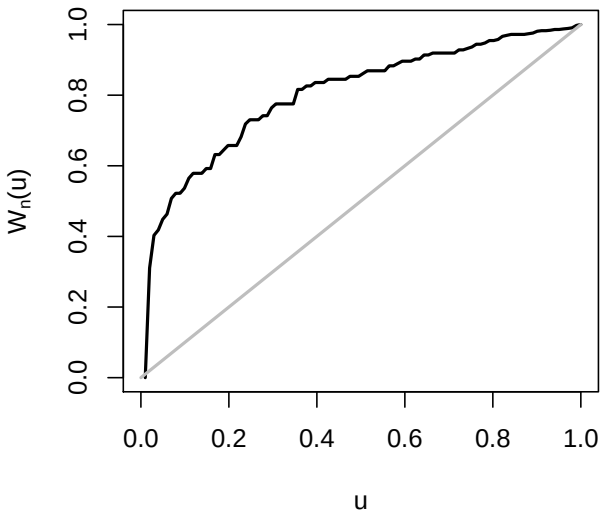

Figura 8: TTT para psi31.

Estimación de parámetros

Con el fin de estimar los parámetros de forma y de escala de la distribución BS, α y β, respectivamente, se han implementado los tres métodos descritos en la sección 2.2, los que están disponibles en ambos paquetes, tanto `bs` 1.0 como `bs` 2.0. A continuación se presenta un ejemplo relacionado al uso de estas funciones.

Ejemplo 3. Birnbaum & Saunders (1969*a*) introdujeron un conjunto de datos de fatiga de materiales de $n = 101$ hojas de aluminio del tipo 6061-T6. Estas hojas fueron cortadas en un ángulo paralelo al sentido de rotación y fueron expuestas a una presión con un máximo de tensión de 31.000 psi (del inglés "pounds per square inch" correspondiente a libras por pulgada cuadrada). Todas las hojas fueron probadas hasta que fallaran, así que no existen datos censurados. La instrucción

`est1bs(psi31)` calcula las estimaciones de VM de α y β, utilizando el estimador "media-media" como valor inicial para el procedimiento numérico, obteniendo

```
$beta.start
[1] 131.8193
$alpha
[1] 0.1703
$beta
[1] 131.8188
$converge
[1] "TRUE"
$iteration
[1] 2
```

Note que para ocupar la función `est1bs(psi31)` es necesario ejecutar previamente el comando `data(psi31)`. Las estimaciones de α y β pueden guardarse en una variable `R` como `estimate <- est1bs(psi31)`, de modo que las estimaciones de α y β pueden obtenerse mediante las instrucciones `alpha <- estimate$alpha` y `beta <- estimate$beta`. Además, por la propiedad de invarianza de los estimadores de VM, se pueden obtener estimaciones de la media, la varianza y los CV, CS y CC de la distribución BS mediante la instrucción `indicatorsbs(psi31)` obteniendo

```
The MLE's are:
Alpha = 0.1703
Beta = 131.8188
Mean = 143.0487
Variance = 522.7530
Coefficient of variation = 0.1710
Coefficient of skewness  = 0.5103
Coefficient of kurtosis  = 3.4329
```

Funciones de bondad de ajuste

Para determinar si un conjunto de datos proviene desde una distribución BS, se pueden usar diferentes métodos de bondad de ajuste. La distribución de los datos se puede juzgar mediante un histograma. Sin embargo, la información que nos arroja el histograma no es concluyente. Métodos de bondad de ajuste más precisos que están implementados en los paquetes `bs` 1.0 y `bs` 2.0 son los gráficos de probabilidad versus probabilidad (PP) y cuantil versus cuantil (QQ) y la prueba KS. Por tanto, un análisis que determine si un conjunto de datos proviene desde una distribución BS puede realizarse usando las funciones mencionadas anteriormente, las cuales son el histograma, la prueba KS y los gráficos PP y QQ.

Para ilustrar las funciones de bondad de ajuste implementadas en los paquetes `bs` 1.0 y `bs` 2.0, se utilizan los mismos datos del ejemplo 3, es decir, los datos `psi31`. La función `histbs()` genera un histograma de los datos. Esta función, si es que el usuario lo determinara necesario, puede generar simultáneamente un gráfico de caja (box-plot) que permite determinar la simetría de los datos y si alguno de

éstos es atípico. Además, usando nuevamente la propiedad de invarianza de los estimadores de VM, se puede agregar a este histograma la f.d.p. estimada de la distribución BS para este conjunto de datos usando las estimaciones de VM de α y β. La instrucción `histbs(psi31, boxPlot = TRUE, densityLine = TRUE)` genera la figura 9 (izquierda). Continuando con el análisis distribucional de los datos, la instrucción

```
ksbs(psi31, graph = FALSE, alternative = "two.sided")
```

realiza la prueba KS para la distribución BS basado en `psi31` obteniendo

```
One-sample Kolmogorov-Smirnov test
Data: psi31
D = 0.085
p-value = 0.4594
Alternative hypothesis: two-sided
```

Además, la instrucción `ppbs(psi31, line = TRUE)` proporciona el gráfico PP basado en el modelo BS mostrado en la figura 9 (centro). Finalmente, se presenta una función que sólo está en la versión avanzada del paquete `bs` 1.0, esto es, la versión `bs` 2.0. La función `MdBs()` puede determinar si un dato es atípico para la distribución BS basada en la transformación dada en (1). Esta transformación permite aproximar la distancia de Mahalanobis usando las estimaciones de los parámetros de la distribución BS. Así, se puede calcular esta distancia para cada uno de los datos, graficarla y estableciendo un punto de corte, que es el percentil 95 de la distribución $\chi^2(1)$, permitiendo establecer si el dato es atípico o no. Para los datos `psi31`, la instrucción `MdBs(psi31)` permite graficar su distancia de Mahalanobis dada en la figura 9 (derecha) identificando cuatro datos atípicos.

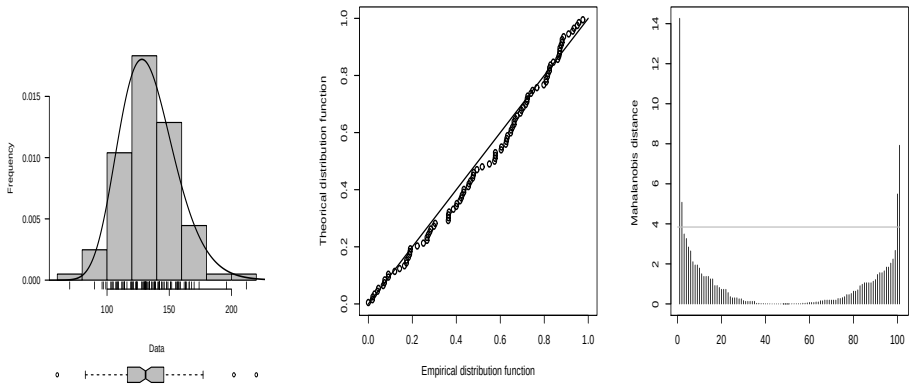

FIGURA 9: Histograma, boxplot y f.d.p. (izquierda), gráfico PP (centro) y distancia de Mahalanobis (derecha) para psi31.

Apéndice B.
Implementación de la distribución BSsum

Aquí, discutimos la implementación del paquete R `bssum`. Se han implementado funciones de probabilidad, un generador de números aleatorios y un método de estimación de parámetros para esta distribución. El paquete `bssum` se ha desarrollado con la finalidad de implementar una metodología de cartas de control para la media basadas en la distribución BS. La implementación está desarrollada para k variables aleatorias BS desde $k = 1, \ldots, 24$, ya que más allá de este valor, se podría aplicar el teorema del límite central pudiendo obtener cartas de control usando la aproximación normal.

Funciones básicas

Para calcular la f.d.p., la f.d.a. y la f.q. de la distribución BSsum con parámetros α y β se usan las funciones `dbssum()`, `pbssum()` y `qbssum()`, respectivamente. En la tabla 7 se encuentran las instrucciones que ilustran estas funciones.

Tabla 7: Funciones de probabilidad básicas BSsum.

Función	Instrucción	Resultado
f.d.p.	`dbssum(2.0, k = 4, alpha = 1.0, beta = 1.0, log = F)`	0.062
f.d.a.	`pbssum(7.0, k = 4, alpha = 1.0, beta = 1.0, log = F)`	0.699
f.q.	`qbssum(0.6, k = 6, alpha = 1.0, beta = 1.0, log.q = F)`	9.305

Un punto importante en la implementación de la distribución BSsum es la generación de números aleatorios, ya que esto ayuda a analizar diferentes formas de la distribución y es también útil para estudios de simulación. A continuación se ilustra la función `rbssum()`:

```
> rbssum(n = 6, k = 8, alpha = 1.0, beta = 1.0)
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
[1,] 11.36413 8.061949 8.519679 8.749674 16.54859 8.264141
```

La figura 10 muestra el histograma de cuatro conjuntos de números aleatorios. Se ve que al aumentar el número de v.a. que se suman en la distribución, los valores de t generados aumentan y la forma de la distribución es más simétrica.

Estimación de parámetros

Los parámetros de la distribución BSsum son k, α y β. Esta distribución se considera una distribución en sí misma, olvidando que viene de una suma de k v.a., tal como ocurre con la suma de v.a. independientes exponenciales que producen una distribución gamma con parámetro de forma entero (también conocida como distribución Erlang). Por este motivo, el parámetro k debe estimarse. Sin embargo, como la idea es usar esta distribución para implementar cartas de control para la distribución BS, el valor de k se asume conocido. Esto se debe a que cuando se

monitorea el tiempo de ocurrencia de fallas, la cantidad de fallas se fija de antemano. Por tanto, los parámetros que se deben estimar en este caso son sólo α y β. El método de estimación ocupado es el de VM. La función `optim()` permite optimizar la función de verosimilitud encontrando los valores de α y β que maximizan esta función. Como se mencionó en la subsección 4.2.1, el comando implementado para esta estimación se denomina `estbssum()` y arroja las estimaciones de VM de los parámetros de la distribución de la suma de k v.a. BS. A continuación se proporciona un ejemplo para esta función usando los datos `valve` que se llaman mediante la instrucción `data(valve)`. El número de términos que se suman en la distribución es $k = 5$ y las estimaciones son:

```
> estbssum(k = 5, valve)
$alpha
[1] 0.3581214
$beta
[1] 3.306289
```

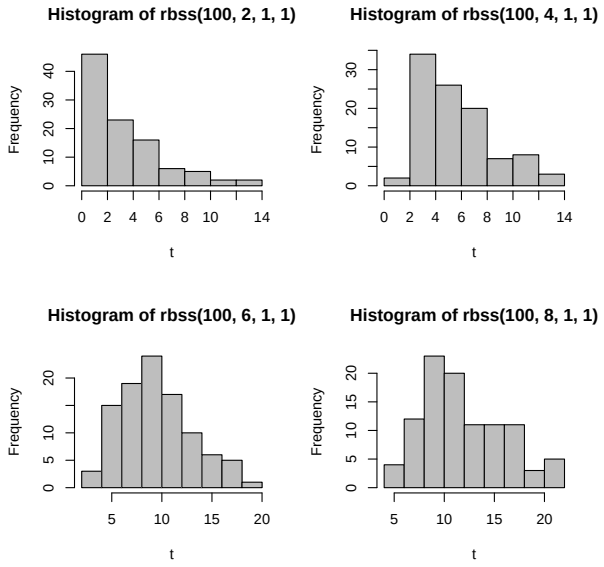

Figura 10: Histogramas de cuatro conjuntos de números aleatorios BSsum.

Revista Colombiana de Estadística
Junio 2011, volumen 34, no. 1, pp. 177 a 195

On the Student-t Mixture Inverse Gaussian Model with an Application to Protein Production

Sobre el modelo gaussiano inverso mezclado t-Student y una aplicación a producción de proteínas

Antonio Sanhueza[1,a], Víctor Leiva[2,b], Liliana López-Kleine[3,c]

[1]Departamento de Matemática y Estadística, Universidad de La Frontera, Temuco, Chile

[2]Departamento de Estadística, CIMFAV, Universidad de Valparaíso, Valparaíso, Chile

[3]Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

In this article, we introduce a mixture inverse Gaussian (MIG) model based on the Student-t distribution and apply it to bacterium-based protein production for food industry. This model is mainly useful to describe data that follow positively skewed distributions and accommodate atypical observations in a better way than its classical version. Specifically, we present a characterization of the MIG-t distribution. In addition, we carry out a hazard analysis of this distribution centered mainly on its hazard rate. Furthermore, we discuss the maximum likelihood method, which produces–in this case–robust parameter estimates. Moreover, to evaluate the potential influence of atypical observations, we produce a diagnostic analysis for the model. Finally, we apply the obtained results to novel bacterium-based protein production data and statistically compare two types of protein producers using the likelihood ratio test based on the MIG-t model as an alternative methodology to the procedures available until now. This fact is very important, since the evaluation of protein production using both constructions allows practitioners to choose the most productive one before the bacterial culture is scaled to an industrial level.

***Key words*:** Distribution mixture, Length-biased, Likelihood methods, distributions, R computer language.

[a]Professor. E-mail: asanhueza@ufro.cl
[b]Professor. E-mail: victor.leiva@uv.cl
[c]Assistant professor. E-mail: llopezk@unal.edu.co

Resumen

En este artículo, introducimos un modelo Gaussiano inverso (MIG) mezclado basado en la distribución t-Student y lo aplicamos a la producción de proteínas basada en bacterias para la industria de alimentos. Este modelo es especialmente útil para describir datos que siguen una distribución con sesgo positivo ya que permite acomodar observaciones atípicas de mejor forma que su versión clásica. Específicamente, presentamos una caracterización de la distribución MIG-t y realizamos un análisis de confiabilidad de esta distribución centrado principalmente en la tasa de fallas. También, discutimos el método de verosimilitud máxima, el cual proporciona en este caso estimaciones robustas de los parámetros del modelo. Con el fin de evaluar la influencia potencial de observaciones atípicas, proponemos un análisis de diagnóstico para la distribución. Finalmente, aplicamos los resultados obtenidos al análisis de datos nuevos de producción de proteína basada en bacterias utilizada en la industria de alimentos y comparamos estadísticamente dos tipos de bacterias productoras usando la prueba de razón de verosimilitudes basada en el modelo MIG-t como una metodología alternativa a los procedimientos disponibles a la fecha. Este punto es muy importante, ya que la evaluación de producción de proteínas usando dos construcciones distintas permite a los investigadores escoger el tipo más productivo antes de proceder al cultivo industrial a gran escala.

***Palabras clave*:** distribuciones de largo sesgado, lenguaje de computación R, métodos de verosimilitud, mezcla de distribuciones.

1. Introduction

The normal distribution has been a reference model in statistics for over one hundred years. Its attractive properties are well-known and widely used in statistical theory and practice. However, inference upon normality is vulnerable to atypical data, which are found in several fields. Specifically, the parameter estimators of the normal model obtained with the maximum likelihood (ML) method are sensitive to atypical observations. Lange, Little & Taylor (1989) proposed to use the Student-t distribution for solving this problem of sensitivity, since it has greater kurtosis than the normal distribution. Thus, such atypical cases could be accommodated in a better way by using the t model than the normal model. Moreover, as it can be seen in Figure 1, the degree of kurtosis of the t model is flexible and then it can appropriately model different quantities and magnitudes of atypical data. For these reasons, the t model has been used as an alternative to the normal model to obtain qualitatively robust estimates, which is a first concept of robustness. See Lucas (1997) and Montgomery, Peck & Vining (2001, pp. 381-413). Specifically, robustness studies the sensitivity of the results of a statistical analysis to deviations in the assumptions that validate this analysis.

A random variable (r.v.) X following the t distribution with ν degrees of freedom, denoted by $X \sim t(\nu)$, has probability density function (p.d.f.) and

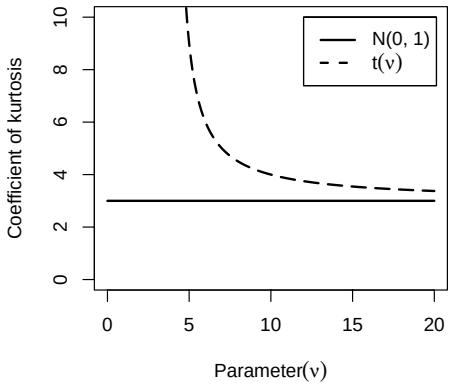

FIGURE 1: Coefficients of kurtosis of the normal and t distributions.

cumulative distribution function (c.d.f.) respectively given by

$$\phi_t(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\,\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\ \left[1+\frac{x^2}{\nu}\right]^{-\frac{\nu+1}{2}} \text{ and } \Phi_t(x) = \frac{1}{2}\left[1+\mathrm{I}_r\left(\tfrac{1}{2},\tfrac{1}{2}\,\nu\right)\right],\ x\in\mathbb{R},\,\nu>0$$

where $\mathrm{I}_r(a,b) = [\int_0^x t^{a-1}[1-t]^{b-1}\,\mathrm{d}t]/[\int_0^1 t^{a-1}[1-t]^{b-1}\,\mathrm{d}t]$ is the beta incomplete function ratio, with $r = x^2/[x^2+\nu]$. See Johnson, Kotz & Balakrishnan (1994, pp. 364). Special cases of the t distribution are the Cauchy distribution, when $\nu = 1$, and the normal distribution, when $\nu \to \infty$. The normal and t models are symmetric distributions in the seto of real numbers. However, many phenomena present data whose distributions are asymmetrical, such as occurs frequently in biotechnology and industry data.

A very popular, positively skewed, asymmetric probability model is the inverse Gaussian (IG) distribution, which is also known as the first passage time distribution of the Brownian motion with positive drift. See Schrödinger (1915), Wald (1947) and Tweedie (1957). The inverse Gaussian (IG) and normal distributions are very similar, although these distributions describe different types of data. In fact, Folks (2007) provided a table that contains 42 analogies between these two distributions. The IG distribution has been widely studied. Several books devoted to this distribution have appeared within the last 30 years. See Jörgensen (1982), Chhikara & Folks (1989), Seshadri (1993, 1999) and Johnson et al. (1994, pp. 259-297). Specifically, the IG model is characterized by the mean (μ) and scale (λ) parameters, denoted by $T \sim \mathrm{IG}(\mu,\lambda)$. An r.v. T with IG distribution has p.d.f. and c.d.f. respectively given by

$$f_T(t) = \phi\left(a_t\right)\frac{\sqrt{\lambda}}{\sqrt{t^3}} \text{ and } F_T(t) = \Phi\left(a_t\right) + \exp\left(\frac{2\lambda}{\mu}\right)\Phi\left(-b_t\right),\ t>0,\,\mu>0,\,\lambda>0$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denotes the N(0,1) p.d.f. and c.d.f. respectively, and

$$a_t(\mu,\lambda) = \frac{\sqrt{\lambda}[t-\mu]}{\mu\sqrt{t}} \quad \text{and} \quad b_t(\mu,\lambda) = \frac{\sqrt{\lambda}[t+\mu]}{\mu\sqrt{t}} \tag{1}$$

According to (1) and for simplicity, we sometimes use the notations $a_t(\mu,\lambda) = a_t$ and $b_t(\mu,\lambda) = b_t$ along the paper. However, when we need to emphasize the

dependence of the functions a_t and b_t on μ and λ, we use the notations $a_t(\mu, \lambda)$ and $b_t(\mu, \lambda)$. Some properties of the IG distribution are $cT \sim \text{IG}(c\,\mu, c\,\lambda)$, with $c > 0$, and to be a member of the exponential family.

Length-biased distributions are particular cases of the weighted distributions and have interesting properties; see Gupta & Kirmani (1990), Patil (2002) and Leiva, Sanhueza & Angulo (2009). The length-biased inverse Gaussian (LBIG) distribution was presented by Gupta & Akman (1995). However, this result was previously postulated by Jörgensen, Seshadri & Whitmore (1991), although with a distinct denomination, since this was called the distribution of the complementary reciprocal of the IG distribution. Specifically, if $T \sim \text{IG}(\mu, \lambda)$, then the r.v. $L = \mu^2/T$ has a LBIG distribution. In this case, the p.d.f. is given by $f_L(l) = \phi\,(a_l)\,\sqrt{\lambda}/[\mu\sqrt{l}]$, for $l > 0$, $\mu > 0$ and $\lambda > 0$.

Mixture distributions provide powerful and popular tools for generating flexible distributions with attractive statistical and probabilistic properties. See McLachlan & Peel (2000). Specifically, if $0 < p < 1$ is a mixing parameter and $f_{X_1}(x)$ and $f_{X_2}(x)$ are the densities of the variates X_1 and X_2, respectively, then the p.d.f. of the r.v. X expressed by the mixture between X_1 and X_2 is $f_X(x) = [1-p]\,f_{X_1}(x) + pf_{X_2}(x)$, for $x > 0$. Thus, an r.v. M with mixture inverse Gaussian (MIG) distribution obtained from the mixture of the IG and LBIG models has p.d.f. given by

$$f_M(m) = \phi\,(a_m)\,\frac{\sqrt{\lambda}}{\sqrt{m^3}}\left[1 - p + \frac{p\,m}{\mu}\right], \quad m > 0,\ \mu > 0,\ \lambda > 0,\ 0 < p < 1 \qquad (2)$$

This is denoted by $M \sim \text{MIG}(\mu, \lambda, p)$. For more details about the MIG distribution and some extensions, see Gupta & Akman (1995), Balakrishnan, Leiva, Sanhueza & Cabrera (2009) and Kotz, Leiva & Sanhueza (2010). We note from (2) that the MIG model is related to the normal model. Thus, by using this relationship, we can define a MIG distribution based on the t model, which we call the MIG-t distribution and should be highly flexible admitting different degrees of kurtosis and asymmetry. In addition, this distribution has parameter estimates that are often non-sensitive to atypical data. Therefore, the MIG-t model can be considered in place of the classic MIG model to produce robust estimation such as occurs with the t and normal models. See Lange et al. (1989). This methodology avoids the use of robust estimation procedures in their classical way, such as Sanhueza, Sen & Leiva (2009) and Leiva, Sanhueza, Sen & Araneda (2010) proposed, by the utilization of the t model in the construction of the MIG distribution.

The IG, LBIG, MIG distributions have been applied in diverse areas, such as actuarial science, agricultural, biotechnology, business and industry, demography, earth sciences, economy and finance, engineering sciences, internet, linguistics, medical sciences, and social and behavior sciences. For more details about these applications, see Chhikara & Folks (1989, pp. 159-184) and Seshadri (1999, pp. 167-316). As mentioned, applications of the IG model in biotechnology and industry have been considered. In this study, we propose a new application to these two fields, which considers bacterium-based protein production for food industry. In general, bacteria are used for the production of proteins with industrial purposes. Simões Barbosa, Abreu, Silva-Neto, Gruss & Langella (2004) and Le-Loir,

Nouaille, Commissaire, Bretigny, Gruss & Langella (2001) investigated the potential of a lactic acid bacteria called *Lactococcus lactis*, which is a microorganism primarily used in the dairy food industry to produce and secret proteins. Such bacteria can also be used for industrial processes such as meat, wine and dairy. Then, to characterize and compare protein production in different strains and constructions is important by using appropriate statistical distributions and tests. We should explored this aspect because the evaluation of protein production by employing several constructions allows practitioners to choose the most productive one before the bacterial culture is scaled to an industrial level.

The aims of this paper are: (*i*) to introduce the MIG-t distribution as a model that can fit data with high kurtosis, such as it could occur in biotechnology and industry, (*ii*) to carry out a hazard analysis for this distribution centered mainly on the hazard rate (h.r.) and (*iii*) to apply the obtained results to protein production data of *Lactococcus lactis*.

The rest of this article is organized as follows. In Section 2, we provide a probabilistic characterization of the MIG-t distribution and carry out an analysis of its h.r. In Section 3, we estimate the parameters of the MIG-t distribution and make inference about them by using the ML method. Also, in this section, we produce a diagnostic analysis for this distribution to evaluate the potential influence of atypical observations. In Section 4, we apply the obtained results to novel bacterium-based protein production data. In addition, in this section, we statistically compare two types of proteins using the MIG-t distribution by the likelihood ratio (LR) test as an alternative methodology to the classical techniques proposed so far. Finally, in Section 5, we drawn some conclusions.

2. The MIG-t Distribution

In this section, we present and characterize the MIG-t probabilistic model.

2.1. The Probabilistic Model

An r.v. T follows the MIG-t distribution with parameters $\mu > 0$, $\lambda > 0$, $0 < p < 1$ and $\nu > 0$ if and only if its p.d.f. is given by

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\lambda}}{\sqrt{\nu\,\pi}\,\Gamma\left(\frac{\nu}{2}\right)\sqrt{t^3}}\left[1+\frac{a_t^2}{\nu}\right]^{-\frac{\nu+1}{2}}\left[1-p+\frac{p\,t}{\mu}\right], \quad t>0$$

The notation $T \sim \text{MIG-}t(\mu, \lambda, p, \nu)$ is used in this case. The following theorem presents some properties of this model.

Theorem 1. *Let $T \sim \text{MIG-}t(\mu, \lambda, p, \nu)$. Then,*

(i) $c\,T \sim \text{MIG-}t(c\,\mu, c\,\lambda, p, \nu)$*, with $c > 0$, i.e., the MIG-t distribution belongs to the scale family.*

(ii) $1/T \sim \text{MIG-}t(1/\mu, \lambda/\mu^2, 1-p, \nu)$*, i.e., the MIG-t distribution belongs to the family closed under reciprocation.*

(iii) The c.d.f. of T is $F_T(t) = \Phi_t(a_t) + [1-2p]\int_{b_t}^{\infty}\phi_t\left(\sqrt{u^2 - 4\,\lambda/\mu}\right)du$.

(iv) $U = \frac{\lambda}{\mu}\left[\frac{T}{\mu} + \frac{\mu}{T} - 2\right] \sim \mathcal{F}(1,\nu)$, *i.e., U follows a Fisher distribution with 1 and* ν *degrees of freedom.*

The following theorems present the mode, denoted by t_m, and the moments of the studied distribution.

Theorem 2. *Let* $T \sim$ MIG-$t(\mu,\lambda,p,\nu)$. *Then, the mode of* T *is given by the solution to*

$$-\frac{\nu+1}{[\nu+a_{t_m}^2]} = \frac{\mu^2 t_m}{\lambda[t_m^2-\mu^2]}\frac{[3(1-p)\mu+p\,t_m]}{[(1-p)\mu+p\,t_m]}.$$

Theorem 3. *Let* $T \sim$ MIG-$t(\mu,\lambda,p,\nu)$. *Then, the first four non-central moments of* T *are given by*

(i) $\mathrm{E}[T] = \mu + p\,\frac{\mu^2}{\lambda}\,\frac{\nu}{[\nu-2]}$;

(ii) $\mathrm{E}[T^2] = \mu^2 + [1+2p]\,\frac{\mu^3}{\lambda}\,\frac{\nu}{[\nu-2]} + p\,\frac{\mu^3}{\lambda^2}\,\frac{3\nu^2}{[\nu-2][\nu-4]}$;

(iii) $\mathrm{E}[T^3] = \mu^3 + 3[1+p]\,\frac{\mu^4}{\lambda}\,\frac{\nu}{[\nu-2]} + [1+4p]\,\frac{\mu^4}{\lambda^2}\,\frac{3\nu^2}{[\nu-2][\nu-4]} + p\,\frac{\mu^4}{\lambda^3}\,\frac{15\nu^3}{[\nu-2][\nu-4][\nu-6]}$;

(iv) $\mathrm{E}[T^4] = \mu^4 + 2[3+2p]\frac{\mu^5}{\lambda}\frac{\nu}{[\nu-2]} + 5[1+2p]\frac{\mu^5}{\lambda^2}\frac{3\nu^2}{[\nu-2][\nu-4]} + [1+6p]\frac{\mu^5}{\lambda^3}\frac{15\nu^3}{[\nu-2][\nu-4][\nu-6]} + p\frac{\mu^5}{\lambda^4}\frac{105\nu^4}{[\nu-2][\nu-4][\nu-6][\nu-8]}$

***Note* 1.** Observe that if $T \sim$ MIG-$t(\mu,\lambda,p,\nu)$ and $p = 0, 0.5$ and 1, then we have the IG, Birnbaum-Saunders (BS) and LBIG distributions obtained from the $t(\nu)$ model, respectively. In addition, as mentioned, recall that the standard normal model is obtained as a limiting distribution of the $t(\nu)$ model when $\nu \to \infty$. Thus, the mentioned particular cases correspond to the classical IG, BS and LBIG distributions when $\nu \to \infty$ and $p = 0, 0.5$ and 1, respectively.

2.2. Hazard Analysis and Order Statistics

A hazard is a dangerous event that could conduct to an emergency or disaster. Thus, a hazard is a potential and not an actual possibility, i.e., it can be statistically evaluated, for example, by a useful descriptor known as the hazard rate. This rate for an r.v. $T > 0$ with p.d.f. $f_T(\cdot)$ and c.d.f. $F_T(\cdot)$, is given by

$$h_T(t) = \lim_{\triangle t \to 0} \frac{\mathbb{P}(t < T < \Delta t | T > t)}{\Delta t} = \frac{f_T(t)}{S_T(t)} = -\frac{d\log(S_T(t))}{dt}, \quad t > 0, \qquad (3)$$

with $0 < S_T(t) < 1$, where $S_T(t) = \mathbb{P}(T \geq t) = 1 - F_T(t) = \int_t^\infty f_T(u)\,du$, for $t > 0$, is the survival function. In addition, another useful descriptor is the mean residual, which is given by $\mu(x) = \mathbb{E}[T|T > x] = x + [\int_x^\infty S_T(t)\ dt]/S_T(x)$, for $x > 0$ and $S_T(x) > 0$, with $\mu(x) = \mu = \mathrm{E}[T]$, when $x = 0$. For more details about these descriptors, see Johnson, Kotz & Balakrishnan (1995, pp. 640-650), Marshall & Olkin (2007) and Saunders (2007). Note from (3) and below this equation that

all of these functions can be expressed by means of the h.r. Therefore, we carry out a hazard analysis based on this rate.

A h.r. function $h_T(t)$ can be increasing, decreasing or constant in t. In particular, if $h_T(t) = \lambda > 0$, for all $t > 0$, then we have that the r.v. T follows an exponential distribution with parameter λ. However, there are distributional families with non-monotone h.r. In this case, an important value for hazard analysis is the change point of the h.r. of T. Within the class of distributions with a non-monotone h.r., we can identify concave or convex hazard rates, i.e., $\cap$-shaped or $\cup$-shaped h.r., respectively. Particularly, for the $\cap$-shaped case, we also have two cases, when the h.r. is initially increasing until its change point and then (i) it decreases to zero, as in the case of the lognormal distribution, or (ii) it decreases until that becomes stabilized in a positive constant, as in the case of the IG and BS distributions. For this reason, when we study distributional families with non-monotone h.r., change point and limit behavior analyses are necessary.

The following theorems provide the MIG-t h.r., its change point, denoted by t_c, and its limiting behavior.

Theorem 4. *Let $T \sim$ MIG-$t(\mu, \lambda, p, \nu)$. Then, the h.r. of T is given by*

$$h_T(t) = \frac{\sqrt{\lambda}\phi_t(a_t)}{\sqrt{t^3}\left[\Phi_t(-a_t)+(2p-1)J(b_t)\right]}\left[1-p+\frac{p\,t}{\mu}\right], \quad t>0$$

where $J(b_t) = \int_{b_t}^{\infty} \phi_t(\sqrt{u^2 - 4\lambda/\mu})du$.

Theorem 5. *Let $T \sim$ MIG-$t(\mu, \lambda, p, \nu)$. Then, the change point of the h.r. of T is obtained as the solution to*

$$\Phi_t(-a_{t_c}) + [2p-1]J(b_{t_c}) = \frac{\frac{\sqrt{\lambda}}{\sqrt{t^3}}\,\phi_t(a_{t_c})\left[1-p+\frac{p\,t_c}{\mu}\right]}{\frac{\nu+1}{2[\nu+a_{t_c}^2]}\,\frac{\lambda[t_c^2-\mu^2]}{\mu^2 t_c^2}+\frac{3\mu[1-p]+p\,t_c}{2t_c[\mu(1-p)+p\,t_c]}}.$$

Theorem 6. *Let $h_T(t)$ be the h.r. of $T \sim$ MIG-$t(\mu, \lambda, p, \nu)$, with ν known. Then,* $\lim_{t\to\infty} h_T(t) = 0$.

Order statistics are useful in several statistical procedures. Thus, if $T_1, \ldots, T_n$ are i.i.d. variates, associated order statistics are denoted by $T_{(1)}, \ldots, T_{(j)}, \ldots, T_{(n)}$, where $T_{(1)}$, $T_{(n)}$ and $T_{(j)}$ denote the minimum, maximum and jth order statistic of the variates $T_1, \ldots, T_n$, respectively. For more details about order statistics, see Arnold, Balakrishnan & Nagaraja (1992).

The following theorem provides the p.d.f. of order statistics for the MIG-t distribution.

Theorem 7. *Let $T_1, \ldots, T_n$ be i.i.d. variates, where $T_i \sim$ MIG-$t(\mu, \lambda, p, \nu)$, for $i = 1, \ldots, n$. Then, for the indicated order statistic, its p.d.f. is given by*

(i) $f_{T_{(1)}}(t) = n\,\phi_t(a_t)\frac{\sqrt{\lambda}}{\sqrt{t^3}}\left[1-p+\frac{p\,t}{\mu}\right]\left[\Phi_t(-a_t)+(2p-1)\,J(b_t)\right]^{n-1}$, $t>0$

(ii) $f_{T_{(n)}}(t) = n\,\phi_t(a_t)\frac{\sqrt{\lambda}}{\sqrt{t^3}}\left[1-p+\frac{p\,t}{\mu}\right]\left[\Phi_t(a_t)+(1-2p)\,J(b_t)\right]^{n-1}$, $t>0$

(iii) $f_{T_{(j)}}(t) = \frac{n!\,\phi_t(a_t)}{(j-1)!(n-j)!}\frac{\sqrt{\lambda}}{\sqrt{t^3}}\left[1-p+\frac{p\,t}{\mu}\right]\left[\Phi_t(a_t)+(1-2p)J(b_t)\right]^{j-1}$
$\times\,[\Phi_t(-a_t)+(2p-1)J(b_t)]^{n-j}, t>0$

3. Inference and Diagnostics in the MIG-t Model

In this section, we present estimation, inference and diagnostics useful to estimate the mean protein production and detect the potential influence of atypical data. In problems with this type of data, generally one has enough amount of them to apply asymptotic results. Inference in small samples is not direct, which presents a challenge for a further study.

3.1. ML Estimation, Information Matrix and Inference

Before we find the ML estimators of the MIG-t model parameters, to discuss how one should handle the parameter ν of this model is important. The question is whether ν should be estimated. Several authors treated this topic for the t distribution and models associated with it. See Lange et al. (1989), Lucas (1997), Leiva, Riquelme, Balakrishnan & Sanhueza (2008) and references therein. These authors noticed problems of unbounded and local maximum in the likelihood function, in addition to lack of robustness, when ν is estimated. For this reason, to fix ν is better and assume that it is a known value or, otherwise, acquire information for it from the data. Thus, once the optimum value of ν is found, the parameters μ, λ and p of the MIG-t distribution are estimated as described next.

3.1.1. ML Estimation

The log-likelihood function for $\boldsymbol{\theta} = (\mu, \lambda, p)^\top$, based on a random sample $T_1, \ldots, T_n$, where $T_i \sim$ MIG-$t(\mu, \lambda, p, \nu)$, for $i = 1, \ldots, n$, is expressed as $\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta})$, where

$$\ell_i(\boldsymbol{\theta}) \propto \frac{n}{2}\log(\lambda) - \frac{[\nu+1]}{2}\log\left(\nu + a_{t_i}^2\right) + \log\left(\mu[1-p] + p\,t_i\right) - \log(\mu) \qquad (4)$$

The score vector of first derivatives of the log-likelihood function is given by

$$\dot{\ell}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = (\dot{\ell}_{\theta_1}), \quad \text{with } \theta_1 = \mu, \lambda, \text{ or } p \qquad (5)$$

where

$$\dot{\ell}_\mu = \frac{\lambda[\nu+1]}{\mu^3}\sum_{i=1}^n\left\{\frac{t_i-\mu}{\nu+a_{t_i}^2}\right\} + \sum_{i=1}^n\left\{\frac{1-p}{\mu[1-p]+pt_i}\right\} - \frac{n}{\mu},$$

$$\dot{\ell}_\lambda = \frac{n}{2\lambda} - \frac{[\nu+1]}{2\lambda}\sum_{i=1}^n\left\{\frac{a_{t_i}^2}{\nu+a_{t_i}^2}\right\} \quad \text{and} \quad \dot{\ell}_p = \sum_{i=1}^n\left\{\frac{t_i-\mu}{\mu[1-p]+p\,t_i}\right\}$$

The ML estimates of the parameters μ, λ and p are solutions to the equations $\dot{\ell}_\mu = 0$, $\dot{\ell}_\lambda = 0$ and $\dot{\ell}_p = 0$. However, these equations do not provide analytical solutions, so that an iterative numerical method is necessary to find the roots. As starting values for this iterative method, we propose considering the ML estimates of μ, λ and p of the MIG distribution. See Seshadri (1999, pp. 145).

To select the value of ν, we propose looking for the value that maximizes the likelihood function for $\nu \in [1, 100]$ using an optimal search of ν by means of the following algorithm:

(A1) For $\nu = 1$ to $\nu = 100$ by 1:

- **(A1.1)** Estimate the parameters μ, λ and p of the MIG-t model considering the ML estimates of μ, λ and p of the MIG distribution starting values for the numerical iterative procedure;
- **(A1.2)** Compute the corresponding likelihood function;

(A2) Choose the value of ν that maximizes this likelihood function and then consider the ML estimates of μ, λ and p the result.

***Note* 2.** Based on the invariance property of the ML estimators, we can estimate different functions of the parameter $\boldsymbol{\theta}$. For example, the mean protein production can be estimated by using the mean of MIG-t distribution given in Theorem 3 (i).

3.1.2. Information Matrix

The observed information matrix is obtained as $\mathcal{J}(\boldsymbol{\theta}) = -\ddot{\ell}$. Here, $\ddot{\ell}$ is the Hessian matrix of second derivatives of the log-likelihood function given by

$$\ddot{\ell}(\boldsymbol{\theta}) = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top} = (\ddot{\ell}_{\theta_1\,\theta_2}), \quad \text{with } \theta_1, \theta_2 = \mu, \lambda, \text{ or } p \tag{6}$$

where

$$\begin{aligned}
\ddot{\ell}_{\mu\mu} &= -\frac{3\lambda[\nu+1]}{\mu^4}\sum_{i=1}^{n}\left\{\frac{t_i-\mu}{\nu+a_{t_i}^2}\right\} - \frac{\lambda[\nu+1]}{\mu^3}\sum_{i=1}^{n}\left\{\frac{1}{\nu+a_{t_i}^2}\right\} \\
&\quad -\frac{2\sqrt{\lambda^3}(\nu+1)}{\mu^5}\sum_{i=1}^{n}\left\{\frac{[t_i-\mu]a_{t_i}\sqrt{t_i}}{[\nu+a_{t_i}^2]^2}\right\} - \sum_{i=1}^{n}\left\{\frac{1-p}{\mu[1-p]+p\,t_i}\right\}^2 \\
\ddot{\ell}_{\mu\lambda} &= \frac{[\nu+1]}{\mu^3}\sum_{i=1}^{n}\left\{\frac{t_i-\mu}{\nu+a_{t_i}^2}\right\} - \frac{[\nu+1]}{\mu^3}\sum_{i=1}^{n}\left\{\frac{[t_i-\mu]a_{t_i}^2}{[\nu+a_{t_i}^2]^2}\right\}, \;\; \ddot{\ell}_{\lambda p} = 0 \\
\ddot{\ell}_{\mu p} &= -\sum_{i=1}^{n}\left\{\frac{t_i}{[\mu[1-p]+p\,t_i]^2}\right\}, \ddot{\ell}_{pp} = -\sum_{i=1}^{n}\left\{\frac{t_i-\mu}{\mu[1-p]+p\,t_i}\right\}^2 \\
\ddot{\ell}_{\lambda\lambda} &= -\frac{n}{2\lambda^2} - \frac{\nu[\nu+1]}{2\lambda^2}\sum_{i=1}^{n}\left\{\frac{a_{t_i}^2}{[\nu+a_{t_i}^2]^2}\right\} + \frac{[\nu+1]}{2\lambda^2}\sum_{i=1}^{n}\left\{\frac{a_{t_i}^2}{\nu+a_{t_i}^2}\right\}
\end{aligned}$$

3.1.3. Inference

Inference for $\boldsymbol{\theta}$ can be based on the asymptotic behavior of the ML estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\lambda}, \widehat{p})^\top$ given by $\sqrt{n}\,[\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}] \xrightarrow{\text{d}} \text{N}_3(\mathbf{0}, \boldsymbol{\Sigma}_{\widehat{\theta}})$, where "$\xrightarrow{\text{d}}$" means convergence in distribution. Here, $\widehat{\boldsymbol{\theta}}$ is the ML estimator of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}_{\widehat{\theta}}$ is the variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$, which can be obtained from the expected information matrix, namely $\mathcal{I}(\boldsymbol{\theta}) = \text{E}[\mathcal{J}(\boldsymbol{\theta})] = -\text{E}[\ddot{\ell}]$, as $\boldsymbol{\Sigma}_{\widehat{\theta}} = \mathcal{I}(\boldsymbol{\theta})^{-1}$. Thus, the standard errors of the ML estimators can be computed by using the square roots of the diagonal elements of $\mathcal{I}(\boldsymbol{\theta})^{-1}$. Their estimated standard errors can be obtained by evaluating $\boldsymbol{\theta}$ at its ML estimate $\widehat{\boldsymbol{\theta}}$.

***Note* 3.** Instead of the expected information matrix, its observed version could be used to approximate the standard errors of the ML estimators. These errors can be computed by using the square roots of the diagonal elements of $\mathcal{J}^{-1}(\boldsymbol{\theta})$. Once again, their estimated standard errors can be obtained by evaluating $\boldsymbol{\theta}$ at its ML estimate $\widehat{\boldsymbol{\theta}}$. For more details about the use of the observed information matrix instead of its expected value, see Efron & Hinkley (1978).

A confidence region for $\boldsymbol{\theta}$ may be constructed by using the asymptotic normal distribution of $\widehat{\boldsymbol{\theta}}$ above mentioned. Thus, an approximate $(1-\alpha)100\%$ confidence region for $\boldsymbol{\theta}$, with $0 < \alpha < 1$, is given by $\mathcal{R} = \{\boldsymbol{\theta} \in \mathbb{R}^3\colon (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \boldsymbol{\Sigma}_{\widehat{\theta}}^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq \chi^2_{1-\alpha}(3)\}$, where $\chi^2_{1-\alpha}(3)$ denotes the $(1-\alpha)$th quantile of the χ^2 distribution with three degrees of freedom.

3.2. Influence Diagnostics

Case deletion is a common way to assess the effect of an observation on the estimation procedure. This is a global influence analysis, since the effect of a case is evaluated by dropping it from the data set. Alternatively, local influence is based more on geometric differentiation than the elimination of observations. In this last case, a differential comparison of estimators is used before and after perturbing the data or the model. We implement the local influence method for evaluating possible atypical cases in the protein production data. As in Cook (1986), we use the likelihood displacement to evaluate the local influence. Next, we present global and local influence techniques that may be useful for detecting atypical protein production data and studying the suitability of the MIG-t model to such data.

3.2.1. Global Influence

Cook's distance is an interesting diagnostics technique of the global influence method. See Cook & Weisberg (1982). A generalization of this distance is expressed as $D_i = [(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)})^\top \widehat{\boldsymbol{\Sigma}}_{\widehat{\theta}}^{-1} (\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)})]/k$, for $i = 1, \ldots, n$, where k is the number of parameters and $\widehat{\boldsymbol{\Sigma}}_{\widehat{\theta}}$ is an estimator of the covariance matrix of $\widehat{\boldsymbol{\theta}}$, which, as mentioned, can be approximated by $-\ddot{\ell}^{-1}$ evaluated at $\widehat{\boldsymbol{\theta}}$, such that $D_i = [(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)})^\top (-\ddot{\ell})(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)})]/k$. If we use an approximation of first order, we obtain $\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)} \approx [\ddot{\ell}_{(i)}]^{-1} \dot{\ell}_{(i)}$, with $\dot{\ell}_{(i)}$ being the score vector and $\ddot{\ell}_{(i)}$ the Hessian matrix

without considering the ith case. Thus, $D_i \approx [(\dot{\ell}_{(i)})^\top (\ddot{\ell}_{(i)})^{-1}(-\ddot{\ell})^{-1}(\ddot{\ell}_{(i)})^{-1}\dot{\ell}_{(i)}]/k$, where a high value for D_i indicates a high impact case on the ML estimator of $\boldsymbol{\theta}$. For the MIG-t model, in D_i, $k = 3$ and $\dot{\ell}_{(i)}$ and $\ddot{\ell}_{(i)}$ are analogously defined as those given in (5) and (6), respectively.

3.2.2. Local Influence

From (4), we can note that the contributions $\ell_i(\boldsymbol{\theta})$ are equally weighted. A perturbed log-likelihood function can be defined by $\ell(\boldsymbol{\theta} \mid \boldsymbol{\omega}) = \sum_{i=1}^n \omega_i \ell_i(\boldsymbol{\theta})$, with $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ being the vector of weights of the contributions from each case to the likelihood function and $\boldsymbol{\omega}_0 = \mathbf{1}_n = (1, \ldots, 1)^\top$ being the non-perturbed point, that is, $\ell(\boldsymbol{\theta} \mid \boldsymbol{\omega}_0) = \ell(\boldsymbol{\theta})$. This scheme of perturbation is useful for evaluating whether the contribution of cases representing to protein production data with different weights influence the ML estimator of $\boldsymbol{\theta}$. Specifically, let $\widehat{\boldsymbol{\theta}}_\omega$ be the ML estimator of $\boldsymbol{\theta}$ obtained from the perturbed likelihood function. The influence of the perturbation $\boldsymbol{\omega}$ on the ML estimator may be checked by means of the likelihood displacement given by $\mathrm{LD}(\boldsymbol{\omega}) = 2[\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_\omega)]$. Cook (1986) postulated studying the local behavior of $\mathrm{LD}(\boldsymbol{\omega})$ around $\boldsymbol{\omega}_0$ employing the normal curvature C_l of $\mathrm{LD}(\boldsymbol{\omega})$ at $\boldsymbol{\omega}_0$ and in the direction of some unitary vector $\boldsymbol{l}$. He showed that $C_l = 2 \mid \boldsymbol{l}^\top \boldsymbol{\Delta}^\top \ddot{\ell}^{-1} \boldsymbol{\Delta} \boldsymbol{l} \mid$, with $\|\boldsymbol{l}\| = 1$, where $\ddot{\ell}$ is as defined in (6) and $\boldsymbol{\Delta}$ is a $3 \times n$ perturbation matrix expressed as $\boldsymbol{\Delta} = [\boldsymbol{\Delta}_1(\boldsymbol{\theta}), \ldots, \boldsymbol{\Delta}_n(\boldsymbol{\theta})]$, both evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega}_0$. For the MIG-t distribution, the elements of $\boldsymbol{\Delta}$ are

$$\boldsymbol{\Delta}_i(\boldsymbol{\theta}) = (\Delta_i(\mu), \Delta_i(\lambda), \Delta_i(p))^\top = \left(\frac{\partial^2 \ell(\mu \mid \boldsymbol{\omega})}{\partial \mu \partial \omega_i}, \frac{\partial^2 \ell(\lambda \mid \boldsymbol{\omega})}{\partial \lambda \partial \omega_i}, \frac{\partial^2 \ell(p \mid \boldsymbol{\omega})}{\partial p \, \partial \omega_i} \right)^\top$$

for $i = 1, \ldots, n$, where

$$\begin{aligned} \Delta_i(\mu) &= \frac{\lambda\,[\nu+1]\,[t_i - \mu]}{\mu^3\,[\nu + a_{t_i}^2]} + \frac{1-p}{\mu[1-p] + p\,t_i} - \frac{1}{\mu} \\ \Delta_i(\lambda) &= \frac{1}{2\lambda} - \frac{[\nu+1]\,a_{t_i}^2}{2\lambda\,[\nu + a_{t_i}^2]} \quad \text{and} \quad \Delta_i(p) = \frac{t_i - \mu}{\mu[1-p] + p\,t_i} \end{aligned}$$

Let $\boldsymbol{l}_{\max}$ be the direction of the maximum normal curvature, which corresponds to the perturbation that reaches the greatest local change in $\widehat{\boldsymbol{\theta}}$. The most influential cases in the protein production data may be identified by their large components of the vector $\boldsymbol{l}_{\max}$. In addition, $\boldsymbol{l}_{\max}$ is the eigenvector associated with the largest eigenvalue of $\boldsymbol{B} = \boldsymbol{\Delta}^\top \ddot{\ell}^{-1} \boldsymbol{\Delta}$. Another interesting direction is $\boldsymbol{l} = \boldsymbol{e}_{in}$, which is the ith unitary vector of $\mathbb{R}^n$. In this case, the normal curvature is given by $C_i = 2|b_{ii}|$, with b_{ii} being the ith diagonal element of $\boldsymbol{B}$. Thus, C_i can be useful to detect the total local influence of the ith case of protein production using as benchmark $C_i > 2\,\overline{C}$, where $\overline{C} = \frac{1}{n}\sum_{i=1}^n C_i$, for indicating whether such a case is potentially influential.

4. Application to Real Data

In this section, for illustrative purposes, we apply some of the obtained results for the MIG-t distribution to novel data corresponding to the production of a secreted protein by *Lactococcus lactis*, where initially just one data set is analyzed as follows. First, an implementation in R code of the MIG-t model is discussed. Second, the problem upon analysis is described. Third, the data set is provided. Fourth, an exploratory data analysis (EDA) of this set is produced. Fifth, the parameters of the MIG-t distribution are estimated by using the ML method. Later, we carry out a brief diagnostic analysis in order to establish the potential influence of some protein production data. Then, goodness-of-fit is presented for studying the suitability of the MIG-t distribution to such data. Finally, we compare the production between two different constructions, one of them corresponding to the analyzed data set. The constructions are bacterial strains genetically engineered to produce and secret a protein of interest. In this concluding analysis, by using the invariance property of the ML estimators of the parameters of the MIG-t distribution, we estimate the mean of two populations (constructions) and conducted a statistical comparison between these mean values by using the LR test.

4.1. Implementation in R Code

R language is an open-source software package for statistical computing and graphics that can be obtained from `http://www.R-project.org`; see R Development Core Team (2009). Several R packages for analyzing data from different distributions are available and can be downloaded from `http://CRAN.R-project.org`. We have developed an R code to analyze data from the MIG-t model. This code contains diverse indicators of the MIG-t distribution and allows us to compute ML estimates of its parameters.

4.2. The Problem upon Analysis

Lactococcus lactis is a lactic acid microorganism corresponding to a well-characterized gram-positive bacterium that can be used for food industry. This microorganism can be genetically modified to allow the production of proteins and secretion of proteins into the culture media. These proteins can be purified and used for several purposes in food industry. Depending on the success of the genetic construction, the yield protein will vary among constructs. For this reason, once genetic constructions are finished, *Lactococcus lactis* is reproduced by experiments *in vitro* at a laboratory before produce it at an industrial scale. At this stage, protein production is measured to study its feasibility and stability, and compare production levels among different constructs. Once the production of proteins from this bacterium has reached a level with a small variation among essays, i.e., it has been stabilized, then such proteins can be produced at big scale in a fermenter where cultures of several liters are produced. *Lactococcus lactis* may yield many types of proteins (see Simões Barbosa et al. 2004), although this bacterium does

not secret an important amount of proteins. Therefore, genetical constructs allowing the production and secretion of proteins of industrial interest in this bacterium have been a major research point. All secreted proteins carry a signal peptide that directs them to the extracellular culture media. Best results have been obtained when a native peptide (belonging to *Lactococcus lactis*) is used even when the produced protein does not belong to this bacterium. See Le-Loir et al. (2001). Specifically, the authors postulated a model for protein secretion based on *Lactococcus lactis* using the staphylococcal nuclease (NucB), a non native protein, and replacing the signal peptide by a native signal peptide (from USp45). The protein production is higher using the *Lactococcus lactis* signal peptides than other peptides, particularly for Usp45 with classical tests for which protein production data do not meet the assumptions. The question that arises here is whether this level of production is transferable to other signal peptides from secreted proteins as YvjB and how this could be addressed with an adequate distribution and an appropriate test for the data. In the application that we make in this study, we analyze data on protein production from secreted NucB possessing the YvjB signal peptide (called "Group 2") and the native signal peptide (called "Group 1"). After analyzing the first data set by the MIG-t distribution, we statistically compare both groups by using this distribution, which may be useful for modeling this kind of protein production data as an alternative procedure to the traditional ones. As mentioned, this fact is very important to determinate the most productive construction before the bacterial culture is scaled to an industrial level.

4.3. The Data Set

As mentioned, the data set corresponds to protein production data (expressed in ng/ml) from *Lactococcus lactis*, which are: 165, 123, 123, 128, 129, 135, 156, 165, 169, 178, 178, 198, 206, 207, 208, 213, 115, 119, 225, 236, 236, 156, 287, 189, 295, 296, 302, 324, 356, 389, and that we simply call `lactis`.

4.4. Exploratory Data Analysis

Table 1 presents a descriptive summary of `lactis`, while Figure 3 (left) shows the histogram and boxplot of these data. An EDA of `lactis` based on Table 1 and Figure 3 shows a positively skewed distribution with an atypical data. We propose the MIG-t model for describing these data.

Table 1: Descriptive statistics for lactis (in ng/ml)

Median	Mean	SD	CV	CS	CK	Range	Min.	Max.	n
193.5	206.867	74.796	36.156%	0.741	2.542	274	115	389	30

4.5. Estimation and Model Checking

To estimate the parameters μ, λ, p and ν of the MIG-t distribution, we use the ML method described in Subsection 3.1.1. As mentioned there, we suggest to fix ν

and assume that it is a known value or, otherwise, get information for it from the data. Thus, to estimate μ, λ and p of the MIG-t_ν model, we fix integer values for ν from 1 to 100 by 1, choosing the value of ν that maximizes the likelihood function. The command `mleMIGt()` has been implemented in the software R to carry out the procedure described in Subsection 3.1.1. The instruction `mleMIGt(lactis)` automatically chooses the value of ν that maximizes the likelihood function and computes the ML estimates of μ, λ and p of the MIG-t model according to (A1)-(A2). The function `optim` is used to solve the corresponding iterative numerical procedure, which is available in the software R.

***Note* 4.** The function `optim` employs the `L-BFGS-B` method developed by Byrd, Lu, Nocedal & Zhu (1995) to carry out the corresponding numerical optimization. This method allows having a "box constraint" and so each variable of the optimization procedure can have a lower or upper bound. The `L-BFGS-B` method uses a limited-memory modification of the quasi-Newton method.

Based on `lactis`, the obtained results for these estimates are $\widehat{\mu} = 204.109$, $\widehat{\lambda} = 1692.646$ and $\widehat{p} = 0.090$, with $-\ell(\widehat{\boldsymbol{\theta}}) = 168.523$ being the negative value of the log-likelihood function evaluated at these estimates.

Next, we detect the effect of potentially influential observations on the ML estimates for `lactis`. These observations are chosen by using the local influence method described in Subsection 3.2.2 by means of the total local influence index plot (Figure 2). From this figure (left), we can note a potential influence of the cases #29 and #30 on the ML estimates of the classical MIG distribution. However, as expected, this potential influence is less pronounced for the MIG-t distribution (see Figure 2, right).

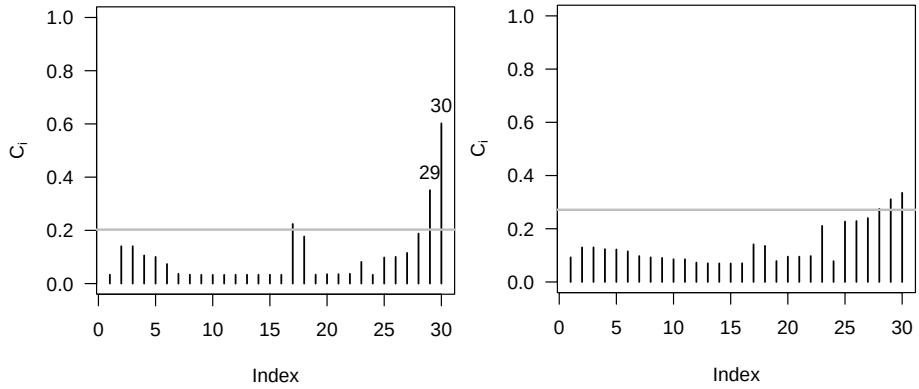

FIGURE 2: Total local influence index plot for the MIG (left) and MIG-t (right) models.

***Note* 5.** Since the purpose of this article is to illustrate the use of the MIG-t model in the context of protein production data and not to conduct an influence analysis, the information provided by the model checking is sufficient for us. In future studies, a more deeper analysis should be carried out on these atypical cases. Also, comparison of the obtained results in this application with other distributions

usually employed in hazard analysis, as well as the analysis of lifetime data by this model, will be addressed in future studies.

Once the MIG-t model parameters have been estimated and the influence analysis conducted, a natural question that arises is how good the fit of the model to `lactis` is. For this purpose, we can calculate the Kolmogorov-Smirnov (KS) distance between the empirical c.d.f. $F_n(\cdot)$, and the MIG-t c.d.f., $F_T(\cdot)$, given by

$$\mathrm{KSD}_i = |F_n(t) - F_T(t)|, \quad i = 1, \ldots, n$$

To compute this distance, we replace the parameters in the MIG-t c.d.f. by their respective ML estimates. Once all the n KS distances are calculated, we determine the maximum value of such distances and then compare it to the $(1-\alpha)$th quantile of the KS distribution to evaluate the suitability of the MIG-t model to `lactis`. The p-value of the KS test is 0.910, which strongly supports the hypothesis that the MIG-t distribution fit `lactis` in a very good way. To visually verify this fact, we use the invariance property of the ML estimators for determining the MIG-t p.d.f. and c.d.f., which are shown in Figure 3 on the histogram and empirical c.d.f. of the data, respectively. These graphs show the excellent agreement between the MIG-t model and `lactis`. Other goodness-of-fit tests, such as the Anderson-Darling test or those for normality as the Lillieford and Shapiro-Wilk test adapted to `lactis` could be also applied, but we consider the information provided by the KS test concluding.

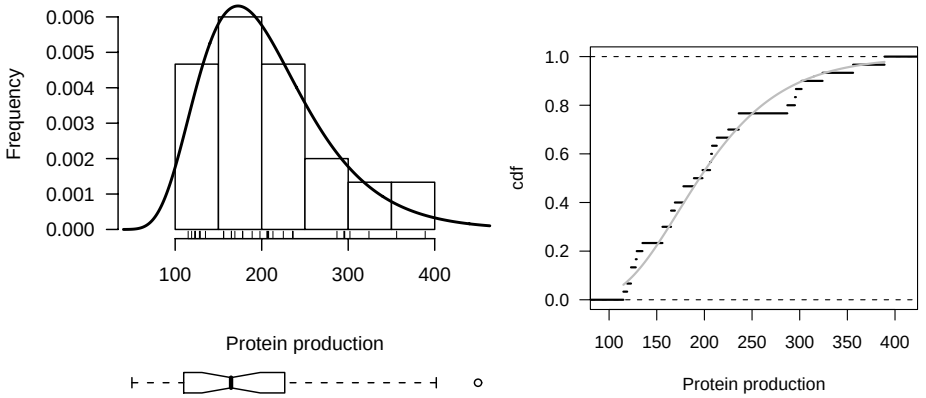

FIGURE 3: Histogram and boxplot with estimated MIG-t p.d.f. (left) and empirical c.d.f. versus estimated MIG-t c.d.f. (right) for lactis.

4.6. A Comparative Analysis

Once the MIG-t distribution has been chosen for fitting `lactis`, we can estimate the mean of this distribution. To do this, we use the ML estimates of μ, λ and p, the optimum value for ν, the invariance property of these estimators and Theorem 3(i). We estimate the mean $\mathrm{E}[T] = \mu + [p\,\mu^2\,\nu]/[\lambda\,(\nu - 2)]$ to detect

the protein production based on the MIG-t distribution using `lactis`, which is $\widehat{\mathrm{E}[T]} = 206.370$ ng/ml.

As mentioned, for practitioners, to compare two constructs is important. Let us to denote these constructs as distributions F and G. On the basis of two independent samples $T_{i1}, \ldots, T_{in_i}$, for $i = 1, 2$, each one randomly extracted from its respective population, we assume that $T_{ij} \sim \text{MIG-}t(\mu_i, \lambda, p, \nu)$, for $i = 1, 2$. We want to test H_0: $\mathrm{E}[T_1] = \mathrm{E}[T_2]$ against H_1: $\mathrm{E}[T_1] \neq \mathrm{E}[T_2]$, where $\mathrm{E}[T_i]$ is the mean of the ith population. For testing H_0 against H_1, we use the LR test, whose statistic is given by

$$\mathrm{LR} = \prod_{j=1}^{n_1} \left[\frac{1 + \frac{1}{\nu} a^2_{t_{1j}}(\widehat{\mu}, \widehat{\lambda})}{1 + \frac{1}{\nu} a^2_{t_{1j}}(\widehat{\mu_1}, \widehat{\lambda})} \right]^{-\frac{\nu+1}{2}} \prod_{j=1}^{n_2} \left[\frac{1 + \frac{1}{\nu} a^2_{t_{2j}}(\widehat{\mu}, \widehat{\lambda})}{1 + \frac{1}{\nu} a^2_{t_{2j}}(\widehat{\mu_1}, \widehat{\lambda})} \right]^{-\frac{\nu+1}{2}} \tag{7}$$

By using the LR statistic defined in (7), we compare the protein production from *Lactococcus lactis* based on the MIG-t distribution for two groups: NucB (Group 1) and PSYvjB (Group 2), which estimated mean values are $\widehat{\mathrm{E}}[T_1] = 206.370$ ng/ml and $\widehat{\mathrm{E}}[T_2] = 262.167$ ng/ml. The p-value for the LR test is < 0.001, which provides enough evidence for rejecting the hypothesis of equality of means, so that PSYvjB statistically produces a greater amount of proteins than NucB. Therefore, we recommend it as microorganism for producing proteins at big scale in the dairy food industry. The found results agree with those obtained in previous studies, where the native peptide allows providing higher amounts of protein.

5. Concluding Remarks

In this article, we have derived a mixture inverse Gaussian model based on the Student-t distribution and applied it to bacterium-based protein production for food industry. This model is very flexible in kurtosis and skewness, and has a kurtosis levels greater than that of its usual version. The mixture inverse Gaussian-t model is mainly useful to describe data that follow positively skewed distributions and accommodate atypical observations in a better way than its usual version. We have provided several statistical, hazard, probabilistic and computational aspects of the mixture inverse Gaussian-t distribution. Specifically, for this distribution, we have carried out a hazard analysis based on the hazard rate, discussed maximum likelihood estimation and evaluated the potential influence of atypical observations by a diagnostic analysis. Thus, we have introduced a statistical distribution that can be useful for modeling different types of data and, particularly, those of protein production from a lactic acid bacterium called *Lactococcus lactis*, which is a microorganism used primarily in dairy food industry. In problems of bacterium-based protein production, generally one has enough amount of data to apply asymptotic results. Inference in small samples for the mixture inverse Gaussian-t model is not direct so that this presents a challenge for a future study. We have applied the obtained results to novel bacterium-based protein production data and statistically compared two types of protein producers using the proposed

distribution by the likelihood ratio test as an alternative methodology to the procedures available so far. This application showed the utility of the mixture inverse Gaussian-t distribution.

Acknowledgements

The authors wish to thank the editors Leonardo Trujillo, Ph.D. and Piedad Urdinola, Ph.D., and anonymous referees for their constructive comments on an earlier version of this manuscript which resulted in this improved version. This study was partially supported by FONDECYT 1090265 grant from Chile.

[Recibido: octubre de 2010 — Aceptado: marzo de 2011]

References

Arnold, B. C., Balakrishnan, N. & Nagaraja, H. N. (1992), *A First Course in Order Statistics*, John Wiley and Sons, New York.

Balakrishnan, N., Leiva, V., Sanhueza, A. & Cabrera, E. (2009), 'Mixture inverse Gaussian distribution and its transformations, moments and applications', *Statistics* **43**, 91–104.

Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. (1995), 'A limited memory algorithm for bound constrained optimization', *SIAM Journal on Scientific Computing* **16**, 1190–1208.

Chhikara, R. S. & Folks, J. L. (1989), *The Inverse Gaussian Distribution*, Marcel Dekker, New York.

Cook, R. D. (1986), 'Assessment of local influence (with discussion)', *Journal of The Royal Statistical Society Series B–Statistical Methodology* **48**, 133–169.

Cook, R. D. & Weisberg, S. (1982), *Residuals and Influence in Regression*, Chapman & Hall, London.

Efron, B. & Hinkley, D. (1978), 'Assessing the accuracy of the maximum likelihood estimator: Observed vs. expected Fisher information', *Biometrika* **65**, 57–487.

Folks, J. L. (2007), Inverse Gaussian distribution, *in* S. Kotz, C. B. Read, N. Balakrishnan & B. Vidakovic, eds, 'The Encyclopedia of Statistical Sciences', Vol. 6, John Wiley & Sons, New York, pp. 3681–3682.

Gupta, R. C. & Akman, H. O. (1995), 'On the reliability studies of the weighted inverse Gaussian model', *Journal of Statistical Planning and Inference* **48**, 69–83.

Gupta, R. C. & Kirmani, S. (1990), 'The role of weighted distributions in stochastic modeling', *Communications in Statistics: Theory and Methods* **19**, 3147–3162.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994), *Continuous Univariate Distributions*, Vol. 1, John Wiley and Sons, New York.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995), *Continuous Univariate Distributions*, Vol. 2, John Wiley and Sons, New York.

Jörgensen, B. (1982), *Statistical Properties of the Generalized Inverse Gaussian Distribution*, Springer, Heidelberg.

Jörgensen, B., Seshadri, V. & Whitmore, G. (1991), 'On the mixture of the inverse Gaussian distribution with its complementary reciprocal', *Scandinavian Journal of Statistics* **18**, 77–89.

Kotz, S., Leiva, V. & Sanhueza, A. (2010), 'Two new mixture models related to the inverse Gaussian distribution', *Methodology and Computing in Applied Probability* **12**, 199–212.

Lange, K. L., Little, J. A. & Taylor, M. G. J. (1989), 'Robust statistical modeling using the t distribution', *Journal of the American Statistical Association* **84**, 881–896.

Le-Loir, Y., Nouaille, S., Commissaire, J., Bretigny, L., Gruss, A. & Langella, P. (2001), 'Signal peptide and propeptide optimization for heterologous protein secretion in lactococcus lactis', *Applied and Environmental Microbiology* **67**, 4119–2127.

Leiva, V., Riquelme, M., Balakrishnan, N. & Sanhueza, A. (2008), 'Lifetime analysis based on the generalized Birnbaum-Saunders distribution', *Computational Statistics and Data Analysis* **21**, 2079–2097.

Leiva, V., Sanhueza, A. & Angulo, J. M. (2009), 'A length-biased version of the Birnbaum-Saunders distribution with application in water quality', *Stochastic Environmental Research and Risk Assessment* **23**, 299–307.

Leiva, V., Sanhueza, A., Sen, P. K. & Araneda, N. (2010), 'M-procedures in the general multivariate nonlinear regression model', *Pakistan Journal of Statistics* **26**, 1–13.

Lucas, A. (1997), 'Robustness of the student t based m-estimator', *Communications in Statistics: Theory and Methods* **26**, 1165–1182.

Marshall, A. W. & Olkin, I. (2007), *Life Distributions*, Springer Verlag, New York.

McLachlan, G. J. & Peel, D. (2000), *Finite Mixture Models*, John Wiley and Sons, New York.

Montgomery, D. C., Peck, E. A. & Vining, G. G. (2001), *Introduction to Linear Regression Analysis*, third edn, John Wiley and Sons, New York.

Patil, G. P. (2002), Weighted distributions, *in* A. H. El-Shaarawi & W. W. Piegorsch, eds, 'Encyclopedia of Environmetrics', Vol. 4, John Wiley & Sons, Chichester, pp. 2369–2377.

R Development Core Team (2009), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org

Sanhueza, A., Sen, P. K. & Leiva, V. (2009), 'A robust procedure in nonlinear models for repeated measurements', *Communications in Statistics: Theory and Methods* **38**, 138–155.

Saunders, S. C. (2007), *Reliability, Life Testing and Prediction of Services Lives*, Springer, New York.

Schrödinger, E. (1915), 'Zur theorie der fall-und steigversucheand teilchen mit brownscher bewegung', *Physikalische Zeitschrift* **16**, 289–95.

Seshadri, V. (1993), *The Inverse Gaussian Distribution: A Case Study in Exponential Families*, Clarendon Press, New York.

Seshadri, V. (1999), *The Inverse Gaussian Distribution: Statistical Theory and Applications*, Springer, New York.

Simões Barbosa, A., Abreu, H., Silva-Neto, A., Gruss, A. & Langella, P. (2004), 'A food-grade delivery system for lactococcus lactis and evaluation of inducible gene expression', *Applied Microbiology and Biotechnology* **65**, 61–67.

Tweedie, M. C. K. (1957), 'Statistical properties of the inverse Gaussian distribution - I', *Annals of Mathematics Statistical* **28**, 362–377.

Wald, A. (1947), *Sequential Analysis*, John Wiley and Sons, New York.

Revista Colombiana de Estadística
Junio 2011, volumen 34, no. 1, pp. 197 a 209

Comparación de intervalos de confianza para la función de supervivencia con censura a derecha

Comparison of Confidence Intervals for the Survival Function in the Presence of Right Censoring

Javier Ramírez[a]

Departamento de Matemáticas y Estadística, Facultad de Ciencias Básicas, Universidad de Córdoba, Montería, Colombia

Resumen

En este trabajo se comparan métodos para encontrar intervalos de confianza para la función de supervivencia, como los métodos de remuestreo *Bootstrap* aplicado a los estimadores de Kaplan-Meier y Nelson-Aalen. También, mediante las transformaciones log, $\log(-\log)$ y arc sen que pueden resultar en muchos casos más efectivos. Además, se muestra el comportamiento que presentan los intervalos de confianza no paramétricos frente a los paramétricos.

***Palabras clave*:** bootstrap, censura a derecha, estimador Kaplan-Meier, función de supervivencia, intervalos de confianza.

Abstract

This work compares methods to find confidence interval for the survival function such as the resampling methods *Bootstrap*, applied to the Kaplan-Meier and Nelson-Aalen estimators. Also through log, $\log(-\log)$ and arcsin transformations that can result more effectives in many cases. The behavior of nonparametric confidence intervals against parametric ones is also shown.

***Key words*:** Bootstrap, Confidence intervals, Kaplan-Meier estimator, Survival function, Right censoring.

1. Introducción

En la actualidad existen diferentes métodos para encontrar intervalos de confianza para la función de supervivencia en un tiempo de interés, con censura a

[a]Profesor asistente. E-mail: javierramirez@sinu.unicordoba.edu.co

derecha, tales como los métodos tradicionales utilizando los estimadores de Kaplan-Meier y Nelson-Aalen, también mediante remuestreo *Bootstrap* aplicado a estos estimadores y a través de las transformaciones log, log(− log) y arc sen, que pueden resultar en muchos casos más efectivos. Así, este trabajo compara de manera simultánea todos estos métodos utilizando diferentes porcentajes de censura y tamaños de muestra, a través de diferentes modelos generadores. Además, proporciona criterios para establecer qué intervalos de confianza no paramétricos utilizar para estimar la función de supervivencia en un tiempo de interés, con censura a derecha y se muestra el comportamiento de los estimadores no paramétricos frente a los paramétricos.

En la sección 2 se muestran los estimadores no paramétricos para la función de supervivencia utilizados en este estudio, como son los estimadores de Kaplan-Meier y Nelson-Aalen. En la sección 3 se presentan los intervalos de confianza utilizados en la comparación como son los tradicionales, *Bootstrap* y mediante transformaciones log, log(− log) y arc sen, luego en la sección 4 se presentan los criterios de comparación de los intervalos de confianza y finalmente en la sección 5 se muestra una descripción de los escenarios de simulación, así como sus resultados y conclusiones en la sección 6.

Al final del documento, en los anexos, se encuentran los resultados mediante el índice de comparación de intervalos de confianza propuesto por Correa & Sierra (2003), con base en las distribuciones Weibull y Exponencial para los tiempos de falla/censura.

2. Estimadores utilizados

2.1. Estimador de Kaplan-Meier

Kaplan & Meier (1958) propusieron una modificación de $\widehat{S}(t)$ a la cual denominaron, estimador del producto límite (EPL) de la función de supervivencia. En efecto, supónganse que existen observaciones de n individuos y que hay $k \leq n$ tiempos distintos en los cuales la muerte ocurre, esto es, $t_1 < t_2 < \cdots < t_k$.

Se admite la posibilidad de tener más de una muerte en t_j y d_j representará el número de muertes en t_j. Además, existen los tiempos de censura t_c para individuos cuyo tiempo de vida no es observado.

El estimador del producto límite, $\widehat{S}(t)$, se define como:

$$\widehat{S}(t) = \prod_{j:t_j<t} \frac{n_j - d_j}{n_j} \tag{1}$$

donde n_j es el número de individuos en riesgo en t_j, es decir, el número de individuos vivos y no censurados justo antes de t_j. Cualquier individuo con tiempo de censura registrado igual a t_j será incluido en el conjunto de n_j individuos en riesgo en t_j, como individuos que murieron en t_j. Esta convención es razonable, puesto que un individuo censurado en el tiempo t_c casi ciertamente sobrevive después de t_c.

2.1.1. Varianza del estimador de Kaplan-Meier

Para evaluar los resultados eficazmente cuando se usa el estimador del producto límite, es conveniente tener un estimador de la varianza de $\widehat{S}(t)$, una de las utilizadas en este trabajo es la fórmula de Greenwood (1926), como

$$\widehat{Var}(\widehat{S}(t)) = \widehat{S}^2(t) \sum_{j:t_j<t} \frac{d_j}{n_j[n_j - d_j]} \tag{2}$$

2.2. Estimador de Nelson-Aalen

Nelson (1969), sugieren otra alternativa para estimar la función de supervivencia basándose en la estimación de la función Hazard Acumulada como

$$\widehat{H}(t) = \sum_{j:t_j<t} \frac{d_j}{n_j} \tag{3}$$

este estimador es muy utilizado en los casos cuando se tienen tamaños de muestra pequeños, donde

$$\widehat{S}(t) = \exp(-\widehat{H}(t))$$

2.2.1. Varianza del estimador de Nelson-Aalen

Una forma de calcular la varianza la función de supervivencia, basándose en la función Hazard Acumulada es

$$\widehat{Var}(\widehat{H}(t)) = \sum_{j:t_j<t} \frac{d_j(n_j - d_j)}{n_j^3} \tag{4}$$

otra forma alternativa para varianza es

$$\widehat{Var}(\widehat{H}(t)) = \sum_{j:t_j<t} \frac{d_j}{n_j^2} \tag{5}$$

Aalen & Johansen (1978) proponen una alternativa para la estimación de la varianza de $\widehat{S}(t)$, como

$$\widehat{Var}(\widehat{S}(t)) = \widehat{S}^2(t) \sum_{j:t_j<t} \frac{d_j}{n_j^2} \tag{6}$$

3. Intervalos de confianza

3.1. Intervalo de confianza tradicional Kaplan-Meier

Teniendo en cuenta la normalidad asintótica de los estimadores de máxima verosimilitud, los intervalos de $100(1-\alpha)\,\%$ de confianza de la función de super-

vivencia en cada tiempo fijo $t_i = 0.2$ pueden calcularse de la siguiente manera

$$\widehat{S}(0.2) \pm Z_{(1-\frac{\alpha}{2})}\widehat{S}(0.2)\sqrt{\sum_{j:t_j<0.2}\frac{d_j}{n_j[n_j-d_j]}} \tag{7}$$

donde $Z_{(1-\alpha/2)}$ es el valor que se excede con probabilidad $(1-\alpha/2)$ para una distribución normal estándar.

3.2. Intervalo de confianza tradicional Nelson-Aalen

Un intervalo de confianza $100(1-\alpha)\,\%$ para $\widehat{S}(0.2)$, mediante el estimador de Nelson-Aalen, está dado por

$$\widehat{S}(0.2) \pm Z_{(1-\frac{\alpha}{2})}\sqrt{Var[\widehat{S}(0.2)]} \tag{8}$$

donde, $Var[\widehat{S}(0.2)]$, está dado en la ecuación 6

3.3. Método de remuestreo *Bootstrap* aplicado a datos de supervivencia

Un problema involucra determinar los límites de confianza para la función de supervivencia teórica o parámetros que describen esta función. Akritas (1986), propone utilizar el método *Bootstrap* para estimar la supervivencia utilizando el estimador de Kaplan-Meier.

Para estimar los intervalos de confianza para función de supervivencia utilizando el estimador de Kaplan-Meier a través del intervalo de confianza (7), mediante el remuestreo *Bootstrap*, consiste en lo siguiente:

1. Dada la muestra de tamaño n, estimar $\widehat{S}(0.2)$ mediante (1). La distribución de esta muestra se considera equivalente a la distribución de la población y $\widehat{S}(0.2)$ es el estimador muestral del parámetro poblacional $S(0.2)$.
2. Generar B muestras *Bootstrap* de tamaño n mediante muestreo con reemplazamiento de la muestra original, asignando a cada tiempo una probabilidad $1/n$ y calcular los correspondientes valores $\widehat{S}(0.2)^{*1}, \widehat{S}(0.2)^{*2}, \ldots, \widehat{S}(0.2)^{*B}$ para cada una de las B muestras *Bootstrap*.
3. Estimar el error estándar del parámetro estimado $\widehat{S}(0.2)$ calculando la desviación estándar de las B réplicas *Bootstrap*.

Así, obtenemos que el error estándar es

$$\sigma^*_{S(0.2)} = \sqrt{\frac{\sum_{b=1}^{B}\left(S(0.2)^{b^*} - \overline{\widehat{S}}(0.2)^*\right)^2}{(B-1)}} \tag{9}$$

donde $\overline{\widehat{S}}(0.2)^*$ corresponde al promedio de la estimación de la función de supervivencia evaluada en el tiempo $t = 0.2$ de las muestras *Bootstrap*.

3.4. Intervalos de confianza mediante transformaciones

En muchos casos resulta de interés calcular intervalos de confianza mediante transformaciones, como son log, log(− log) y arc sen, con el hecho principal ya sea de simetrizar la distribución de un parámetro cualquiera, estabilizar la varianza, etc.

3.4.1. Intervalos de confianza mediante la transformación log

Estos intervalos de confianza fueron sugeridos inicialmente por Kalbfleisch & Prentice (1980); luego un intervalo de confianza $100(1-\alpha)\,\%$, mediante esta transformación, para $S(0.2)$, es

$$\widehat{S}(0.2)\exp\left\{\frac{\pm Z_{(1-\frac{\alpha}{2})}\sigma_s(0.2)}{\widehat{S}(0.2)}\right\} \tag{10}$$

donde

$$\sigma_s^2(0.2)=\frac{\widehat{Var}[\widehat{S}(0.2)]}{\widehat{S}^2(0.2)} \tag{11}$$

es decir que $\sigma_s(0.2)$ corresponde a la raíz de la suma en la formula de Greenwood (1926) (2).

3.4.2. Intervalos de confianza mediante la transformación log(− log)

Para encontrar intervalos de confianza $100(1-\alpha)\,\%$ mediante esta transformación, se deben encontrar los límites

$$\widehat{S}(0.2)\exp\left\{\pm Z_{(1-\frac{\alpha}{2})}\frac{\sigma_s(0.2)}{\log(\widehat{S}(0.2))}\right\} \tag{12}$$

los intervalos de confianza mediante la transformación log(− log) son muy utilizados en la práctica, debido a sus propiedades asintóticas.

3.4.3. Intervalos de confianza mediante la transformación arc sen

Otra alternativa para calcular intervalo de confianza $100(1-\alpha)\,\%$ para la función de supervivencia es mediante la transformación arc sen, sugeridos inicialmente por Nair (1984), dado por

$$\text{sen}^2\left\{\text{máx}\left\langle 0,\text{arc sen}(\sqrt{\widehat{S}(0.2)})\pm 0.5Z_{(1-\frac{\alpha}{2})}\sigma_s(0.2)\sqrt{\frac{\widehat{S}(0.2)}{1-\widehat{S}(0.2)}}\right\rangle\right\} \tag{13}$$

4. Criterio de comparación de los intervalos de confianza

Hay dos conceptos importantes que se deben de considerar al evaluar los intervalos de confianza: la precisión indicada por la longitud del intervalo y la probabilidad de cobertura $P(LI \leq S(t) \leq LS)$. Estos criterios no se pueden analizar por separado porque de poco nos sirve un intervalo con probabilidad de cobertura alta si su longitud es muy grande o un intervalo con una longitud muy pequeña, pero con probabilidad de cobertura muy baja.

Idealmente se quiere ver que los intervalos sean cortos y tengan probabilidad de cobertura muy cercana al nivel de confianza nominal, que los procedimientos que se utilicen para construir los intervalos de confianza den intervalos tales que sus longitudes sean pequeñas, pero diferentes de cero y que la probabilidad de cobertura no sea inferior al nivel de confianza nominal, Correa & Sierra (2003).

En este trabajo se calcula para cada método las tasas de error (TE), la longitud promedio del intervalo de confianza (LPI) para la función de supervivencia y la propuesta de Correa & Sierra (2003), del índice de comparación de intervalos de confianza (I), el cual tiene en cuenta simultáneamente el nivel de confianza real (NR), el nivel de confianza nominal (NN), la longitud promedio del intervalo (LPI) como

$$I = \frac{2 - LPI}{2} \times \frac{NR}{NN} \tag{14}$$

donde el nivel de confianza real (NR) corresponde a la proporción de intervalos simulados que cubre el verdadero valor de $S(t)$. Por lo tanto, mientras mayor sea el índice, mejor será el método, luego:

$$T.E = \frac{\#\ de\ I.C\ que\ no\ cubren\ el\ verdadero\ valor\ S(t)}{N}$$

$$NR = 1 - T.E$$

$$LPI = \sum_{i=1}^{N} \frac{(LS_i - LI_i)}{N}$$

donde N corresponde al número de simulaciones, LS y LI representa el límite superior e inferior respectivamente.

5. Estudio de simulación

En esta sección se presentan los resultados obtenidos al comparar los intervalos de confianza para la función de supervivencia en un tiempo de interés $t_i = 0.2$, este tiempo de interés fue escogido teniendo en cuenta la literatura y con el propósito de ser comparables con los resultados de Borgan & Liestøl (1990), mediante los estimadores de Kaplan-Meier y Nelson-Aalen, a través de los métodos de remuestreo *Bootstrap* aplicado a los estimadores, transformaciones log, $\log(-\log)$ y arc sen. Estos métodos se compararon utilizando un algoritmo en R Development Core Team (2008).

5.1. Resultados simulación

Se utilizaron diferentes combinaciones de modelos generadores de los tiempos de supervivencia (Time) y de censura (Cens), como son el modelo exp(λ) y $Weib(\lambda, \beta)$, donde $\lambda = 2$, $\lambda = 1$ y $\beta = 0.5$ con porcentajes de censura tipo I, de 0 %, 15 %, 25 %, 35 %, 45 %, 55 % y tamaños de muestra $n = 25, 50, 75$ y 100.

La razón de utilizar los porcentajes de censura y los tamaños de muestra se debe a la finalidad de comparar los resultados con los resultados de algunos autores referenciados. Con el propósito de abreviar los títulos de las tablas se presenta las siguientes siglas, I.C: Intervalos de confianza, TE: Tasas de Error, LPI: Longitud promedio del intervalo (valores en paréntesis), NN: Nivel de confianza nominal, T.KM: I.C mediante el estimador de Kaplan-Meier, Boot.KM y Boot.NA: corresponden a los I.C de los estimadores de Kaplan-Meier y Nelson-Aalen, mediante el remuestreo *Bootstrap*. Adicionalmente log, log(− log) y arc sen corresponden a las transformaciones de las funciones de supervivencia definidas en (10), (12) y (13), respectivamente.

Para determinar la efectividad de utilizar intervalos de confianza no paramétricos para la función de supervivencia frente a los paramétricos, los valores de referencia para el caso exponencial exp(2) la $S(t) = 0.67$ y exp(1) la $S(t) = 0.82$, mientras que para el caso $weib(1, 0.5)$ la $S(t) = 0.64$. Además, para el caso exp / exp la $LPI = 0.53$, exp /$weib$ la $LPI = 0.53$, $weib$/ exp la $LPI = 0.32$ y para $weib/weib$ la $LPI = 0.32$, los resultados para $\widehat{S}_{KM}(0.2)$ y $\widehat{S}_{NA}(0.2)$ correspondiente al estimador de Kaplan-Meier y Nelson-Aalen, respectivamente, son:

5.2. Resultados para $n = 25$

Tabla 1: TE (LPI) para un NN de 95 % con 0 % de censura y $n = 25$.

Estimador $\widehat{S}_{KM}(0.2)$					
Time/Cens	T.KM	Boot.KM	log	log(− log)	arc sen
exp / exp	0.05(0.37)	0.08(0.39)	**0.02**(0.63)	0.08(0.37)	0.05(0.36)
exp /$weib$	0.03(0.32)	0.09(0.32)	**0.01**(0.42)	0.07(0.33)	0.07(0.31)
$weib$/ exp	0.10(0.37)	0.08(0.38)	**0.02**(0.64)	0.05(0.37)	0.06(0.36)
$weib/weib$	0.10(0.37)	0.08(0.38)	**0.02**(0.64)	0.05(0.37)	0.06(0.36)
Estimador $\widehat{S}_{NA}(0.2)$					
Time/Cens	T.NA	Boot.NA	log	log(− log)	arc sen
exp / exp	0.05(0.36)	0.11(0.36)	**0.02**(0.62)	0.08(0.37)	0.05(0.36)
exp /$weib$	0.03(0.31)	0.10(0.30)	**0.01**(0.41)	0.07(0.33)	0.07(0.30)
$weib$/ exp	0.08(0.37)	0.10(0.36)	**0.02**(0.62)	0.05(0.37)	0.04(0.36)
$weib/weib$	0.08(0.37)	0.10(0.36)	**0.02**(0.62)	0.05(0.37)	0.04(0.36)

En los resultados para $n = 25$ se nota que cuando no hay observaciones censuradas, los I.C mediante la transformación $\log(S(t))$, presentan mejor comportamiento, esto se debe a que resultan ser más amplios, independientemente del estimador que se utilice y el modelo generador. Además a medida que se aumenta el porcentaje de censura los I.C para la función de supervivencia mediante la transformación $\log(-\log(S(t)))$ resultan ser mejores.

Tabla 2: TE y LPI de NN 95 % con 25 % de censura para $n = 25$.

Estimador $\widehat{S}_{KM}(0.2)$					
Time/Cens	T.KM	Boot.KM	log	log(− log)	arc sen
exp / exp	0.11(0.35)	0.16(0.37)	**0.01**(0.52)	0.02(0.36)	0.08(0.34)
exp /*weib*	0.06(0.29)	0.14(0.31)	0.05(0.36)	**0.04**(0.31)	0.05(0.29)
weib/ exp	0.17(0.35)	0.11(0.32)	0.11(0.53)	**0.04**(0.36)	0.11(0.34)
weib/*weib*	0.15(0.36)	0.13(0.33)	0.12(0.54)	**0.05**(0.36)	0.12(0.34)
Estimador $\widehat{S}_{NA}(0.2)$					
Time/Cens	T.NA	Boot.NA	log	log(− log)	arc sen
exp / exp	0.11(0.35)	0.17(0.34)	0.11(0.51)	**0.02**(0.36)	0.09(0.34)
exp /*weib*	0.057(0.29)	0.15(0.29)	0.05(0.35)	**0.02**(0.32)	0.05(0.28)
weib/ exp	0.19(0.35)	0.16(0.33)	0.12(0.51)	**0.04**(0.36)	0.11(0.34)
weib/*weib*	0.17(0.35)	0.17(0.34)	0.13(0.53)	**0.05**(0.37)	0.12(0.34)

Tabla 3: TE y LPI de NN 95 % con 45 % de censura para $n = 25$.

Estimador $\widehat{S}_{KM}(0.2)$					
Time/Cens	T.KM	Boot.KM	log	log(− log)	arc sen
exp / exp	0.31(0.33)	0.33(0.34)	0.26(0.43)	**0.08**(0.34)	0.21(0.32)
exp /*weib*	0.17(0.26)	0.17(0.29)	0.17(0.31)	**0.04**(0.30)	0.14(0.26)
weib/ exp	0.39(0.33)	0.41(0.35)	0.34(0.44)	**0.15**(0.34)	0.33(0.32)
weib/*weib*	0.37(0.34)	0.37(0.36)	0.34(0.46)	**0.14**(0.35)	0.30(0.33)
Estimador $\widehat{S}_{NA}(0.2)$					
Time/Cens	T.NA	Boot.NA	log	log(− log)	arc sen
exp / exp	0.32(0.32)	0.36(0.32)	0.29(0.42)	**0.08**(0.34)	0.23(0.31)
exp /*weib*	0.17(0.26)	0.18(0.27)	0.17(0.30)	**0.04**(0.30)	0.14(0.26)
weib/ exp	0.43(0.32)	0.43(0.33)	0.35(0.43)	**0.15**(0.34)	0.34(0.31)
weib/*weib*	0.39(0.33)	0.42(0.34)	0.35(0.45)	**0.14**(0.36)	0.32(0.32)

5.3. Resultados para $n = 50$

Tabla 4: TE y LPI de NN 95 % con 0 % de censura para $n = 50$.

Estimador $\widehat{S}_{KM}(0.2)$					
Time/Cens	T.KM	Boot.KM	log	log(− log)	arc sen
exp / exp	0.06(0.26)	0.04(0.27)	**0.02**(0.42)	0.08(0.26)	0.06(0.26)
exp /*weib*	0.04(0.22)	0.06(0.22)	**0.02**(0.28)	0.07(0.22)	0.08(0.22)
weib/ exp	0.08(0.26)	0.13(0.27)	**0.03**(0.41)	0.03(0.26)	0.05(0.26)
weib/*weib*	0.08(0.26)	0.13(0.27)	**0.03**(0.41)	0.03(0.26)	0.05(0.26)
Estimador $\widehat{S}_{NA}(0.2)$					
Time/Cens	T.NA	Boot.NA	log	log(− log)	arc sen
exp / exp	0.06(0.26)	0.06(0.26)	**0.02**(0.41)	0.05(0.26)	0.06(0.26)
exp /*weib*	0.04(0.22)	0.07(0.21)	**0.02**(0.28)	0.07(0.23)	0.04(0.21)
weib/ exp	0.08(0.26)	0.12(0.26)	**0.03**(0.41)	0.03(0.26)	0.04(0.26)
weib/*weib*	0.08(0.26)	0.12(0.26)	**0.03**(0.41)	0.03(0.26)	0.04(0.26)

Tabla 5: TE y LPI de NN 95 % con 25 % de censura para $n = 50$.

Estimador $\widehat{S}_{KM}(0.2)$					
Time/Cens	T.KM	Boot.KM	log	log(− log)	arc sen
exp / exp	0.17(0.25)	0.20(0.25)	0.11(0.35)	**0.08**(0.25)	0.15(0.24)
exp /*weib*	0.11(0.20)	0.17(0.21)	0.11(0.25)	**0.03**(0.21)	0.10(0.20)
weib/ exp	0.29(0.25)	0.26(0.24)	0.21(0.35)	**0.16**(0.25)	0.26(0.24)
weib/*weib*	0.27(0.25)	0.27(0.25)	0.19(0.36)	**0.16**(0.25)	0.23(0.25)
Estimador $\widehat{S}_{NA}(0.2)$					
Time/Cens	T.NA	Boot.NA	log	log(− log)	arc sen
exp / exp	0.17(0.25)	0.23(0.24)	0.12(0.34)	**0.09**(0.25)	0.16(0.24)
exp /*weib*	0.11(0.20)	0.17(0.20)	0.11(0.24)	**0.03**(0.21)	0.10(0.20)
weib/ exp	0.32(0.25)	0.25(0.30)	0.23(0.35)	**0.16**(0.25)	0.27(0.24)
weib/*weib*	0.29(0.25)	0.25(0.31)	0.21(0.35)	**0.16**(0.26)	0.25(0.25)

Tabla 6: TE y LPI de NN 95 % con 45 % de censura para $n = 50$.

Estimador $\widehat{S}_{KM}(0.2)$					
Time/Cens	T.KM	Boot.KM	log	log(− log)	arc sen
exp / exp	0.58(0.23)	0.59(0.24)	0.46(0.29)	**0.39**(0.23)	0.50(0.22)
exp /*weib*	0.33(0.18)	0.31(0.19)	0.32(0.21)	**0.09**(0.20)	0.23(0.18)
weib/ exp	0.73(0.23)	0.80(0.23)	0.62(0.29)	**0.57**(0.23)	0.67(0.22)
weib/*weib*	0.67(0.24)	0.74(0.24)	0.55(0.31)	**0.48**(0.24)	0.61(0.23)
Estimador $\widehat{S}_{NA}(0.2)$					
Time/Cens	T.NA	Boot.NA	log	log(− log)	arc sen
exp / exp	0.59(0.23)	0.60(0.23)	0.47(0.29)	**0.40**(0.23)	0.53(0.22)
exp /*weib*	0.34(0.18)	0.32(0.19)	0.33(0.21)	**0.09**(0.20)	0.25(0.18)
weib/ exp	0.74(0.23)	0.81(0.22)	0.65(0.29)	**0.57**(0.23)	0.71(0.22)
weib/*weib*	0.69(0.23)	0.74(0.24)	0.57(0.30)	**0.49**(0.24)	0.63(0.23)

Tabla 7: TE y LPI de NN 95 % con 0 % de censura para $n = 100$.

Estimador $\widehat{S}_{KM}(0.2)$					
Time/Cens	T.KM	Boot.KM	log	log(− log)	arc sen
exp / exp	0.06(0.18)	0.04(0.18)	**0.01**(0.28)	0.06(0.18)	0.06(0.18)
exp /*weib*	0.05(0.15)	0.07(0.15)	**0.02**(0.19)	0.06(0.16)	0.05(0.15)
weib/ exp	0.09(0.19)	0.19(0.18)	**0.04**(0.28)	0.06(0.18)	0.06(0.18)
weib/*weib*	0.09(0.19)	0.19(0.18)	**0.04**(0.28)	0.06(0.18)	0.06(0.18)
Estimador $\widehat{S}_{NA}(0.2)$					
Time/Cens	T.NA	Boot.NA	log	log(− log)	arc sen
exp / exp	0.06(0.18)	0.04(0.18)	**0.01**(0.28)	0.04(0.19)	0.06(0.18)
exp /*weib*	0.05(0.15)	0.07(0.15)	**0.02**(0.19)	0.06(0.16)	0.05(0.15)
weib/ exp	0.09(0.18)	0.18(0.18)	**0.04**(0.28)	0.06(0.19)	0.09(0.18)
weib/*weib*	0.09(0.18)	0.18(0.18)	**0.04**(0.28)	0.06(0.19)	0.09(0.18)

5.4. Resultados para $n = 100$

Se nota que los resultados para $n = 50$ son similares a $n = 25$. Es de resaltar que los I.C para la función de supervivencia mediante el remuestreo *Bootstrap* para Kaplan-Meier y Nelson-Aalen arrojan resultados inapropiados, esto se debe a que se utilizó la opción type = "norm", el cual es sensible. Por otra parte en el último escenario la comparación de estos intervalos de confianza para la función

Tabla 8: TE y LPI de NN 95 % con 25 % de censura para $n = 100$.

Estimador $\widehat{S}_{KM}(0.2)$					
Time/Cens	T.KM	Boot.KM	log	log(− log)	arc sen
exp / exp	0.36(0.17)	**0.23**(0.19)	**0.23**(0.24)	0.26(0.18)	0.31(0.17)
exp /*weib*	0.19(0.14)	0.23(0.15)	0.14(0.17)	**0.09**(0.15)	0.14(0.14)
weib/ exp	0.60(0.17)	0.50(0.17)	**0.44**(0.24)	0.50(0.18)	0.57(0.17)
weib/*weib*	0.54(0.18)	0.56(0.19)	**0.38**(0.25)	0.45(0.18)	0.51(0.18)
Estimador $\widehat{S}_{NA}(0.2)$					
Time/Cens	T.NA	Boot.NA	log	log(− log)	arc sen
exp / exp	0.37(0.17)	0.33(0.20)	**0.24**(0.24)	0.27(0.18)	0.32(0.17)
exp /*weib*	0.20(0.14)	0.21(0.21)	0.15(0.17	**0.09**(0.15)	0.14(0.14)
weib/ exp	0.62(0.17)	0.62(0.22)	**0.45**(0.24)	0.52(0.18)	0.59(0.17)
weib/*weib*	0.56(0.18)	0.62(0.22)	**0.40**(0.24)	0.46(0.18)	0.53(0.18)

Tabla 9: TE y LPI de NN 95 % con 45 % de censura para $n = 100$.

Estimador $\widehat{S}_{KM}(0.2)$					
Time/Cens	T.KM	Boot.KM	log	log(− log)	arc sen
exp / exp	0.89(0.16)	0.89(0.16)	**0.80**(0.20)	0.83(0.16)	0.87(0.16)
exp /*weib*	0.53(0.13)	0.54(0.13)	0.45(0.15)	**0.35**(0.13)	0.48(0.13)
weib/ exp	0.98(0.16)	0.94(0.16)	**0.93**(0.20)	0.96(0.16)	0.97(0.16)
weib/*weib*	0.94(0.17)	0.87(0.17)	**0.88**(0.21)	0.91(0.17)	0.93(0.16)
Estimador $\widehat{S}_{NA}(0.2)$					
Time/Cens	T.NA	Boot.NA	log	log(− log)	arc sen
exp / exp	0.90(0.16)	0.91(0.16)	**0.81**(0.20)	0.83(0.16)	0.88(0.16)
exp /*weib*	0.54(0.13)	0.56(0.13)	0.51(0.15)	**0.35**(0.13)	0.48(0.13)
weib/ exp	0.98(0.16)	**0.93**(0.16)	0.94(0.20)	0.96(0.16)	0.98(0.16)
weib/*weib*	0.95(0.16)	**0.87**(0.16)	0.89(0.21)	0.91(0.17)	0.94(0.16)

de supervivencia en muestras de tamaño $n = 100$, los resultados respaldan lo mencionado con los demás tamaños de muestra, sin embargo es de resaltar que a medida que se aumenta el porcentaje de censura al caso más extremo, las tasas de error son bastantes altas en particular al aumentar de 35 % a 45 % observaciones censuradas en la muestra simulada, resultando similar a los resultados de Borgan & Liestøl (1990) para $t = 0.2$, sin embargo los autores no incluyen los I.C mediante la transformación log(− log) que en este trabajo resultan ser mejores.

Por otra parte los resultados de los I.C mediante la transformación log(− log) cuando se presenta altos porcentajes de censura coinciden con los resultados de Anderson, Bernstein & Pike (1982). Teniendo en cuenta los resultados anteriores en algunos casos se presentan confusiones para la escogencia del IC que presenta mejores resultados por lo que se implementó la metodología de Correa & Sierra (2003) para comparar dichos intervalos a través del índice propuesto por dichos autores, modificando el nivel de confianza nominal (NN = 0.9, 0.95 y 0.99), lo anterior se muestra tomando la distribución de fallas/censura exp / exp, esto con el fin de presentar el funcionamiento del índice presentado en el anexo.

Es de resaltar que mientras mayor sea el índice mejor será el intervalo de confianza, sin embargo éste índice fue propuesto para comparar intervalos de confianza para diferencia de proporciones, lo que cambiaría el rango de valores resultantes,

pero la analogía de la interpretación se mantiene. Cabe resaltar que se realizaron 2000 simulaciones para la comparación y $B = 1000$ remuestras *Bootstrap*, estos valores se consideraron teniendo en cuenta que no se presentaron cambios significativos para un número mayor de simulaciones en las estimaciones.

6. Conclusiones

Cuando no se presentan observaciones censuradas en los datos los I.C mediante la transformación log poseen menor tasas de error, independientemente de la distribución de falla/censura y tamaños de muestra, como también el estimador utilizado, esto se debe a que resultan ser más amplios.

A medida que se aumenta el porcentaje de censura, los IC mediante la transformación $\log(-\log)$ resultan ser más efectivos, en general para los modelos generadores, tamaños de muestra y estimador de supervivencia utilizado.

Cuando las falla/censura se distribuyen exp / exp los resultados de la estimación de $S(0.2)$ resultan ser mejores que las demás combinaciones de distribución de falla/censura, en particular utilizando el estimador de KM.

[Recibido: septiembre de 2010 — Aceptado: marzo de 2011]

Referencias

Aalen, O. & Johansen, S. (1978), 'An empirical transition matrix for nonhomogeneus Markov chains based on censored observations', *Scandinavian Journal of Statistics* **5**(3), 141–150.

Akritas, M. (1986), 'Bootstrapping the Kaplan-Meier estimator', *American Statistical Association* **81**(396), 1032–1038.

Anderson, J., Bernstein, L. & Pike, M. (1982), 'Confidence intervals for probabilities of survival and quantiles in life-table analysis', *Biometrics* **38**(2), 407–416.

Borgan, Ø. & Liestøl, K. (1990), 'A note on confidence intervals and bands for the survival function based on transformations', *Scandinavian Journal of Statistics* **17**(1), 35–41.

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

Greenwood, M. (1926), 'The natural duration of cancer', *Reports on Public Health and Medical Subjects* (33), 1–26.

Kalbfleisch, J. & Prentice, R. (1980), *The Statistical Analysis of Failure Time Data*, Wiley, New York, United States.

Kaplan, E. & Meier, P. (1958), 'Estimation from incomplete observations', *American Statistical Association* **53**(282), 457–481.

Nair, V. (1984), 'Confidence bands for survival functions with censored data', *Technometrics* **26**(3), 265–275.

Nelson, W. (1969), 'Hazard plotting for incomplete failure data', *Journal of Quality Technology* **61**(1), 27–52.

R Development Core Team (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

Apéndice
Tablas

Tabla 10: Índice de IC para $\widehat{S}_{KM}(t)$ con 15 % de censura.

NN	Trad.K-M	Boot.K-M	log	log(− log)	arc sen
		n=25			
90 %	0.85	0.84	0.87	0.89	0.84
95 %	0.78	0.78	0.79	0.81	0.80
99 %	0.73	0.73	0.73	0.74	0.73
		n=50			
90 %	0.83	0.86	0.89	0.90	0.84
95 %	0.82	0.82	0.85	0.87	0.83
99 %	0.80	0.79	0.81	0.82	0.81
		n=75			
90 %	0.83	0.85	0.90	0.88	0.86
95 %	0.83	0.83	0.87	0.87	0.85
99 %	0.82	0.83	0.84	0.85	0.84
		n=100			
90 %	0.89	0.82	0.90	0.85	0.85
95 %	0.84	0.83	0.85	0.88	0.85
99 %	0.83	0.82	0.83	0.86	0.84

TABLA 11: Índice de IC para $\widehat{S}_{NA}(t)$ con 15 % de censura.

NN	Trad.N-A	Boot.N-A	log	log(− log)	arc sen
		n=25			
90 %	0.80	0.81	0.87	0.88	0.85
95 %	0.78	0.77	0.79	0.81	0.79
99 %	0.73	0.75	0.73	0.74	0.73
		n=50			
90 %	0.83	0.86	0.89	0.90	0.84
95 %	0.82	0.85	0.84	0.86	0.82
99 %	0.79	0.80	0.81	0.82	0.81
		n=75			
90 %	0.82	0.86	0.89	0.90	0.82
95 %	0.82	0.82	0.87	0.87	0.84
99 %	0.82	0.82	0.84	0.85	0.83
		n=100			
90 %	0.88	0.80	0.89	0.89	0.80
95 %	0.84	0.74	0.85	0.88	0.85
99 %	0.82	0.71	0.84	0.86	0.84

Información para los autores

La **Revista Colombiana de Estadística** publica artículos originales de carácter teórico o aplicado en cualquiera de las ramas de la estadística. Se consideran también artículos divulgativos de gran calidad de exposición sobre metodologías o técnicas estadísticas aplicadas en diferentes campos del saber. Únicamente se publican artículos en español e inglés, si el autor escribe en una lengua diferente a la nativa debe enviar un certificado de un traductor oficial o de un corrector de estilo que haya revisado el texto.

El Comité Editor únicamente acepta trabajos para evaluación que no han sido publicados previamente y que no están siendo propuestos simultáneamente para publicación en otros medios, ni lo serán sin previo consentimiento del Comité, a menos que, como resultado de la evaluación, se decida no publicarlos en la Revista. Se supone además que cuando los autores hacen entrega de un documento con fines de publicación en la **Revista Colombiana de Estadística**, conocen las condiciones anteriores y que están de acuerdo con ellas.

Material

Los artículos remitidos a la **Revista Colombiana de Estadística** deben ser presentados en archivo PDF o PS, con textos, gráficas y tablas en color negro y, además, los autores deben agregar una versión del artículo sin nombres ni información de los autores, que se utilizará para el arbitraje. Se debe enviar una carta firmada por cada uno de los autores, donde manifiesten estar de acuerdo con someter el artículo y con las condiciones de la Revista. Si un artículo es aceptado, los autores deben poner a disposición del Comité Editorial los archivos: fuente en LaTeX y de gráficas en formato EPS en blanco y negro.

Para facilitar la preparación del material publicado se recomienda utilizar MiKTeX[1], usando los archivos de la plantilla y del estilo *revcoles* disponibles en la página Web de la Revista[2] y siguiendo las instrucciones allí incorporadas.

Todo artículo debe incluir:

- Título en español y su traducción al inglés.
- Los nombres completos y el primer apellido, la dirección postal o electrónica y la afiliación institucional de cada autor.
- Un resumen con su versión en inglés (*abstract*). El resumen en español no debe pasar de 200 palabras y su contenido debe destacar el aporte del trabajo en el tema tratado.
- Palabras clave (*Key words*) en número entre 3 y 6, con su respectiva traducción al inglés, siguiendo las recomendaciones del *Current Index to Statistics* (CIS)[3].

[1] http://www.ctan.org/tex-archive/systems/win32/miktex/
[2] http://www.estadistica.unal.edu.co/revista
[3] http://www.statindex.org/CIS/homepage/keywords.html

- Cuando el artículo se deriva de una tesis o trabajo de grado debe indicarse e incluirse como una referencia.
- Si se deriva de un proyecto de investigación, se debe indicar el título del proyecto y la entidad que lo patrocina.
- Referencias bibliográficas, incluyendo solamente las que se hayan citado en el texto.

Referencias y notas al pie de página

Para las referencias bibliográficas dentro del texto se debe utilizar el formato autor-año, dando el nombre del autor seguido por el año de la publicación dentro de un paréntesis. La plantilla LaTeX suministrada utiliza, para las referencias, los paquetes BibTeX y Harvard[4]. Se recomienda reducir el número de notas de pie de página, especialmente las que hacen referencia a otras notas dentro del mismo documento y no utilizarlas para hacer referencias bibliográficas.

Tablas y gráficas

Las tablas y las gráficas, con numeración arábiga, deben aparecer referenciadas dentro del texto mediante el número correspondiente. Las tablas deben ser diseñadas en forma que se facilite su presentación dentro del área de impresión de la Revista. En este sentido, los autores deben considerar en particular la extensión de las tablas, los dígitos representativos, los títulos y los encabezados. Las gráficas deben ser visualmente claras y debe ser posible modificar su tamaño. Cuando el artículo sea aceptado para su publicación, los autores deben poner la versión definitiva a disposición del Comité Editorial. Todos los elementos como barras, segmentos, palabras, símbolos y números deben estar impresos en color negro.

Responsabilidad legal

Los autores se hacen responsables por el uso de material con propiedad intelectual registrada como figuras, tablas, fotografías, etc.

Arbitraje

Los artículos recibidos serán revisados por el Comité Editorial y sometidos a arbitraje por pares especializados en el tema respectivo. El arbitraje es "doble ciego" (árbitros anónimos para los autores y viceversa). El Comité Editorial decide aceptar, rechazar o solicitar modificaciones a los artículos con base en las recomendaciones de los árbitros.

[4] http://tug.ctan.org/tex-archive/macros/latex/contrib/harvard

La Revista Colombiana de Estadística agradece a las personas, que no integran los Comités Editorial y Científico, por su colaboración en el volumen 33 (2010).

Alejandro C. Frery, Ph.D.
Alex Rojas, Ph.D.
Alexandre de Bustamente Simas, Ph.D.
Ana Pérez Espartero, Ph.D.
Andres Gutiérrez, Ms.C.
Arnaldo Goitia, Ph.D.
Arturo Ruiz, M.Sc.
Aurea Grané, Ph.D.
Beatriz Bernabé Loranca, Ph.D.
Bernardo Lanza, Ph.D.
Carlos Diniz, Ph.D.
Carlos Javier Barrera, Ph.D.
Carlos Mario Lopera, Ph.D.
Carlos Maté, Ph.D.
Daniel Barraez, M.D.
Dave Marx, Ph.D.
Edilberto Cepeda, Ph.D.
Elkin Castaño, M.Sc.
Enrique Villa, Ph.D.
Eva Cristina Manotas, Ph.D.
Francisco Montes, Ph.D.
Francisco Palomino, Ph.D.
Frederico Zanqueta Poleto, Ph.D.
Heather Tierney, Ph.D.
Heleno Bolfarine, Ph.D.
Hiram Beltran-Sanchez, Ph.D.
Holger Capa, Ph.D.
Humberto Llinas, Ph.D.
Ignacio Méndez, Ph.D.
Jaime Abel Huertas, M.Sc.
Jairo Andrés Rendón, Ph.D.
Jesús Orbe, Pregrado
José A. Villaseñor, Ph.D.
Jorge Luis Bazán, Ph.D.
Jorge Luis Romeu, Ph.D.
Jorge Mateu Mahiques, Ph.D.
Juan Carlos García, M.Sc.
Juan Kalemkerian, Ph.D.
Julio Singer, Ph.D.
Leonardo Duarte Vergara, M.Sc.
Luis Guillermo Díaz, M.Sc.
Luis Medina, Ph.D.
Ludovic Lebart, Ph.D.
Marco Scavino, Ph.D.
Maurizio Zevallos, Ph.D.
Mario de Castro, Ph.D.
Mark Reckase, Ph.D.
Miren I. Portilla, Ph.D.
Norberto Rodríguez, M.Sc.
Oscar Soto, M.Sc.
Pablo Martínez, Ph.D.
Pedro Monterrey, M.Sc.
Raúl Pérez, M.Sc.
Raúl Ramírez, Pregrado
Silvia María Freitas, Ph.D.
Sveli Mingoti, Ph.D.
Thimothy E. Hanson, Ph.D.
Yolanda Villacampa, M.Sc.
Víctor Aguirre Torres, Ph.D.
Víctor López, Ph.D.
Víctor M. Gonzalez, M.Sc.