Revista Colombiana de Estadística

Colombian Journal of Statistics

Volumen 37. Número 1 - junio - 2014 **ISSN 0120 - 1751**

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

Revista Colombiana de Estadística

Colombian Journal of Statistics

http://www.estadistica.unal.edu.co/revista
http://es.wikipedia.org/wiki/Revista_Colombiana_de_Estadistica
http://www.emis.de/journals/RCE/
revcoles_fcbog@unal.edu.co

Indexada en: Ulrichsweb, 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. 10878
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: Proceditor, Tel. 2204275-2204276, Bogotá, D.C.

Revista Colombiana de Estadística	Bogotá	Vol. 37	Nº 1
ISSN 0120 - 1751	COLOMBIA	junio-2014	Págs. 1-245

Contenido

Editorial

Leonardo Trujillo[a]

Department of Statistics, Universidad Nacional de Colombia, Bogotá, Colombia

Welcome to the first issue of the 37th volume of Revista Colombiana de Estadistica (Colombian Journal of Statistics). We are very proud to announce that the Colombian Journal of Statistics have maintained its categorization as an A1 Journal by Publindex (Colciencias) which ranges the journals in the country, being A1 the maximum category. Thanks to all the Editorial and Scientific Committees and Patricia Chávez, our assistant in the Journal, as this is a result of the continuous help obtained from all of them. More information available at `http://201.234.78.173:8084/publindex/EnIbnPublindex/resultados.do`

In this issue, we are having the highest number of papers in the history of the journal so far: fifteen (15) papers in total with the particularity of all being written in English language. We think we can keep this minimum number of fifteen (15) for posterior issues in order to make the journal more readable and useful for people interested in Statistics.

The topics in this current issue range over diverse areas of statistics: four papers in Probability Distributions by Al-Zahrani and Sagor; Ameli, Jarrahiferiz and Borzadaran; Gomez, Gomez and Bolfarine; and Jafari, Tahmasebi and Alizadeh; three papers in Survey Sampling by Ashgar, Sanaullah and Hanif; Shabbir, Haq and Gupta; and Yan and Xue; two papers in Multivariate Analysis by Abril, Gavilan and Velasco-Morente; and Riaz, Munir and Ashgar; two papers in Regression Analysis by Acosta, Cabrera, Vega and Cabrera; and Martinez and Barrera; one paper in Biostatistics by Velez and Correa; one paper in Experimental Design by Gaviria and Lopez-Rios; one paper in Nonparametric Statistics by Martinez-Camblor, Carleos and Corral; and finally, one paper in Statistical Quality Control by Guevara, Vargas and Linero.

A Special Issue in Current Topics in Statistical Graphics will be published in December this year. This special issue has the purpose of bringing together current advances and uses of well-known and novel graphical methods from different research areas so that the reader finds potential applications to his/her own research field. Many thanks to our guest editor, Professor Fernando Marmolejo (`fernando.marmolejoramos@adelaide.edu.au`) for this hard task in order to compile high level papers in Statistics and with potential applications in engineering, manufacturing, process/chemical industry, physical sciences, social sciences, and agricultural industries.

[a] Editor in Chief
E-mail: ltrujilloo@unal.edu.co

The XIII CLAPEM (Latin American Congress of Probability and Mathematical Statistics) keeps growing in the number of participant institutions and its organization. This event will be held for the first time in Colombia at the city of Cartagena with the help of the Latin American Chapter of the Bernoulli Society. CLAPEM is the largest conference gathering scientists in the particular areas of Probability and Mathematical Statistics in the region and takes place every two/three years. It has already been organized in Argentina, Brazil, Chile, Cuba, Mexico, Peru, Uruguay and Venezuela. The CLAPEM activities include lectures held by invited researchers, satellite meetings, sessions of oral and poster contributions, short courses, and thematic sessions. The XIII CLAPEM is organized by the Bernoulli Society, National University of Colombia, Universidad de los Andes, Universidad Central, Universidad Industrial de Santander, Universidad Antonio Nariño, Universidad EAFIT, Universidad Sergio Arboleda, Universidad Pedagógica y Tecnológica de Colombia and Universidad de Cartagena. The Scientific Committee is as follows: Alejandro Jara (Chile), Antonio Galves (Brazil), Graciela Boente (Argentina), José Rafael Leon (Venezuela), Karine Bertin (Chile), Leonardo Trujillo (Colombia), Pablo Ferrari (Argentina), Paola Belmolen (Uruguay), Ramón Giraldo (Colombia), Serguei Popov (Brazil), Victor Perez Abreu (Mexico). We are confident of the success of this event to be held from September 22nd to the 26th this year. Three courses will take place in this event: Confidence Distribution (CD): A New approach in Distributional Inference and its Applications by Professor Regina Liu from Rutgers University, USA; Stochastic Models of Population Genetics by Professor Alison Etheridge from Oxford University, UK; Topics in Quantitative Risk Management by Professor Paul Embrechts from ETH Zurich, Switzerland. If you are interested you can also get more details with Ricardo Fraiman (president of the XIII CLAPEM, `fraimanricardo@gmail.com`), Leonardo Trujillo (`ltrujilloo@unal.edu.co`) or in the official website `www.clapem.unal.edu.co`

Also in Colombia, as it has been tradition every year, the XXIV Colombian Symposium in Statistics will be held from July 24th to the 26th in the AR Hotel in Bogotá, Colombia. Five international speakers will be offering short courses and conferences: Adriana Pérez from The University of Texas (USA) in Biostatistics, Daniel Thorburn from Stockholm University (Sweden) in Official Statistics, Nikos Tzavidis from the University of Southampton (UK) in Survey Sampling, Thibaut Jombart from the Imperial College of London (UK) in Statistical Software and Victor Guerrero from the ITAM (Mexico) in Time Series Analysis.

I would like to congratulate our colleagues in Russia who have established the Russian Statistical Association on April this year. The aim is "to facilitate the collaboration between the national statistical office, business and universities, contribute to efficient government decisions based on transparency and reliable scientific data, unite all people interested in promoting the role of statistics in country development, and engage Russian statisticians in ISI activities". We are still awaiting for our Colombian Statistical Association to be reestablished after these long years of inactivity. Clearly the experience of our Russian colleagues could help us to take this path back soon in order to integrate DANE (the national statistical office), interested particulars and the growing number of universities with programs and researchers in Statistics around the country.

Editorial

Leonardo Trujillo[a]

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

Bienvenidos al primer número del volumen 37 de la Revista Colombiana de Estadística (Colombian Journal of Statistics). Estamos muy gratos en anunciar que la Revista Colombiana de Estadística ha mantenido su categoría A1 ante Publindex (Colciencias) que categoriza las revistas a nivel nacional y siendo esta la máxima categoría de calidad para revistas nacionales. Gracias a todos los Comités Científico y Editorial y a Patricia Chávez, la asistente de la Revista, pues este es el resultado de la continua ayuda obtenida por parte de todos ellos. Mas información disponible en la página web `http://201.234.78.173:8084/publindex/EnIbnPublindex/resultados.do`

En este número, tendremos la cantidad más alta de artículos en la historia de la revista: quince (15) artículos en total con la particularidad de estar todos escritos en idioma inglés. Estamos seguros de poder mantener este número mínimo de quince (15) artículos para posteriores ediciones con el fin de hacer la revista más visible y útil para investigadores interesados en la estadística.

Los temas del presente número varían a través de diversas áreas de la estadística: cuatro artículos en Distribuciones de Probabilidad de Al-Zahrani y Sagor; Ameli, Jarrahiferiz y Borzadaran; Gómez, Gómez y Bolfarine; Jafari, Tahmasebi y Alizadeh; tres artículos en Muestreo de Ashgar, Sanaullah y Hanif; Shabbir, Haq y Gupta; Yan y Xue; dos artículos en Análisis Multivariado de Abril, Gavilán y Velasco-Morente; Riaz, Munir y Ashgar; dos artículos en Análisis de Regresión de Acosta, Cabrera, Vega y Cabrera; Martínez y Barrera; un artículo en Bioestadística de Vélez y Correa; un artículo en Diseño Experimental de Gaviria y López-Ríos; un artículo en Control de Calidad de Guevara, Vargas y Linero; y finalmente, un artículo en Estadística no Paramétrica de Martínez-Camblor, Carleos y Corral.

Un número especial en “Current Topics in Statistical Graphics” será publicado en Diciembre de este año. Este número especial tiene el propósito de integrar avances recientes y el uso de métodos gráficos bien conocidos en diferentes áreas de investigación. Muchas gracias a nuestro editor invitado, profesor Fernando Marmolejo (`fernando.marmolejoramos@adelaide.edu.au`) por esta ardua labor de compilar artículos de alto nivel en estadística con gran potencial de aplicación en ingeniería, manufactura, ciencias físicas, ciencias sociales y agricultura, entre otros.

La XIII CLAPEM (Conferencia Latinoamericana de Probabilidad y Estadística Matemática) continúa creciendo en términos de instituciones participantes y organización. Este evento será organizado por primera vez en Colombia en la ciudad

[a] Editor General
E-mail: ltrujilloo@unal.edu.co

de Cartagena de Indias con la ayuda del Capitulo Latinoamericano de la Sociedad Bernoulli. CLAPEM es la principal y mayor conferencia que reúne científicos en las áreas de Probabilidad y Estadística Matemática en la región y toma lugar cada dos o tres años. Se ha llevado a cabo anteriormente en Argentina, Brasil, Chile, Cuba, México, Perú, Uruguay y Venezuela. Las actividades del CLAPEM incluyen charlas a cargo de investigadores internacionales invitados, cursos cortos, reuniones satélites, sesiones de contribuciones orales y posters y sesiones temáticas. La XIII CLAPEM será organizada por la Sociedad Bernoulli, la Universidad Nacional de Colombia, Universidad de los Andes, Universidad Central, Universidad Industrial de Santander, Universidad Antonio Nariño, Universidad EAFIT, Universidad Sergio Arboleda, Universidad Pedagógica y Tecnológica de Colombia y Universidad de Cartagena. El Comité Científico esta conformado por: Alejandro Jara (Chile), Antonio Galves (Brasil), Graciela Boente (Argentina), José Rafael León (Venezuela), Karine Bertín (Chile), Leonardo Trujillo (Colombia), Pablo Ferrari (Argentina), Paola Belmolen (Uruguay), Ramón Giraldo (Colombia), Serguei Popov (Brasil), Víctor Pérez Abreu (México). Estamos seguros del éxito de este evento a realizarse de Septiembre 22 al 26, 2014. Tres cursos tomarán lugar en este evento: Confidence Distribution (CD): A New approach in Distributional Inference and its Applications por la Profesora Regina Liu de Rutgers University, USA; Stochastic Models of Population Genetics por la Profesora Alison Etheridge de la Universidad de Oxford, Inglaterra; Topics in Quantitative Risk Management por el Profesor Paul Embrechts de ETH Zurich, Suiza. Si esta interesado puede obtener mas detalles acerca del evento puede contactar a Ricardo Fraiman (presidente del XIII CLAPEM, `fraimanricardo@gmail.com`), con Leonardo Trujillo (`ltrujilloo@unal.edu.co`) o en la página oficial `www.clapem.unal.edu.co.`

También en Colombia y como ha sido tradición anualmente, el XXIV Simposio Internacional de Estadística tendrá lugar de Julio 24 al 26 en el Hotel AR en Bogotá. Cinco conferencistas internacionales ofrecerán conferencias y cursos cortos: Adriana Perez de University of Texas (USA) en Bioestadística, Daniel Thorburn de Stockholm University (Suecia) en Estadísticas Oficiales, Nikos Tzavidis de la Universidad de Southampton (Inglaterra) en Muestreo, Thibaut Jombart del Imperial College of London (UK) en Software Estadístico y Víctor Guerrero del ITAM (México) en Análisis de Series Temporales.

Quisiera finalizar enviando una sentida felicitación a nuestros colegas en Rusia quienes han establecido la Asociación Rusa de Estadística desde Abril de este año. El objetivo como ellos mismos manifiestan es facilitar la colaboración entre la oficina de estadísticas nacionales, diversas empresas y la academia, contribuir a la toma de decisiones gubernamentales eficientes basadas en la transparencia y en datos científicos confiables, congregar a todas las personas interesadas en promover el uso de la estadística para el desarrollo del país, y promover a los estadísticos rusos en las actividades del ISI. Aun continuamos a la espera de un restablecimiento de nuestra Sociedad Colombiana de Estadística después de muchos años de inactividad. Claramente la experiencia de nuestros colegas rusos podrían ayudarnos en regresar a este camino con el fin de integrar al DANE (oficina nacional de estadística), particulares interesados y el reciente número en ascenso de universidades con programas o investigadores en estadística alrededor del país.

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 1 a 24

On the Performance Evaluation of Different Measures of Association

Evaluación de diferentes medidas de asociación

Muhammad Riaz[1,a], Shahzad Munir[2,b], Zahid Asghar[2,c]

[1] Department of Mathematics and Statistics, King Fahad University of Petroleum and Minerals, Dhahran, Saudi Arabia

[2] Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan

Abstract

In this article our objective is to evaluate the performance of different measures of associations for hypothesis testing purposes. We have considered different measures of association (including some commonly used) in this study, one of which is parametric and others are non-parametric including three proposed modifications. Performance of these tests are compared under different symmetric, skewed and contaminated probability distributions that include Normal, Cauchy, Uniform, Laplace, Lognormal, Exponential, Weibull, Gamma, t, Chi-square, Half Normal, Mixed Weibull and Mixed Normal. Performances of these tests are measured in terms of power. We have suggested appropriate tests which may perform better under different situations based on their efficiency grading(s). It is expected that researchers will find these results useful in decision making.

***Key words*:** Measures of association, Non-Normality, Non-Parametric methods, Normality, Parametric methods, Power.

Resumen

En este articulo el objetivo es evaluar el desempeño de diferentes medidas de asociación para pruebas de hipótesis. Se consideran diferentes medidas, algunas paramétricas y otras no paramétricas, así como tres modificaciones propuestas por los autores. El desempeño de estas pruebas se evalúa considerando distribuciones simétricas, sesgadas y contaminadas incluyendo la distribución normal, Cauchy, uniforme, Laplace, lognormal, exponencial, Weibull, Gamma, t, Chi-cuadrado, medio normal, Weibull mezclada y normal mezclada. El desempeño se evalúa en términos de la potencia de los tests. Se sugieren tests apropiados que tienen un mejor desempeño bajo

[a] Professor. E-mail: riaz76qau@yahoo.com
[b] Professor. E-mail: farhan.saif@qau.edu.pk
[c] Professor. E-mail: g.zahid@gmail.com

diferentes niveles de eficiencia. Se espera que los investigadores encuentren estos resultados útiles en la toma de decisiones.

***Palabras clave*:** medidas de asociación, no normalidad, métodos no paramétricos, métodos paramétricos, potencia.

1. Introduction

It is indispensable to apply statistical tests in almost all the observational and experimental studies in the fields of agriculture, business, biology, engineering etc. These tests help the researchers to reach at the valid conclusions of their studies. There are number of statistical testing methods in literature meant for different objectives, for example some are designed for association dispersion, proportion and location parameter(s). Each method has a specific objective with a particular frame of application. When more than one method qualifies for a given situation, then choosing the most suitable one is of great importance and needs extreme caution. This mostly depends on the properties of the competing methods for that particular situation. From a statistical viewpoint, power is considered as an appropriate criterion of selecting the finest method out of many possible ones. In this paper our concern is with the methods developed for measuring and testing the association between the variables of interest defined on a some population(s). For the sake of simplicity we restrict ourselves with the environment of two correlated variables i.e. the case of bivariate population(s).

The general procedural framework can be laid down as follows: Let we have two correlated random variables of interest X and Y defined on a bivariate population with their association parameter denoted by ρ. To test the hypothesis $H_0 : \rho = 0$ (i.e. no association) vs. $H_1 : \rho \neq 0$, we have a number of statistical methods available depending upon the assumption(s) regarding the parent distribution(s). In parametric environment the usual Pearson correlation coefficient is the most frequent choice (cf. Daniel 1990) while in non parametric environment we have many options. To refer the most common of these: Spearman rank correlation coefficient introduced by Spearman (1904); Kendall's tau coefficient proposed by Kendall (1938); a modified form of Spearman rank correlation coefficient which is known as modified rank correlation coefficient proposed by Zimmerman (1994); three Gini's coefficients based measures of association given by Yitzhaki (2003) (two of which are asymmetrical measures and one is symmetrical). We shall refer all the aforementioned measures with the help of notations given in Table 1 throughout this chapter.

This study is planned to investigate the performance of different measures of association under different distributional environments. The association measures covered in the study include some existing and some proposed modifications and performance is measured in terms of power under different probability models. The organization of the rest of the article is as: Section 2 provides description of different existing measures of association; Section 3 proposes some modified measures of association; Section 4 deals with performance evaluations of these measures; Section 5 offers a comparative analysis of these measures; Section 6 includes an illustrative example; Section 7 provides summary and conclusions of the study.

Table 1: Notations.

r_P	The usual Pearson Product Moment Correlation Coefficient (cf. Daniel 1990) proposed by Karl Pearson
r_S	Spearman Rank Correlation Coefficient (cf. Spearman 1904)
r_M	Modified Rank Correlation Coefficient (cf. Zimmerman 1994)
r_{g1}	Gini Correlation Coefficient between X and Y (asymmetric) (cf. Yitzhaki 2003)
r_{g2}	Gini Correlation Coefficient between Y and X (asymmetric) (cf. Yitzhaki 2003)
r_{g3}	Gini Correlation Coefficient between X and Y or between Y and X (symmetric) (cf. Yitzhaki 2003)
τ	Kendall's Tau (cf. Kendall 1938)

2. Measures of Association

In order to define and describe the above mentioned measures, let we have two dependent random samples in the form of pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ drawn from a bivariate population (with the association parameter ρ) under all the assumptions needed for a valid application of all the association measures under consideration. The description of the above mentioned measures along with their main features and their respective test statistics are provided below:

Pearson Product Moment Correlation Coefficient (r_P): It is a measure of the relative strength of the linear relationship between two numerical variables of interest X and Y. The mathematical definition for this measure (denoted by r_P) is given as:

$$r_P = \frac{\operatorname{cov}(X, Y)}{SD(X)SD(Y)} \tag{1}$$

where $\operatorname{cov}(X, Y)$ refers to the covariance between X and Y; $SD(X)$ and $SD(Y)$ are the standard deviations of X and Y respectively.

The value of r_P ranges from -1 to $+1$ implying perfect negative and positive correlation respectively. A value of zero for r_P means that there is no linear correlation between X and Y. It requires the data on at least interval scale of measurement. It is a symmetric measure that is invariant of the changes in location and scale. Geometrically it is defined as the cosine of the angle between the two regression lines (Y on X and X on Y). It is not robust to the presence of outliers in the data. To test the statistical significance of r_P we may use the usual t-test (under normality) and even under non-normality t-test may be a safe approximation.

Spearman Rank Correlation Coefficient (r_S): It is defined as the Pearson product moment correlation coefficient between the ranked information of X and Y rather than their raw scores. The mathematical definition for this measure (denoted by r_S) is given as:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2 - 1)} \tag{2}$$

where n is the sample size; $\sum_{i=1}^{n} D_i^2$ is the sum of the squares of the differences between the ranks of two samples after ranking the samples individually. It is a non-parametric measure that lies between -1 to $+1$ (both inclusive) referring to perfect negative and positive correlations respectively. The sign of r_S indicates the direction of relationship between the actual variables of interest. A value of zero for r_S means that there is no interdependency between the original variables. It requires the data on at least ordinal scale. Using normal approximation, the statistical significance of r_S may tested using the usual t-test. Modified Rank Correlation Coefficient (r_M): It is a modified version of Spearman rank correlation coefficient based on transformations of X and Y into standard scores and then using the concept of ranking.The mathematical definition for this measure (denoted by r_M) is given as:

$$r_M = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{3}$$

where d is the difference between the ranks assigned transforming the values of X and Y separately into standard scores, assigning the ranks to standard scores collectively and then make separate groups of the ranks according to their respective random samples. Now defines the difference between the ranks and $\sum_{i=1}^{n} d_i^2$ in (3) is the sum of the squares of the differences between the ranks.

It is also a non-parametric measure that may take zero value for no correlation, positive value and negative values for negative and positive correlations respectively, as in the above case. A value of -1 refers to the perfect correlations among the variables of interest.

Gini Correlation Coefficient (Asymmetric and Symmetric): These correlation measures are based on the covariance measures between the original variables X and Y and their cumulative distribution functions $F_X(X)$ and $F_Y(Y)$. We consider here three measures of association based on Gini's coefficients (two of which are asymmetrical measures and one is symmetrical). These measures of association, denoted by r_{g1}, r_{g2} and r_{g3}, are defined as:

$$r_{g1} = \frac{\operatorname{cov}(X, F_Y(Y))}{\operatorname{cov}(X, F_X(X))} \tag{4}$$

$$r_{g2} = \frac{\operatorname{cov}(Y, F_X(X))}{\operatorname{cov}(Y, F_Y(Y))} \tag{5}$$

$$r_{g3} = \frac{G_X r_{g1} + G_Y r_{g2}}{G_X + G_Y} \tag{6}$$

where $\operatorname{cov}(X, F_Y(Y))$ is the covariance between X and cumulative distribution function of Y; $\operatorname{cov}(Y, F_X(X))$ is the covariance between X and its cumulative distribution function; $\operatorname{cov}(Y, F_X(X))$ is the covariance between Y and cumulative distribution function of X; $\operatorname{cov}(Y, F_Y(Y))$ is the covariance between Y and its cumulative distribution function; $G_X = 4\operatorname{cov}(X, F_X(X))$ and $G_Y = 4\operatorname{cov}(Y, F_Y(Y))$.

In the above mentioned measures given in (4)-(6), r_{g1} and r_{g2} are the asymmetric Gini correlation coefficients while r_{g3} is the symmetric Gini correlation coefficient. Here are some properties of Gini correlation coefficients (cf. Yitzhaki 2003): The Gini coefficient is bounded, such that $+1 \geq r_{g_{js}} \geq -1 (j, s = X, Y)$. If X

and Y are independent then; $r_{g1} = r_{g2} = 0$; r_{g2} is not sensitive to a monotonic transformation of Y. In general, $r_{g_{js}}$ need not be equal to $r_{g_{sj}}$ and they may even have different signs. If the random variables Z_j and Z_s are exchangeable up to a linear transformation, then $r_{g_{js}} = r_{g_{sj}}$.

Kendall's Tau (τ): It is a measure of the association between two measured variables of interests X and Y. It is defined as the rank correlation based on the similarity orderings of the data with ranked setup. The mathematical definition for this measure (denoted by τ) is given as:

$$\tau = \frac{S}{\frac{n(n-1)}{2}} \tag{7}$$

where n is the size of sample and S is defined as the difference between the number of pairs in natural and reverse natural orders. We may define S more precisely as arranging the observations (X_i, Y_i) (where $i = 1, 2, \ldots, n$) in a column according to the magnitude of the $X's$, with the smallest X first, the second smallest second and so on. Then we say that the $X's$ are in natural order. Now in equation (7), S is equal to $P - Q$, where P is the number of pairs in natural order and Q is number of pairs in reverse order of random variable Y.

This measure is non-parametric being free from the parent distribution. It takes values between $+1$ and -1 (both inclusive). A value equal to zero indicates no correlation, $+1$ means perfect positive and -1 means perfect negative correlation. It requires the data on at least ordinal scale. Under independence its mean is zero and variance $2(2n + 5)/9n(n - 1)$.

3. Proposed Modifications

Taking the motivations from the aforementioned measures as given in equation (1)-(7) we suggest here three modified proposals to measure association. In order to define r_M in equation (3), Zimmerman (1994) used mean as an estimate of the location parameter to convert the variables into standard scores. Mean as a measure of location is able to produce reliable results when data is normal or at least symmetrical because it is highly affected by the presence of outliers as well as due to the departure from normality. It means that the sample mean is not a robust estimator and hence cannot give trustworthy outcomes. To overcome this problem, we may use median and trimmed mean as alternative measures. The reason being that in case of non-normal distributions and/or when outliers are present in the data median and trimmed mean exhibit robust behavior and hence the results based on them are expected to become more reliable than mean.

Based on the above discussion we now suggest here three modifications/proposals to measure the association. These three proposals are modified forms of Spearman rank correlation coefficient, namely i) trimmed mean rank correlation by using standard deviation about trimmed mean; ii) median rank correlation by using standard deviation about median; iii) median rank correlation by using mean deviation about median. These three proposals are based on Spearman rank correlation coefficient in which we shall transform the variables into standard scores (like in Zimmerman (1994) using the measures given in (i)-(iii) above. We shall refer the three proposed modifications with the help of notations given in Table 2 throughout this chapter.

Table 2: Notations Table for the Proposed Modifications.

r_T	Trimmed Rank Correlation Coefficient
r_{MM}	Median Rank Correlation Coefficient by using Mean Deviation about Median
r_{MS}	Median Rank Correlation Coefficient by using Standard Deviation about Median

Keeping intact the descriptions of equation (1)-(7) we now provide the explanation of the three proposed modified measures. Before that we defined here few terms used in the definitions of r_T, r_{MM} and r_{MS}. These terms include Standard Deviation by using Trimmed Mean (denoted by $SD_1(X)$ and $SD_1(Y)$ for X and Y respectively), Mean Deviation about Median (denoted by $MDM(X)$ and $MDM(Y)$ for X and Y respectively) and Standard Deviation by using Median (denoted by $SD_2(X)$ and $SD_2(Y)$ for X and Y respectively). These terms are defined as under:

$$SD_1(X) = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X}_t)^2}{n-1}} \qquad \text{and} \qquad SD_1(Y) = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_t)^2}{n-1}} \tag{8}$$

In equation (8), $\overline{X}_t$ and $\overline{Y}_t$ are the trimmed means of X and Y respectively.

$$MDM(X) = \frac{\sum_{i=1}^{n}|X_i - \tilde{X}|}{n} \qquad \text{and} \qquad MDM(Y) = \frac{\sum_{i=1}^{n}|Y_i - \tilde{Y}|}{n} \tag{9}$$

In equation (9), $\tilde{X}$ and $\tilde{Y}$ are the medians of X and Y respectively.

$$SD_2(X) = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \widetilde{X}_t)^2}{n-1}} \qquad \text{and} \qquad SD_2(Y) = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \widetilde{Y}_t)^2}{n-1}} \tag{10}$$

In equation (10), all the terms are as defined earlier.

Based on the above definitions we are now able to define r_T, r_M and r_{MS} as under:

$$r_T = 1 - \frac{6\sum_{i=1}^{n} d_{i,T}^2}{n(n^2-1)} \tag{11}$$

For equation (11); first we separately transform the values of random variables X and Y into standard scores by using their respective trimmed means and standard deviation about trimmed means of their respective random sample from (X,Y), assign the ranks to standard scores collectively and then separate the ranks according to their random samples. Now in equation (11), $\sum_{i=1}^{n} d_{i,T}^2$ is the sum of the squares of the differences between the ranks. It is to be mentioned that we have trimmed 2 values from each sample, so the percentages of trimming in our computations are 33%, 25%, 20%, 17%, 13%, 10% and 7% of samples 6, 8, 10, 12, 16, 20 and 30 respectively.

$$r_M = 1 - \frac{6\sum_{i=1}^{n} d_{i,MS}^2}{n(n^2-1)} \tag{12}$$

For equation (12); first we separately transform the values of random variables X and Y into standard scores by using their respective medians and standard deviation about medians of their respective random variables from X and Y, assign the ranks to standard scores collectively and then separate the ranks according to their random samples. Now in equation (12) $\sum_{i=1}^{n} d_{i,MS}^2$ is the sum of the squares of the differences between the ranks.

$$r_{MM} = 1 - \frac{6\sum_{i=1}^{n} d_{i,MM}^2}{n(n^2-1)} \tag{13}$$

For equation (13); first we separately transform the values of random variables X and Y into standard scores by using their respective medians and mean deviation about medians of the respective random sample from (X, Y), assign the ranks to standard scores collectively and then separate the ranks according to their random samples. Now in equation (13), $\sum_{i=1}^{n} d_{i,MM}^2$ is the sum of the squares of the differences between the ranks.

All the existing measures given in equation (1)-(7) and the proposed modifications given in equation (11)-(13) are nonparametric except the one given in equation (1). The existing measures as given equation (1)-(7) have many attractive properties in their own independent capacities (e.g. see Spearman 1904, Kendall 1938, Zimmerman 1994, Gauthier 2001, Yitzhaki 2003, Mudelsee 2003, Walker 2003, Maturi & Elsayigh 2010). But it is hard to find articles in the existing literature which compare the performance of these measures simultaneously under different distributional environments. The same is one of the motivations of this study. Additionally we plan to investigate the performances (in terms of power) of our proposed modifications under different probability models and also compare them with the existing counter parts. Although there are some other tests available to serve the purpose but the reason to choose these ten out of many is their novelty.

There are different ways to use the information (such as ratio, interval, ordinal and count) and each test has its own strategy to exploit this information. The tests considered here cover almost all of these common approaches. Although the results for the usual ones may be readily available but their comparisons in a broader frame will provide useful and interesting results. Actually the main objective of this study is to investigate the performance of these different methods/measures and see which of these have optimal efficiency under different distributional environments of the parent populations following line of action of Munir, Asghar & Riaz (2011).

This investigation would help us to grade the performance of these different methods for measuring and testing the association parameter under different parent situations. Consequently practitioners may take benefit out of it by picking up the most appropriate measure(s) to reach at the correct decision in a given situation. Practitioners generally prefer statistical measure(s) or method(s) which has higher power and they use it for their research proposals (cf. Mahoney & Magel 1996), so the findings of this research would be of great value for them for their future studies.

4. Performance Evaluations

Power is an important measure for the performance of a testing procedure. It is the probability of rejecting H_0 when it is false and it is the probability that a statistical measure(s)/procedure(s) will lead to a correct decision. In this section we intend to evaluate the power of the ten association measures under consideration in this study and find out which of them have relatively higher power(s) than the others under different parent situations. To calculate the power of different methods of measuring and testing the association under study we have followed the following procedure for power evaluations.

Let X and Y be the two correlated random variables referring to the two inter dependent characteristics of interest from where we have a random sample of n pairs in the form of (x_1, y_1), (x_2, y_2),…,(x_n, y_n) from a bivariate population. To get the desire level of correlation between X and Y the steps are listed as:

- Let X and Y be independent random variables and Y be a transformed random variable defined as: $Y = a(X) + b(W)$;
- The correlation between X and Y is given as: $r_{XY} = \frac{a}{\sqrt{a^2+b^2}}$, where a and b are unknown constants;
- The expression for a in the form of b and r_{XY} may be written as $a = \frac{b(r_{XY})}{\sqrt{1-r_{XY}^2}}$,
- If b=1 then we have: $a = \frac{r_{XY}}{\sqrt{1-r_{XY}^2}}$, and by putting the desire level of correlation in this equation we get the value of a;
- For the above mentioned values of a and b we can now obtain the variables X and Y having our desired correlation level.

Hypotheses and Testing Procedures: For our study purposes we state the null and alternative hypotheses as: $H_0 : \rho = 0$ versus H_1 i.e. $\rho > 0$. This is a one sided version of the hypothesis that may be easily defined for two sided case. It is supposed that the samples are drawn under all the assumptions needed for a valid application of all the methods related with the association measures of this study. We compute the values of our test statistics for association measures by using all the ten methods for different choices of ρ (on positive side only because of right sided alternative hypothesis) and calculate their chances of rejecting H_0 by comparing them with their corresponding critical values. These probabilities under H_0 refer to the significance level while under H_1 this will be power of the test. It is to be mentioned that to test the aforementioned H_0 vs. H_1, we have converted all the coefficients of association (except Kendall's tau) into the following statistic:

$$t_a = \frac{r_a\sqrt{n-2}}{\sqrt{1-r_a^2}} \tag{14}$$

where in equation (14), t_a is the statistic of student t-distribution with $n - 2$ degrees of freedom (i.e. t_{n-2}); r_a is the correlation coefficient calculated by any of the association methods of this study.

Distributional Models: In order to cover the commonly used practical models of parent distributions, we have considered (in bivariate setup) Normal, Uniform, Laplace, Lognormal, Exponential, Weibull, Gamma, Half Normal, Mixed Weibull, and Mixed Normal distributions as some representative parent distributions for our study. We also include Gamma, Exponential and Weibull distributions with outliers (contamination) in our study. For the choices of the distributions of X and Y, we have the following particular parameter selections to create bivariate environments: $N(0,1)$ for Normal; $U(0,1)$ for Uniform; $L(0.5,3)$ for Laplace; $LN(0,1)$ for Lognormal; $Exp(0.5)$ for Exponential; $W(1,2)$ for Weibull; $G(1,2)$ for Gamma; $HN(0,1)$ for Half Normal; $W(0.5,3)$ with probability 0.95 and $W(1,2)$ with probability 0.05 for Mixed Weibull; $N(0,1)$ with probability 0.95 and $N(0,400)$ with probability 0.05 for Mixed Normal; $G(0.5,3)$ with 5% outliers from $G(4,10)$ for contaminated Gamma; $W(1,2)$ with 5% outliers from $W(50,100)$ for contaminated Wiebull; $\exp(0.5)$ with 5% outliers from $\exp(4)$ for contaminated Exponential.

Computational Details of Experimentation: We have computed powers of the ten methods of measuring and testing the association by fixing the significance level at α using a simulation code developed in `MINITAB`. The critical values at a given α are obtained from the table of t_{n-2} for all the measures given in Equation ((1)-(7) and (11)-(13)) and their corresponding test statistics given in Equation (14), except for Kendall's coefficient given in Equation (7). For the Kendall's tau coefficient (τ) we have used the true critical values as given in Daniel (1990). The reason being that for all other cases the approximation given in Equation (14) is able to work fairly good but for the Kendall's tau coefficient it is not the case (as we here observed in our computations). The change in shape of the parent distribution demands an adjustment in the corresponding critical values. This we have done by our simulation algorithm for these ten methods to achieve the desired value of α. For different choices of $\rho = 0$, 0.25, 0.5 and 0.75 powers are obtained with the help of our simulation code in `MINITAB` at α significance level.

We have considered thirteen representative bivariate environments mentioned above for $n = 6, 8, 10, 12, 16, 20, 30$ at varying values of α. For these choices of n, α we have run our `MINITAB` simulation code (developed for the ten methods under investigation here) 10,000 times for power computations. The resulting power values are given in the tables given in Appendix for all the thirteen probability distributions and the ten methods under study for some selective choices from the above mentioned values of n at $\alpha = 0.05$. For the sake of brevity we omit the results at other choices of α like 0.01 and 0.005.

5. Comparative Analysis

This section presents a comparative analysis of the existing and proposed association measures. For ease in discussion and comparisons, the power values mentioned above are also displayed graphically in the form of power curves for all the aforementioned thirteen probability distributions by taking particular sample sizes and ten methods of association for some selective cases. These graphs are shown in Figures 1-13 where different values of $\rho = 0$, 0.25, 0.5 and 0.75 are taken on horizontal axis and the powers on vertical axis. Each figure is for a different parent distribution with different sample sizes and contains the power curves of all the ten methods. Labeling of the power curves in these figures is according to the notations given in Tables 1 and 2.

It is advocated from the above power analysis (cf. Table A1-A13 and Figures 1-13) that:

- With an increase in the value of n and/or ρ, power efficiency of all the association measures improves for all distributions.
- In general, Pearson correlation coefficient is superior to the Spearman rank correlation, Kendall's tau, modified rank correlation coefficient and proposed methods in normal distribution. However in some cases of normal distribution Gini correlation coefficients work better than the Pearson correlation coefficient.
- In non-normal distributions and in the case of outliers (contamination) the Pearson correlation coefficient grant a smaller amount of power than Spearman rank correlation, modified rank correlation coefficient and proposed methods except half normal, uniform, mixed normal and Laplace distributions. But Gini correlation coefficients r_{g1} and r_{g2} in general remain better in terms of power than Pearson correlation coefficient.
- Among the three Gini correlation coefficients r_{g1} performs better than r_{g2} and r_{g3}.
- The proposed three modifications grant improved power than the Spearman correlation coefficient, in general, for all the distributional environments. But in contaminated distributions the median rank correlation coefficient by using mean deviation about median works better than modified rank correlation coefficient for all sample sizes.
- Kendall's tau has inferior power than that of the Spearman rank correlation coefficient, modified rank correlation coefficient and the proposed methods. In Weibull, Mixed Weibull and Lognormal distributions, Kendall's tau has superior amount of power than the Gini mean correlation coefficient r_{g2}. But for these three distributions, if the sample size is greater than ten Kendall's tau has superior power performance than the Pearson correlation coefficient and Gini mean correlation coefficient r_{g3}. In the outlier cases, if the sample is moderate then Kendall's tau is superior to Pearson correlation coefficient and the two Gini mean correlation coefficients (r_{g2} and r_{g3}) for moderate sample sizes.
- From the analysis above, it is pertinent to note that the Gini mean correlation coefficient r_{g1} is the best choice for measuring and testing the association than Spearman rank correlation coefficient, Kendall's tau, modified rank correlation coefficient and the proposed methods in normal, non-normal and contaminated distributions.

- The powers of r_{MM}, r_{MS}, r_T and r_M slightly differ from each others in all the distributional environments. It means that these are close competitors to each other.

It is to be mentioned that other testing measures may also be evaluated on the similar lines but we think that the options we have chosen cover the most practical ones.

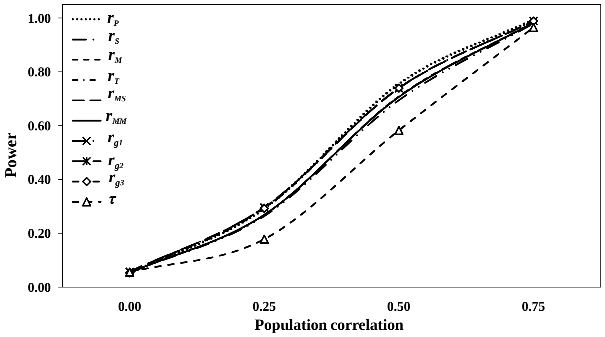


Figure 1: Normal distribution ($n = 20$).

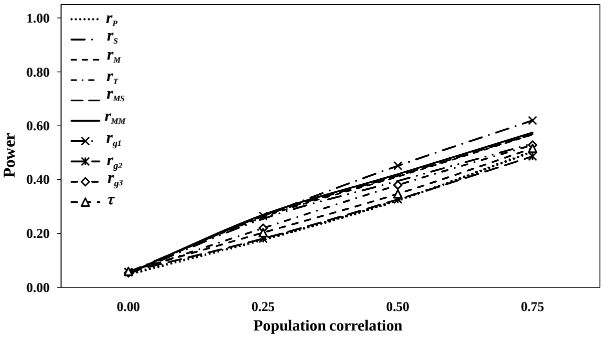


Figure 2: Weibull distribution ($n = 8$).

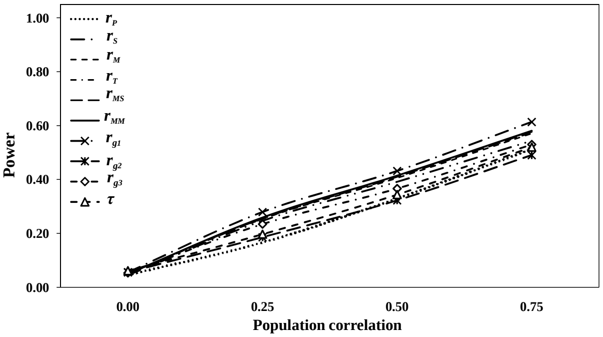


Figure 3: Mixed Weibull distribution ($n = 8$).

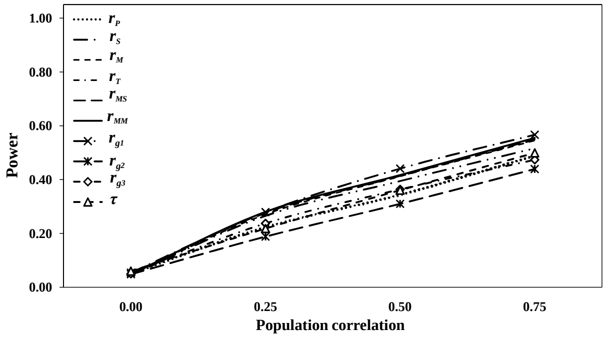


FIGURE 4: Lognormal distribution ($n = 8$).

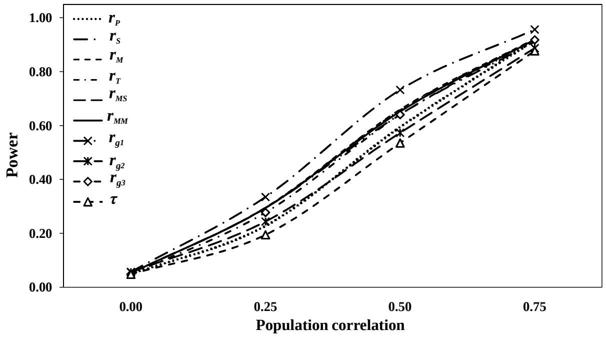


FIGURE 5: Exponential distribution ($n = 16$).

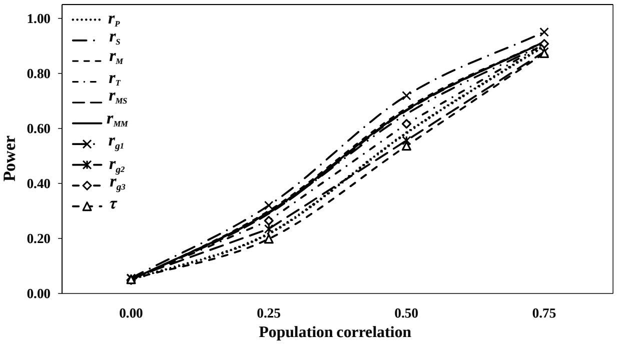


FIGURE 6: Gamma distribution ($n = 16$).

6. Numerical Illustration

Besides the evidence in terms of statistical efficiency it is very useful to test a technique on some real data for their practical implications. For this purpose we consider here a data set from Zimmerman (1994) on two variables of scores. The data set is given in Table 3 which contains eight pair of scores as reported by Zimmerman (1994).

TABLE 3: Eight pairs of Scores.

		Pair#							
		1	2	3	4	5	6	7	8
Scores	X	3.02	15.7	9.88	20.53	17.1	18.15	17.52	1.7
	Y	43.02	52.84	54.25	57.99	52.35	47.4	55.37	49.52

We state our null hypothesis as: There is no association between the two variables (i.e. $H_0 : \rho = 0$) versus the alternative hypothesis $H_1 : \rho > 0$. By fixing the level of significance at $\alpha = 0.05$, we apply all the ten methods and see what decisions they grant for the data set given in Table 3. The values of test statistic and their corresponding decisions are given in Table 4. The critical value used are: 0.571 for Kendall's tau and 1.94 for all the other tests.

TABLE 4: Values of the test statistics t_{journ} and their corresponding decisions.

t_P	t_S	t_M	t_T	t_{MS}
1.96 (Reject H_0)	1.41 (Don't reject H_0)	1.95 (Reject H_0)	1.95 (Reject H_0)	1.45 (Don't reject H_0)
t_{MM}	t_{g1}	t_{g2}	t_{g3}	τ
1.4 (Don't reject H_0)	1.91 (Don't reject H_0)	1.52 (Don't reject H_0)	1.74 (Don't reject H_0)	0.36 (Don't reject H_0)

It is obvious from the analysis of Table 4 that t_P, t_M and t_T reject H_0 while all others do not reject H_0. This is, in general, in accordance in the findings of Section 3. We may, therefore, sum up that this study will be of great use for the practitioners and researchers who make use of these measures frequently in their research projects.

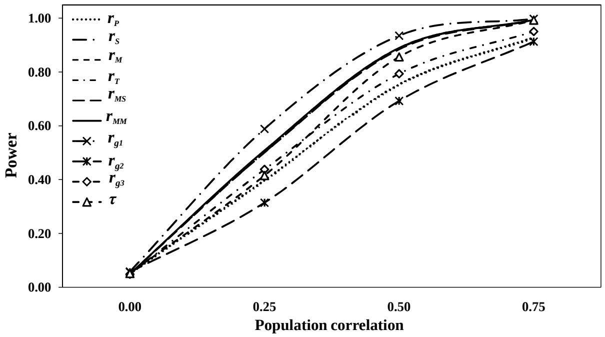


FIGURE 7: Exponential distribution with outliers ($n = 30$).

7. Summary and Conclusions

This study has evaluated the performance of different association measures including some existing and few newly suggested modifications. One of these measures is parametric and the others non-parametric ones. Performance evaluations (in terms of power) and comparisons are carried out under different symmetric, skewed and contaminated probability distributions including Normal, Cauchy, Uniform, Laplace, Lognormal, Exponential, Weibull, Gamma, t, Chi-square, Half Normal, Mixed Weibull and Mixed Normal.

Power evaluations of this study revealed that in normal distribution the Pearson correlation coefficient is the best choice to measure association. Further we have observed that the Pearson correlation coefficient and Gini's correlation coefficients (r_{g2} and r_{g3}) have superior power performances than the Spearman rank correlation, The modified rank correlation and the proposed correlation coefficients for symmetrical and low peaked distributions. But in non-symmetrical and high peaked distributions the Spearman rank correlation, modified rank correlation and the proposed correlation coefficients worked with supreme power than the Pearson correlation coefficient and the two Gini's correlation coefficients (r_{g2} and r_{g3}).

In contaminated distributions, r_{MM} exhibited better performance than the modified rank correlation coefficient. The Gini's correlation coefficient (r_{g1}) performed better than the Spearman rank correlation, modified rank correlation, Kendall's tau and the proposed correlation *coefficie nts* in symmetrical, asymmetrical, low peaked, highly peaked and contaminated distributions.

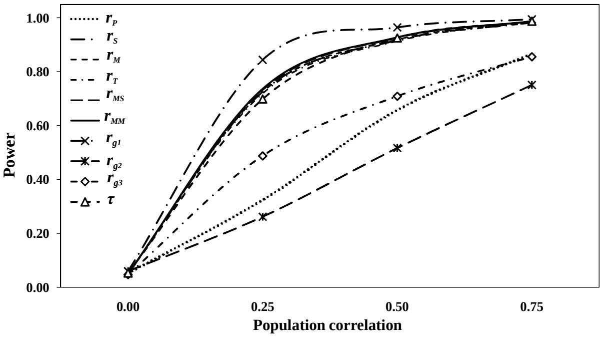


Figure 8: Weibull distribution with outliers ($n = 30$).

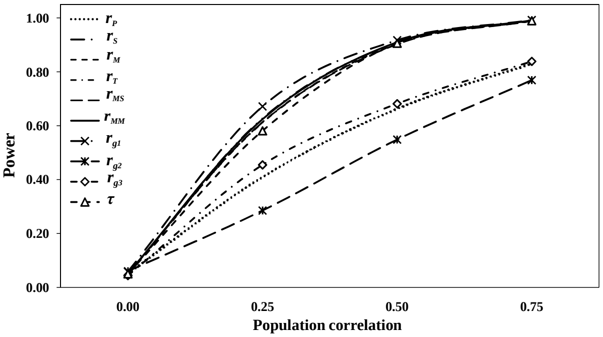


Figure 9: Gamma distribution with outliers ($n = 30$).

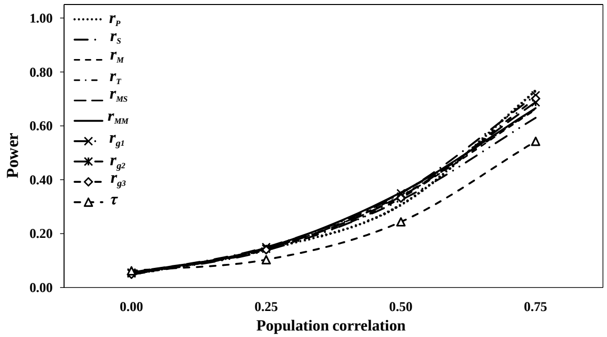


FIGURE 10: Halfnormal distribution ($n = 8$).

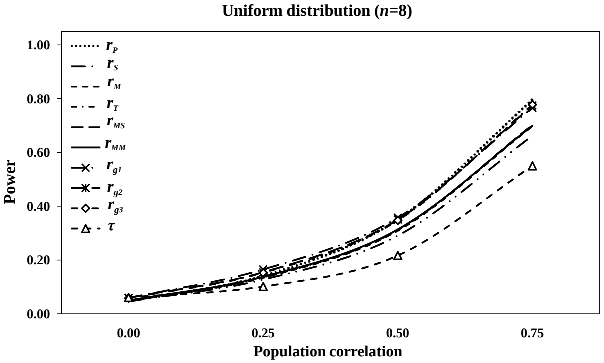


FIGURE 11: Uniform distribution ($n = 8$).

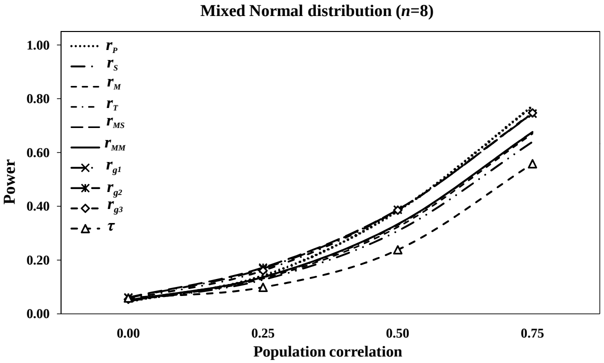


FIGURE 12: Mixed Normal distribution ($n = 8$).

Acknowledgments

The authors are thankful to the anonymous reviewers for their valuable comments on the previous version of the article. The author Muhammad Riaz is indebted to King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia for providing excellent research facilities under project #IN111059.

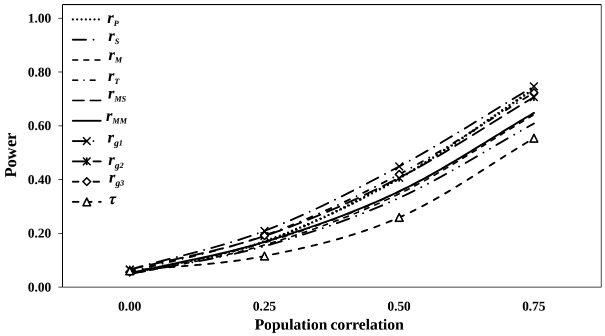


FIGURE 13: Laplace distribution ($n = 8$).

[Recibido: agosto de 2012 — Aceptado: noviembre de 2013]

References

Daniel, W. W. (1990), *Applied Nonparametric Statistics*, Duxbury Classic Series, New York.

Gauthier, T. D. (2001), 'Detecting the trends using the Spearman's rank correlation coefficient', *Environmental Forensics* **2**, 359–362.

Kendall, M. G. (1938), 'A new measure of rank correlation', *Biometrika* **5**, 81–93.

Mahoney, M. & Magel, R. (1996), 'Estimation of the power of the Kruskal-Wallis test', *Biometrical Journal* **38**, 613–630.

Maturi, T. A. & Elsayigh, A. (2010), 'A comparison of correlation coefficients via a three-step bootstrap approach', *Journal of Mathematics Research* **2**, 3–10.

Mudelsee, M. (2003), 'Estimating Pearson's correlation coefficient with bootstrap confidence interval from serially dependent time series', *Mathematical Geology* **35**, 651–665.

Munir, S., Asghar, Z. & Riaz, M. (2011), 'Performance evaluation of different tests for location parameters', *Communications in Statistics-Simulation and Computation* **40**(6), 839–853.

Spearman, C. (1904), 'The proof and measurement of association between two things', *American Journal of Psychology* **15**, 73–101.

Walker, D. A. (2003), 'JMASM9: Converting Kendall's tau for correlational or meta-analytic analyses', *Journal of Modern Applied Statistical Methods* **2**, 525–530.

Yitzhaki, S. (2003), 'Gini mean difference: A superior measure of variability for non normal distribution', *Metron-International Journal of Statistics* **LXI**, 285–316.

Zimmerman, D. W. (1994), 'A note on modified rank correlation', *Journal of Educational and Behavioral Statistics* **19**, 357–362.

Appendix

Table A1: Probability of rejecting the null hypothesis of independence for $N(0, 1)$.

n	ρ	r_P	r_S	r_M	r_T	r_{MS}	r_{MM}	r_{R1}	r_{R2}	r_{R3}	τ
6	0	0.0478	0.0476	0.0431	0.0461	0.0528	0.0525	0.059	0.0589	0.0526	0.054
	0.25	0.1236	0.1131	0.104	0.1126	0.1206	0.1234	0.1366	0.1362	0.1264	0.0761
	0.5	0.2772	0.2262	0.2211	0.2343	0.246	0.2511	0.2894	0.292	0.274	0.0755
	0.75	0.6096	0.4606	0.4917	0.5049	0.5219	0.5264	0.5681	0.5653	0.5669	0.161
8	0	0.0457	0.046	0.0489	0.0498	0.0521	0.0511	0.0555	0.0597	0.0528	0.0603
	0.25	0.1458	0.1315	0.1354	0.1409	0.1414	0.1402	0.1639	0.1667	0.1595	0.0974
	0.5	0.3795	0.3067	0.3278	0.3328	0.3345	0.333	0.3866	0.3893	0.3813	0.2339
	0.75	0.7702	0.6406	0.6723	0.6752	0.6745	0.6751	0.75	0.7429	0.7509	0.5562
10	0	0.0489	0.0524	0.0512	0.0503	0.0522	0.0523	0.0619	0.0631	0.0584	0.0496
	0.25	0.1773	0.1711	0.1669	0.1693	0.1674	0.1669	0.1958	0.1946	0.188	0.0889
	0.5	0.4613	0.4096	0.4115	0.412	0.4118	0.4109	0.4585	0.4607	0.4544	0.2577
	0.75	0.8633	0.7992	0.7995	0.8014	0.8001	0.7991	0.8508	0.8508	0.8548	0.637
12	0	0.0503	0.0475	0.0485	0.0474	0.0476	0.0478	0.0565	0.0568	0.0519	0.0653
	0.25	0.1909	0.1805	0.184	0.1826	0.1822	0.1826	0.2129	0.2148	0.2086	0.1274
	0.5	0.5395	0.473	0.487	0.4876	0.4829	0.483	0.5405	0.5401	0.5393	0.3742
	0.75	0.9262	0.8691	0.8795	0.8816	0.8794	0.8801	0.9121	0.9119	0.9139	0.8003
16	0	0.0493	0.0514	0.0507	0.0502	0.0511	0.0496	0.0585	0.0599	0.056	0.0536
	0.25	0.2448	0.2208	0.2257	0.2235	0.2238	0.2247	0.2519	0.2495	0.2424	0.1333
	0.5	0.6613	0.6	0.6129	0.6114	0.6081	0.607	0.654	0.6561	0.6551	0.4614
	0.75	0.9753	0.9478	0.9528	0.9541	0.952	0.9508	0.9708	0.9715	0.9739	0.9039
20	0	0.0518	0.0532	0.0534	0.0526	0.0535	0.0532	0.0573	0.0561	0.0526	0.0553
	0.25	0.2937	0.2635	0.268	0.2684	0.2686	0.2682	0.2964	0.2956	0.2923	0.1778
	0.5	0.7562	0.6942	0.7088	0.709	0.7066	0.7061	0.7399	0.7384	0.7396	0.5822
	0.75	0.994	0.9797	0.9839	0.9838	0.9832	0.9829	0.9889	0.9893	0.9886	0.965
30	0	0.0533	0.0517	0.0527	0.0528	0.0524	0.0514	0.056	0.0575	0.0549	0.0533
	0.25	0.3839	0.3523	0.3583	0.3572	0.3564	0.3567	0.3938	0.3916	0.3935	0.251
	0.5	0.8999	0.8576	0.861	0.8602	0.8598	0.8601	0.8875	0.885	0.8884	0.776
	0.75	0.9998	0.9992	0.9992	0.999	0.999	0.9991	0.9988	1	1	0.9969

Table A2: Probability of rejecting the null hypothesis of independence for $W(0.5, 3)$.

n	ρ	r_P	r_S	r_M	r_T	r_{MS}	r_{MM}	r_{R1}	r_{R2}	r_{R3}	τ
6	0	0.0513	0.0477	0.043	0.0482	0.051	0.0531	0.0494	0.0531	0.0448	0.0538
	0.25	0.16	0.1933	0.1878	0.1989	0.2049	0.2115	0.1997	0.14	0.16	0.1427
	0.5	0.2837	0.2925	0.3121	0.3219	0.3288	0.335	0.3131	0.2249	0.2487	0.2388
	0.75	0.4355	0.3752	0.4311	0.44	0.4507	0.4552	0.4286	0.3268	0.3453	0.3585
8	0	0.0509	0.0489	0.0522	0.0536	0.0519	0.0527	0.0565	0.0576	0.0538	0.0597
	0.25	0.1791	0.2545	0.2605	0.2648	0.265	0.269	0.265	0.1812	0.2199	0.2032
	0.5	0.3244	0.3951	0.411	0.4169	0.4143	0.4202	0.4513	0.3266	0.3798	0.3473
	0.75	0.5048	0.5342	0.5671	0.5674	0.569	0.5745	0.6195	0.4865	0.5286	0.5164
10	0	0.0499	0.0492	0.0473	0.0483	0.0494	0.0507	0.0556	0.0547	0.0494	0.0508
	0.25	0.2027	0.3144	0.3017	0.3032	0.3022	0.3058	0.3513	0.2109	0.2713	0.2189
	0.5	0.3684	0.4996	0.4978	0.4953	0.494	0.4969	0.5685	0.3771	0.4472	0.3948
	0.75	0.578	0.6709	0.6759	0.6731	0.6777	0.6819	0.7339	0.563	0.6126	0.5753
16	0	0.0521	0.052	0.0517	0.0521	0.0529	0.0527	0.0571	0.0513	0.0455	0.0536
	0.25	0.2435	0.4444	0.4507	0.4471	0.4517	0.4545	0.5226	0.2223	0.3333	0.3853
	0.5	0.4877	0.6849	0.7042	0.6982	0.6984	0.703	0.7755	0.4373	0.5523	0.6457
	0.75	0.7283	0.8592	0.8723	0.8696	0.8738	0.877	0.9175	0.6653	0.7432	0.8446

Table A3: Probability of rejecting the null hypothesis of independence for mixed Weibull distribution (i.e. $W(0.5, 3)$ with probability 0.95 and $W(1, 2)$ with probability 0.05.

n	ρ	r_P	r_S	r_M	r_T	r_{MS}	r_{MM}	r_{R1}	r_{R2}	r_{R3}	τ
6	0	0.0568	0.0499	0.0474	0.0506	0.052	0.0549	0.0502	0.0521	0.0451	0.0534
	0.25	0.1611	0.1856	0.1833	0.193	0.1998	0.2051	0.202	0.1383	0.165	0.1368
	0.5	0.2867	0.2952	0.3147	0.3227	0.3284	0.3348	0.315	0.2318	0.254	0.2413
	0.75	0.4322	0.3732	0.4342	0.4438	0.4533	0.4578	0.4361	0.334	0.3576	0.3584
8	0	0.0471	0.0448	0.0497	0.05	0.05	0.0497	0.0553	0.0555	0.0533	0.0611
	0.25	0.1673	0.2466	0.2534	0.2537	0.2548	0.2589	0.279	0.1857	0.2342	0.1969
	0.5	0.3305	0.3914	0.4054	0.4104	0.4095	0.4144	0.4315	0.3224	0.3663	0.3437
	0.75	0.5141	0.5437	0.5708	0.5739	0.5767	0.5808	0.6135	0.4904	0.5297	0.52
10	0	0.05	0.0526	0.0506	0.0528	0.0515	0.0541	0.0527	0.0543	0.0465	0.0483
	0.25	0.1983	0.3191	0.3127	0.3103	0.3117	0.3176	0.3426	0.2051	0.2635	0.2139
	0.5	0.3854	0.4847	0.4867	0.4837	0.4885	0.4932	0.5607	0.369	0.4396	0.396
	0.75	0.5862	0.6624	0.6711	0.6672	0.6728	0.676	0.7339	0.5531	0.6077	0.5861
16	0	0.051	0.0457	0.0472	0.0466	0.0462	0.0456	0.0583	0.0547	0.046	0.0519
	0.25	0.2387	0.4488	0.4536	0.4503	0.4486	0.4547	0.5263	0.2328	0.3441	0.3707
	0.5	0.4906	0.6933	0.7076	0.7015	0.7045	0.7093	0.7749	0.428	0.5459	0.6325
	0.75	0.7362	0.8507	0.8655	0.8603	0.8626	0.8643	0.9175	0.6653	0.7331	0.8411

Table A4: Probability of rejecting the null hypothesis of independence for $LG(5,4)$.

n	ρ	r_P	r_S	r_M	r_T	r_{MS}	r_{MM}	r_{R1}	r_{R2}	r_{R3}	τ
6	0	0.0513	0.0521	0.0468	0.0539	0.0523	0.0555	0.0561	0.0531	0.0554	0.0555
	0.25	0.1927	0.2067	0.2043	0.2131	0.2184	0.2252	0.2488	0.1686	0.2066	0.1519
	0.5	0.3033	0.3	0.3196	0.3316	0.3356	0.436	0.3476	0.2504	0.2833	0.2481
	0.75	0.4154	0.3758	0.4358	0.4405	0.4431	0.449	0.4375	0.336	0.358	0.3553
8	0	0.0515	0.0452	0.0473	0.0487	0.0485	0.0484	0.0537	0.0483	0.05	0.0597
	0.25	0.2235	0.2651	0.2705	0.2753	0.2762	0.28	0.279	0.1882	0.2371	0.2179
	0.5	0.3433	0.3946	0.4118	0.4135	0.4135	0.4171	0.4406	0.31	0.3635	0.3604
	0.75	0.4882	0.5165	0.545	0.5473	0.5492	0.5551	0.5666	0.4393	0.4736	0.4992
10	0	0.0549	0.0514	0.0513	0.0523	0.0512	0.0514	0.0523	0.0503	0.0504	0.0457
	0.25	0.2565	0.3379	0.3316	0.3321	0.3294	0.3338	0.3494	0.217	0.2853	0.2351
	0.5	0.4089	0.5066	0.5015	0.4991	0.5056	0.5062	0.5136	0.3543	0.419	0.4072
	0.75	0.5591	0.6546	0.6476	0.6455	0.6548	0.6576	0.665	0.496	0.5441	0.5821
16	0	0.0538	0.0495	0.0495	0.0501	0.0499	0.0487	0.0511	0.0526	0.0475	0.0478
	0.25	0.2884	0.4856	0.4827	0.4785	0.4807	0.4872	0.544	0.2385	0.3549	0.4272
	0.5	0.4801	0.696	0.6899	0.6898	0.6959	0.7022	0.7388	0.3854	0.4914	0.6813
	0.75	0.6492	0.8412	0.8389	0.838	0.8424	0.8457	0.867	0.567	0.6269	0.844

Table A5: Probability of rejecting the null hypothesis of independence for $Exp(0.5)$.

n	ρ	r_P	r_S	r_M	r_T	r_{MS}	r_{MM}	r_{R1}	r_{R2}	r_{R3}	τ
6	0	0.0502	0.0504	0.0464	0.0492	0.0539	0.0555	0.0573	0.057	0.0496	0.0572
	0.25	0.1162	0.1328	0.1245	0.1327	0.1414	0.1458	0.1617	0.1477	0.1407	0.0869
	0.5	0.2613	0.2611	0.262	0.2714	0.2849	0.2929	0.3047	0.2758	0.2761	0.1777
	0.75	0.5209	0.4224	0.4594	0.4731	0.4895	0.493	0.5205	0.4753	0.486	0.361
8	0	0.0508	0.0493	0.0507	0.0516	0.0541	0.0549	0.0533	0.0563	0.0535	0.0614
	0.25	0.1488	0.1613	0.1671	0.1702	0.1729	0.1725	0.1852	0.163	0.1664	0.1174
	0.5	0.3574	0.3521	0.3675	0.3731	0.3737	0.3779	0.4103	0.3471	0.367	0.2724
	0.75	0.6692	0.6099	0.6427	0.6435	0.6456	0.6465	0.6928	0.6325	0.6553	0.5422
10	0	0.0507	0.0564	0.0543	0.0544	0.0548	0.0552	0.0537	0.0571	0.0492	0.0472
	0.25	0.1535	0.2072	0.2001	0.2003	0.1969	0.1996	0.2165	0.1814	0.1891	0.1163
	0.5	0.3948	0.4487	0.4491	0.4479	0.4443	0.4472	0.5066	0.4201	0.4526	0.3124
	0.75	0.6721	0.7347	0.7431	0.7436	0.7427	0.7447	0.8005	0.718	0.7495	0.6182
16	0	0.05	0.0523	0.0505	0.051	0.0527	0.0522	0.0566	0.0556	0.0478	0.0479
	0.25	0.2296	0.294	0.2957	0.2947	0.2932	0.2942	0.3348	0.2438	0.2771	0.1943
	0.5	0.5962	0.6413	0.6595	0.6576	0.6527	0.652	0.7324	0.5731	0.6404	0.535
	0.75	0.9189	0.9106	0.9218	0.9212	0.9188	0.919	0.9565	0.8875	0.9179	0.877

Table A6: Probability of rejecting the null hypothesis of independence for $G(1,2)$.

n	ρ	r_P	r_S	r_M	r_T	r_{MS}	r_{MM}	r_{R1}	r_{R2}	r_{R3}	τ
6	0	0.049	0.0514	0.0461	0.0495	0.0545	0.0565	0.0563	0.0554	0.0484	0.0569
	0.25	0.1209	0.131	0.1251	0.1327	0.1417	0.1455	0.1436	0.1318	0.1268	0.0912
	0.5	0.2666	0.259	0.2619	0.2729	0.72874	0.295	0.298	0.2684	0.2701	0.187
	0.75	0.5131	0.4271	0.4703	0.4823	0.4977	0.502	0.5074	0.4661	0.4753	0.3585
8	0	0.0513	0.0509	0.0547	0.0577	0.0574	0.0571	0.0585	0.0582	0.0543	0.0612
	0.25	0.139	0.1581	0.1613	0.1659	0.1702	0.1708	0.195	0.1695	0.1703	0.1182
	0.5	0.3385	0.3423	0.3597	0.3649	0.3647	0.3689	0.4033	0.3493	0.3651	0.2536
	0.75	0.6598	0.6051	0.6343	0.6392	0.6444	0.6426	0.7066	0.6459	0.6688	0.5309
10	0	0.0508	0.055	0.0524	0.0528	0.0522	0.0527	0.0518	0.0538	0.0476	0.0468
	0.25	0.1611	0.2048	0.2035	0.2033	0.2017	0.2032	0.2199	0.1862	0.1938	0.117
	0.5	0.4018	0.4441	0.4442	0.4457	0.4464	0.4499	0.5049	0.4112	0.449	0.3078
	0.75	0.7492	0.7305	0.742	0.7394	0.7406	0.7393	0.8061	0.7278	0.7601	0.6269
16	0	0.0516	0.0496	0.0477	0.0483	0.0484	0.0494	0.0553	0.0546	0.0486	0.0514
	0.25	0.2193	0.2928	0.3009	0.2967	0.2894	0.2953	0.3199	0.2348	0.2646	0.1985
	0.5	0.5849	0.6523	0.6738	0.6679	0.664	0.667	0.7193	0.5561	0.6167	0.5366
	0.75	0.9017	0.9074	0.9162	0.9165	0.9153	0.9153	0.9505	0.8787	0.9076	0.8734

Table A7: Probability of rejecting the null hypothesis of independence for contaminated Exponential (i.e. exp(0.5) with 5% outliers from exp(4)).

n	ρ	r_P	r_S	r_M	r_T	r_{MS}	r_{MM}	r_{R1}	r_{R2}	r_{R3}	τ
6	0	0.0512	0.0536	0.0458	0.05	0.0545	0.0569	0.0517	0.0528	0.0476	0.0519
	0.25	0.1461	0.1431	0.135	0.1392	0.1483	0.1546	0.1565	0.1315	0.1367	0.0996
	0.5	0.2881	0.2621	0.2681	0.2754	0.2905	0.2962	0.3026	0.2638	0.2725	0.1905
	0.75	0.5212	0.429	0.4658	0.4762	0.4919	0.4964	0.497	0.4505	0.4626	0.3647
8	0	0.0507	0.0497	0.0499	0.0498	0.053	0.0535	0.0507	0.052	0.0521	0.0607
	0.25	0.1676	0.17	0.1767	0.182	0.1801	0.1804	0.1985	0.1638	0.1766	0.1253
	0.5	0.3504	0.3496	0.3652	0.3704	0.3694	0.3722	0.4112	0.3493	0.3758	0.2796
	0.75	0.6112	0.5953	0.6266	0.6275	0.6269	0.6277	0.6665	0.5978	0.6231	0.5418
10	0	0.0509	0.052	0.0496	0.0504	0.0502	0.0506	0.0533	0.0523	0.0478	0.0436
	0.25	0.1998	0.2257	0.2239	0.2208	0.2189	0.2208	0.2563	0.1967	0.2222	0.1264
	0.5	0.4251	0.4579	0.4588	0.4582	0.4541	0.4566	0.5209	0.4263	0.4615	0.3227
	0.75	0.7097	0.7281	0.7339	0.7323	0.7322	0.7346	0.777	0.6866	0.7178	0.621
12	0	0.0526	0.0527	0.0521	0.0522	0.0535	0.0555	0.0545	0.0578	0.0521	0.0606
	0.25	0.2149	0.242	0.2507	0.252	0.2477	0.2461	0.2975	0.2045	0.2445	0.1945
	0.5	0.483	0.5151	0.5264	0.5275	0.5234	0.5285	0.5991	0.4666	0.5158	0.4561
	0.75	0.7649	0.7929	0.8016	0.7998	0.7979	0.8005	0.8574	0.7445	0.7793	0.7668
16	0	0.0544	0.0484	0.0509	0.0512	0.0497	0.049	0.0547	0.0567	0.0506	0.053
	0.25	0.258	0.3191	0.321	0.3197	0.3157	0.3191	0.3685	0.2446	0.2998	0.215
	0.5	0.5678	0.6474	0.6584	0.6582	0.6532	0.6545	0.7273	0.5415	0.611	0.556
	0.75	0.8368	0.903	0.9067	0.9052	0.9066	0.91	0.941	0.8103	0.849	0.8692
20	0	0.0562	0.0523	0.051	0.0514	0.0514	0.0518	0.0583	0.0587	0.0518	0.0571
	0.25	0.3111	0.373	0.3759	0.3729	0.3728	0.3772	0.4473	0.2821	0.3508	0.2895
	0.5	0.6519	0.7378	0.7458	0.7445	0.7415	0.7459	0.8176	0.5943	0.677	0.6702
	0.75	0.8862	0.9577	0.958	0.959	0.9574	0.9593	0.9776	0.8673	0.905	0.903
30	0	0.0523	0.0494	0.0488	0.0504	0.048	0.0493	0.0579	0.0561	0.0483	0.0512
	0.25	0.3965	0.5011	0.508	0.5059	0.4997	0.5081	0.5882	0.314	0.4378	0.4151
	0.5	0.7544	0.8872	0.8899	0.8895	0.8835	0.8889	0.935	0.6925	0.7935	0.8559
	0.75	0.9281	0.993	0.9931	0.9931	0.9926	0.9936	0.9976	0.9129	0.9506	0.9927

Table A8: Probability of rejecting the null hypothesis of independence for contaminated Weibull (i.e. $W(0.5, 3)$ with 5% outliers from $W(50, 100)$).

n	ρ	r_P	r_S	r_M	r_T	r_{MS}	r_{MM}	r_{R1}	r_{R2}	r_{R3}	τ
6	0	0.0563	0.0495	0.046	0.0507	0.05	0.054	0.0546	0.0595	0.0515	0.0536
	0.25	0.1961	0.2037	0.2032	0.2134	0.216	0.224	0.2388	0.155	0.1866	0.1477
	0.5	0.3217	0.2982	0.3185	0.3313	0.334	0.3422	0.3358	0.2338	0.265	0.246
	0.75	0.4379	0.377	0.4256	0.4351	0.4491	0.4536	0.4303	0.3343	0.3538	0.3604
8	0	0.053	0.0472	0.0537	0.0558	0.055	0.0559	0.052	0.0513	0.051	0.0621
	0.25	0.2091	0.2595	0.2674	0.2699	0.2681	0.2715	0.2878	0.1903	0.2486	0.2158
	0.5	0.3449	0.3958	0.4105	0.412	0.4158	0.4193	0.4391	0.3224	0.3734	0.3559
	0.75	0.4918	0.5282	0.5584	0.5593	0.562	0.5642	0.5971	0.4593	0.5004	0.5103
10	0	0.056	0.0529	0.0543	0.0545	0.0536	0.0542	0.0546	0.05	0.0468	0.0439
	0.25	0.2245	0.3157	0.3159	0.3144	0.3178	0.3194	0.348	0.2076	0.278	0.2278
	0.5	0.3897	0.4948	0.4944	0.4923	0.4935	0.499	0.5287	0.3585	0.4245	0.4066
	0.75	0.5662	0.6507	0.6523	0.649	0.6548	0.6575	0.6933	0.52	0.5715	0.5827
12	0	0.0543	0.047	0.0494	0.0486	0.0485	0.0487	0.0541	0.0526	0.0499	0.0616
	0.25	0.2444	0.3695	0.3706	0.3651	0.3703	0.3758	0.4141	0.2037	0.291	0.3282
	0.5	0.4294	0.5665	0.5735	0.5709	0.5747	0.5803	0.6293	0.3733	0.458	0.5457
	0.75	0.632	0.7293	0.7405	0.7376	0.7445	0.7495	0.7476	0.542	0.573	0.7277
16	0	0.0551	0.0473	0.0486	0.048	0.046	0.0465	0.0583	0.0556	0.0466	0.051
	0.25	0.2573	0.474	0.4759	0.4675	0.4748	0.4796	0.5768	0.2373	0.3601	0.403
	0.5	0.4808	0.6826	0.6908	0.6864	0.6938	0.6978	0.7853	0.427	0.5396	0.6537
	0.75	0.6946	0.8457	0.8518	0.8478	0.8587	0.8609	0.909	0.617	0.595	0.8397
20	0	0.0564	0.0494	0.0478	0.0467	0.0483	0.0478	0.0561	0.0553	0.0421	0.0524
	0.25	0.2922	0.5503	0.5525	0.5436	0.5526	0.5592	0.6545	0.2324	0.3813	0.5066
	0.5	0.5422	0.7837	0.7879	0.783	0.7933	0.7966	0.855	0.4396	0.5745	0.7711
	0.75	0.7668	0.9111	0.9201	0.914	0.921	0.9232	0.9549	0.6459	0.7369	0.9212
30	0	0.0536	0.0489	0.0514	0.0507	0.0506	0.0508	0.0587	0.0591	0.0471	0.0526
	0.25	0.3259	0.7257	0.7253	0.7191	0.7315	0.7364	0.8438	0.2616	0.4875	0.6984
	0.5	0.6594	0.9165	0.919	0.9132	0.9253	0.9272	0.9643	0.5166	0.7088	0.925
	0.75	0.8675	0.9824	0.9837	0.9825	0.9862	0.9868	0.9944	0.751	0.8554	0.9874

Table A9: Probability of rejecting the null hypothesis of independence for contaminated Gamma (i.e. $G(0.5, 3)$ with 5% outliers from $G(4, 10)$).

n	ρ	r_P	r_S	r_M	r_T	r_{MS}	r_{MM}	r_{R1}	r_{R2}	r_{R3}	τ
6	0	0.0558	0.0511	0.0441	0.0488	0.0517	0.0532	0.058	0.0566	0.0505	0.059
	0.25	0.19	0.1675	0.1591	0.1692	0.1753	0.1797	0.1523	0.1316	0.129	0.125
	0.5	0.3134	0.273	0.282	0.2907	0.2992	0.3073	0.2931	0.2613	0.2716	0.2153
	0.75	0.4823	0.3927	0.444	0.4511	0.459	0.4694	0.5315	0.5214	0.522	0.3621
8	0	0.0529	0.0468	0.0469	0.0508	0.0511	0.0513	0.052	0.0484	0.049	0.0568
	0.25	0.2071	0.2201	0.2256	0.2268	0.2282	0.2306	0.2465	0.181	0.2228	0.1691
	0.5	0.3542	0.3778	0.3964	0.4007	0.3973	0.3997	0.4075	0.3271	0.361	0.3331
	0.75	0.5516	0.5551	0.5834	0.5872	0.5885	0.5828	0.604	0.5069	0.5346	0.5147
10	0	0.05	0.0516	0.0511	0.0506	0.0505	0.0506	0.0542	0.0567	0.055	0.0451
	0.25	0.2311	0.2757	0.2692	0.2698	0.2682	0.2727	0.3208	0.2124	0.2766	0.1862
	0.5	0.4032	0.4814	0.4774	0.4768	0.4736	0.4796	0.5083	0.3796	0.4294	0.3687
	0.75	0.6091	0.6786	0.683	0.6836	0.6855	0.6884	0.7269	0.5859	0.6264	0.5994
12	0	0.0537	0.0492	0.0497	0.0523	0.0488	0.0495	0.0567	0.0563	0.056	0.0608
	0.25	0.2629	0.3188	0.3159	0.3134	0.312	0.3142	0.3794	0.2291	0.3031	0.2638
	0.5	0.4416	0.534	0.5377	0.533	0.5336	0.5404	0.6092	0.4245	0.4915	0.5029
	0.75	0.6671	0.7588	0.7627	0.7605	0.7649	0.7715	0.8115	0.63	0.6743	0.7449
16	0	0.0512	0.048	0.0465	0.0475	0.0473	0.047	0.0583	0.0577	0.0549	0.0557
	0.25	0.3053	0.4014	0.3947	0.391	0.3895	0.3965	0.4578	0.2491	0.3418	0.3181
	0.5	0.5198	0.6707	0.6731	0.6677	0.6677	0.6781	0.7086	0.4612	0.5372	0.613
	0.75	0.7337	0.8729	0.8735	0.8728	0.8705	0.878	0.909	0.686	0.7425	0.8595
20	0	0.0543	0.0529	0.0525	0.054	0.053	0.052	0.0552	0.058	0.05	0.0533
	0.25	0.3422	0.4829	0.472	0.4699	0.4691	0.4771	0.517	0.2521	0.3698	0.4049
	0.5	0.5784	0.7569	0.7588	0.7547	0.752	0.7604	0.7905	0.4822	0.578	0.7328
	0.75	0.7558	0.9349	0.9339	0.9332	0.932	0.9371	0.9521	0.7193	0.7757	0.9307
30	0	0.0537	0.0476	0.0516	0.0505	0.0522	0.0513	0.0578	0.0599	0.0436	0.0503
	0.25	0.4102	0.6268	0.6133	0.6109	0.6127	0.6239	0.6714	0.2853	0.4545	0.581
	0.5	0.6641	0.9085	0.9049	0.9024	0.9024	0.9096	0.9171	0.5486	0.6815	0.9059
	0.75	0.8317	0.9894	0.9891	0.9887	0.9886	0.9909	0.992	0.769	0.8385	0.9898

Table A10: Probability of rejecting the null hypothesis of independence for $HN(0,1)$.

n	ρ	r_P	r_S	r_M	r_T	r_{MS}	r_{MM}	r_{R1}	r_{R2}	r_{R3}	τ
6	0	0.0527	0.055	0.0474	0.05	0.0572	0.0569	0.0547	0.0587	0.0502	0.0489
	0.25	0.1081	0.1107	0.1053	0.1129	0.1219	0.1241	0.1279	0.1255	0.1125	0.0719
	0.5	0.2726	0.2383	0.2354	0.245	0.2601	0.2619	0.2794	0.2628	0.2608	0.1603
	0.75	0.5664	0.4319	0.4686	0.4788	0.4972	0.5033	0.5393	0.5163	0.518	0.3526
8	0	0.0501	0.0476	0.0477	0.0488	0.0507	0.0502	0.0523	0.0556	0.049	0.062
	0.25	0.1431	0.1374	0.1417	0.1429	0.1478	0.1473	0.1498	0.1439	0.14	0.1034
	0.5	0.3078	0.3244	0.3394	0.3422	0.35	0.3526	0.3495	0.3381	0.332	0.2437
	0.75	0.7315	0.6296	0.6601	0.6632	0.6649	0.6655	0.7133	0.6879	0.7012	0.5427
10	0	0.0513	0.0565	0.0545	0.0541	0.056	0.0549	0.0538	0.0501	0.0481	0.0498
	0.25	0.16	0.1748	0.1704	0.1715	0.1682	0.1666	0.1784	0.1693	0.1661	0.0946
	0.5	0.4419	0.4141	0.4185	0.4176	0.4165	0.4168	0.455	0.4234	0.4336	0.2748
	0.75	0.8075	0.7678	0.7817	0.7811	0.781	0.7791	0.8225	0.7998	0.8116	0.6435
16	0	0.0553	0.0483	0.0463	0.0465	0.0467	0.0475	0.0585	0.057	0.0535	0.0491
	0.25	0.2478	0.2443	0.2473	0.2468	0.2452	0.2443	0.2796	0.2526	0.2591	0.1532
	0.5	0.6544	0.6136	0.633	0.6339	0.624	0.6253	0.695	0.6351	0.6625	0.4969
	0.75	0.9678	0.9398	0.9496	0.9486	0.9462	0.9465	0.9681	0.952	0.9603	0.896

Table A11: Probability of rejecting the null hypothesis of independence for $U(0,1)$.

n	ρ	r_P	r_S	r_M	r_T	r_{MS}	r_{MM}	r_{R1}	r_{R2}	r_{R3}	τ
6	0	0.0504	0.0472	0.0433	0.0452	0.0513	0.0503	0.0657	0.0623	0.0572	0.0607
	0.25	0.1215	0.1108	0.1071	0.1127	0.1216	0.1228	0.1418	0.1389	0.1287	0.0726
	0.5	0.2433	0.2029	0.2059	0.2118	0.2259	0.2276	0.2719	0.258	0.2523	0.1367
	0.75	0.602	0.4506	0.4673	0.4821	0.5034	0.5062	0.5819	0.569	0.5703	0.3385
8	0	0.0472	0.0455	0.0458	0.0472	0.049	0.0492	0.058	0.0591	0.0577	0.0602
	0.25	0.1422	0.127	0.1331	0.1344	0.1376	0.1375	0.1646	0.1541	0.1524	0.1009
	0.5	0.3514	0.2906	0.3082	0.3103	0.3123	0.3136	0.3572	0.3501	0.3472	0.2165
	0.75	0.7977	0.6619	0.6967	0.7004	0.7025	0.6989	0.766	0.7746	0.7777	0.5501
10	0	0.048	0.0502	0.0491	0.0506	0.0493	0.0481	0.0532	0.0546	0.0502	0.0464
	0.25	0.1683	0.1614	0.1581	0.1577	0.1561	0.1564	0.1838	0.1834	0.1773	0.0844
	0.5	0.4359	0.389	0.3944	0.3956	0.3928	0.3905	0.4428	0.4507	0.4433	0.244
	0.75	0.9033	0.8261	0.8388	0.8403	0.8367	0.836	0.8745	0.8883	0.8876	0.6622
16	0	0.0494	0.0501	0.0469	0.0485	0.0492	0.0479	0.0542	0.0544	0.0529	0.0522
	0.25	0.2319	0.2183	0.2167	0.2163	0.2135	0.2129	0.2358	0.2358	0.2308	0.1308
	0.5	0.6541	0.585	0.605	0.604	0.5972	0.5959	0.6205	0.6535	0.6394	0.4423
	0.75	0.9904	0.9732	0.9779	0.978	0.9759	0.9761	0.9818	0.9874	0.9871	0.9407

Table A12: Probability of rejecting the null hypothesis of independence for Mixed Normal (i.e. $N(0,1)$ with probability 0.95 and $N(0,400$ with probability 0.05).

n	ρ	r_P	r_S	r_M	r_T	r_{MS}	r_{MM}	r_{R1}	r_{R2}	r_{R3}	τ
6	0	0.0522	0.0512	0.0471	0.0499	0.0537	0.0543	0.0662	0.0675	0.0628	0.0575
	0.25	0.1265	0.1104	0.1037	0.1106	0.121	0.1222	0.1382	0.1362	0.1287	0.0784
	0.5	0.2723	0.2205	0.2197	0.2307	0.2459	0.2487	0.2811	0.2837	0.2702	0.1593
	0.75	0.611	0.4506	0.471	0.4927	0.5139	0.5145	0.5607	0.5623	0.5584	0.3533
8	0	0.0463	0.0468	0.0476	0.0493	0.0499	0.0499	0.0591	0.0607	0.0539	0.0586
	0.25	0.142	0.1262	0.1309	0.135	0.1365	0.1361	0.168	0.1717	0.1606	0.0984
	0.5	0.3834	0.3081	0.3226	0.3296	0.3326	0.3328	0.3867	0.3871	0.3852	0.2378
	0.75	0.7721	0.6394	0.6697	0.6769	0.6762	0.6756	0.7445	0.7452	0.7473	0.558
10	0	0.0472	0.0517	0.0491	0.0485	0.0484	0.049	0.0587	0.0615	0.0564	0.0484
	0.25	0.1734	0.1606	0.1589	0.1594	0.1623	0.1609	0.1949	0.1946	0.1868	0.0929
	0.5	0.4634	0.4077	0.4045	0.408	0.4086	0.4062	0.4658	0.4627	0.4564	0.2627
	0.75	0.8661	0.7901	0.7991	0.8016	0.7986	0.7944	0.8536	0.8557	0.8569	0.6404
16	0	0.05	0.0528	0.053	0.0533	0.0538	0.0544	0.0582	0.0595	0.0561	0.0502
	0.25	0.238	0.2148	0.2168	0.2158	0.216	0.2146	0.2521	0.2513	0.247	0.1399
	0.5	0.6723	0.6033	0.6161	0.6174	0.6122	0.6135	0.6501	0.6524	0.6528	0.4655
	0.75	0.9775	0.951	0.9569	0.9568	0.9552	0.9549	0.9709	0.971	0.9731	0.9109

Table A13: Probability of rejecting the null hypothesis of independence for $L(0.5,3)$.

n	ρ	r_P	r_S	r_M	r_T	r_{MS}	r_{MM}	r_{R1}	r_{R2}	r_{R3}	τ
6	0	0.0503	0.0515	0.0435	0.0503	0.0536	0.0548	0.0628	0.065	0.0575	0.0588
	0.25	0.1396	0.1259	0.1141	0.1252	0.1335	0.1362	0.1661	0.1534	0.1452	0.0853
	0.5	0.3145	0.2439	0.2422	0.2604	0.2753	0.2768	0.3361	0.3201	0.317	0.1764
	0.75	0.6045	0.4345	0.4734	0.4919	0.5058	0.5058	0.5602	0.5498	0.5543	0.3735
8	0	0.0528	0.0481	0.0465	0.0497	0.0506	0.0514	0.0652	0.0646	0.0559	0.0602
	0.25	0.1712	0.152	0.1567	0.165	0.1668	0.1681	0.209	0.1898	0.1914	0.1151
	0.5	0.4058	0.3317	0.3469	0.3542	0.3566	0.3558	0.4487	0.4065	0.4192	0.2589
	0.75	0.7375	0.6085	0.641	0.6483	0.6479	0.6473	0.7465	0.7065	0.723	0.5538
10	0	0.0495	0.0487	0.0447	0.047	0.0468	0.0462	0.0666	0.0682	0.0585	0.0467
	0.25	0.1904	0.1895	0.1845	0.1904	0.1899	0.1892	0.2397	0.2089	0.21	0.1074
	0.5	0.4806	0.4303	0.4328	0.4359	0.4363	0.4359	0.5409	0.4761	0.5031	0.2983
	0.75	0.8313	0.7533	0.7566	0.7561	0.7583	0.7573	0.8397	0.8053	0.8213	0.6283
16	0	0.0491	0.0458	0.0438	0.0445	0.0455	0.046	0.0667	0.0661	0.059	0.0515
	0.25	0.2535	0.2572	0.2585	0.2588	0.259	0.263	0.3378	0.2773	0.2969	0.174
	0.5	0.656	0.6127	0.6236	0.6223	0.625	0.6227	0.7213	0.6377	0.7661	0.5104
	0.75	0.9504	0.9197	0.9225	0.9234	0.9235	0.9234	0.9602	0.9386	0.9495	0.8766

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 25 a 34

A New Extension of the Exponential Distribution

Una nueva extensión de la distribución exponencial

Yolanda M. Gómez[1,a], Heleno Bolfarine[1,b], Héctor W. Gómez[2,c]

[1] Instituto de Matematica e Estatística, Universidade de São Paulo, São Paulo, Brazil

[2] Departamento de Matemáticas, Facultad de Ciencias Básicas, Universidad de Antofagasta, Antofagasta, Chile

Abstract

The present paper considers an extension of the exponential distribution based on mixtures of positive distributions. We study the main properties of this new distribution, with special emphasis on its moments, moment generator function and some characteristics related to reliability studies. We also discuss parameter estimation considering the maximum likelihood and moments approach. An application reveals that the model proposed can be very useful in fitting real data. A final discussion concludes the paper.

***Key words*:** Exponential distribution, Mixtures of distribution, Likelihood.

Resumen

En el presente paper se considera una extensión de la distribución exponencial basada en mezclas de distribuciones positivas. Estudiamos las principales propiedades de esta nueva distribución, con especial énfasis en sus momentos, función generadora de momentos, y algunas características relacionadas a estudios de confiabilidad. También se analiza la estimación de parámetros a través de los métodos de momentos y de máxima verosimilitud. Una aplicación muestra que el modelo propuesto puede ser muy útil para ajustar datos reales. Una discusión final concluye el artículo.

***Palabras clave*:** distribución exponencial, mezcla de distribuciones, verosimilitud.

[a] Professor. E-mail: ymgomez@ime.usp.br
[b] Professor. E-mail: hbolfar@ime.usp.br
[c] Professor. E-mail: hector.gomez@uantof.cl

1. Introduction

In lifetime data analysis it is usually the case that models with monotone risk functions are preferred as is the case of the gamma distribution. For some models there are no closed form risk functions (such as the Gamma model) and numerical integration might be required for its computation. In recent statistical literature modified extensions of the exponential distributions have been proposed to contour such difficulties. For example, Gupta & Kundu (1999) and Gupta & Kundu (2001) introduced an extension of the exponential distribution typically called the generalized exponential (GE) distribution. Therefore, it is said that the random variable X follows the GE distribution if its density function is given by

$$g_1(x;\alpha,\beta)=\alpha\beta e^{-\alpha x}(1-e^{-\alpha x})^{\beta-1}$$

where $x>0$, $\alpha>0$ and $\beta>0$. We use the notation $X\sim GE(\alpha,\beta)$ for a random variable with such distribution.

More recently, Nadarajah & Haghighi (2011) introduced another extension of the exponential model, so that a random variable X follows the Nadarajah and Haghighi's exponential distribution (NHE) if its density function is given by

$$g_2(x;\alpha,\beta)=\alpha\beta(1+\alpha x)^{\beta-1}e^{\{1-(1+\alpha x)^{\beta}\}}$$

where $x>0$, $\alpha>0$ and $\beta>0$. We use the notation $X\sim NHE(\alpha,\beta)$.

Both distributions have the exponential distribution (E) with scale parameter α, as a special case when $\beta=1$, that is,

$$g_1(x;\alpha,\beta=1)=g_2(x;\alpha,\beta=1)=\alpha e^{-\alpha x}$$

where $x>0$, $\alpha>0$ with the notation $X\sim E(\alpha)$. Other extensions of the exponential model in the survival analysis context are considered in the Marshall & Olkin's (2007) book.

The main object of this paper is to present yet another extension for the exponential distribution that can be used as an alternative to the ones mentioned above. We discuss some properties for this new distribution like its moments and moment generating function which can be used for parameter estimation as starting values for computing maximum likelihood estimators.

The paper is organized as follows. Section 2 delivers the density and distribution functions, moments, moment generating function, asymmetry and kurtosis coefficients and hazard function. Section 3 is devoted to parameter estimation based on maximum likelihood and moments approach. It is recommended that the moment estimators are used to initialize the maximum likelihood approach. In Section 4 an application to a real data set is presented and comparisons between the proposed model and other extensions of the exponential distribution are reported. The main conclusion is that the new model can perform well in applied situations.

2. Density and Properties

A random variable X is distributed according to the extended exponential distribution (EE) with parameters α and β if its density function is given by

$$f(x;\alpha,\beta)=\frac{\alpha^2(1+\beta x)e^{-\alpha x}}{\alpha+\beta} \tag{1}$$

where $x>0$, $\alpha>0$ and $\beta\geq 0$. We use the notation $X\sim EE(\alpha,\beta)$.

Figures 1 and 2 depict the behavior of the distribution for some parameter values.

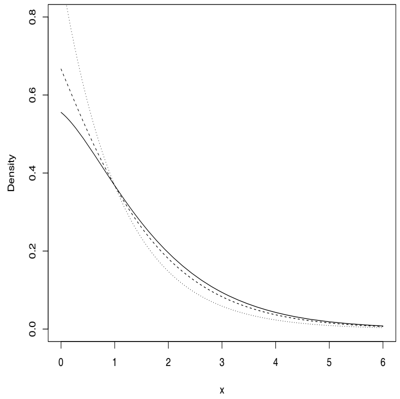


FIGURE 1: Plots of the $EE(1,0.8)$ (solid line), $EE(1,0.5)$ (dashed line), $EE(1,0.1)$ (dotted line).

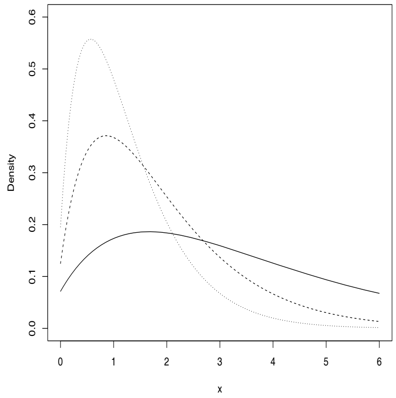


FIGURE 2: Plots of the $EE(0.5,3)$ (solid line), $EE(1,7)$ (dashed line), $EE(1.5,10)$ (dotted line).

2.1. Properties

Let $X \sim EE(\alpha, \beta)$, $Y \sim E(\alpha)$ and $Z \sim Gamma(2, \alpha)$. Then, the distribution function for the random variable X is given by

$$F_X(x) = \frac{\alpha + \beta - (\beta + \alpha + \alpha\beta x)e^{-\alpha x}}{\alpha + \beta} \tag{2}$$

The expectation and variance are given by

$$E(X) = \frac{\alpha + 2\beta}{\alpha(\alpha + \beta)}$$

$$Var(X) = \frac{\alpha^3 + 5\alpha^2\beta + 6\alpha\beta^2 + 2\beta^3}{\alpha^5 + 3\alpha^4\beta + 3\alpha^3\beta^2 + \alpha^2\beta^3}$$

The moment generation function can also be obtained in closed form and is given by

$$M_X(t) = \frac{\alpha^2(\alpha + \beta - t)}{(\alpha + \beta)(t - \alpha)^2} \tag{3}$$

It also follows that its density can be obtained as a mixture of two positive ones, namely,

$$f_X(x; \alpha, \beta) = \frac{\alpha}{\alpha + \beta} f_Y(x; \alpha) + \frac{\beta}{\alpha + \beta} f_Z(x; \alpha) \tag{4}$$

Using the representation as a mixture of two positive densities, we can provide a general representation for the distribution moments, namely,

$$E(X^r) = \frac{\alpha}{\alpha + \beta} E(Y^r) + \frac{\beta}{\alpha + \beta} E(Z^r) = \frac{r\Gamma(r)}{\alpha^r(1 + \beta)} \left[\alpha + (1 + r)\beta\right], \quad r = 1, 2, ..., \tag{5}$$

where $\Gamma(\cdot)$ is the usual gamma function.

Using the moments above for the EE model, we can compute asymmetry ($\sqrt{\beta_1}$) and kurtosis (β_2) coefficients, which are given by

$$\sqrt{\beta_1} = \frac{2(\alpha + 2\beta)^3 - 12\beta^2(\alpha + \beta)}{(\alpha^2 + 4\alpha\beta + 2\beta^2)^{3/2}} \tag{6}$$

$$\beta_2 = \frac{3(\alpha + 2\beta)^2(3\alpha^2 + 12\alpha\beta + 8\beta^2) - 72\beta^2(\alpha + \beta)^2}{(\alpha^2 + 4\alpha\beta + 2\beta^2)^2} \tag{7}$$

Lemma 1. *Note that as $\beta \to 0$, then $\sqrt{\beta_1} \to 2$ and $\beta_2 \to 9$ which correspond to the asymmetry and kurtosis respectively for the exponential model. General coefficients of asymmetry and kurtosis are such that $\sqrt{2} < \sqrt{\beta_1} \leq 2$ and $6 < \beta_2 \leq 9$, respectively, as shown in Figures 3 and 4.*

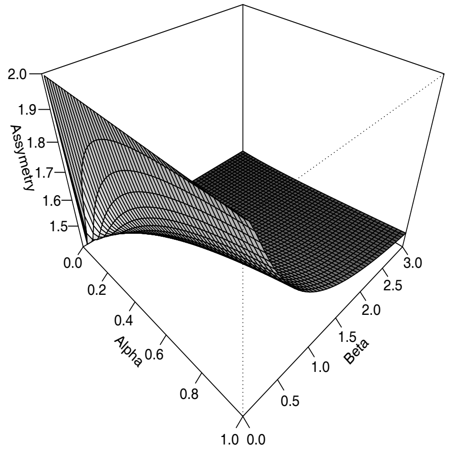


FIGURE 3: Graphs for asymmetry coefficient for the EE model.

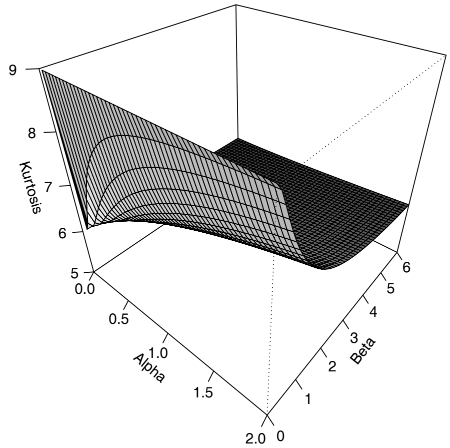


FIGURE 4: Graphs for the kurtosis coefficient for the EE model.

The hazard function for the random variable $X \sim EE(\alpha, \beta)$ is given by

$$h(x) = \frac{f(x;\alpha,\beta)}{1 - F_X(x)} = \frac{\alpha^2(1+\beta x)}{\beta + \alpha(1+\beta x)}$$

i) If $\beta = 0$, then $h(x) = \alpha$, is the hazard function for the exponential model $\forall x \in \mathbb{R}$.

ii) $\forall \beta$, $h(x)$ is monotonically increasing with $h(0) = \frac{\alpha^2}{\alpha+\beta}$.

iii) $\forall \beta$, $h(x) \to \alpha$, as $x \to \infty$.

iv) $h(x)$ is bounded, that is, $\frac{\alpha^2}{\alpha+\beta} < h(x) < \alpha$.

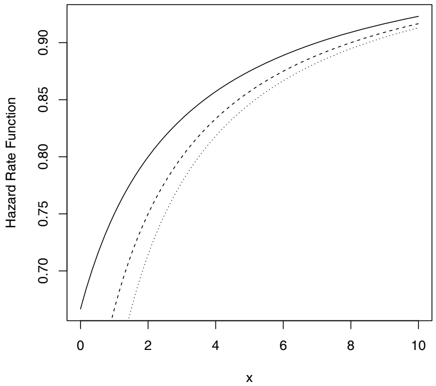


Figure 5: Plots for the hazard function for $\alpha = 1$ and $\beta = 0.5$ (solid line), $\beta = 1$ (dashed line), $\beta = 2$ (dotted line).

The Figure 5 illustrates the behavior of the hazard function for some parameter values.

3. Inferential Considerations

In this section, we consider inference for the EE using moments and the maximum likelihood approach.

3.1. Method of Moments

The moment estimators for the parameters α and β are obtained by solving

$$\begin{aligned} \frac{\alpha + 2\beta}{\alpha(\alpha+\beta)} &= \overline{x} \\ \frac{2\alpha + 6\beta}{\alpha^2(\alpha+\beta)} &= \overline{x}^2 \end{aligned} \tag{8}$$

From the first equation we obtain the moment estimators for $\beta(\widetilde{\beta})$ as a function of the moment estimator for $\alpha(\widetilde{\alpha})$.

$$\widetilde{\beta} = \frac{\widetilde{\alpha}(1-\widetilde{\alpha}\overline{x})}{\widetilde{\alpha}\overline{x} - 2}, \qquad \widetilde{\alpha} \in \left(\frac{1}{\overline{x}}, \frac{2}{\overline{x}}\right) \tag{9}$$

using (9) and the second equation for the system given in (8) we obtain the moment estimator for α.

$$\widetilde{\alpha} = \frac{2\overline{x} \pm \sqrt{4\overline{x}^2 - 2\overline{x^2}}}{\overline{x^2}} \tag{10}$$

Therefore, $\widetilde{\alpha}$ from (10) replacing α in (9) we obtain $\widetilde{\beta}$. These estimators will be used as initial parts to get the maximum likelihood estimation in the next section.

3.2. Maximum Likelihood

Let $x_1, x_2, \ldots, x_n$ a random sample from $X \sim EE(\alpha, \beta)$, so that we obtain the log-likelihood function

$$l(\alpha, \beta) = 2n \log(\alpha) - n \log(\alpha + \beta) - \alpha \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \log(1 + \beta x_i) \tag{11}$$

Differentiating the log-likelihood function with respect to α and β, the following equations follow:

$$\frac{\partial l}{\partial \alpha} = \frac{2n}{\alpha} - \frac{n}{\alpha + \beta} - \sum_{i=1}^{n} x_i = 0 \tag{12}$$

$$\frac{\partial l}{\partial \beta} = -\frac{n}{\alpha + \beta} + \sum_{i=1}^{n} \frac{x_i}{1 + \beta x_i} = 0 \tag{13}$$

From (12) we obtain

$$\widehat{\beta} = \frac{\widehat{\alpha}(1 - \widehat{\alpha}\overline{x})}{\widehat{\alpha}\overline{x} - 2}, \qquad \widehat{\alpha} \in \left(\frac{1}{\overline{x}}, \frac{2}{\overline{x}}\right) \tag{14}$$

and the maximum likelihood estimator for α is obtained by resolving numerically the following equation

$$\sum_{i=1}^{n} \frac{x_i}{1 - (1 - \overline{x}\widehat{\alpha})(\widehat{\alpha} x_i - 1)} = \frac{n}{\widehat{\alpha}} \tag{15}$$

The estimator $\widehat{\alpha}$ is the solution to the equation (15), and replacing it in (14) we obtain $\widehat{\beta}$. This algorithm leads to the maximum likelihood estimators for α and β.

4. Real Data Illustration

We consider a data set of the life of fatigue fracture of Kevlar 373/epoxy that are subject to constant pressure at the 90% stress level until all had failed, so we have complete data with the exact times of failure. For previous studies with the data sets see Andrews & Herzberg (1985) and Barlow, Toland & Freeman (1984). These data are:

0.0251,0.0886,0.0891,0.2501,0.3113,0.3451,0.4763,0.5650,0.5671,0.6566,0.6748,0.6751,
0.6753,0.7696,0.8375,0.8391,0.8425,0.8645,0.8851,0.9113,0.9120,0.9836,1.0483,1.0596,
1.0773,1.1733,1.2570,1.2766,1.2985,1.3211,1.3503,1.3551,1.4595,1.4880,1.5728,1.5733,
1.7083,1.7263,1.7460,1.7630,1.7746,1.8275,1.8375,1.8503,1.8808,1.8878,1.8881,1.9316,
1.9558,2.0048,2.0408,2.0903,2.1093,2.1330,2.2100,2.2460,2.2878,2.3203,2.3470,2.3513,
2.4951,2.5260,2.9911,3.0256,3.2678,3.4045,3.4846,3.7433,3.7455,3.9143,4.8073,5.4005,
5.4435,5.5295,6.5541,9.0960.

Using results from Section 3.1, moment estimators were computed leading to the following values: $\widetilde{\alpha} = 0.889$ and $\widetilde{\beta} = 2.563$, which were used as initial estimates for the maximum likelihood approach.

Table 1 presents basic descriptive statistics for data set. We use the notation $\sqrt{b_1}$ and b_2 to represent sample asymmetry and kurtosis coefficients.

Table 1: Descriptive statistics for rupture time.

Data set	n	$\overline{X}$	S	$\sqrt{b_1}$	b_2
Kevlar	76	1.959	1.574	2.019	8.600

Table 2: Parameter estimates for GE, NHE and EE models for the stress-rupture life data set.

Parameter estimates	GE	NHE	EE
α	0.703	0.195	0.954
β	1.709	2.007	6.366
AIC	248.488	253.476	247.3

For comparing model fitting, Akaike (1974), namely

$$AIC = -2 * \hat{\ell}(\cdot) + 2 * k$$

where k is the number of parameters in the model under consideration. The AIC specifies that the model that best fits the data is the one with the smallest AIC value.

Table 2 shows parameter estimators for distributions GE, NHE and EE using maximum likelihood (MLE) approach and the corresponding Akaike information criterion (AIC). For these data, AIC shows a better fit for the EE model. Figure 6 reveals model fitting for the three models, and Figure 7 compares the distribution functions for the three models with the empirical distribution function.

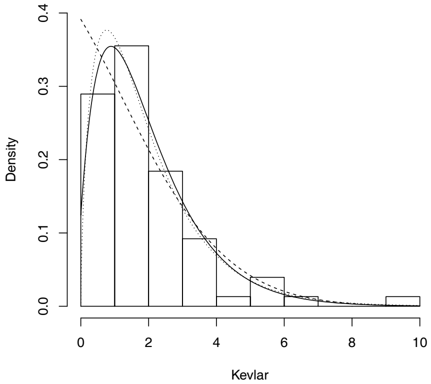


Figure 6: Models fitted by the maximum likelihood approach for the stress-rupture data set: $EE(\hat{\alpha}, \hat{\beta})$ (solid line), $NHE(\hat{\alpha}, \hat{\beta})$ (dashed line) and $GE(\hat{\alpha}, \hat{\beta})$ (dotted line)

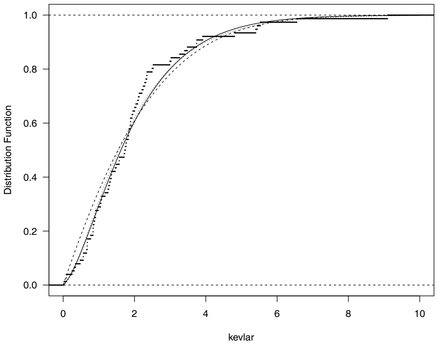


Figure 7: Empirical c.d.f. with estimated EE c.d.f. (solid line), estimated NHE c.d.f. (dashed line) and estimated GE c.d.f. (dotted line).

5. Concluding Remarks

This paper introduces a new model positive data. It is shown that the model can be represented as the mixture of two distributions. The scale-exponential distribution can be seen as a particular case of the new model. It is shown that the distribution function, hazard function and moment generating function can be obtained in closed form. Moment estimators are derived and maximum likelihood estimators can be computed using Newton-Raphson type algorithms. The moment estimators can be used as starting values for the maximum likelihood estimators. Asymmetry and kurtosis coefficients are derived and their ranges are plotted. It is illustrated the fact that the model proposed has more flexibility in terms of coefficients of asymmetry and kurtosis. A real data application has demonstrated that the model studied is quite useful for dealing with real data and behaves better in terms of fitting than other models proposed in the literature such as the GE and NHE models.

Acknowledgments

We acknowledge two referees for comments and suggestions that substantially improved the presentation. The research of Yolanda M. Gómez was supported by Becas-Chile of the Chilean government. The research of Heleno Bolfarine was supported by CNPq-Brasil.

[Recibido: enero de 2013 — Aceptado: diciembre de 2013]

References

Akaike, H. (1974), 'A new look at statistical model identification', *IEEE Transaction on Automatic Control* **AC-19**(6), 716–723.

Andrews, D. F. & Herzberg, A. M. (1985), *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, Springer Series in Statistics, New York.

Barlow, R. E., Toland, R. H. & Freeman, T. (1984), A Bayesian analysis of stress-rupture life of kevlar 49/epoxy spherical pressure vessels, *in* 'Proc. Conference on Applications of Statistics', Marcel Dekker, New York.

Gupta, R. D. & Kundu, D. (1999), 'Generalized exponential distributions', *Australian and New Zealand Journal of Statistics* **41**(2), 173–188.

Gupta, R. D. & Kundu, D. (2001), 'Exponentiated exponential family: An alternative to Gamma and Weibull distribution', *Biometrical Journal* **43**(1), 117–130.

Marshall, A. W. & Olkin, I. (2007), *Life Distributions: Structure of Nonparametric, Semiparametric and Parametric Families.*

Nadarajah, S. & Haghighi, F. (2011), 'An extension of the exponential distribution', *Statistics: A Journal of Theoretical and Applied Statistics* **45**(6), 543–558.

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 35 a 43

Discrete Likelihood Ratio Order for Power Series Distribution

Orden de la razón de verosimilitud discreta para la distribución de series de potencias

Narjes Ameli[1,a], Jalil Jarrahiferiz[2,b],
Gholam Reza Mohtashami-Borzadaran[2,c]

[1]Department of Sciences, Payam nour University of Mashhad, Mashhad, Iran
[2]Department of Statistics, School of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran

Abstract

It is well-known that some discrete distributions belong to the power series distribution (PSD) family, so it seems useful to study conditions to establish the discrete likelihood ratio order for this family. In this paper, conditions to some cases of PSD family under which the discrete likelihood ratio order we have looked at the holds. Also, we study the discrete version of the proportional likelihood ratio as an extension of the likelihood ratio order. Then we compare some members of the PSD family by discrete proportional likelihood ratio order.

***Key words*:** Binomial distribution, Geometric distribution, Logarithmic series distribution, Negative binomial distribution, Poisson distribution, Proportional likelihood ratio order.

Resumen

Es bien conocido en la literatura que algunas distribuciones discretas pertenecen a la familia de distribuciones de series de potencias (PSD, power series distributions por sus siglas en inglés). Por lo tanto, es útil estudiar algunas condiciones para establecer el orden de la razón de verosimilitud para esta familia. En este artículo, se estudian las condiciones para algunos casos de la familia PSD bajo las cuales se mantiene el orden de la razón de verosimilitud. Otros autores han introducido y estudiado el orden de la razón de verosimilitud proporcional como una extensión del orden de razón de verosimilitud para variables aleatorias continuas. Aquí, se presenta el

[a]M.Sc. E-mail: ameli_na83@yahoo.com
[b]Ph.D Student. E-mail: jarrahi.jalil@yahoo.com
[c]Professor. E-mail: grmohtashami@um.ac.ir

orden de razón de verosimilitud proporcional para variables aleatorias discretas y se estudian para la familia PSD.

***Palabras clave*:** distribución binomial, distribución binomial negativa, distribución de series logarítmicas, distribución geométrica, distribución Poisson, orden de la razón de verosimilitud proporcional.

1. Introduction

Recently, many papers have been devoted to compare random variables according to stochastic orderings in particular likelihood ratio order. Most of the contributions are for the continuous random variables. We refer to Shanthikumar & Yao (1986), Lillo, Nanda & Shaked (2001), Hu, Nanda, Xie & Zhu (2003), Shaked & Shanthikumar (2007), Misra, Gupta & Dhariyal (2008), Blazej (2008), Navarro (2008) and Bartoszewicz (2009) for more details.

Ramos-Romero & Sordo-Diaz (2001) introduced a new stochastic order between two continuous and non-negative random variables and called it proportional likelihood ratio (PLR) order, which is closely related to the usual likelihood ratio order. Belzunce, Ruiz & Ruiz (2002), extended hazard rate and reversed hazard rate orders to proportional state in the same manner and called them proportional (reversed) hazard rate orders. So, they studied their properties, preservations and relations with other orders. In general, the proportional versions are stronger orderings and easy to verify in many situations, so they are helpful to check what components are more reliable, and consequently systems formed from them.

In the next section, we recall the discrete likelihood ratio order and then compare some members of PSD family. Then we present discrete proportional likelihood ratio order and study it for PSD family at the last section of this paper.

2. Discrete Likelihood Ratio Order for Power Series Distribution Family

We obtain the conditions under which the discrete likelihood ratio order is established for some cases of the power series distribution family.

Definition 1. Let X and Y be discrete non-negative random variables with probability functions $P_X(x)$ and $P_Y(x)$ respectively. X is said to be smaller than Y in the discrete likelihood ratio order (denoted by $X \leq_{lr} Y$), if

$$\frac{P_Y(x)}{P_X(x)} \text{ is increasing in } x \in N. \tag{1}$$

Noack (1950) defined a random variable X taking non-negative integer values with probabilities

$$P(X = x) = \frac{a_x \theta^x}{b(\theta)},\ a_x \geq 0,\ x = 0, 1, 2, \dots \tag{2}$$

He called the discrete probability distribution given by (2) a power series distribution and derived some of its properties relating its moments, cumulants, etc. Patil (1961, 1962) studied the generalized power series distribution (GPSD) family with probability function like (2), whose support is any non-empty and enumerable set of non-negative integers.

Note that the Poisson, negative binomial and geometric distributions belong to PSD family and binomial and logarithmic distributions are in the GPSD family.

Suppose that X and Y have probability functions $P(X = x) = \frac{\alpha_x \theta_1^x}{b(\theta_1)}$ and $P(Y = x) = \frac{\beta_x \theta_2^x}{b(\theta_2)}$ respectively. So, using Definition 1, $X \leq_{lr} Y$ if $\frac{P_Y(x)}{P_X(x)} \leq \frac{P_Y(x+1)}{P_X(x+1)}$ for all x, or equivalently

$$(\frac{\alpha_{x+1}}{\alpha_x})(\frac{\beta_x}{\beta_{x+1}}) \leq \frac{\theta_2}{\theta_1}. \tag{3}$$

Now, we check equation (3) for some members of the PSD family:

Poisson Distribution: In equation (2), $a_x = \frac{1}{x!}$ and $b(\lambda) = e^{\lambda}$, leads to the Poisson distribution with parameter λ. Also, we get

$$\frac{P_X(x+1)}{P_X(x)} = \frac{\lambda}{1+x}.$$

Now, if X and Y possess Poisson distribution with parameters λ_1 and λ_2 respectively, then, using (3), $X \leq_{lr} Y$ if and only if $\lambda_1 \leq \lambda_2$.

Binomial Distribution: Suppose that X has binomial distribution with parameters n_1 and p_1 and Y has binomial distribution with parameters n_2 and p_2, for all $n_1 < n_2$. Using (3) and after simplification,

$$\left(\frac{n_1 - x}{n_2 - x}\right)\left(\frac{p_1}{1-p_1}\right)\left(\frac{1-p_2}{p_2}\right) \leq 1,\ x = 0, 1, \ldots, n_1 - 1$$

the left side of the above inequality gets its maximum at $x = 0$, so, if $n_1 < n_2$ and $\frac{n_1 p_1}{1-p_1} \leq \frac{n_2 p_2}{1-p_2}$ then $X \leq_{lr} Y$.

Negative Binomial Distribution: Suppose that X has negative binomial distribution with parameters r_1 and p_1 and Y has negative binomial distribution with parameters r_2 and p_2. Using (3)

$$\left(\frac{r_1 + x}{r_2 + x}\right)\left(\frac{1-p_1}{1-p_2}\right) \leq 1,\ x = 0, 1, \ldots$$

if $r_2 \leq r_1$ then, $\frac{r_1+x}{r_2+x} \leq 1$ is decreasing in $x \in N$, so gets maximum at $x = 0$. Therefore, $r_2 \leq r_1$ and $r_1(1-p_1) \leq r_2(1-p_2)$ imply that $X \leq_{lr} Y$.

Geometric Distribution: If X and Y are random variables of geometric distribution with parameters p_1 and p_2 respectively, then $p_2 \leq p_1$ implies that $X \leq_{lr} Y$ (it is evident that the geometric distribution is obtained from the negative binomial distribution where $r = 1$).

Logarithmic Series Distribution: For random variables X and Y with logarithmic series distribution with parameters θ_1 and θ_2 respectively, if $\theta_1 \leq \theta_2$ then $X \leq_{lr} Y$.

Binomial Distribution versus Poisson Distribution: If X is binomial distribution with parameters n and p and Y is Poisson distribution with parameter λ, then $X \leq_{lr} Y$ if

$$\left(\frac{p}{1-p}\right)\left(\frac{n-x}{\lambda}\right) \leq 1,\ x = 0, 1, 2, \dots, n$$

Also, maximum of the left side expression of the above inequality are given at $x = 0$, so, if $np \leq \lambda(1-p)$ then $X \leq_{lr} Y$.

Poisson Distribution versus Negative Binomial distribution: Consider random variable X having Poisson distribution with parameter λ and Y having negative binomial distribution with parameters r and p. Since $\frac{1}{r+x}$ is decreasing in x, then $\lambda \leq r(1-p)$ leads to $X \leq_{lr} Y$.

Poisson Distribution versus Geometric distribution: If X is Poisson distribution with parameter λ and Y is geometric distribution with parameter p, then, $X \leq_{lr} Y \Longleftrightarrow \lambda \leq 1 - p$.

Poisson Distribution versus Logarithmic Series Distribution: Let X and Y be random variables of Poisson and logarithmic series distributions with parameters θ_1 and θ_2 respectively. So, $X \leq_{lr} Y \Longleftrightarrow \theta_1 \leq \theta_2$.

Negative Binomial versus Logarithmic Series Distribution: The random variable X of negative binomial with parameters r and p is smaller in sense of likelihood ratio order than Y of logarithmic series distribution with parameter θ in the likelihood ratio order if $\theta \geq (1-p)(r+1)$.

Table 1: Necessary conditions for establishment discrete likelihood ratio order.

$X \leq_{lr} Y$	Conditions
$X \sim Poi(\lambda_1)$ and $Y \sim Poi(\lambda_2)$	$\lambda_1 \leq \lambda_2$
$X \sim Bin(n_1, p_1)$ and $Y \sim Bin(n_2, p_2)$	$n_1 \leq n_2$ and $\frac{n_1 p_1}{1-p_1} \leq \frac{n_2 p_2}{1-p_2}$
$X \sim Nb(r_1, p_1)$ and $Y \sim Nb(r_2, p_2)$	$r_2 \leq r_1$ and $r_2(1-p_2) \geq r_1(1-p_1)$
$X \sim Ge(p_1)$ and $Y \sim Ge(p_2)$	$p_1 \geq p_2$
$X \sim Ls(\theta_1)$ and $Y \sim Ls(\theta_2)$	$\theta_1 \leq \theta_2$
$X \sim Bin(n, p)$ and $Y \sim Poi(\lambda)$	$np \leq \lambda(1-p)$
$X \sim Poi(\lambda)$ and $Y \sim Nb(r, p)$	$\lambda \leq r(1-p)$
$X \sim Poi(\lambda)$ and $Y \sim Ge(p)$	$\lambda \leq (1-p)$
$X \sim Poi(\lambda)$ and $Y \sim Ls(\theta)$	$\lambda \leq \theta$
$X \sim Nb(r, p)$ and $Y \sim Ls(\theta)$	$\theta \geq (r+1)(1-p)$

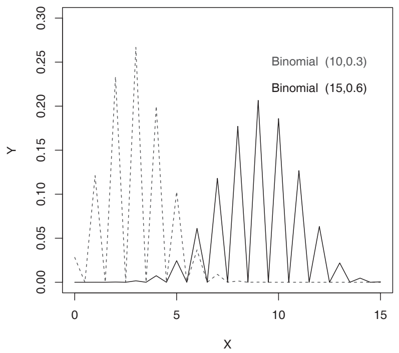


Figure 1: The Dot-Dot line shows the Binomial distribution with parameters $n_1 = 10$ and $p_1 = 0.3$ and the stretch shows the Binomial distribution with parameters $n_2 = 15$ and $p_2 = 0.6$.

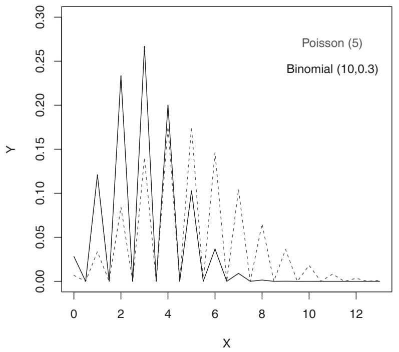


Figure 2: The Dot-Dot line shows the Poisson distribution with parameter $\lambda = 5$ and the stretch shows the Binomial distribution with parameters $n = 10$ and $p = 0.3$.

3. Discrete Proportional Likelihood Ratio Order for Power Series Distribution Family

Ramos-Romero & Sordo-Diaz (2001) studied proportional likelihood ratio order as extension of the likelihood ratio order for non-negative absolutely continuous random variables. They obtained various properties and applications of the proportional likelihood ratio order. In this section, discrete proportional likelihood ratio order is studied. Also, we looked the conditions under which this ordering is hold for PSD.

Definition 2. For two discrete non-negative random variables X and Y with probability functions $P_X(x)$ and $P_Y(x)$ respectively, if

$$\frac{P_Y([\lambda x])}{P_X(x)} \text{ is increasing in } x \in N \tag{4}$$

where $\lambda \leq 1$ is any positive constant and $[\cdot]$ denote the integer part function. Then, we say that X is smaller than Y in the discrete proportional likelihood ratio order (denoted by $X \leq_{plr} Y$).

Definition 3. We say that the discrete non-negative random variables X has increasing likelihood ratio order (denoted by $X \in IPLR$) if $\frac{p_X([\lambda x])}{p_X(x)}$ for $0 \leq \lambda \leq 1$ in increasing.

Theorem 1. *Let X and Y be two discrete non-negative random variables with probability functions $P_X(x)$ and $P_Y(x)$ respectively. If $X \leq_{lr} Y$ and $Y \in IPLR$, then $X \leq_{plr} Y$.*

***Proof*.** Since

$$\frac{p_Y([\lambda x])}{p_X(x)} = \frac{p_Y(x)}{p_X(x)} \frac{p_Y([\lambda x])}{p_Y(x)}$$

the proof is clear. □

Let X and Y be discrete non-negative random variables with probability functions $P(X = x) = \frac{\alpha_x \theta_1^x}{b(\theta_1)}$ and $P(Y = x) = \frac{\beta_x \theta_2^x}{b(\theta_2)}$ respectively. So, using Definition 2, $X \leq_{plr} Y$ if and only if

$$\left(\frac{\alpha_{[\lambda x+\lambda]}}{\alpha_{[\lambda x]}}\right)\left(\frac{\beta_x}{\beta_{x+1}}\right) \geq \frac{\theta_2}{{\theta_1}^{[\lambda x+\lambda]-[\lambda x]}}. \tag{5}$$

Geometric Distribution: Let X and Y having geometric distribution with parameters p_1 and p_2 respectively, using (5), we have $X \leq_{plr} Y$ if

$$\frac{P_Y([\lambda x])}{P_X(x)} = \frac{q_2^{[\lambda x]-1} p_2}{q_1^{x-1} p_1}$$

is increasing in x. That is

$$\frac{q_2^{[\lambda x]-1} p_2}{q_1^{x-1} p_1} \leq \frac{q_2^{[\lambda x+\lambda]-1} p_2}{q_1^{x} p_1}$$

that is equivalent to $q_1 \leq q_2^{[\lambda x+\lambda]-[\lambda x]}$. If $[\lambda x + \lambda] = [\lambda x]$, then $q_1 \leq 1$. If $[\lambda x + \lambda] = [\lambda x] + 1$, then $q_1 \leq q_2$. So, $X \leq_{plr} Y$ if and only if $p_1 \geq p_2$.

Poisson Distribution: Let X having Poisson distribution with parameter θ. If

$$\frac{x!}{[\lambda x]!}\theta^{[\lambda x]-x} \leq \frac{(x+1)!}{[\lambda x+\lambda]!}\theta^{[\lambda x+\lambda]-x-1}$$

then,

$$\frac{P_X([\lambda x])}{P_X(x)} = \frac{x!}{[\lambda x]!}\theta^{[\lambda x]-x}$$

is increasing. If $[\lambda x + \lambda] = [\lambda x]$, then $x!\theta^{[\lambda x]-x} \leq (x+1)!\theta^{[\lambda x]-x-1}$, so, $\theta \leq x+1$, that by increasing $h(x) = x+1$, it implies that $\theta \leq 1$. But if $[\lambda x + \lambda] = [\lambda x] + 1$, then

$$\frac{x!}{[\lambda x]!}\theta^{[\lambda x]-x} \leq \frac{(x+1)!}{([\lambda x]+1)!}\theta^{([\lambda x]+1)-x-1}$$

that is $[\lambda x + 1] \leq x + 1$, which always is true. Therefore, if X and Y having Poisson distribution with parameters θ_1 and θ_2 respectively and $\theta_1 \leq \theta_2 \leq 1$, then $X \leq_{plr} Y$.

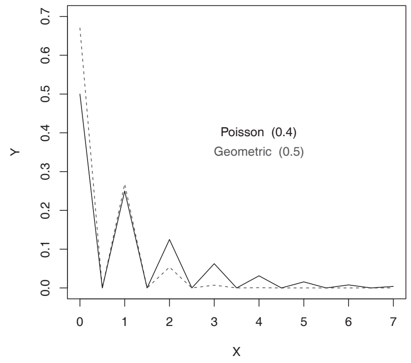


Figure 3: The Dot-Dot line shows the Geometric distribution with parameter $p = 0.5$ and the stretch shows the Poisson distribution with parameter $\lambda = 0.4$.

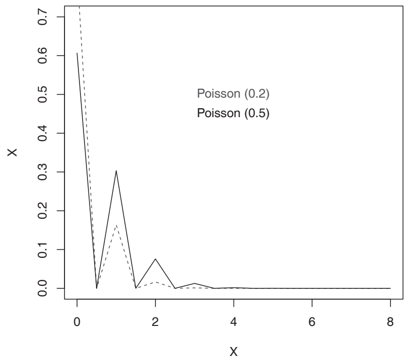


Figure 4: The Dot-Dot line shows the Poisson distribution with parameter $\lambda_1 = 0.2$ and the stretch shows the Poisson distribution with parameter $\lambda_2 = 0.5$.

Binomial Distribution: Consider X having binomial distribution with parameters n and p, then,

$$\frac{P_X([\lambda x])}{P_X(x)} = \frac{x!}{[\lambda x]!}\frac{(n-x)!}{(n-[\lambda x])!}\left(\frac{p}{q}\right)^{[\lambda x]-x}$$

is increasing in x if

$$\frac{x!}{[\lambda x]!}\frac{(n-x)!}{(n-[\lambda x])!}\left(\frac{p}{q}\right)^{[\lambda x]-x} \leq \frac{(x+1)!}{[\lambda x+\lambda]!}\frac{(n-x-1)!}{(n-[\lambda x+\lambda])!}\left(\frac{p}{q}\right)^{[\lambda x+\lambda]-x-1}$$

If $[\lambda x + \lambda] = [\lambda x]$, we have

$$\frac{x!}{(x+1)!}\frac{(n-x)!}{(n-x-1)!} \leq \frac{q}{p}$$

that means $\frac{n-x}{x+1} \leq \frac{q}{p}$. The function $h(x) = \frac{n-x}{x+1}$ is decreasing in x. So, $q \geq np$.

If $[\lambda x + \lambda] = [\lambda x] + 1$, then,

$$\frac{n-x}{n-[\lambda x]} \leq \frac{x+1}{[\lambda x]+1}$$

that is $n[\lambda x] - x \leq nx - [\lambda x]$ which always is true. Therefore, if X having binomial distribution with parameters n_1 and p_1 and Y having binomial distribution with parameters n_2 and p_2, which $n_1 < n_2$ respectively. If $\frac{n_1 p_1}{1-p_1} \leq \frac{n_2 p_2}{1-p_2} \leq 1$, then, $X \leq_{plr} Y$.

Table 2: Necessary conditions for establishment discrete proportional likelihood ratio order.

$X \leq_{plr} Y$	Conditions
$X \sim Poi(\lambda_1)$ and $Y \sim Poi(\lambda_2)$	$\lambda_1 \leq \lambda_2 \leq 1$
$X \sim Bin(n_1, p_1)$ and $Y \sim Bin(n_2, p_2)$	$n_1 < n_2$ and $\frac{n_1 p_1}{1-p_1} \leq \frac{n_2 p_2}{1-p_2} \leq 1$
$X \sim Ge(p_1)$ and $Y \sim Ge(p_2)$	$p_1 \geq p_2$

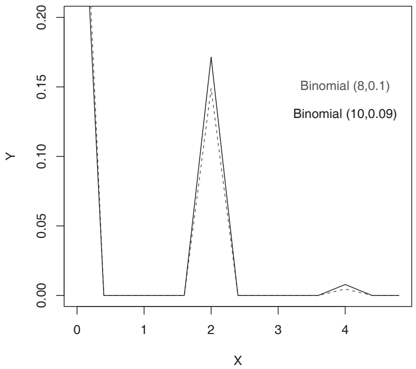


Figure 5: The Dot-Dot line shows the Binomial distribution with parameters $n_1 = 8$ and $p_1 = 0.1$ and the stretch shows the Binomial distribution with parameters $n_2 = 10$ and $p_2 = 0.09$.

At the end of paper and in order to better understand, some distributions of the PSD family are simulated satisfying in the above conditions.

4. Conclusions

In this paper, we compare some members of the PSD family due to discrete likelihood ratio order. Then we presented the discrete version of proportional likelihood ratio order as an extension of the discrete likelihood ratio order and studied it for the PSD family.

[Recibido: abril de 2013 — Aceptado: diciembre de 2013]

References

Bartoszewicz, J. (2009), 'On a represervation of weighted distributions', *Statistics and Probability Letters* **79**, 1690–1694.

Belzunce, F., Ruiz, J. M. & Ruiz, C. (2002), 'On preservation of some shifted and proportional orders by systems', *Statistics and Probability Letters* **60**, 141–154.

Blazej, P. (2008), 'Reservation of classes of life distributions under weighting with a general weight function', *Statistics and Probability Letters* **78**, 3056–3061.

Hu, T., Nanda, A. K., Xie, H. & Zhu, Z. (2003), 'Properties of some stochastic orders: A unified study', *Naval Research Logistic* **51**, 193–216.

Lillo, R. E., Nanda, A. K. & Shaked, M. (2001), 'Preservation of some likelihood ratio stochastic orders by order statistics', *Statistics and Probability Letters* **51**, 111–119.

Misra, N., Gupta, N. & Dhariyal, I. (2008), 'Preservation of some aging properties and stochastic orders by weighted distributions', *Communications in Ststistics-Theory and Methods* **37**, 627–644.

Navarro, J. (2008), 'Likelihood ratio ordering of order statistics, mixture and systems', *Statistical of Planning and Inference* **138**, 1242–1257.

Noack, A. (1950), 'A class of random variables with discrete distributions', *Annals of Mathematical Statistics* **21**, 127–132.

Patil, G. P. (1961), Contributions to estimation in a class of discrete distributions, Ph.D thesis, University of Michigan.

Patil, G. P. (1962), 'Certain properties of the generalized power series distributions', *Annals of the Statistical Mathematics* **14**, 179–182.

Ramos-Romero, H. M. & Sordo-Diaz, M. A. (2001), 'The proportional likelihood ratio order and applications', *Questiio* **25**, 211–223.

Shaked, M. & Shanthikumar, J. G. (2007), *Stochastic Orders*, 1 edn, Academic Press, New York.

Shanthikumar, J. G. & Yao, D. D. (1986), 'The preservation of likelihood ratio ordering under convolutions', *Stochastic Processes and their Applications* **23**, 259–267.

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 45 a 67

Cramér-Von Mises Statistic for Repeated Measures

El estadístico de Cramér-Von Mises para medidas repetidas

Pablo Martínez-Camblor[1,2,a], Carlos Carleos[2,b], Norberto Corral[2,c]

[1]Oficina de Investigación Biosanitaria (OIB), FICYT, Oviedo, Spain
[2]Departamento Estadística e IO y DM, Universidad de Oviedo, Asturias, Spain

Abstract

The Cramér-von Mises criterion is employed to compare whether the marginal distribution functions of a k-dimensional random variable are equal or not. The well-known Donsker invariance principle and the Karhunen-Loéve expansion is used in order to derive its asymptotic distribution. Two different resampling plans (one based on permutations and the other one based on the general bootstrap algorithm, gBA) are also considered to approximate its distribution. The practical behaviour of the proposed test is studied from a Monte Carlo simulation study. The statistical power of the test based on the Cramér-von Mises criterion is competitive when the underlying distributions are different in location and is clearly better than the Friedman one when the sole difference among the involved distributions is the spread or the shape. Both resampling plans lead to similar results although the gBA avoids the usual required interchangeability assumption. Finally, the method is applied on the study of the evolution inequality incomes distribution between some European countries along the years 2000 and 2011.

***Key words*:** Asymptotic Distribution, Bootstrap, Cramér-von Mises statistic, Hypothesis testing, Permutation test, Repeated Measures.

Resumen

El criterio de Cramér-von Mises es empleado para comparar la igualdad entre las distribuciones marginales de una variable aleatoria k-dimensional. El conocido principio de invaranza de Donsker y la expansión de Karhunen-Loéve se usan para derivar su distribución asintótica. Dos planes de remuestreo diferentes (uno basado en permutaciones y el otro basado en el algoritmo bootstrap general, gBA) son usados para aproximar su distribución. El comportamiento práctico del test propuesto es estudiado mediante simulaciones de Monte Carlo. La potencia estadística del test basado en el criterio

[a] Biostatistician. E-mail: pmcamblor@hotmail.com
[b] Professor. E-mail: carleos@uniovi.es
[c] Lecturer. E-mail: norbert@uniovi.es

de Cramér-von Mises es competitiva cuando la distribuciones subyacentes difieren en el parámetro de localización. Este test es claramente superior al de Friedman cuando las únicas diferencias son en la dispersión o la forma. Ambos planes de remuestreo obtienen resultados similares aunque el gBA evita la hipótesis de intercambiabilidad. Finalmente, el método propuesto es aplicado al estudio de la evolución de las desigualdades en los ingresos entre algunos países Europeos entre los años 2000 y 2011.

***Palabras clave*:** Bootstrap, distribución asintótica, estadístico de Cramér-von Mises, medidas repetidas, test de hipótesis, test de permutaciones.

1. Introduction

The comparison of the equality among the marginal distribution functions of a k-dimensional random variable is a common problem in statistical inference (for example, in biomedicine, in problems of comparing diagnostic procedures or bioequivalence (Freitag, Czado & Munk 2007). In practice, most frequent cases are the study of one feature measured on the same subjects at different time moments (analysis of repeated measures) and matched studies. Despite of, there exists a number of methods of comparing the equality among k-distributions from independent samples, the k-sample problem for dependent data has not been as widely studied and, the traditional parametric (ANOVA) and nonparametric (Friedman test) repeated measures procedures are the usual used techniques to solve these problems.

In this context, several rank tests have been proposed. In a non exhaustive revision: Ciba-Geigy & Olsson (1982) developed a specific one for comparing dispersion in paired samples design; Lam & Longnecker (1983) introduced modifications which improve the power of the classical Wilcoxon rank sum test for this topic; Munzel (1999*a*) used the normalized version of distribution functions to derive an asymptotic theory for rank statistics including ties and considered a mixed model which permits almost arbitrary dependences; Munzel (1999*b*) studied different nonparametric permutation methods for repeated measures problems in a two sample framework; most recently, Freitag et al. (2007) proposed a test based on the Mallows distance with this goal. Other authors as Govindarajulu (1995) Govindarajulu (1997) or Podgor & Gastwirth (1996) also dealt with this topic from different approaches.

Although the use of bootstrap on multivariate problems is straightforward in order to build confidence intervals and related estimates, the way to resampling under the null (in particular, the way to involve this assumption on the resampling) for preserving the original data structure is not direct and the use of bootstrap on hypothesis testing (which involve paired design) is not so clear. It is not trivial how to involve the (null) hypothesis of equality of the k marginal distributions of a multivariate random variable. The most common procedure, the *permutation* test (see, for example, Good 2000, Munzel 1999*b*), implies that the different components of the k-dimensional random vector must be *interchangeable* (see Venkatraman & Begg 1996 or, most recently, Nelsen 2007). Under the null, this is not a very strong assumption to compare two samples (most of the previous cited works engage, exclusively, on this particular case) but, for three or more samples it means that the relationship between each pair must be the same (it is also known as *sphericity hypothesis*) and, in spite of for most of the usual statistics, in practice, the permutation test has demonstrated its robustness with respect to this assumption, it is usually violated.

In this paper, the authors deal with the problem of comparing the equality among the k marginal distribution functions from a typical multivariate problem. With this goal, the traditional Cramér-von Mises criterion is considered. The Donsker invariance principle and the classical Gaussian processes theory, in particular, the Karhunen-Loève expansion, are used in order to obtain (a not explicit version of) the asymptotic distribution for the Cramér-von Mises statistic when the samples are from the same subjects. The properties of this statistic allow to develop a resampling procedure which does not need the (usual) *interchangeability* (or sphericity) assumption. This method is described and its consistency is proved. We think it is worth mentioning that, the considered procedures (the asymptotic, permutation and the bootstrap ones), are simple, useful and easily to implement. A simulation study is carried out (Section 3); its results suggest that the Cramér-von Mises criterion obtains good results in all considered situations and it is clearly better than the Friedman test when distributions differ mainly in their spread or shape. These results are the usual ones when the Cramér-von Mises criterion is used in other context (see, for example, Martínez-Camblor & Uña-Álvarez (2009) or Martínez-Camblor (2011)). Finally, the proposed method is applied on the study of the inequality incomes between thirty European countries during the years 2000 and 2011 (Section 4).

During the revision process of this paper, it has been published the work of Quessy & Éthier (2012) (QE) which, from a slight different approach, deals with the same problem. The main results of the present manuscript had been developed around 2008-2009 and, of course, independently of the previously cited work. In order to keep this independence and, in spite of several reported results are overlapping with the obtained by QE, we have maintained them in the appendix.

2. Cramér-von Mises Statistic for Repeated Measures

The well-known Cramér-von Mises criterion introduced, separately by Harald Crámer and Richard Edler von Mises (Cramér 1928, Von Mises 1991), was originally to compare the goodness of fit of a probability distribution F^* and a fixed distribution function F_0 and is given by

$$W^2 = \int (F^*(t) - F_0(t))^2 dF_0(t)$$

In the immediately one-sample applications, F_0 is the theoretical cumulative distribution function (CDF) and F^* is the empirical cumulative distribution function (ECDF), $\hat{F}_n$. Csörgo & Faraway (1996) derived the exact distribution for this statistic and proposed a correction for its asymptotic distribution. Anderson (1962) derived the asymptotic distribution for the two-sample case. A standard k-dimensional generalization was proposed by Kiefer (1959) which considered the expression

$$W_k^2 = \sum_{i=1}^{k} n_i \int (\hat{F}_{n_i}(X_i, t) - \hat{F}_N(X, t))^2 d\hat{F}_N(X, t)$$

where $N = \sum_{i=1}^{k} n_i$ and $\hat{F}_{n_i}(X_i, t)$ and $\hat{F}_n(X, t)$ are the ECDF referred to the ith sample $(1 \leq i \leq k)$ and to the pooled sample, respectively. Brown (1982) also dealt with the k-sample problem, he studied the asymptotic distribution and introduced a permutation test based on the same criterion. In Martínez-Camblor & Uña-Álvarez (2009), the statistical power of this statistic was considered in a simulation study (joint with other six statistics based on the ECDF and four more based on the kernel density estimator). The W_k^2 test obtained very competitive results in the eight considered models (four symmetrical and four asymmetrical).

In this section we study the different approximations for the distribution of W_k^2 when data are from a multivariate variable i.e., in our case, we have a k-dimensional random sample $\boldsymbol{X} = (X_1, \ldots, X_k)$ with $X_i = (x_{i1}, \ldots, x_{in})$ (n subjects have been collected) for $i \in 1, \ldots, k$, from a k-dimensional random variable $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)$. For each $\boldsymbol{u} = (u_1, \ldots, u_k)$ with $u_1, \ldots, u_k \in \mathbb{R}$ the k dimensional functions

$$\begin{aligned} \hat{\boldsymbol{F}}_{\boldsymbol{n}}(\boldsymbol{X}, \boldsymbol{u}) =& (\hat{F}_{n,1}(X_1, u_1), \ldots, \hat{F}_{n,k}(X_k, u_k)) \\ \boldsymbol{F}(\boldsymbol{u}) =& (F_1(u_1), \ldots, F_k(u_k)) \end{aligned}$$

denote the vectors with the ECDFs and the theoretical cumulative distribution functions (CDFs), respectively. In Theorem 1, it is proved (we must remark; a non explicit version of) the asymptotic distribution for the statistic

$$W_k^2(n) = \sum_{i=1}^{k} n \int (\hat{F}_{n,i}(X_i, t) - \hat{F}_{n,\bullet}(\boldsymbol{X}, t))^2 d\hat{F}_{n,\bullet}(\boldsymbol{X}, t)$$

where $\hat{F}_{n,i}(X_i, t)$ $(1 \leq i \leq k)$ is the ECDF referred to the ith sample and $\hat{F}_{n,\bullet}(\boldsymbol{X}, t) = k^{-1} \sum_{i=1}^{k} \hat{F}_{n,i}(X_i, t)$, when the (null) hypothesis

$$H_0 : F_1 = \cdots = F_k \; (= F) \tag{1}$$

is true.

Theorem 1. *Let $\boldsymbol{\xi}$ be a k-dimensional random vector and let $\boldsymbol{X}$ be a random sample from $\boldsymbol{\xi}$ (with size n), by using the above notation, if $F_1 = \cdots = F_k \; (= F)$ (null hypothesis), it is hold the (weak) convergence*

$$W_k^2(n) \xrightarrow{\mathcal{L}}_n \sum_{i=1}^{k} \sum_{l \in \mathbb{N}} \lambda_{i,l} M_{i,l}^2$$

where $\{\boldsymbol{M}_l = (M_{1,l}, \ldots, M_{k,l})\}_{l \in \mathbb{N}}$ is a sequence of k-dimensional, normal distributed random variables whose marginals follow a $\mathcal{N}(0, 1)$ distribution and $\{\{\lambda_{i,l}\}_{i=1}^{k}\}_{l \in \mathbb{N}}$ are non negative constants satisfying $\sum_{l \in \mathbb{N}} \lambda_{i,l}^2 < \infty$ for $1 \leq i \leq k$.

The above Theorem guarantees the consistency and gives the convergence rate for the studied statistic. However, strictly speaking, this result does not provide its distribution in full. In order to build asymptotic critical regions, the explicit values for the $\{\{\lambda_{i,l}\}_{i=1}^{k}\}_{l\in\mathbb{N}}$ coefficients must been known (eigenvalues and eigenfunctions must be computed). However, we want to note that this is a non-trivial problem which involves complex (and sometimes, for some readers, cumbersome analysis see, for example, Deheuvels 2005). In addition, these eigenvalues depend on the covariance data structure and they should be computed particularly for each problem. The following remark is devoted to point out some comments about the eigenvalues calculation in the two-sample case.

***Note* 1.** In the two sample-case, the asymptotic distribution of $W_k^2(n)$ under the null is equivalent to the distribution of $W^2 = \int(\mathcal{W}_1\{t\} - \mathcal{W}_2\{t\})^2 dt$, where $\mathcal{W}_i\{t\}$ $(i \in 1, 2)$ is a standard Brownian bridge. Eigenvalues and eigenfunctions are the non zero solutions to the Fredholm type integral equation

$$\lambda_j e_j(u) = \int \mathbb{C}(u, v) e_(v) dv$$

with the above restrictions on the eigenfunctions, e_j (i.e. orthonormality). In this particular setting,

$$\begin{aligned}\mathbb{C}(u, v) =&\mathbb{E}[(\mathcal{W}_1\{u\} - \mathcal{W}_2\{u\})(\mathcal{W}_1\{v\} - \mathcal{W}_2\{v\})]\\ =&2\,(u \wedge v - uv) - (f(u, v) + f(v, u))\end{aligned}$$

where $f(s, t) = \mathbb{E}[\mathcal{W}_1\{s\}\mathcal{W}_2\{t\}]$ (note that $f(s, t) = 0$ for independent samples). Obviously, the particular solutions depend on the function f. For instance, assuming $f(u, v) + f(v, u) = u \wedge v - uv$, functions $\sin(j\pi u)$ and $\cos(j\pi u)$ $(j \in \mathbb{N})$ are possible solutions which lead to eigenvalues in the form $\lambda_j = (j\pi)^{-2}$ (see, for instance, Van der Vaart 1998). □

Usually, in order to approximate the asymptotic distribution, the largest eigenvalue is taken and the other ones are ignored i.e., by using the coefficients properties (see the Theorem 1 proof in the Appendix, in particular, equation (8)), it is obtained the approximation

$$\sum_{i=1}^{k}\sum_{l\in\mathbb{N}} \lambda_{i,l} M_{i,l}^2 = \sum_{i=1}^{k}\left(\sum_{l\in\mathbb{N}} \lambda_{i,l}\left(M_{i,l}^2 - 1\right) + C_i\right) \sim \sum_{i=1}^{k}\left(\lambda_{i,1}\left(M_{i,1}^2 - 1\right) + C_i\right)$$

Unfortunately, the first eigenvalue is also unknown. However, for each $i \in 1, \dots, k$, we can approximate the first (and, therefore, the biggest) eigenvalue by

$$\lambda_{i,1}^2 \sim \tilde{\lambda}_{i,1}^2 = \iint \mathbb{C}_{i,i}(s, t)^2 dF(s) dF(t)$$

Note that it is known that $\tilde{\lambda}_{i,1} \geq \lambda_{i,1}$ $(i \in 1, \dots, k)$ and the equality is true only when $\lambda_{i,l} = 0 \quad \forall l > 1$. Finally, in order to save the relationship among the different involved samples, we build $\boldsymbol{M}_1 = (M_{1,1}, \dots, M_{k,1})$ such that, for $1 \leq i, j \leq k$

$$C_i C_j \mathbb{E}\left[M_{i,1}^2 M_{j,1}^2\right] = \mathbb{E}\left[\int \{Y_{F_i}^i(t) - \bar{Y}_{F,\bullet}(t)\}^2 dF(t) \int \{Y_{F_j}^j(t) - \bar{Y}_{F,\bullet}(t)\}^2 dF(t)\right]$$

We can work, without loss of generality, with $\mathbb{E}\left[M_{1,1}M_{2,1}\right]$. It is easy to check that $2\,\mathbb{E}[M_{1,1}M_{2,1}]^2 = \mathbb{E}[M_{1,1}^2 M_{2,1}^2] - 1$ and

$$\begin{aligned}
\mathbb{E}\left[M_{1,1}^2 M_{2,1}^2\right] =& \frac{1}{C_1C_2}\mathbb{E}\left[\int \{Y_{F_1}^1(t) - \bar{Y}_{F,\bullet}(t)\}^2 dF(t) \int \{Y_{F_2}^2(t) - Y_{F,\bullet}^2(t)\}^2 dF(t)\right] \\
=& \frac{1}{C_1C_2}\mathbb{E}\left[\iint \{Y_{F_1}^1(t) - \bar{Y}_{F,\bullet}(t)\}^2 \{Y_{F_2}^2(s) - \bar{Y}_{F,\bullet}(s)\}^2 dF(t)dF(s)\right] \\
=& \frac{1}{C_1C_2}\iint \mathbb{E}\left[(\boldsymbol{Y_F(t)a}_1^t)^2(\boldsymbol{Y_F(s)a}_2^t)^2\right] dF(t)dF(s)
\end{aligned}$$

With some additional computes and taking into account that, for $1 \leq i,j \leq k$, $F_{i,j}(u,v) = F_{j,i}(v,u)$, it is obtained

$$\begin{aligned}
C_{1,2}(s,t) =& \mathbb{E}\left[(\boldsymbol{Y_F(s)a}_1^t)(\boldsymbol{Y_F(t)a}_2^t)\right] \\
=& F_{1,2}(s,t) - \bar{F}_{1,\cdot}(s,t) - \bar{F}_{2,\cdot}(t,s) + \bar{F}_{\cdot,\cdot}(s,t) \qquad (2)
\end{aligned}$$

then,

$$\mathbb{E}\left[M_{1,1}^2 M_{2,1}^2\right] = \frac{1}{C_1C_2}\iint \left(2\,C_{1,2}^2(s,t) + C_1(s)C_2(t)\right) dF(s)dF(t) \qquad (3)$$

and the asymptotic distribution can be approximated by

$$C_A = \sum_{i=1}^{k}\left(\tilde{\lambda}_{i,1}\left(M_{i,1}^2 - 1\right) + C_i\right) \qquad (4)$$

We compute $\tilde{\lambda}_{i,1}$ $(1 \leq i \leq k)$ by using the estimation of some parameters of the statistic and, unfortunately, from this method we cannot estimate any other eigenvalue. In the independent case, the quality of this approximation has been checked via simulations (see, for instance, Martínez-Camblor, Carelos & Corral (2012), and references therein).

Note that both the expected value and the variance of C_A are equal to the $W_k^2(n)$ ones. The (theoretical) unknown parameters which are involved in the equation (4) can be estimated by putting the respective ECDFs instead of the theoretical ones (typical plug-in method) in their explicit expressions (equations (2) and (3)). At this point, it is worth to remember that, under the null hypothesis, all the marginal functions are equal. Once these values are computed, the asymptotic distribution under the null might be approximated by using some bound for the quadratic forms (see, for example, Alkarni & Siddiqui 2001) or by using the Monte Carlo method: Generating T independent samples (with the original sample size) from the k-dimensional normal distribution (previously we must compute its correlation matrix by using the corresponding equations) and computing the respective T asymptotic values of the statistic by using (4). In Section 3, the latter possibility is employed in the simulation study.

On the other hand, the Cramér-von Mises statistic properties allow to propose an useful resampling plan in order to approximate its distribution for paired samples in small size problems. The following subsection is devoted to develop a bootstrap approximation in the current context.

2.1. Bootstrap Approximation

The *bootstrap*, introduced and explored in detail by Bradley Efron (Efron 1979, Efron 1982), is an (not only but mainly nonparametric) intensive computer-based method of statistical inference which is often used in order to solve many real questions without the need of knowing the underlying mathematical formulas. Besides, under regularity conditions, the distribution bootstrap estimation is asymptotically minimax among all possible estimates (Beran 1982).

Despite of the bootstrap method has received a great deal of attention and popularity, its use on statistical hypothesis testing has received considerable, although minor attention (Martin 2007). Following Hall & Wilson (1991), many authors such as Westfall & Young (1993) have promoted null resampling as critical to the proper construction of bootstrap tests. However, in related sample distribution comparison, it is not straightforward how to resample under the null and the permutation tests (Good 2000) are the usual ones employed with this goal. In order to guarantee the consistency of the last method, *exchangeability* among the different components must be assumed (Venkatraman & Begg 1996) and, although for most of the statistics, in practice, this technique has proved its robustness with respect to this assumption, in k-dimensional problems ($k > 2$) it is usually violated. Recently, Martínez-Camblor et al. (2012) proposed a general resampling plan which focus its use on hypothesis testing without the need of assuming additional conditions. In particular, for the present problem, under the null, it is easy to prove that:

Theorem 2. *Under the Theorem 1 assumptions and by using the same notation. Let $\boldsymbol{X}^* = (X_1^*, \dots, X_k^*)$ be an independent random sample generated from $\hat{\boldsymbol{F}}_{\boldsymbol{n}}(\boldsymbol{X}, \cdot)$ (multivariate ECDF referred to the random sample $\boldsymbol{X}$). If*

$$\begin{aligned} W_k^{2,*}(n) = \sum_{i=1}^{k} n \int \{\hat{F}_{n,i}^*(X_i^*, t) - \hat{F}_{n,i}(X_i, t)\}^2 d\hat{F}_{n,\bullet}^*(\boldsymbol{X}^*, t) \\ - nk \int \{\hat{F}_{n,\bullet}^*(\boldsymbol{X}^*, t) - \hat{F}_{n,\bullet}(\boldsymbol{X}, t)\}^2 d\hat{F}_{n,\bullet}^*(\boldsymbol{X}^*, t) \end{aligned}$$

where for each $i \in 1, \dots, k$, $\hat{F}_{n,i}^*(X_i^*, t)$ is the ECDF referred to X_i^* and $\hat{F}_{n,\bullet}^*(\boldsymbol{X}^*, t) = k^{-1} \sum_{i=1}^{k} \hat{F}_{n,i}^*(X_i^*, t)$. Under the null, it is held,

$$\left\{\mathcal{P}_{\boldsymbol{X}}\left(\mathcal{W}_{\|}^{\in,*}(\backslash) \leq \sqcap\right) - \mathcal{P}\left(\mathcal{W}_{\|}^{\in}(\backslash) \leq \sqcap\right)\right\} \longrightarrow_n 0 \qquad a.s.$$

where $\mathcal{P}_{\boldsymbol{X}}$ denotes probability conditionally on sample $\boldsymbol{X}$.

The above result proves the punctual convergence (for each, fixed, $u \in \mathbb{R}$) of the bootstrap method. Uniform convergence can also be derived (under mild and usual conditions) from general theory of U and V statistics (see, for example, Arcones & Gine 1992). Theorem 2 guarantees that the distribution of $W_k^2(n)$ can be approximated by the $W_k^{2,*}(n)$ one and, as usual, this distribution can be approximated by using the Monte Carlo method following the algorithm:

B_1. From the original sample, $\boldsymbol{X}$, compute $W_k^2(n)$.

B_2. From the multivariate cumulative empirical distribution function, $\hat{\boldsymbol{F}}_{\boldsymbol{n}}(\boldsymbol{X}, t)$, draw B independent k-dimensional random samples with size n,

$$\boldsymbol{X^{*,b}} = (X_1^*, \dots, X_k^*), \quad 1 \le b \le B$$

B_3. For $b \in 1, \dots, B$ compute $W_k^{2,*,b}(n)$, from $\boldsymbol{X^{*,b}}$.

B_4. The distribution of $W_k^2(n)$ is approximated by $\{W_k^{2,*,1}(n), \dots, W_k^{2,*,B}(n)\}$ i.e., the final P-value is given by

$$P = \frac{1}{B} \sum_{b=1}^{B} I\{W_k^{2,*,b}(n) > W_k^2(n)\}$$

The main difference between this algorithm and the *classical* bootstrap is that, in the proposed method, the null hypothesis (and only the null hyporthesis) is used in order to compute the statistic (bootstrap) values instead of to be used to draw the bootstrap samples. We do not resampling from the null and this fact, allows to preserve the original data structure.

Permutation method is based on the idea that within the same subject, each value can be located in any position. For this claim, not only the null must be true but the interchangeability it also must be hold. Although, in practice, the permutation method has proved its robustness for a wide variety of statistics, let us to go to an extreme. We consider a three-sample problem (sample size n) where the first and second variables are the same and the third one is independent from the other two. In this setting it is derived the equality

$$\begin{aligned} W_k^2(n) =& n \int \left[\frac{1}{9} \{\hat{F}_{n,1}(X_1, t) - \hat{F}_{n,3}(X_3, t)\}^2 \right] d\hat{F}_{n,\bullet}(\boldsymbol{X}, t) \\ &+ n \int \left[\frac{1}{9} \{\hat{F}_{n,1}(X_1, t) - \hat{F}_{n,3}(X_3, t)\}^2 \right] d\hat{F}_{n,\bullet}(\boldsymbol{X}, t) \\ &+ n \int \left[\frac{4}{9} \{\hat{F}_{n,3}(X_3, t) - \hat{F}_{n,1}(X_1, t)\}^2 \right] d\hat{F}_{n,\bullet}(\boldsymbol{X}, t) = S_{n,1} + S_{n,2} + S_{n,3}. \end{aligned}$$

It is obvious that the value of the difference between the $\hat{F}_{n,1}$ and $\hat{F}_{n,3}$ has not the same weigth in the three summands. However the permutation algorithm assumes that the summands have the same distribution, in particular the same expected value. Table 1 depicts the means of $S_{n,i}$ (labelled as $\bar{S}_{n,i}$) for $i \in 1, 2, 3$ in 2,000 Monte Carlo (MC) simulations and when the permutation (P) and the proposed bootstrap (B) are used. The observed rejection proportion ($\alpha = 0.05$) and the value of $W_k^2(n)$ are also included. The underlying distributions are uniforms on $[0, 1]$ and $n = 50$.

Although the $W_k^2(n)$ is, in general, well estimated by the permutation method (the mean is similar to the expected one), the total value is evenly distributed

TABLE 1: Means for the three summands and for $W_k^2(n)$ in the problem above described for the Monte Carlo approximation (MC) (2,000 iterations) and for the Bootstrap and Permutation methods. Observed rejection percentages ($\alpha = 0.05$) are also included.

	$\bar{S}_{n,1}$	$\bar{S}_{n,2}$	$\bar{S}_{n,3}$	$W_k^2(n)$	% Rejection
MC	0.0378	0.0378	0.1456	0.2213	5.0%
P	0.0753	0.0744	0.0749	0.2247	8.8%
B	0.0368	0.0368	0.1473	0.2210	5.4%

among the three summands. In spite of for the $W_k^2(n)$ the results are, in general, good (the observed rejection proportion is too big, but this is an extreme and bit realistic problem), permutation method does not reflect the data structure and this fact can drive to mistake when different weighting are considered for the involved summands or, for instance, in presence of different missing data frameworks. In summary, the correct performance of the permutation method can not be guaranteed in absence of the exchangeability hyptothesis.

3. Simulation Study

In order to investigate the practical behaviour of the proposed methodology, a Monte Carlo simulation study has been carried out. We estimate the statistical power ($\alpha = 0.05$) from 2,000 Monte Carlo replications for different problems. For the Cramér-von Mises test, asymptotic approximation, C_A (the P-value is approximated from 499 Monte Carlo replications following the approximation given in (4)), bootstrap approximation, C_B ($B = 499$ in algorithm B_1-B_4) and the permutation method, C_P (the P-value is also approximated from $T = 499$ replications) are considered. Although the number of random combinations is small to obtain a good estimation for a particular P-value, it is enough to obtain a good estimation for the statistical power. Note that, here, we are not interested in the result for each particular problem but in the final rejection proportion. The classical non-parametric Friedman (F_R) test is also included as the reference one. Let be $\boldsymbol{Z} = (Z_1, Z_2, Z_3)$ a three dimensional random vector from a $\mathcal{N}_3(\mathbf{0}, \boldsymbol{\Sigma})$ distribution where $\mathbf{0} = (0, 0, 0)$ and the components of the covariance matrix are $\sigma_{i,j} = 1$ if $i = j$ and $\sigma_{1,2} = \sigma_{1,3} = 1/4$ and $\sigma_{2,3} = b$ (cases $b = 1/4$ and $b = 3/4$ are considered) and let be N_i, $1 \leq i \leq 4$, independent random variables with standard normal distribution. A three dimensional random sample with size n, $\boldsymbol{X} = (X_1, X_2, X_3)$, is drawn from the following symmetrical models (MD):

0-I. $X_1 \equiv Z_1$, $X_2 \equiv Z_2$, $X_3 \equiv Z_3$ (Null hypothesis).
1-I. $X_1 \equiv Z_1$, $X_2 \equiv Z_2$, $X_3 \equiv (1-a) * Z_3 + a * 3Z_3$.
2-I. $X_1 \equiv Z_1$, $X_2 \equiv Z_2$, $X_3 \equiv (1-a) * Z_3 + a * (Z_3 + 1)$.
3-I. $X_1 \equiv Z_1$, $X_2 \equiv Z_2$, $X_3 \equiv (1-a) * Z_3 + a * (\sqrt{3} Z_3 + 1)$.

The following asymmetrical models are also considered:

0-II. $X_1 \equiv Z_1^2 + N_1^2 + N_2^2$, $X_2 \equiv Z_2^2 + N_1^2 + N_3^2$, $X_3 \equiv Z_3^2 + N_2^2 + N_3^2$ (Null hypothesis).
1-II. $X_1 \equiv Z_1^2 + N_1^2 + N_2^2$, $X_2 \equiv Z_2^2 + N_1^2 + N_3^2$, $X_3 \equiv (1-a) * (Z_3^2 + N_2^2 + N_3^2) + a * (Z_3 + 3)$.
2-II. $X_1 \equiv Z_1^2 + N_1^2 + N_2^2$, $X_2 \equiv Z_2^2 + N_1^2 + N_3^2$, $X_3 \equiv (1-a) * (Z_3^2 + N_2^2 + N_3^2) + a * Z_3^2$.

3-II. $X_1 \equiv Z_1^2 + N_1^2 + N_2^2$, $X_2 \equiv Z_2^2 + N_1^2 + N_3^2$, $X_3 \equiv (1-a) * (Z_3^2 + N_2^2 + N_3^2) + a * (Z_3^2 + \sum_{i=1}^4 N_i^2)$.

Where $a = 3/4$ and $M = (1-b) * X + b * Y$ denotes a mixture which takes values on X with probability $(1-b)$ and on Y otherwise. A graphical representation for the respective density functions is shown in Figure 1.

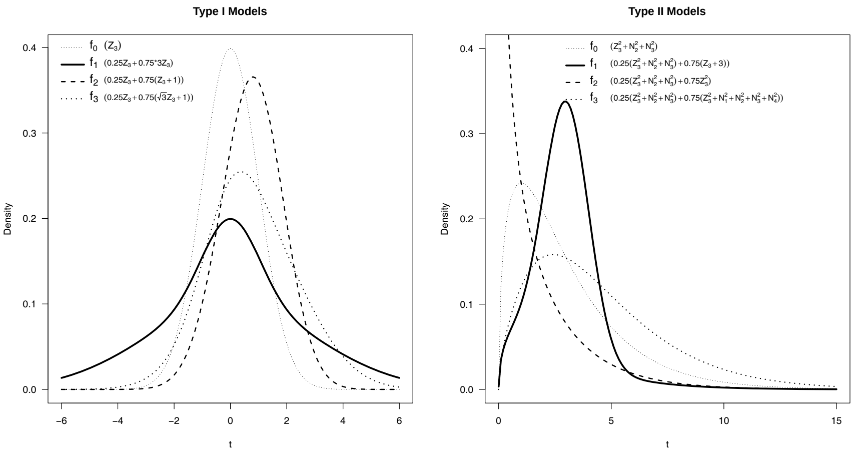


FIGURE 1: Density functions for the different considered models.

Table 2 shows the observed rejection proportions for type I (symmetrical) model for two different sample sizes ($n = 25, 50$). Figure 2 depicts the observed statistical power for the type I models against sample size (sample sizes of 10, 25, 40, 50, 65 and 75 were considered). Despite of the rejected observed percentages are bit larger than the expected ones (in special for the C_P approximation for $b = 3/4$) the nominal level is, in general, well respected. On the other hand, the Cramér-von Mises test obtains better results than the Friedman one even when the difference among the distributions is only in the position parameter while the variance-covariance structure is the same. Approximations by permutation and bootstrap obtain quite similar results, although the permutation one is a bit better for $\sigma_{2,3} = 3/4$. The asymptotic approximation obtains worst results for small sample sizes but quite similar than the other ones for middle sample sizes ($n > 40$).

Table 3 and Figure 3 are analogous to Table 2 and Figure 2, respectively, for type II (asymmetrical) models. The nominal level is well respected in all considered cases. For type II models (Figure 3) the Cramér-von Mises criterion is still the best when the main difference is not the location parameter (model 1-II). When the location parameter is the main difference among the curves, the Friedman test is the best one in model 3-II and the Cramér-von Mises test is the best for model 2-II, although both tests obtain quite similar results. In this scheme the approximation to the asymptotic distribution for the Cramér-von Mises test is slow and, in general, its results are not competitive for $n \leq 50$.

Table 2: Observed rejection probabilities for type I (symmetrical ones) models. The nominal level is $\alpha = 0.05$.

		$b = 1/4$				$b = 3/4$			
	n	C_B	C_P	C_A	F_R	C_B	C_P	C_A	F_R
0-I	**25**	0.059	0.060	0.037	0.051	0.065	0.074	0.042	0.068
	50	0.051	0.052	0.051	0.048	0.048	0.057	0.045	0.039
1-I	**25**	0.319	0.335	0.219	0.059	0.365	0.435	0.255	0.058
	50	0.713	0.733	0.699	0.050	0.804	0.856	0.778	0.049
2-I	**25**	0.793	0.780	0.772	0.711	0.875	0.889	0.815	0.913
	50	0.980	0.980	0.980	0.938	1.000	1.000	1.000	1.000
3-I	**25**	0.576	0.578	0.500	0.410	0.645	0.682	0.528	0.585
	50	0.874	0.873	0.874	0.688	0.985	0.990	0.980	0.877

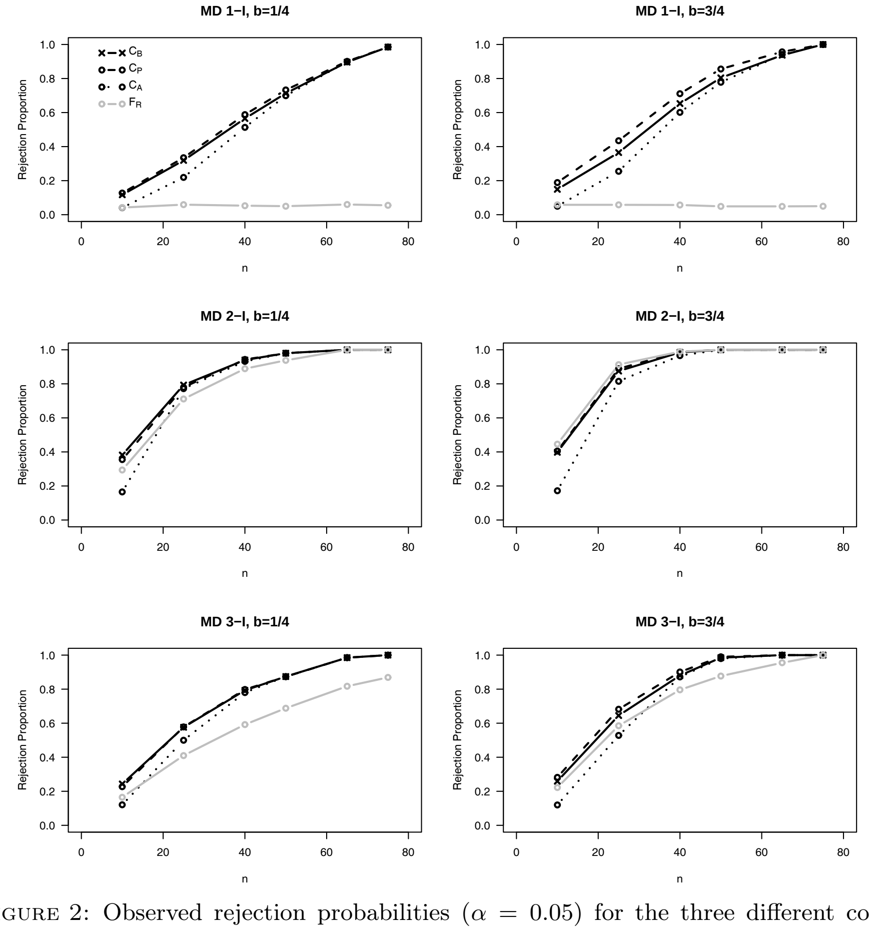


Figure 2: Observed rejection probabilities ($\alpha = 0.05$) for the three different considered approximations of the Crámer-von Mises statistic distribution and the Friedman test against sample size for the symmetrical (type I) models.

Table 3: Observed rejection probabilities for type II (asymmetrical ones) models. The nominal level $\alpha = 0.05$.

		$b = 1/4$				$b = 3/4$			
	n	C_B	C_P	C_A	F_R	C_B	C_P	C_A	F_R
0-II	**25**	0.052	0.058	0.030	0.058	0.056	0.062	0.032	0.054
	50	0.061	0.063	0.057	0.058	0.054	0.060	0.051	0.054
1-II	**25**	0.293	0.295	0.197	0.106	0.312	0.304	0.207	0.108
	50	0.632	0.628	0.603	0.163	0.619	0.625	0.592	0.159
2-II	**25**	0.902	0.904	0.872	0.772	1.000	1.000	1.000	0.999
	50	0.932	0.931	0.906	0.845	1.000	1.000	1.000	1.000
3-II	**25**	0.726	0.738	0.624	0.847	0.737	0.752	0.632	0.875
	50	0.985	0.985	0.979	1.000	1.000	1.000	0.995	1.000

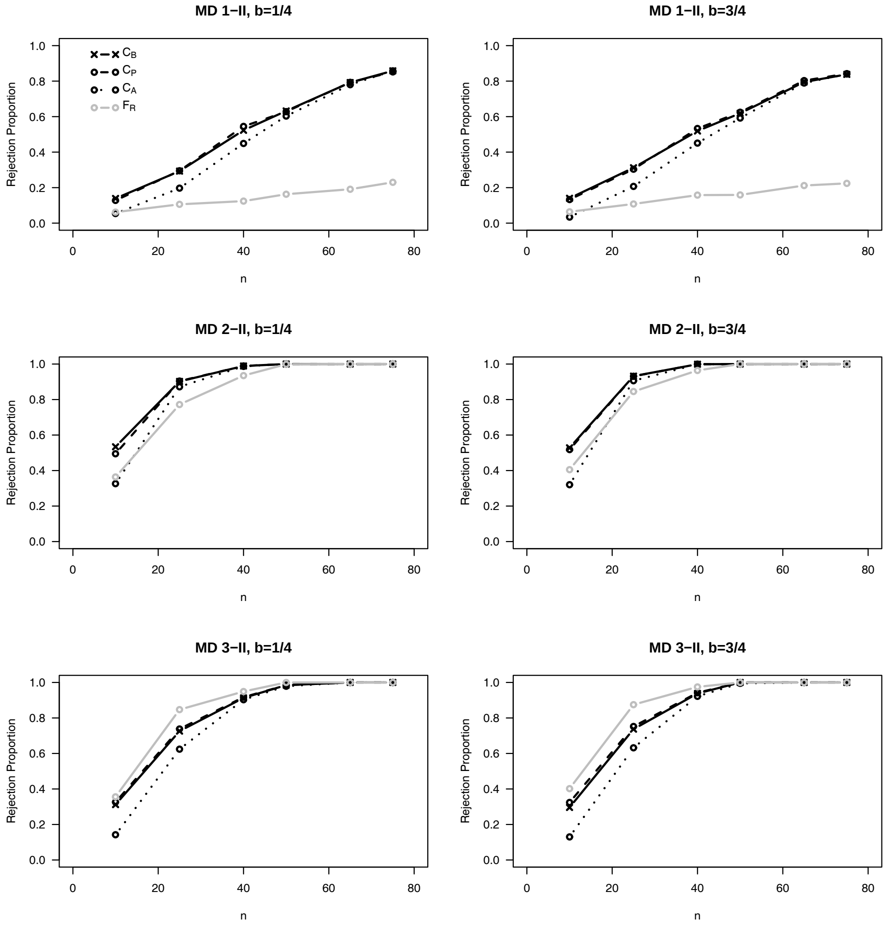


Figure 3: Observed rejection probabilities ($\alpha = 0.05$) for the three different considered approximations of the Crámer-von Mises statistic distribution and the Friedman test against sample size for the asymmetrical (type II) models.

4. Inequality Incomes Analysis

In order to illustrate the practical performance of the proposed method we considered the study of the inequality incomes between thirty European countries. The inequality measure is a complex problem which has been addressed from different approaches (see Cowel (2009) and references therein). Although the Gini index is, probably, the most popular measure of inequality, other approaches have also been considered (see, for instance, Martínez-Camblor 2007). Our objective is to study the (possible) changes in the income distribution inequalities in Europe. With this goal the GDP per capita in PPS (quote from the site `http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&plugin=1&language=en&pcode=tec00114`: *Gross Domestic Product (GDP) is a measure for the economic activity. It is defined as the value of all goods and services produced less the value of any goods or services used in their creation. The volume index of GDP per capita in Purchasing Power Standards (PPS) is expressed in relation to the European Union (EU-27) average set to equal 100. If the index of a country is higher than 100, this country's level of GDP per head is higher than the EU average and vice versa. Basic figures are expressed in PPS, i.e. a common currency that eliminates the differences in price levels between countries allowing meaningful volume comparisons of GDP between countries. Please note that the index, calculated from PPS figures and expressed with respect to EU27 = 100, is intended for cross-country comparisons rather than for temporal comparisons.*) in thirty European countries in the years 2000 and 2011 have been collected (downloaded from the above website). Due to our objective is not to study the incomes distribution but the inequalities of these incomes, we have considered the relative GDP per capita in PPS distribution i.e., the considered variable are 100 times the original values divided by the European Union one (considering the currently twenty-seven countries members) and the particular mean has been sustracted. Figure 4 depicts the empirical cummulative distribution function (ECDF) and the density estimation function for the considered GDP transformations.

The value of the Cramér-von Mises statistics between these two distributions was 0.171. The approximate P-values were 0.012, 0.005 and 0.001 from the asymptotic, bootstrap and permutation algorithms (based on 10,000 replications), respectively. All of them reject the null and it can be concluded the inequality of the incomes does not be equal in 2000 and 2011. The Gini indices were 0.251 and 0.220, respectively.

5. Main Conclusions

The Cramér-von Mises criterion is widely used in order to compare cumulative distribution functions. Despite of different situations have been considered, independent k-sample comparison is the most studied problem. We propose the use of this criterion in a typical k-related sample design. By using the Donsker invariance principle and the Karhunen-Loève decomposition for stochastic Gaussian processes its asymptotic distribution is developed. Although its explicit

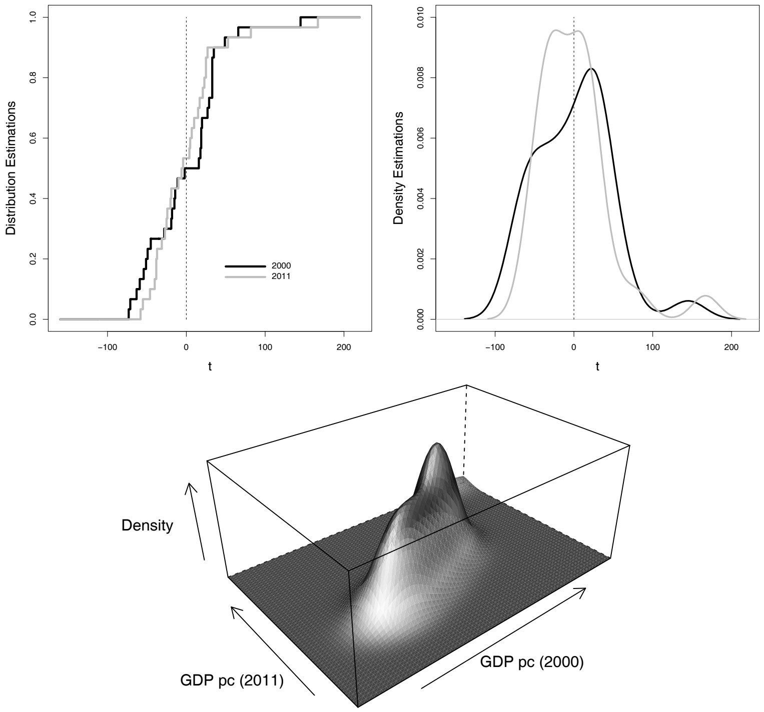


FIGURE 4: Upper, distribution (left) and density (right) estimation functions for the relative GDP per capita in PPS in thirty European countries in the years 2000 (black) and 2011 (gray). Lower, bivariate density estimation for the GDP per capita in PPS in years 2000 and 2011.

asymptotic distribution is still unknown, the obtained results allow to develop an useful approximation. As usual, we also explore two different resampling approximations: the classical and well-known permutation test and the most recent general bootstrap algorithm (gBA).

For independent samples, the Cramér-von Mises statistic is underlying distribution-free, its distribution does not depend on the distribution function where the samples were drawn, and it can be tabulated in order to obtain the P-value for a particular problem. In a paired sample design, the statistic distribution depends on the relationships among the involved variables; this relationship always must be estimated from the sample (therefore, universal eigenvalues do not exist for this topic), increasing the necessary time to compute the given asymptotic approximation. This is the main handicap for using the asymptotic approximation which, in general, obtains good results for moderate sample sizes.

A general bootstrap algorithm (gBA) and the usual permutation method are also studied. The considered bootstrap procedure exploits a particular pivotal function and introduce the null hypothesis at the moment of computing the value of the (bootstrap) statistic instead of in the random bootstrap samples generation process. The main advantage is that the data structure is preserved and no additional assumptions (only the null) are required. Some details of its consistency are also reported, however reader is referred to Martínez-Camblor et al. (2012) for more details. This technique has already been used with success in a paired sample extension of the $\mathcal{AC}$-statistic (Martínez-Camblor 2010) and in inference on a particular ecological diversity index (Martínez-Camblor, Corral & Vicente 2011). In an extreme example is showed how the permutation method can lead to mistakes when the interchangeability assumption is violated, which is the usual situation when $k > 2$. However, the observed statistical power in our simulation study is similar for the three different considered methods. We must remark that the asymptotic and the bootstrap method avoid the exchangeability assumption (Von Mises 1991, Nelsen 2007) and do not increase the methodology complexity.

As in the independent case (see, for example Martínez-Camblor & Uña-Álvarez 2009), the simulation study results suggest that the Cramér-von Mises criterion is clearly better than the (classical) Friedman test when the main difference among the curves is not the location parameter and it obtains very good results otherwise. On the other hand, the proposed asymptotic approximation obtains similar statistical power than the ones based on resampling for moderate sample sizes. Relevant differences between two considered variance-covariance matrix structures ($b = 1/4$ and $b = 3/4$) have not been observed.

We think that the considered practical case is specially good in order to illustrate the use of the proposed methodology. When the focus is not the location parameter but the shape, which is the case of the inequality, the Cramér-von Mises statistics conventionally obtains good powers in order to check the equality of the involved distribution functions. In this context, traditional text like Friedman or the Student T-test do not work but the Cramér-von Mises criterion has proved that is a poweful test and a valuable tool for this kind of goals.

Part of the results provided in this manuscript are overlapped with the ones obtained in Quessy & Éthier (2012). However, the present work has been developed independent and previously to the publication of the Quessy and Éthier one. The main differences between the works are:

(a) Our approach is more practical and, from our point of view more easy to understand for non probabilistic readers.

(b) The permutation method is considered and discussed. A pathological case where this method fails has been provided.

(c) A practical use of the gBA and a simulation study where the quality of the provided approximations can be checked.

(d) The considered practical problem illustrates a situation where the equality among the location parameters is not the hypothesis to be tested.

Acknowledgements

As usual, we are grateful with Susana Díaz Coto for the final manuscript revision. First author was supported by the research Grant MTM2011-23204 of the Spanish Ministerio de Ciencia e Innovación (FEDER support included).

[Recibido: septiembre de 2013 — Aceptado: diciembre de 2013]

References

Adler, R. J. (1990), An introduction to continuity, extrema and related topics for general gaussian processes, *in* 'IMS Lecture Notes-Monograph Series', Vol. 12, Institute of Mathematical Statistics, Hayward, California.

Alkarni, S. H. & Siddiqui, M. M. (2001), 'An upper bound for the distribution function of a positive definite quadratic form', *Journal of Statistical Computation and Simulation* **69**(1), 51–56.

Anderson, T. W. (1962), 'On the distribution of the two-sample cramér-von mises criterion', *Annals of Mathematical Statistics* **33**(3), 1148–1159.

Arcones, M. A. & Gine, E. (1992), 'On the bootstrap of u and v statistics', *Annals of Statistics* **20**(2), 655–674.

Beran, R. (1982), 'Estimated sampling distributions: The bootstrap and competitors', *Annals of Statistics* **10**, 212–225.

Ciba-Geigy, L. S. & Olsson, B. (1982), 'A nearly distribution-free test for comparing dispersion in paired samples', *Biometrika* **69**(2), 484–485.

Cowel, F. A. (2009), Measuring inequality. Accesed 16/04/2013.
*http://darp.lse.ac.uk/papersdb/cowell_measuringinequality3.pdf

Cramér, H. (1928), 'On the composition of elementary errors: II statistical applications', *Skandinavisk Aktuarietidskrift* **11**, 141–180.

Csörgo, S. & Faraway, J. J. (1996), 'The exact and asymptotic distributions of Cramér-von Mises statistics', *Journals of the Royal Statistical Society B* **58**(1), 221–234.

Deheuvels, P. (2005), 'Weighted multivariate Cramér-von Mises-type statistics', *Afrika Statistika* **1**(1), 1–14.

Efron, B. (1979), 'Bootstrap methods: Another look at the jackknife', *Annals of Statistics* **7**, 1–26.

Efron, E. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics.
*http://epubs.siam.org/doi/abs/10.1137/1.9781611970319

Freitag, G., Czado, C. & Munk, A. (2007), 'A nonparametric test for similarity of marginals–with applications to the assessment of population bioequivalence', *Journal of Statistical Planning & Inference* **137**(3), 697–711.

Good, P. (2000), *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, Springer Verlag, New York.

Govindarajulu, Z. (1995), 'A class of asymptotically distribution free test procedures for equality of marginals under multivariate dependence', *American Journal of Mathematical and Management Sciences* **15**, 375–394.

Govindarajulu, Z. (1997), 'A class of asymptotically distribution free tests for equality of marginals in multivariate populations', *Mathematical Methods of Statistics* **6**, 92–111.

Hall, P. & Wilson, S. R. (1991), 'Two guidelines for bootstrap hypothesis testing', *Biometrics* **47**, 757–762.

Horváth, L. & Steinebach, J. (1999), 'On the best approximation for bootstrapped empirical processes', *Statistical & Probability Letters* **41**, 117–122.

Kiefer, J. (1959), 'k-Sample analogues of the Kolmogorov-Smirnov, Cramér-von Mises tests', *Annals of Mathematical Statistis* **30**, 420–447.

Lam, F. C. & Longnecker, M. T. (1983), 'Modified Wilcoxon rank sum test for paired data', *Biometrika* **70**(2), 510–513.

Martin, M. A. (2007), 'Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties', *Computational Statistics & Data Analysis* **51**, 6321–6342.

Martínez-Camblor, P. (2007), 'Central limit theorems for S-Gini and Theil inequality coefficients', *Revista Colombiana de Estadística* **30**(2), 287–300.

Martínez-Camblor, P. (2010), 'Nonparametric k-sample test based on kernel density estimator for paired design', *Computational Statistics & Data Analysis* **54**, 2035–2045.

Martínez-Camblor, P. (2011), 'Testing the equality among distribution functions from independent and right censored samples via Cramér-von Mises criterion', *Journal of Applied Statistics* **38**(6), 1117–1131.

Martínez-Camblor, P., Carelos, C. & Corral, N. (2012), 'Sobre el estadístico de Cramér-von Mises', *Revista de Matemáticas: Teoría y Aplicaciones* **19**, 89–101.

Martínez-Camblor, P., Corral, N. & Vicente, D. (2011), 'Statistical comparison of the genetic sequence type diversity of invasive Neisseria meningitidis isolates in northern Spain (1997-2008)', *Ecological Informatics* **6**(6), 391–398.

Martínez-Camblor, P. & Uña-Álvarez, J. (2009), 'Non-parametric k-sample tests: density functions vs distribution functions', *Computational Statistics & Data Analysis* **53**(9), 3344–3357.

Munzel, U. (1999*a*), 'Linear rank score statistics when ties are present', *Statistics & Probability Letters* **41**, 389–395.

Munzel, U. (1999*b*), 'Nonparametric methods for paired samples', *Statistica Neerlandica* **53**(3), 277–286.

Nelsen, R. B. (2007), 'Extremes of non-exchangeability', *Statistical Papers* **48**, 329–336.

Podgor, M. J. & Gastwirth, J. L. (1996), Efficiency robust rank tests for stratified data, *in* E. Brunner & M. Denker, eds, 'Research Developments in Probability and Statistics', Festschrift in honor of Madan L. Puri. VSP International Science Publishers, Leiden, Netherlands.

Quessy, J. F. & Éthier, F. (2012), 'Cramér-von Mises and characteristic function tests for the two and k-sample problems with dependent data', *Computational Statistics and Data Analysis* **56**, 2097–2111.

Van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Venkatraman, E. S. & Begg, C. B. (1996), 'A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment', *Biometrika* **83**(4), 835–848.

Von Mises, R. E. (1991), *Wahrscheinlichkeitsrechnung*, Deuticke, Vienna.

Westfall, P. H. & Young, S. S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*, Wiley, New York.

Appendix

In this appendix we provide proofs for the two previously enunciated theorems. In particular, Theorem 1 is based on the well-known Donsker invariance principle and on classical Gaussian processes theory. In particular, we used the Karhunen-Loéve decomposition in order to guarantee the existence of the necessary variables and coefficients. These coefficients values (eigenvalues) are not explicitly computed (the respective eigenfunctions are neither computed). These calculus are, in general, cumbersome and complex depending, in the present case, on the data covariance structure (therefore, they are different for each particular problem, universal coefficients do not exist). The following auxiliar result is, directly, derived from the Donsker invariance principle:

Lemma 1. *Let $\boldsymbol{\xi}$ be a k-dimensional random vector and let $\boldsymbol{X}$ be a random sample from $\boldsymbol{\xi}$ (with size n), by using the above notation, it is had the following weak convergence*

$$\sqrt{n}\{\hat{\boldsymbol{F}}_{\boldsymbol{n}}(\boldsymbol{X},\boldsymbol{u}) - \boldsymbol{F}(\boldsymbol{u})\} \xrightarrow{\mathcal{L}}_n \boldsymbol{Y}_{\boldsymbol{F}}(\boldsymbol{u})$$

where $\boldsymbol{Y}_{\boldsymbol{F}}(\boldsymbol{u}) = (Y^1_{F_1}(u_1),\dots,Y^k_{F_k}(u_k))$ is a k-dimensional Gaussian process which follows a distribution $\mathcal{N}_k(0,\boldsymbol{\Sigma}(\boldsymbol{u}))$ where

$$\boldsymbol{\Sigma}(\boldsymbol{u}) = \begin{pmatrix} \sigma_{1,1}(u_1,u_1) & \sigma_{1,2}(u_1,u_2)\dots & \sigma_{1,k}(u_1,u_k) \\ \dots\dots\dots\dots\dots\dots\dots\dots & & \dots \\ \sigma_{k,1}(u_k,u_1) & \sigma_{k,2}(u_k,u_2)\dots & \sigma_{k,k}(u_k,u_k) \end{pmatrix} \tag{5}$$

and $\sigma_{i,j}(u_i,u_j) = F_{i,j}(u_i,u_j) - F_i(u_i)F_j(u_j)$ and $F_{i,j}(u_i,u_j) = \mathcal{P}\{\xi_\rangle \leq \sqcap_\rangle \cap \xi_| \leq \sqcap_|\}$ for $1 \leq i,j \leq k$.

Furthermore, for $u,v \in \mathbb{R}$ and $1 \leq i,j \leq k$, $\mathbb{E}[Y^i_{F_i}(u)Y^j_{F_j}(v)] = \sigma_{i,j}(u,v) = F_{i,j}(u,v) - F_i(u)F_j(v)$ where if $(u \wedge v) = \min\{u,v\}$,

$$F_{i,j}(u,v) = \begin{cases} \mathcal{P}\{\xi_\rangle \leq \sqcap \cap \xi_| \leq \sqsubseteq\} & \text{if } i \neq j \\ \mathcal{P}\{\xi_\rangle \leq (\sqcap \wedge \sqsubseteq)\} & \text{if } i = j \end{cases}$$

Under the null given in (1) and if for each $1 \leq i,j \leq k$ and for $u,v \in \mathbb{R}$ we define the functions: $\bar{F}_{\cdot,i}(u) = k^{-1}\sum_{j=1}^k F_{j,i}(u,u)$, $\bar{F}_{\cdot,\cdot}(v) = k^{-1}\sum_{i=1}^k F_{\cdot,i}(v,v)$, $\bar{F}_{\cdot,i}(u,v) = k^{-1}\sum_{j=1}^k F_{j,i}(u,v)$, $\bar{F}_{\cdot,\cdot}(u,v) = k^{-1}\sum_{i=1}^k F_{\cdot,i}(u,v)$, $C_i(u) = F(u) - 2\bar{F}_{\cdot,i}(u) + \bar{F}_{\cdot,\cdot}(u)$ and $C_{i,j}(u,v) = F_{i,j}(u,v) - \bar{F}_{\cdot,i}(v,u) - \bar{F}_{\cdot,j}(u,v) + \bar{F}_{\cdot,\cdot}(u,v)$ we can obtain the following result,

Theorem 3. *By using the Lemma 1 notation, if $F_1 = \cdots = F_k \,(= F)$ (null hypothesis), it is held the (weak) convergence*

$$W_k^2(n) \xrightarrow{\mathcal{L}}_n \sum_{i=1}^k \sum_{l\in\mathbb{N}} \lambda_{i,l} M_{i,l}^2,$$

where $\{\boldsymbol{M}_l = (M_{1,l},\dots,M_{k,l})\}_{l\in\mathbb{N}}$ is a sequence of k-dimensional, normal distributed random variables whose marginals follow a $\mathcal{N}(0,1)$ distribution and $\{\{\lambda_{i,l}\}_{i=1}^k\}_{l\in\mathbb{N}}$ are non negative constants satisfying $\sum_{l\in\mathbb{N}}\lambda_{i,l}^2 < \infty$ for $1 \leq i \leq k$.

Proof. Keeping the Lemma 1 notation, if $\bar{Y}_{F,\bullet}(t) = k^{-1}\sum_{i=1}^k Y^i_{F_i}(t)$ and taking into account the well-known convergence $\sup_{t\in\mathbb{R}}\{\hat{F}_n(X,t) - F(t)\} \longrightarrow_n 0$ (a.s.). It is easy to see that

$$\begin{aligned} W_k^2(n) = \sum_{i=1}^k n \int \{\hat{F}_{n,i}(X_i,t) & \\ - \hat{F}_{n,\bullet}(\boldsymbol{X},t)\}^2 d\hat{F}_{n,\bullet}(\boldsymbol{X},t) & \xrightarrow{\mathcal{L}}_n \sum_{i=1}^k \int \{Y^i_{F_i}(t) - \bar{Y}_{F,\bullet}(t)\}^2 dF(t) \end{aligned} \tag{6}$$

or equivalently, $\forall u \in \mathbb{R}$,

$$\left\{\mathcal{P}\left(\mathcal{W}^{\in}_{\|}(\backslash) \leq \sqcap\right) - \mathcal{P}\left(\sum_{\rangle=\infty}^{\|} \int \{\mathcal{Y}^{\rangle}_{\mathcal{F}_{\rangle}}(\sqcup) - \bar{\mathcal{Y}}_{\mathcal{F},\bullet}(\sqcup)\}^{\in} \lceil \mathcal{F}(\sqcup) \leq \sqcap\right)\right\} \longrightarrow_N 0 \quad a.s.$$

On the other hand, if $\mathcal{X}_{(\rangle)}(\sqcup) = \{\mathcal{Y}^{\rangle}_{\mathcal{F}_{\rangle}}(\sqcup) - \bar{\mathcal{Y}}_{\mathcal{F},\bullet}(\sqcup)\}$ $(1 \leq i \leq k)$ then, under the null,

$$\boldsymbol{\mathcal{X}}(t) = (\mathcal{X}_{(\infty)}(\sqcup), \ldots, \mathcal{X}_{(\|)}(\sqcup))$$

is a centred k-dimensional *Gaussian* process. Moreover, if for $t \in \mathbb{R}$, $\boldsymbol{t} = (t, \ldots, t)$ and $\boldsymbol{\Sigma}(\boldsymbol{t})$ stands for the matrix defined in (2), for symmetry and under the null, for $i \in 1, \ldots, k$ it is obtained,

$$\begin{aligned}\mathbb{E}[\mathcal{X}_{(\rangle)}(\sqcup)^{\in}] =& \boldsymbol{a}_i \boldsymbol{\Sigma}(\boldsymbol{t}) \boldsymbol{a}_i^t \\ =& \sigma_{i,i}(t,t) - \bar{\sigma}_{\cdot,i}(t) - \bar{\sigma}_{i,\cdot}(t) + \bar{\sigma}_{\cdot,\cdot}(t) = F(t) - 2\bar{F}_{\cdot,i}(t) + \bar{F}_{\cdot,\cdot}(t) = C_i(t)\end{aligned}$$

where for $1 \leq i \leq k$, $\boldsymbol{a}_i = (-1/k, \ldots, \overset{(i)}{(k-1)}/k, \ldots, -1/k)$, $\bar{\sigma}_{\cdot,i}(t) = k^{-1}\sum_{j=1}^{k} \sigma_{j,i}(t,t)$, $\bar{\sigma}_{i,\cdot}(t) = k^{-1}\sum_{j=1}^{k} \sigma_{i,j}(t,t)$ and $\bar{\sigma}_{\cdot,\cdot}(t) = k^{-1}\sum_{i=1}^{k} \bar{\sigma}_{i,\cdot}(t,t)$. In addition, it is had the covariance

$$\begin{aligned}\mathbb{C}_{i,j}(s,t) =& \mathbb{E}[\mathcal{X}_{(\rangle)}(\int)\, \mathcal{X}_{(|)}(\sqcup)] \\ =& F_{i,j}(s,t) - \bar{F}_{\cdot,i}(s,t) - \bar{F}_{\cdot,j}(s,t) + \bar{F}_{\cdot,\cdot}(s,t) = C_{i,j}(s,t) \quad (1 \leq i,j \leq k)\end{aligned}$$

and it is easy to check that, for $i \in 1, \ldots, k$,

$$\iint \mathbb{C}_{i,i}(s,t)^2 dF(s) dF(t) < \infty$$

This property allows to apply the Karhunen-Loève decomposition (see, for example, Adler 1990) in order to obtain the representation

$$\mathcal{X}_{(\rangle)}(\sqcup) = \sum_{\updownarrow \in \mathbb{N}} \sqrt{\lambda_{\rangle,\updownarrow}}\, \rceil_{\rangle,\updownarrow}(\sqcup)\, \mathcal{M}_{\rangle,\updownarrow} \qquad \text{(for each } i \in 1, \ldots, k) \tag{7}$$

where (for each $i \in 1, \ldots, k$) $\{e_{i,l}(t)\}_{l \in \mathbb{N}}$ is a *convergent orthonormal sequence* (also known as *eigenfunctions*) i.e.,

$$\int e_{i,p}(t) e_{i,q}(t) dF(t) = \begin{cases} 0 & \text{if } p \neq q \\ 1 & \text{if } p = q \end{cases}$$

$\{\boldsymbol{M}_l = (M_{1,l}, \ldots, M_{k,l})\}_{l \in \mathbb{N}}$ are k-dimensional random vectors which marginal distributions follow a $\mathcal{N}(0,1)$ and, for $1 \leq i \leq k$ $\{\{\lambda_{i,j}\}_{i=1}^{k}\}_{j \in \mathbb{N}}$, are non negative constants (also known as *eigenvalues*) satisfying

$$\lambda_{i,1} \geq \cdots \geq \lambda_{i,l} \geq \cdots \geq 0 \qquad \forall i \in 1, \ldots, k.$$

From (7), it is straightforward that, for $i \in 1, \dots, k$,

$$\begin{aligned}
\int \mathcal{X}_{(\rangle)}(\sqcup)^{\in}\lceil\mathcal{F}(\sqcup) &= \int \left(\sum_{l\in\mathbb{N}} \sqrt{\lambda_{i,l}}\, e_{i,l}(t)\, M_{i,l} \right)^2 dF(t) \\
&= \int \sum_{l\in\mathbb{N}} \sum_{j\in\mathbb{N}} \sqrt{\lambda_{i,l}}\, \sqrt{\lambda_{i,j}}\, e_{i,l}(t) e_{i,j}(t) M_{i,l} M_{i,j} dF(t) \\
&= \sum_{l\in\mathbb{N}} \sum_{j\in\mathbb{N}} \sqrt{\lambda_{i,l}}\, \sqrt{\lambda_{i,j}} M_{i,l} M_{i,j} \int e_{i,l}(t) e_{i,j}(t) dF(t) \qquad (8) \\
&= \sum_{l\in\mathbb{N}} \lambda_{i,l} M_{i,l}^2 \qquad (9)
\end{aligned}$$

Therefore, from the Fubini Theorem, for $i \in 1, \dots, k$,

$$\begin{aligned}
\sum_{l\in\mathbb{N}} \lambda_{i,l} = \mathbb{E}\left(\int \mathcal{X}_{(\rangle)}(\sqcup)^{\in}\lceil\mathcal{F}(\sqcup) \right) &= \int \mathbb{E}\left(\mathcal{X}_{(\rangle)}(\sqcup)^{\in} \right) dF(t) \\
&= \int C_i(t) dF(t) = C_i
\end{aligned} \qquad (10)$$

and

$$\sum_{l\in\mathbb{N}} \lambda_{i,l}^2 = \iint \mathbb{C}_{i,i}(s,t)^2 dF(s) dF(t) < \infty \qquad (11)$$

Now, we are interested in studying the joint distribution of $\boldsymbol{M}_l$ (for each fixed $l \in \mathbb{N}$). We will prove that $\sum_{i=1}^k a_i M_{i,l}$ follows a normal distribution for each $a_1, \dots, a_k \in \mathbb{R}$ and for $l \in \mathbb{N}$. Note that, for each (fixed) $l \in \mathbb{N}$, we have that

$$\boldsymbol{\mathcal{X}}^*(t) = (a_1 \mathcal{X}_{(\infty)}(\sqcup)\, \rceil_{\infty,\updownarrow}(\sqcup), \cdots, \dashv_{\|} \mathcal{X}_{(\|)}(\sqcup)\, \rceil_{\|,\updownarrow}(\sqcup))$$

is a k-dimensional centred *Gaussian* process. From (7), for each $i \in 1, \dots, k$

$$a_i\, e_{i,l}(t) \sum_{j\in\mathbb{N}} \sqrt{\lambda_{i,l}}\, e_{i,j}(t)\, M_{i,j} = a_i \mathcal{X}_{(\rangle)}(\sqcup)\, \rceil_{\rangle,\updownarrow}(\sqcup)$$

hence,

$$a_i \sum_{j\in\mathbb{N}} \sqrt{\lambda_{i,j}} M_{i,j} \int e_{i,l}(t)\, e_{i,j}(t) dF(t) = a_i \int \mathcal{X}_{(\rangle)}(\sqcup)\, \rceil_{\rangle,\updownarrow}(\sqcup) \lceil\mathcal{F}(\sqcup)$$

We can assume that $\lambda_{i,l} \neq 0$ for $1 \leq i \leq k$ (if some $\lambda_{i,l} = 0$ $(1 \leq i \leq k)$, $M_{i,l}$ does not interfere in any definition and we would have freedom to select it independently with the other ones) hence, from the eigenfunctions properties

$$\begin{aligned}
\sum_{i=1}^k a_i M_{i,l} &= \sum_{i=1}^k \frac{a_i}{\sqrt{\lambda_{i,l}}} \int \mathcal{X}_{(\rangle)}(\sqcup)\, \rceil_{\rangle,\updownarrow}(\sqcup) \lceil\mathcal{F}(\sqcup) \\
&= \int \sum_{i=1}^k \frac{a_i}{\sqrt{\lambda_{i,l}}} \mathcal{X}_{(\rangle)}(\sqcup)\, \rceil_{\rangle,\updownarrow}(\sqcup) \lceil\mathcal{F}(\sqcup)
\end{aligned}$$

follows a normal distribution.

From (6), (7) and (8) we know that there exists a sequence of k-dimensional normal distributed random variables, $\{\boldsymbol{M}_j\}_{j\in\mathbb{N}}$ whose marginals follow a $\mathcal{N}(0,1)$ distribution and non negative constants $\{\{\lambda_{i,j}\}_{i=1}^{k}\}_{j\in\mathbb{N}}$ satisfying (9) and (10), such that

$$W_k^2(n) \xrightarrow{\mathcal{L}}_n \sum_{i=1}^{k}\sum_{j\in\mathbb{N}} \lambda_{i,j} M_{i,j}^2$$

and the proof is concluded. □

Theorem 4. *Under the Lemma 1 assumptions and by using the same notation. Let $\boldsymbol{X}^* = (X_1^*, \dots, X_k^*)$ be an independent random sample generated from $\hat{\boldsymbol{F}}_{\boldsymbol{n}}(\boldsymbol{X}, \cdot)$ (multivariate ECDF referred to the random sample $\boldsymbol{X}$). If*

$$\begin{aligned} W_k^{2,*}(n) = \sum_{i=1}^{k} n \int \{\hat{F}_{n,i}^*(X_i^*, t) - \hat{F}_{n,i}(X_i, t)\}^2 d\hat{F}_{n,\bullet}^*(\boldsymbol{X}^*, t) \\ - nk \int \{\hat{F}_{n,\bullet}^*(\boldsymbol{X}^*, t) - \hat{F}_{n,\bullet}(\boldsymbol{X}, t)\}^2 d\hat{F}_{n,\bullet}^*(\boldsymbol{X}^*, t) \end{aligned}$$

where for each $i \in 1, \dots, k$, $\hat{F}_{n,i}^*(X_i^*, t)$ is the ECDF referred to X_i^* and $\hat{F}_{n,\bullet}^*(\boldsymbol{X}^*, t) = k^{-1}\sum_{i=1}^{k} \hat{F}_{n,i}^*(X_i^*, t)$. Under the null, it is derived,

$$\left\{\mathcal{P}_{\boldsymbol{X}}\left(\mathcal{W}_{\|}^{\in,*}(\backslash) \leq \sqcap\right) - \mathcal{P}\left(\mathcal{W}_{\|}^{\in}(\backslash) \leq \sqcap\right)\right\} \longrightarrow_n 0 \qquad a.s.$$

where $\mathcal{P}_{\boldsymbol{X}}$ denotes probability conditionally on sample $\boldsymbol{X}$.

Proof. It is easy to check that,

$$\begin{aligned} W_k^2(n) =& \sum_{i=1}^{k} n \int \{\hat{F}_{n,i}(X_i, t) - \hat{F}_{n,\bullet}(\boldsymbol{X}, t)\}^2 d\hat{F}_{n,\bullet}(\boldsymbol{X}, t) \\ =& \sum_{i=1}^{k} n \int \{\hat{F}_{n,i}(X_i, t) - F_i(t)\}^2 d\hat{F}_{n,\bullet}(\boldsymbol{X}, t) \\ & -nk \int \{\hat{F}_{n,\bullet}(\boldsymbol{X}, t) - F(t)\}^2 d\hat{F}_{n,\bullet}(\boldsymbol{X}, t). \end{aligned}$$

And, directly from the Lemma 1,

$$W_k^2(n) \xrightarrow{\mathcal{L}}_n \sum_{i=1}^{k} \int Y_{F_i}^i(t)^2 dF(t) - k \int \bar{Y}_{F,\bullet}(t)^2 dF(t)$$

Of course, the above equation is equivalent to the one in (6). Also from the Lemma 1 and classical theory of stochastic processes (in particular, Horváth & Steinebach (1999) proved that the expressions $sup_{t\in\mathbb{R}}|\hat{F}_n(X,t) - F(t)|$ and $sup_{t\in\mathbb{R}}|\hat{F}_n^*(X^*,t) - \hat{F}_n(X,t)|$ where X^* is an independent random sample with size n generated from

$\hat{F}_n(X, \cdot)$ (ECDF referred to the random sample X which sample size is n) have the same asymptotic distribution), for each $u \in \mathbb{R}$, it is had the convergence,

$$\left\{ \mathcal{P}_{\boldsymbol{X}} \left(\mathcal{W}_{\|}^{\in,*}(\backslash) \leq \sqcap \right) - \mathcal{P} \left(\sum_{\rangle=\infty}^{\|} \mathcal{I}_{\backslash} \left(\mathcal{Y}^{\rangle}_{\hat{\mathcal{F}}_{\backslash,\rangle}}, \mathcal{Y}_{\hat{\mathcal{F}}_{\backslash,\bullet}}(\sqcup) \right) \leq \sqcap \right) \right\} \longrightarrow_n 0 \qquad a.s.$$

where $Y_{\hat{F}_{n,i}}$ $(1 \leq i \leq k)$ and $\bar{Y}_{\hat{F}_{n,\bullet}}$ are the processes which appear in the Lemma 2.1 and, for $i \in 1, \ldots, k$,

$$I_n \left(Y^i_{\hat{F}_{n,i}}, Y_{\hat{F}_{n,\bullet}}(t) \right) = \int Y^i_{\hat{F}_{n,i}}(t)^2 d\hat{F}_{n,\bullet}(\boldsymbol{X}, t) - k \int \bar{Y}_{\hat{F}_{n,\bullet}}(t)^2 d\hat{F}_{n,\bullet}(\boldsymbol{X}, t)$$

Due, under the null hypothesis, $\forall t \in \mathbb{R}$, it is had the convergence $\hat{\boldsymbol{F}}_{\boldsymbol{n}}(\boldsymbol{X}, t) \longrightarrow_n \boldsymbol{F}(t)$ (a.s.), for each $u \in \mathbb{R}$, it is directly derived that

$$\left\{ \mathcal{P}_{\boldsymbol{X}} \left(\mathcal{W}_{\|}^{\in,*}(\backslash) \leq \sqcap \right) - \mathcal{P} \left(\mathcal{W}_{\|}^{\in}(\backslash) \leq \sqcap \right) \right\} \longrightarrow_n 0 \qquad a.s.$$

□

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 69 a 78

A New Method for Detecting Significant p-values with Applications to Genetic Data

Una nuevo método para la detección de valores p significativos y su aplicación a datos genéticos

Jorge Iván Vélez[1,2,3,a], Juan Carlos Correa[3,4,b], Mauricio Arcos-Burgos[1,2,c]

[1]Genomics and Predicitive Medicine Group, Genome Biology Department, John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia

[2]Group of Neurosciences, University of Antioquia, Medellín, Colombia

[3]Research Group in Statistics, National University of Colombia, Medellín, Colombia

[4]Department of Statistics, National University of Colombia, Medellín, Colombia

Abstract

A new method for detecting significant p-values is described in this paper. This method, based on the distribution of the m-th order statistic of a $U(0,1)$ distribution, is shown to be suitable in applications where $m \to \infty$ independent hypothesis are tested and it is of interest for a fixed type I error probability to determine those being significant while controlling the false positives. Equivalencies and comparisons between our method and others methods based-on p-values are also established, and a graphical representation of the distribution of the test statistic is depicted for different values of m. Finally, our proposal is illustrated with two microarray data sets.

***Key words*:** Extreme values theory, p-value, Type I error probability, Multiple testing, Genetic data.

Resumen

Se describe una nuevo método para la detección de valores p significativos. Este método, basado en el m-ésimo estadístico de orden de la distribución $U(0,1)$, es adecuado en casos en los que se realizan $m \to \infty$ pruebas de hipótesis independientes y es de interés determinar aquellas que son significativas, controlando los falsos positivos, para una probabilidad de error tipo I predeterminada. Adicionalmente, se realiza una comparación con algunas

[a]Ph.D Scholar. E-mail: jorge.velez@anu.edu.au
[b]Associate professor. E-mail: jccorrea@unal.edu.co
[c]Associate professor. E-mail: mauricio.arcos-burgos@anu.edu.au

pruebas clásicas y se grafica la distribución del estadístico de prueba para diferentes valores de m. Finalmente se ilustra el uso de la metodología con dos conjuntos de datos provenientes de estudios con microarreglos.

***Palabras clave*:** teoría de valores extremos, valor-p, probabilidad de error tipo I, comparaciones múltiples, datos genéticos.

1. Introduction

Genome-wide association studies (GWAS) are aimed at identifying genetics variants associated with a trait (Manolio 2010). For this, hundred of thousands participants with and without a particular disease (or trait) are required, and hundred of thousand of genetic variants, i.e., single nucleotide polymorphisms (SNPs), are read using SNPs arrays. Associated variants are further determined after performing (not necessarily) independent statistical tests comparing either the allele frequency or the distribution of the genotypes of these SNPs between cases and controls. Further, the correspondent p-value for each SNP is used to determine whether it is associated with the disease.

As a total of $m \to \infty$ independent SNPs are being tested in a typical GWAS, the problem of determining which variants are associated with the specific trait can be reduced to a multiple testing problem (for a review see Shaffer 1995) and so the family-wise error rate (FWER), i.e., the probability that one or more of the significance tests results in a type I error, must be controlled at level α. For such purpose, several methods can be applied (Bonferroni 1935, Shaffer 1995, Benjamini & Hochberg 1995, Nyholt 2004, Liu et al. 2010). In general terms, these methods use the p-values for each SNP and compare with a (adaptative) threshold, such that the SNPs associated with the trait are those for which the p-value is grater (or lower) than that threshold.

Here we describe a new method to detect p-values while controlling the FWER at level α. This method is heavily based on extreme values theory and considers the distribution of m-th order statistic of a $U(0, 1)$. We derive the test statistic, show its equivalency with Bonferroni's method, and provide asymptotic results for its limiting distribution. In addition, we report preliminary results of a simulation study in which, under the null hypothesis, i.e., $p \sim U(0, 1)$, the limiting distribution and the simulated values are depicted for different values of m. Finally, we apply our method to two well-known microarray data sets (Golub et al. 1999, Mootha et al. 2003).

2. Describing the Method

2.1. Background

Suppose that $m \to \infty$ independent hypotheses of the form

$$H_{0,i} : \theta_i \in \Theta \quad \text{vs.} \quad H_{1,i} : \theta_i \notin \Theta \qquad i = 1, 2, \dots, m \tag{1}$$

are tested, with θ_i some parameter of interest and Θ the parameter space. Let $\alpha \in (0,1)$ be the type I error probability at which the ith hypothesis is tested and

$$P_i = 1 - G(T_i) \qquad i = 1, 2, \dots, m \tag{2}$$

be its P-value. In (2), T_i is the test statistic for the ith hypothesis and G its cumulative distribution function (*cdf*). Under H_0, $P_1, P_2, \dots, P_m$ is a random sample from a $U(0,1)$ (Sackrowitz & Samuel-Cahn 1999, Murdoch, Tsai & Adcock 2008).

Let V be a random variable with *cdf* F, and let $V_{(m)} = \max\{V_1, V_2, \dots, V_m\}$ be its maximum in a random sample of size m. The exact distribution of $V_{(m)}$ is given by Casella & Berger (2001):

$$P(V_{(m)} \leq t) = \{F(t)\}^m \tag{3}$$

Note that if F is not known, (3) cannot be calculated. However, Serfling (1980, pp. 89) presents an alternative using extreme values theory and asymptotic results. As in a GWAS $m \to \infty$ independent hypothesis are being tested, to build up our methodology on such results seems intuitive.

2.2. The Test

Consider the random variable

$$D_m = (V_{(m)} - a_m)/b_m \tag{4}$$

with $V_{(m)}$ as previously defined. For some choices of constants $\{a_m\}$ and $\{b_m\}$, the limiting distribution of D_m is known (Serfling 1980, pp. 89). It follows from the $U(0,1)$ null distribution of the p-values that $-\log(p)$ has a standard exponential distribution with parameter $\lambda = 1$, and choosing $a_m = \log(m)$ and $b_n = 1$ yields (Serfling 1980, pp. 90)

$$\begin{aligned} F_{D_m}(t) &= P(D_m \leq t) \\ &= P(V_{(m)} - \log(m) \leq t) \\ &\to e^{-e^{-t}}, \quad m \to \infty \end{aligned} \tag{5}$$

making possible the calculation of (3). It is straightforward to show that the limiting density function of D_m is given by

$$\begin{aligned} f_{D_m}(t) &= \frac{d}{dt} F_{D_m} \\ &\to \exp\{-(t + \exp(-t))\}, \quad m \to \infty \end{aligned} \tag{6}$$

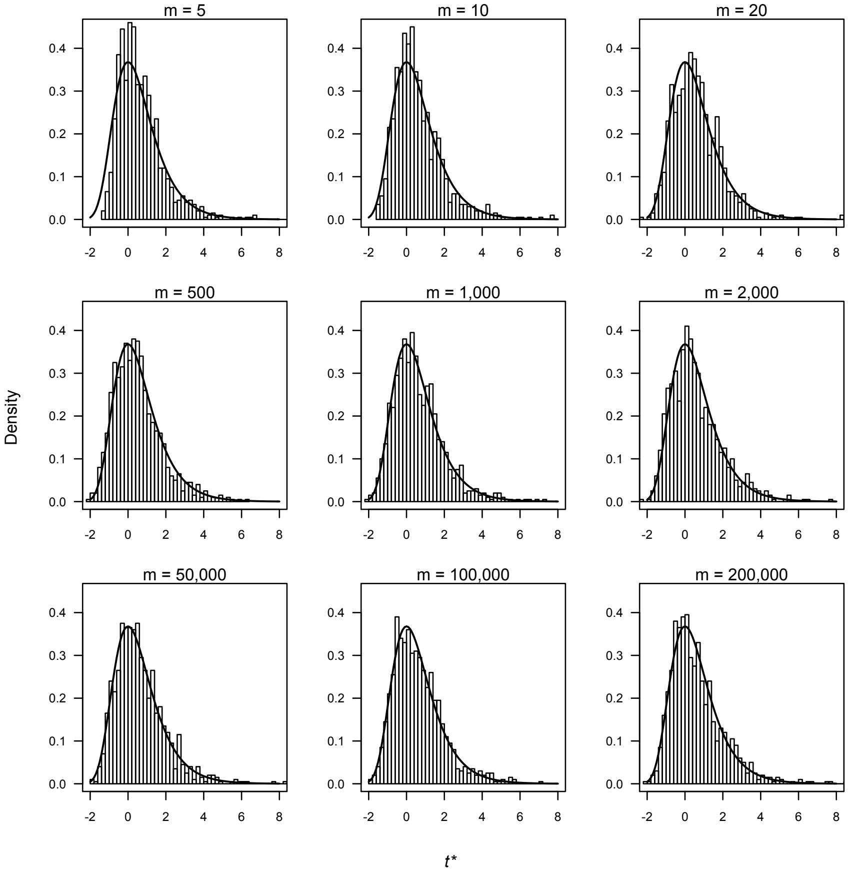


FIGURE 1: Simulation-based distribution of t^* for different values of m when the p-values come from a $U(0,1)$ and $\alpha = 0.05$. Here, the black line corresponds to $f_{D_m}(t)$ in (6).

We shall say that the ith p-value is significant at level α if

$$t_i^* > t_c \qquad i = 1, 2, \dots, m \tag{7}$$

where

$$t_i^* = -\log(-\log(1 - P_i)) \tag{8}$$

is the test statistic and t_c the critical value of the test at level α, e.g., t_c is such that

$$P(V_{(m)} - \log(m) \geq t_c) = \alpha \tag{9}$$

Combining (5) and (9), and solving for t leads to

$$t_c = -\log(-\log(1 - \alpha)) \tag{10}$$

In Figure 1 we depict the simulation-based distribution of t^* when $P_1, P_2, \dots, P_m \overset{iid}{\sim} U(0,1)$ for different values of m.

It is also possible to establish some equivalencies between our proposed method and others. For instance, if the Bonferroni (1935) method is to be applied to control by multiple testing (Shaffer 1995), the critical value

$$t_c^* = t_c + \log(m) \tag{11}$$

should be used instead of (10). This result is particularly useful in situations where a stringent control of the FWER (and hence the false positives) is required.

2.3. Using the Test

The following steps are suggested for detecting those p-values being statistically significant:

1. For each p-value, calculate t_i^* as in (8) and denote them as $t_1^*, t_2^*, \dots, t_m^*$. Here, higher values of t^* indicate strong evidence against H_0 in (1).
2. Determine which $t_i^{*\prime}$s are greater than t_c (or t_c^*).
3. Define the p-values from step 2 as potential candidates.

In order to facilitate the use of our proposal, an implementation of the aforementioned steps in R (R Core Team 2013) is provided in 4. This function takes a vector of p-values as the main argument, calculates the test statistic and the critical value, and prints the number of rejected p-values as well as the rejection rate. Furthermore, an invisible object (a list) with three components is returned; this list contains the actual p-value, the test statistic and the correspondent decision (significant: `TRUE`; not significant: `FALSE`). If necessary, such an object can be used for further analyses.

3. Examples

In this section, we consider two gene expression data sets to illustrate the usefulness of our proposed method for the identification of significant p-values.

3.1. Tumor Data

Golub et al. (1999) present a generic approach to cancer classification based on gene expression monitoring by DNA microarrays. As a test case, the authors use gene expression data from 3,051 genes in 38 tumor mRNA samples from patients with leukemia; 27 samples come from patients with lymphoblastic leukemia (ALL)(cases) and 11 from patients with acute myeloid leukemia (AML)(controls). For analysis, the processed data was obtained from the `multtest` package (Pollard, Gilbert, Ge, Taylor & Dudoit 2011).

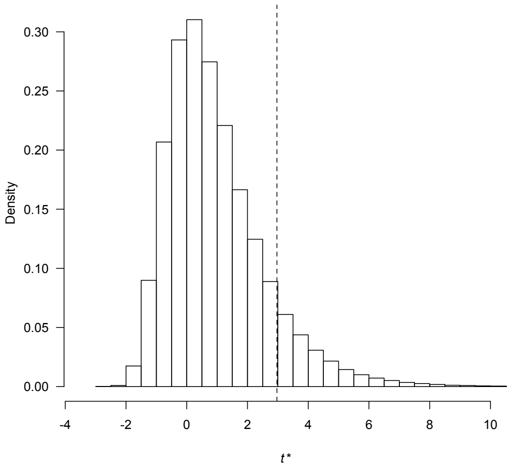


Figure 2: Distribution of t^* for the microarray data in Golub et al (1999). The vertical dotted line represents the critical value of the test for $\alpha = 0.05$ when no correction for multiple testing is applied.

We tested whether the ith gene ($i = 1, 2, \dots, m = 3,051$) was differentially expressed (DE), i.e., if there was any statistical difference between the expression levels in cases and controls. This is equivalent to test

$$H_{0,i} : \mu_{\text{ALL},i} = \mu_{\text{AML},i} \quad \text{vs.} \quad H_{1,i} : \mu_{\text{ALL},i} \neq \mu_{\text{AML},i} \tag{12}$$

As implemented in the `genefilter` package (Gentleman, Carey, Huber & Hahne 2011), we used a two-sample t-test for testing (12) and calculated the p-value for each gene. Further, these p-values were used to calculate (10) and (11).

In Figure 2 we present the distribution of t^* using equation (8) for the m genes. When no correction for multiple testing is applied on the p-values, a total of 1,045 (34.3%, $t_c = 2.97$) genes were found to be DE, which were reduced to 98 (3.2%, $t_c^* = 10.99$) when a Bonferroni correction was applied. On the other hand, when the p-values were FDR-corrected before applying our methodology, 681 (22.3%, $t_c = -5.05$) were found to be DE. Equivalent results were obtained using built-in R function `p.adjust()`.

3.2. Type 2 Diabetes Data

Mootha et al. (2003) presented an analytical strategy for detecting modest but coordinate changes in gene expression using DNA microarray data. This data consists of 22,283 gene expression levels measured in 43 age-matched males skeletal muscle biopsy samples, 17 with normal glucose tolerance (NGT), 8 with impaired glucose tolerance (IGT) and 18 with type 2 diabetes (T2D).

After randomly selecting 1,000 gene expression levels for T2D samples from the original data, the linear correlation coefficient ρ for each pair of genes was calculated. ρ might be seen as a «proxy» of the potential interacting effects between pair of genes.

Table 1: Significant correlation coefficients for pairs of genes in 1,000 randomly selected gene expression levels (Mootha et al. 2003) when only T2D samples are included. Bonferroni correction was applied. CI: Confidence Interval.

Genes	$\hat{\rho}$	95%CI	t-statistic	t_c^*	Raw P-value
G12-G720	0.939	(0.840, 0.977)	10.899	18.621	8.16×10^{-9}
G291-G350	0.938	(0.837, 0.977)	10.777	16.643	9.60×10^{-9}
G490-G698	0.927	(0.812, 0.973)	9.903	17.274	3.14×10^{-8}
G108-G434	-0.921	(-0.971, -0.797)	-9.459	16.642	5.91×10^{-8}
G210-G720	0.920	(0.795, 0.970)	9.409	16.570	6.36×10^{-8}
G293-G308	0.917	(0.787, 0.969)	9.196	16.257	8.69×10^{-8}

A total of $m = 499,500$ hypothesis of the form

$$H_{0,i} : \rho_i = 0 \quad \text{vs.} \quad H_{1,i} : \rho_i \neq 0 \qquad i = 1, 2, \dots, m \tag{13}$$

were tested. For $\alpha = 0.05$, 52,576 (10.53%, $t_c = 2.97$) correlation coefficients were significant when no correction for multiple testing was applied, which reduced to 319 (0.06%, $t_c = 2.97$) and 6 ($\sim$0%, $t_c^* = 16.09$), respectively, when the FDR and Bonferroni corrections were used. Results for the latter are presented in Table 1.

4. Discussion

In this paper, we propose a new method to determine whether a p-value is significant under a multiple testing setting while controlling (or not) the FWER. Our proposal, based on the m-th order statistic of a $U(0, 1)$ distribution, has been shown to give equivalent results to Bonferroni's method while controlling the FWER, and to classical methods while not. Furthermore, under the null hypothesis, the proportion of true null hypothesis being rejected is close to the nominal level α. Observe that, by no means, we are stating that our method is improving any of the other alternatives available in the literature to correct by multiple testing, and which have extensively been applied in the genetics field.

The contribution of this paper can be seen under two perspectives. First, it offers a graphical alternative to represent p-values and the cutoff value beyond which, in the genetic context, we consider that a SNP (or gene in a microarray) is statistically significant. Second, the use of asymptotic statistics and extreme values theory in genetics. In a review of the literature previous to the writing of this paper, we found no mention or application of these two important concepts in genetics. The main advantages of this new approach are the direct calculation of the cutoff value labelling a p-value as significant, the simplicity of its calculations, and how easy it is to graphically represent the results. Computationally, our approach is better than the FDR (Benjamini & Hochberg 1995) as it does not require to store all the p-values.

Although in our applications section we showed how to use our approach to determine significant p-values with GWAS and microarray data, it is not limited, under any circumstance, to these type of data. The main reason for this is that our approach uses the p-values of the hypotheses tested regardless of the type(s) of data on which they have been tested. Future extensions of this methodology include considering correlated tests as those proposed by Benjamini & Yekutieli (2001).

Ackowledgements

We are grateful to Ms. Eva K. Pillai for critical reading of this manuscript. We also thank two anonymous reviewers for their insightful comments. The first author was supported by the Eccles Scholarship in Medical Sciences, the Fenner Merit Scholarship and the Australian National University High Degree Research Scholarship.

[Recibido: noviembre de 2012 — Aceptado: enero de 2014]

References

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society, Series B (Methodological)* **57**(1), 389–300.

Benjamini, Y. & Yekutieli, D. (2001), 'The control of the false discovery rate in multiple testing under dependency', *Annals of Statistics* **29**(4), 1165 – 1188.

Bonferroni, C. E. (1935), 'Il calcolo delle assicurazioni su gruppi di teste', *Studi in Onore del Professore Salvatore Ortu Carboni,* pp. 13–60.

Casella, G. & Berger, R. (2001), *Statistical Inference*, 2 edn, Duxbury Press, United States of America.

Devroye, L. (1986), *Non-Uniform Random Variate Generation*, New York: Spring-Verlang.

Gentleman, R., Carey, V., Huber, W. & Hahne, F. (2011), *genefilter: Methods for filtering genes from microarray experiments.* R package version 1.34.0.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. & Lander, E. (1999), 'Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring', *Science* **286**, 531–537.

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., Hayward, N. K., Montgomery, G. W., Visscher, P. M., Martin, N. G. & Macgregor, S. (2010), 'A versatile gene-based test for genome-wide association studies', *The American Journal of Human Genetics* **87**(1), 139 – 145.

Manolio, T. A. (2010), ‘Genomewide association studies and assessment of the risk of disease’, *New England Journal of Medicine* **363**(2), 166–176.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. & Groop, L. C. (2003), ‘Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes’, *Nature Genetics* **34**(3), 267–73.

Murdoch, D., Tsai, Y. & Adcock, J. (2008), ‘P-values are random variables’, *The American Statistician* **62**(3), 242–245.

Nyholt, D. R. (2004), ‘A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other’, *The American Journal of Human Genetics* **74**(4), 765 – 769.

Pollard, K. S., Gilbert, H. N., Ge, Y., Taylor, S. & Dudoit, S. (2011), *multtest: Resampling-based multiple hypothesis testing.* R package version 2.8.0.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org/

Sackrowitz, H. & Samuel-Cahn, E. (1999), ‘P Values as Random Variables-Expected P Values’, *The American Statistician* **53**(4), 326–331.

Serfling, R. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, United States of America.

Shaffer, J. P. (1995), ‘Multiple hypothesis testing’, *Annual Review of Psychology* **46**, 561–584.

Appendix. Detect significant p-values in R using the proposed method

```
#-----------------------------------------------------------------------------
#                                    ARGUMENTS
#   p           vector of p-values
#   plot        histogram of transformed p-values? (default: TRUE)
#   corrected   correction by multiple testing? (default: TRUE)
#   line        add vertical line indicating critical value?  (default: TRUE)
#   alpha       type I error probability (default: 0.05)
#   ...         additional arguments passed to hist()
#-----------------------------------------------------------------------------
pvaltest <- function(p, plot = TRUE, corrected = TRUE,
                            line = TRUE, alpha = 0.05, ...){
m <- length(p)
ti <-  -log(-log(1 - p))
tc <- -log(-log(1 - alpha))
tcstar <- tc + 1 * corrected * log(m)
total <- sum(ti > tcstar)

if(plot){
hist(ti, breaks = 50, prob = TRUE, las = 1,
   xlab = expression(italic(t*"*")), ylab = "Density", ...)
   abline(v = 1* line * tcstar, col = 1, lty = 2)
}
cat("Number of tests = ", m, "\n")
cat("Critical value = ", round(tcstar, 2), "\n")
cat("Total rejected = ", total, "(", round(100*total/m, 2), "%)", "\n")
invisible(list(p.value = p, statistic = ti, reject = ti > tcstar))
}

## Example
set.seed(123)
p <- c(runif(100, 0, 1e-4), runif(5000))
res <- pvaltest(p, main = "")
# Number of tests =  5000
# Critical value =  11.49
# Total rejected =  7 ( 0.14 %)

str(res)
# List of 3
#  $ p.value  : num [1:5100] 2.88e-05 7.88e-05 4.09e-05 8.83e-05 9.40e-05 ...
#  $ statistic: num [1:5100] 10.46 9.45 10.1 9.33 9.27 ...
#  $ reject   : logi [1:5100] FALSE FALSE FALSE FALSE FALSE TRUE ...
```

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 79 a 94

Three Similarity Measures between One-Dimensional Data Sets

Tres medidas de similitud entre conjuntos de datos unidimensionales

Luis Gonzalez-Abril[1,a], Jose M. Gavilan[1,b], Francisco Velasco Morente[1,c]

[1] Departamento de Economía Aplicada I, Facultad de Ciencias Económicas y Empresariales, Universidad de Sevilla, Sevilla, Spain

Abstract

Based on an interval distance, three functions are given in order to quantify similarities between one-dimensional data sets by using first-order statistics. The Glass Identification Database is used to illustrate how to analyse a data set prior to its classification and/or to exclude dimensions. Furthermore, a non-parametric hypothesis test is designed to show how these similarity measures, based on random samples from two populations, can be used to decide whether these populations are identical. Two comparative analyses are also carried out with a parametric test and a non-parametric test. This new non-parametric test performs reasonably well in comparison with classic tests.

***Key words*:** Data mining, Interval distance, Kernel methods, Non-parametric tests.

Resumen

Basadas en una distancia intervalar, se dan tres funciones para cuantificar similaridades entre conjuntos de datos unidimensionales mediante el uso de estadísticos de primer orden. Se usa la base de datos Glass Identification para ilustrar cómo esas medidas de similaridad se pueden usar para analizar un conjunto de datos antes de su clasificación y/o para excluir dimensiones. Además, se diseña un test de hipótesis no parámetrico para mostrar cómo similaridad, basadas en muestras aleatorias de dos poblaciones, se pueden usar para decidir si esas poblaciones son idénticas. También se realizan dos análisis comparativos con un test paramétrico y un test no paramétrico. Este nuevo test se comporta razonablemente bien en comparación con test clásicos.

***Palabras clave*:** distancia entre intervalos, métodos del núcleo, minería de datos, tests no paramétricos.

[a] Senior lecturer. E-mail: luisgon@us.es
[b] Senior lecturer. E-mail: gavi@us.es
[c] Senior lecturer. E-mail: velasco@us.es

1. Introduction

Today, in many tasks in which data sets are analysed, researchers strive to achieve some way of measuring the features of data sets, for instance, to distinguish between informative and non-informative dimensions. A first step could be to study whether several sets of data are similar. The similarity may be defined as a measure of correspondence between the data sets under study. That is, a function which, given two data sets X and Y, returns a real number that measures their similarity.

In data mining, there exist several similarity measures between data sets: for instance, in Parthasarathy & Ogihara (2000), a similarity is used which compares the data sets in terms of how they are correlated with the attributes in a database. A similar problem, studied in Burrell (2005), is the measurement of the relative inequality of productivity between two data sets using the Gini coefficient (González-Abril, Velasco, Gavilán & Sánchez-Reyes 2010). A similarity measure based on mutual information (Bach & Jordan 2003) is used to determine the similarity between images in Nielsen, Ghugre & Panigrahy (2004). Similarity between molecules is used in Sheridan, Feuston, Maiorov & Kearsley (2004) to predict the nearest molecule and/or the number of neighbours in the training set.

A common problem with the aforementioned similarity measures is that their underlying assumptions are often not explicitly stated. This study aims to use first-order statistics to explain the similarity between data sets. In this paper, the similarity is established in the sense that one-dimensional data sets are similar simply by comparing the statistics of the variables in each data set.

In statistics, other similarity measures between data sets are also available (González, Velasco & Gasca 2005), for instance, those which are used in hypothesis testing. In this way, a non-parametric hypothesis test based on the proposed similarity is presented in this paper and a comparative analysis is carried out with several well-known hypothesis tests.

The remainder of the paper is arranged as follows: In Section 2, we introduce some notation and definitions. Sections 3 and 4 are devoted to give two similarity measures between one-dimensional data sets. An example is presented in Section 5 to show their use. A non-parametric test is derived in Section 6 and experimental results are given to illustrate its behaviour and good features. Finally, some conclusions are drawn and future research is proposed.

2. Concepts

Following Lin (1998), with the purpose of providing a formal definition of the intuitive concept of similarity between two entities X and Y, the intuitions about similarity must be clarified. Thus: i) The similarity is related to their commonality in that the more commonality they share, the more similar they are; ii) The similarity is related to the differences between them, in that the more differences they have, the less similar they are; and iii) The maximum similarity is reached when X and Y are identical.

Let us denote a similarity measure between X and Y by $K(X,Y)$. Ideally this function must satisfy the following properties:

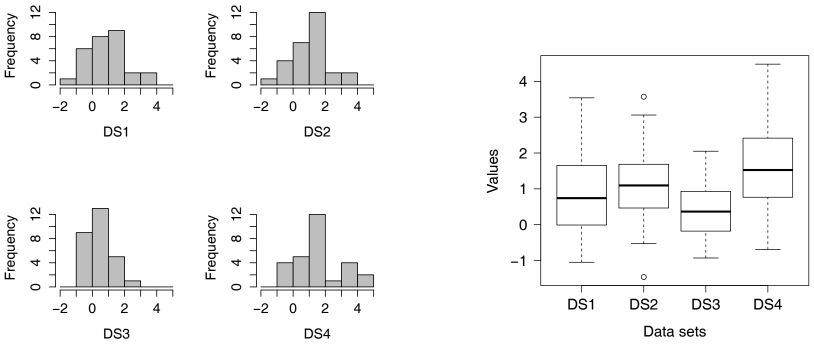
1. Identity: $K(X,Y)$ at its maximum corresponds to the fact that the two entities are identical in all respects;

2. Distinction: $K(X,Y) = 0$ corresponds to the fact that the two entities are distinct in all respects; and

3. Relative Ordinality: If $K(X,Y) > K(X,Z)$, then it should imply that X is more similar to Y than it is to Z.

Hence, certain similarities are defined in this paper which are consistent with the above intuitions and properties.

Let us consider four one-dimensional data sets, DS_1, DS_2, DS_3 and DS_4 (see the Appendix), where the DS_1 and DS_2 data sets are taken from a N(1,1) distribution, the DS_3 data set from a N(0.5,1) distribution, and the DS_4 data set from a N(1.5, 1.25) distribution, where a N(μ,σ) distribution is a Normal distribution with mean μ and standard deviation σ. In practice, comparison of these data sets involves: a) plotting graphical summaries, such as histograms and boxplots, next to each other; b) simply comparing the means and variances (see Figure 1); or c) calculating correlation coefficients (if items of data are appropriately paired). These methods are straightforward to interpret and explain. Nevertheless, these approaches contain a major drawback since the interpretation is subjective and the similarities are not quantified.



	DS_1	DS_2	DS_3	DS_4
Mean	0.989	1.046	0.427	1.632
Variance	1.221	1.152	0.569	1.956

Figure 1: Histograms, boxplots, means, and variances of the data sets of the Appendix.

Let us introduce the concept of interval distance. Given an open interval (similarly for another kind of interval) of finite length, there are two main ways to represent that interval: using the extreme points as *(a,b)* (classic notation) or as an open ball $\mathrm{B}_r(\mathrm{c})$ (Borelian notation) where $c = (a+b)/2$ (centre) and $r = (b-a)/2$ (radius). Using Borelian notation, the following distance between intervals given in González, Velasco, Angulo, Ortega & Ruiz (2004) is considered:

Definition 1. Let $I_1 = (c_1 - r_1, c_1 + r_1)$ and $I_2 = (c_2 - r_2, c_2 + r_2)$ be two real intervals. A distance between these intervals is defined as follows:

$$d_{\mathbf{W}}(I_1, I_2) = \sqrt{(\Delta c \ , \ \Delta r)\mathbf{W}\begin{pmatrix} \Delta c \\ \Delta r \end{pmatrix}} \tag{1}$$

where $\Delta c = c_2 - c_1$, $\Delta r = r_2 - r_1$, and $\mathbf{W}$ is a symmetrical and positive-defined 2×2 matrix, called weight-matrix.

It is clear from matrix algebra that $\mathbf{W}$ can be written as[1] $\mathbf{W} = \mathbf{P}^t\mathbf{P}$, where $\mathbf{P}$ is a non-singular 2×2 matrix, and hence $d_{\mathbf{W}}(I_1, I_2) = \|\mathbf{P}\,(\Delta c\,,\,\Delta r)^t\|$, where $\|\cdot\|$ is the quadratic norm in $\mathbb{R}^2$, and therefore $d_W(\cdot,\cdot)$ is an ℓ_2-distance. It can be observed that, from the matrix $\mathbf{W}$, the weight assigned to the position of the intervals c, and to their size r, can be controlled. Furthermore, the distance (1) provides more information on the intervals than does the Hausdorff distance (González et al. 2004).

From the distance given in (1), three new similarity measures are defined in this paper.

3. A First Similarity

Definition 2. Given a data set $X = \{x_1, \ldots, x_n\}$ and a parameter $\ell > 1$, the ℓ-associated interval of X, denoted by I_X^ℓ, is defined as follows:

$$I_X^\ell = (\overline{X} - \ell \cdot S_X, \overline{X} + \ell \cdot S_X)$$

where $\overline{X}$ and S_X are the mean and the standard deviation of X, respectively.

It is worth noting that Chebyshev's inequality states that there are at least a $(1 - 1/\ell^2)$ proportion of observations x_i in the interval I_X^ℓ. Hence, the similarity between two data sets X and Y can be quantified from the distance between the intervals I_X^ℓ and I_Y^ℓ. However, it is possible that some instances $z \in X \cup Y$ exist such that $z \notin I_X^\ell \cup I_Y^\ell$. Thus, a penalizing factor (the proportion of instances within I_X^ℓ and I_Y^ℓ) is taken into account in the following similarity measure.

Definition 3. Given two data sets $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_m\}$ and a parameter $\ell > 1$, a similarity measure between X and Y, denoted by $K_{\mathbf{W}}^\ell(X, Y)$, is defined as follows:

$$K_{\mathbf{W}}^\ell(X,Y) = \frac{\#((X \cup Y) \cap (I_X^\ell \cap I_Y^\ell))}{\#(X \cup Y)} \cdot \frac{1}{1 + d_{\mathbf{W}}(I_X^\ell, I_Y^\ell)} \tag{2}$$

where $\#A$ denotes the cardinality of set A.

[1] The notation "$\mathbf{u}^t$" denotes the transposed vector of $\mathbf{u}$.

The function defined, $K^{\ell}_{\mathbf{W}}$, is a similarity measure (Cristianini & Shawe-Taylor 2000) which has been proposed based on distance measurements in Lee, Kim & Lee (1993) and Rada, Mili, Bicknell & Blettner (1989). Furthermore, for any ℓ and $\mathbf{W}$, $K^{\ell}_{\mathbf{W}}$ is a positive, symmetrical function since it is a radial basis function (Skhölkopf & Smola 2002).

It can be proved from (1) that $d^2_{\mathbf{W}}(I^{\ell}_X, I^{\ell}_Y) = w_{11}(\Delta\overline{X})^2 + 2\ell w_{12}\Delta\overline{X}\Delta S + \ell^2 w_{22}(\Delta S)^2$, where $\Delta\overline{X} = \overline{X} - \overline{Y}$, $\Delta S = S_X - S_Y$, and the weight-matrix is $\mathbf{W} = \{w_{ij}\}^2_{i,j=1}$ with $w_{ij} = w_{ji}$ when $i \neq j$.

Thus, the $K^{\ell}_{\mathbf{W}}$ similarity takes into account the following characteristics: i) The position of the whole data set on the real line given by the mean; ii) The spread of the data set around its mean given by the standard deviation multiplied by a parameter $\ell > 1$; iii) The weighted importance of the mean and the standard deviation of each data set, given in the weight-matrix $\mathbf{W}$; and iv) A factor which quantifies, from the number of outlying values, the goodness of fit of the associated intervals.

Example 1. For the data sets of the Appendix, $\ell = 2$ and $\mathbf{W} = \mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, the similarity $K^{\ell=2}_{\mathbf{I}}$ is given in Table 4. It can be seen that the similarities obtained are consistent with the distributions generating the data sets.

After having experimented with different choices of ℓ and $\mathbf{W}$, it is observed that the numerical results differ slightly but the conclusions on their similarities remain the same. ||

4. A Second Similarity

When the size of the data set is large, consideration of only the number of outlying values and the mean and the standard deviation is grossly insufficient to obtain meaningful results. Furthermore, it is clear that these features are not likely to be very helpful outside a normal distribution family (the mean and variance are highly sensitive to heavy tails and outliers, and are unlikely to provide good measures of location, scale or goodness-of-fit in their presence). Hence more characteristics which summarize the information of each data set must be taken into account.

In this framework, the percentiles of the data set are used. Let

$$Q_X = \{p_{1X}, \cdots, p_{qX}\}$$

be a set of q percentiles of a data set X with $p_{iX} \leq p_{(i+1)X}$ and $q \geq 2$. Hence, $q-1$ intervals, denoted by I_{iX}, are considered as follows: $I_{iX} = (p_{iX}, p_{(i+1)X})$, for $i = 1, \ldots, q-1$.

Example 2. Given the DS_1 data set, an example of the Q_{DS_1} set is given by

$$Q_{DS_1} = \{-0.8926,\ -0.0099,\ 0.7376,\ 1.6571,\ 3.5146\}$$

where these values are the percentiles 2.5, 25, 50, 75, and 97.5, respectively, and $q = 5$.

Definition 4. Given a weight-matrix $\mathbf{W}$ and two sets of q percentiles, Q_X and Q_Y, of the data sets X and Y, respectively, a similarity between X and Y, denoted by $K^Q_{\mathbf{W}}(X,Y)$, is defined as follows:

$$K^Q_{\mathbf{W}}(X,Y) = \frac{1}{1 + \frac{1}{q-1}\sum_{i=1}^{q-1} d_{\mathbf{W}}(I_{iX}, I_{iY})} \tag{3}$$

The $K^Q_{\mathbf{W}}$ similarity has the following properties: i) This function is positive and symmetrical; ii) If $X = Y$ then $K^Q_{\mathbf{W}}(X,Y) = 1$; iii) The similarity is low if the percentiles are far from each other; and iv) It is a radial basis function.

In Table 1, several examples of $d_{\mathbf{W}}(I_{iX}, I_{iY})$ can be seen whereby the symmetrical weight-matrix $\mathbf{W} = \{w_{ij}\}^2_{i,j=1}$ is varied. In cases 1 and 2, $\mathbf{W}$ is a non-regular matrix ($det(\mathbf{W}) = 0$) and therefore this situation is inadequate. In case 3, $\mathbf{W} = \mathbf{I}$ is the identity matrix, and case 4 provides a straightforward weight-matrix which presents the cross product between the percentiles.

Table 1: Distance between intervals for different weight-matrices **W**.

Case	w_{11}	w_{12}	w_{22}	$d^2_{\mathbf{W}}(I_{iX}, I_{iY})$
1	1	1	1	$(p_{(i+1)X} - p_{(i+1)Y})^2$
2	1	-1	1	$(p_{iX} - p_{iY})^2$
3	1	0	1	$\frac{1}{2}((p_{(i+1)X} - p_{(i+1)Y})^2 + (p_{iX} - p_{iY})^2)$
4	$\frac{3}{4}$	0	$\frac{1}{4}$	$\frac{1}{4}((p_{(i+1)X} - p_{(i+1)Y})^2 + (p_{iX} - p_{iY})^2 + ..$ $... + (p_{(i+1)X} - p_{(i+1)Y})(p_{iX} - p_{iY}))$

On the other hand, there are many different ways to choose the Q_X set for a fixed data set X; in this paper the discretization process[2] based on equal-frequency intervals (Chiu, Wong & Cheung 1991) is used. Furthermore, in order to obtain a specific value of q, there are several selections based on experience such as Sturges' formula, $q_1 = Int\left[\frac{3}{2} + \frac{Log(n)}{Log(2)}\right]$ and $q_2 = Int[\sqrt{n}]$, where the operator $Int[\cdot]$ is the integer part and n is the size of the data set. Henceforth, $q \equiv q_1$ is considered with $n = \max\{\#X, \#Y\}$ and the set of percentiles Q is obtained such that in each interval I_i. there is the same quantity of items of data.

In the following section, an example is presented to show how these similarities could be used.

5. The Glass Identification

The Glass Identification is obtained from the UC Irvine Machine Learning Repository (Bache & Lichman 2013). This database is often used to study the

[2] A discretization process converts continuous attributes into discrete attributes by yielding intervals in which the attribute value can reside instead of being represented as singleton values.

performance between different classifiers. Its main properties are: 214 instances, 9 continuous attributes and 1 attribute with 6 classes (labels). The number of instances in each class is 70, 76, 17, 13, 9 and 29, respectively.

Suppose that a preliminary analysis of this data set is desired before applying a classifier. Firstly, $K^{Q}_{\mathbf{W}}$ similarities between continuous attributes are given in Table 2 for $\mathbf{W} = \mathbf{Id}$.

TABLE 2: $K^{Q}_{\mathbf{W}}$ similarities between continuous attributes of the Glass data set.

Attr.	2	3	4	5	6	7	8	9
1	0.0777	0.3365	**0.7733**	0.0139	0.4778	0.1209	0.4000	0.4034
2	1	0.0862	0.0771	0.0166	0.0717	0.1780	0.0697	0.0697
3		1	0.3450	0.0141	0.2802	0.1424	0.2523	0.2528
4			1	0.0138	0.5008	0.1195	0.4182	0.4194
5				1	0.0136	0.0154	0.0136	0.0136
6					1	0.1068	**0.6871**	**0.7089**
7						1	0.1026	0.1026
8							1	**0.9153**

It is observed that attributes 1 and 4 are very similar to each other; and attributes 6, 8 and 9 are also very similar, particularly attributes 8 and 9. Hence, it may be a good idea to eliminate some attributes before the implementation of the classifier, for instance attributes 4, 6 and 8.

Let us study the attributes to determine similarities between the same attributes but with different labels. Hence, if the similarity obtained is low, then the classification is straightforward.

The number of instances with label 1 is 70, and with label 2 this is 76, and $K^{Q}_{\mathbf{W}}$ similarities between the nine attributes are given in Table 3. It can be seen that these values are very high, which indicates that the discrimination between these two labels is not easy.

On the other hand, the number of instances with label 3 is 17, and with label 4 this is 13, and $K^{\ell}_{\mathbf{W}}$ similarities between the nine attributes are given in Table 3 for $\ell = 2$. Hence, attributes 3 and 7 are the best in order to separate labels 3 and 4. However the main problem with respect to labels 3 and 4 is that there are very few instances.

TABLE 3: $K^{\ell}_{\mathbf{W}}$ and $K^{Q}_{\mathbf{W}}$ similarities between different labels of the Glass data set.

Labels	Attr.	1	2	3	4	5	6	7	8	9
1 – 2	$K^{Q}_{\mathbf{W}}$	0.9984	0.8991	0.6175	0.8123	0.8710	0.9088	0.6290	1	0.9746
3 – 4	$K^{\ell}_{\mathbf{W}}$	0.8333	0.8844	0	0.7733	0.7488	0.7991	0.4898	0.8807	0.8986

The main conclusion in this brief preliminary analysis is that the classes of the Glass Identification Database are difficult to separate based only on individual features for the given instances. A good classifier is therefore necessary in order to obtain acceptable accuracy for this classification problem.

An experiment[3] is carried out to show that the conclusions of this brief analysis are correct. Thus, the algorithm considered is the standard 1-v-r SVM formulation (Vapnik 1998), by following the recommendation given in Salazar, Vélez & Salazar (2012), and its performance, (in the form of accuracy rate), has been evaluated using the Gaussian kernel, $k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ where two hyperparameters must be set: the regularization term C and the width of the kernel σ. This space is explored on a two-dimensional grid with the following values: $C = \{2^0, 2^1, 2^2, 2^3, 2^4, 2^5\}$ and $\sigma^2 = \{2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1\}$. The criterion used to estimate the generalized accuracy is a ten-fold cross-validation on the whole training data. This procedure is repeated 10 times in order to ensure good statistical behaviour. The optimization algorithm used is the exact quadratic program-solver provided by Matlab software.

The best cross-validation mean rate among the several pairs (C, σ^2) is obtained for $C = 1$ and $\sigma^2 = 1$ with 70.95% accuracy rate when all attributes are used and, when attributes 4, 6 and 8 are eliminated, then the best cross-validation mean rate is obtained for $C = 16$ and $\sigma^2 = 1$ with 68.38% accuracy rate. This experiment indicates that the Glass Identification Database is difficult to separate and that the elimination of attributes 4, 6 and 8 only slightly modifies the accuracy rates.

In the following section, a new hypothesis test is designed and is compared with other similar hypothesis tests.

6. Hypothesis Testing

Definition 5. Let $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_m\}$ be two data sets. Two further data sets X^c and Y^c, called the quasi-typified data sets of X and Y, respectively, are defined as follows:

$$x_i^c = \frac{S_Y}{S_X^2}\left(x_i - \overline{Z}\right), \qquad y_i^c = \frac{S_X}{S_Y^2}\left(y_i - \overline{Z}\right),$$

where $Z = \{x_1, \dots, x_n, y_1, \dots\ y_m\}$. This process is called quasi-typification.

It is straightforward to prove that $\overline{X^c} = mS_Y(\overline{X} - \overline{Y})/(S_X^2(n+m))$, $\overline{Y^c} = nS_X(\overline{Y} - \overline{X})/(S_Y^2(n+m))$, $S_{X^c} = S_Y/S_X$, and $S_{Y^c} = S_X/S_Y$. Therefore, if $\overline{X} = \overline{Y}$ and $S_X = S_Y$, then $\overline{X^c} = \overline{Y^c} = 0$ and $S_{X^c} = S_{Y^c} = 1$.

From Definition 5, a third similarity measure between data sets is given as follows:

Definition 6. Let X and Y be two data sets. A measure of similarity between these sets is defined as: $KC_{\mathbf{W}}^{Q}(X, Y) = K_{\mathbf{W}}^{Q}(X^c, Y^c)$ provided that X^c and Y^c are the quasi-typified data sets of X and Y, $\mathbf{W}$ is a weight-matrix, and the sets of q percentiles are Q_X and Q_Y of the data sets X and Y, respectively.

Example 3. $KC_{\mathbf{I}}^{Q}$ similarities between the data sets DS_1, DS_2, DS_3 and DS_4 are given in Table 4. In Figure 2, each subplot depicts the boxplot of data DS_i,

[3] Most results have been obtained following the experimental framework proposed by Hsu & Lin (2002) and continued in Anguita, Ridella & Sterpi (2004).

DS_j, T_i and T_j where the T_i's are the quasi-typified data sets of DS_i and DS_j for $i, j = 1, 2, 3, 4$ and $i \neq j$.

TABLE 4: $K^{\ell=2}_{\mathbf{Id}}$, $K^{Q}_{\mathbf{Id}}$ and $KC^{Q}_{\mathbf{Id}}$ similarities between the data sets in the Appendix.

$K^{\ell=2}_{\mathbf{Id}}$	DS_2	DS_3	DS_4
DS_1	0.928	0.880	0.862
DS_2	——	0.862	0.863
DS_3	——	——	0.781

$K^{Q}_{\mathbf{Id}}$	DS_2	DS_3	DS_4
DS_1	0.728	0.646	0.557
DS_2	——	0.595	0.623
DS_3	——	——	0.441

$KC^{Q}_{\mathbf{Id}}$	DS_2	DS_3	DS_4
DS_1	0.897	0.574	0.600
DS_2	——	0.495	0.592
DS_3	——	——	0.347

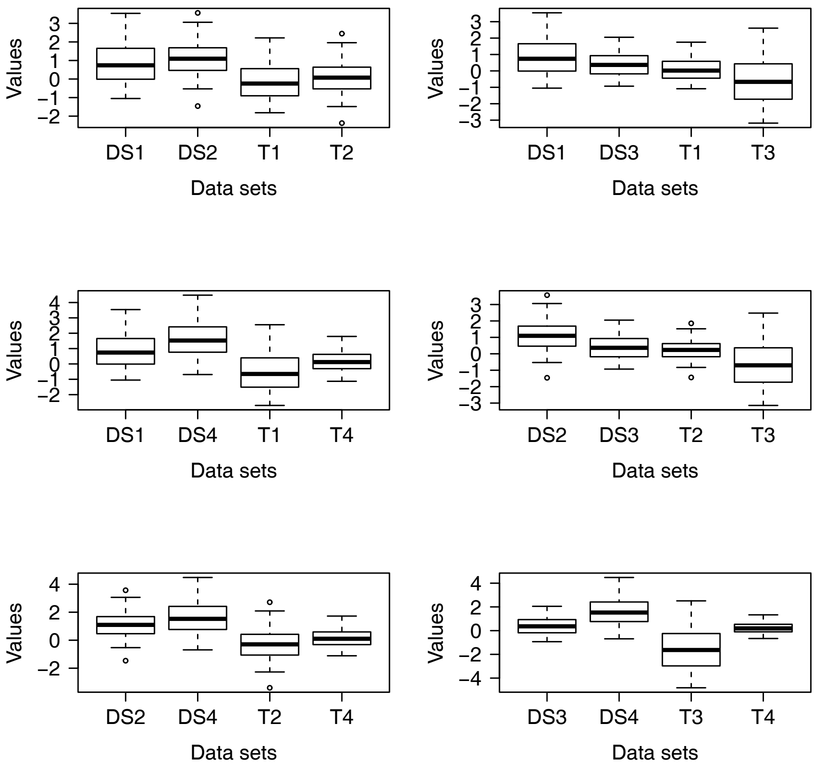


FIGURE 2: Boxplots of each pair of data sets of the Appendix before (DS_i data sets) and after (T_i data sets) applying the quasi-typification.

It is worth noting that all three similarities verify that the similarity between DS_1 and DS_2 is the highest and the similarity between DS_3 and DS_4 is the lowest similarity. Thus, the similarities obtained are consistent with the distribution that generates the data sets. ‖

Several percentiles are obtained from $KC_{\mathbf{I}}^{Q}$ similarities of a simulated distribution between random samples of size 100 from two $N(0,1)$ distributions. The results are shown in Table 5. It is important to point out that the thresholds have been simulated $1,000,000$ times and it is observed that the sensitivity of the thresholds is very low (less than 10^{-5} units). Hence, it is now possible to use these percentiles to construct a hypothesis test.

Table 5: Percentiles of the simulated distribution $KC_{\mathbf{I}}^{Q}$ between two $N(0,1)$ distributions for $n = 100$.

α	0.001	0.01	0.025	0.05	0.10
$P(100,\alpha)$	0.60110	0.65353	0.67917	**0.70166**	0.72802

Definition 7. Let $X = \{X_1, \dots, X_n\}$ and $Y = \{Y_1, \dots, Y_m\}$ be two random samples from populations $\mathcal{F}$ and $\mathcal{F}'$. Let a hypothesis test be $H_0 : \mathcal{F}' = \mathcal{F}$ versus $H_1 : \mathcal{F}' \neq \mathcal{F}$. Let $RC = \{(X,Y) : KC(X,Y) < P(n^*,\alpha)\}$ be the critical region of size α where $P(n^*,\alpha)$ is the percentile α of the simulated distribution $KC_{\mathbf{I}}^{Q}$ between two $N(0,1)$ distributions for $n^* = \min(n,m)$. Henceforth, this test is denoted as the GA-test.

***Note* 1.** It is worth noting that this test is valid for normal or similar populations. If another type of population is given, then the corresponding percentiles should be calculated.

6.1. Comparison with a Parametric Test

Let the following test be: $H_0 : \mathcal{F}' = \mathcal{F}$ versus $H_1 : \mathcal{F}' \neq \mathcal{F}$, where $\mathcal{F} = N(\mu_1,\sigma_1)$, $\mathcal{F}' = N(\mu_2,\sigma_2)$ and where μ_1, μ_2, σ_1 and σ_2 are unknown parameters. In this case, the null hypothesis states that the two normal populations have both identical means and variances.

Let $X = \{X_1, \dots, X_{100}\}$ and $Y = \{Y_1, \dots, Y_{100}\}$ be two random samples from $N(\mu_1,\sigma_1)$ and $N(\mu_2,\sigma_2)$ distributions. A classic test (C-test) is considered, which is a union of two tests. Firstly, a test is performed to determine whether two samples from a normal distribution have the same mean when the variances are unknown but assumed equal. The critical region of size 0.025 is

$$RC_1 = \left\{(X,Y) : \left|\overline{X} - \overline{Y}\right| > 2.2586\sqrt{(S_1^2 + S_2^2)/99}\right\}$$

where 2.258 is the percentile 0.9875 of Student's t distribution with 198 degrees of freedom. Another test is also performed to determine whether two samples from a normal distribution have the same variance. The critical region of size 0.025 is

$$RC_2 = \left\{(X,Y) : S_1^2/S_2^2 < 0.6353,\, S_1^2/S_2^2 > 1.5740\right\}$$

where 0.6353 and 1.5740 are the percentiles 0.0125 and 0.9875 of Snedecor's F distribution, both with 99 degrees of freedom.

In this framework, a comparison is made between the classic test for Normal populations whose critical region of size 0.0494 ($= 1 - 0.975^2$) is $RC = RC_1 \cap RC_2$, versus the GA-test whose critical region of size 0.05 is $\{(X, Y) : KC(X, Y) < 0.70166\}$ (see Table 5). For this comparison, it is considered that one population is $N(20, 4)$ and the other population is $N(\mu, \sigma)$, and the hypothesis test is carried out for 100,000 simulations for each value $\mu = 18$, 19, 20, 21, 22 and $\sigma = 3$, 3.5, 4, 4.5, 5. The results of the experiment are given in Table 6, where the percentage of acceptance of the null hypothesis is shown for the two tests. The best result for each value of the parameters is printed in bold, that is, the minimum of the two values except for the case $\sigma = 4$ and $\mu = 20$ in which the null hypothesis is true and then the maximum of the two values is printed in bold.

The first noteworthy conclusion is that there are no major differences between the two methods and therefore the results of the GA-test are good. As expected, the results are almost symmetrical for equidistant values from the true mean and variance. When only one of the two parameters is the actual value, then the classic test behaves better in general, possibly due to the fact that the classic test is sequential and the other is simultaneous. However, when both parameters only slightly differ from the actual values, then the GA-test performs better. The same holds true for values of the mean that differ from the actual value and for great differences in the variance.

Table 6: Acceptance percentage of the null hypothesis when comparing the $N(20, 4)$ and the $N(\mu, \sigma)$ distributions for different values of the mean and of the standard deviation using the classic test (C) and the GA-test. The desired level of significance is 5% and the best result in each case is printed in bold.

σ	3		3.5		4		4.5		5	
μ	GA	C	GA	C	GA	C	GA	C	GA	C
18	**00.88**	01.03	06.56	**05.47**	13.55	**09.72**	14.44	**12.26**	**08.94**	09.56
19	**14.06**	16.06	**51.94**	52.58	70.32	**66.97**	**61.56**	61.77	**35.94**	38.17
20	27.79	**26.92**	**79.85**	79.93	**94.97**	94.84	83.77	**83.24**	51.02	**49.18**
21	**13.73**	15.71	**52.33**	52.86	70.67	**67.32**	**61.35**	61.53	**35.99**	38.16
22	**00.89**	01.06	06.58	**05.43**	13.95	**10.11**	14.42	**12.34**	**08.80**	09.53

6.2. Comparison with a Non-Parametric Test

In this section, the GA-test is used with non-normal distributions than remains similar to a Normal distribution. At this point, the interest lies in testing $H_0 : \mathcal{F}' = \mathcal{F}$ versus $H_1 : \mathcal{F}' \neq \mathcal{F}$ for a number of populations $\mathcal{F}$ and $\mathcal{F}'$. The GA-test is compared against the Kolmogorov-Smirnov test. In both cases the desired level of significance is 0.05, the hypothesis test is carried out for 10,000 simulations where the populations are $Bi(100, 0.2)$ (Binomial), $Po(20)$ (Poisson) and $N(20, 4)$ (Normal). Figure 3 shows that these distributions are very similar and the size of random samples is 100.

The results of the experiment are given in Table 7 in the form of percentage of acceptance of the null hypothesis. Again, the best result in each case is printed in bold, that is, the minimum of the two values when the null hypothesis is false (values outside the diagonal) and the maximum of the two values when the null hypothesis is true (values in the diagonal). It can be seen that the GA-test can differentiate between the Poisson distribution and the other two better than can the Kolmogorov-Smirnov test. Nevertheless, the Kolmogorov-Smirnov test behaves better than the GA-test in Binomial and Poisson populations under the null hypothesis (the opposite is true for the normal distribution) and when distinguishing between the normal and the binomial distributions.

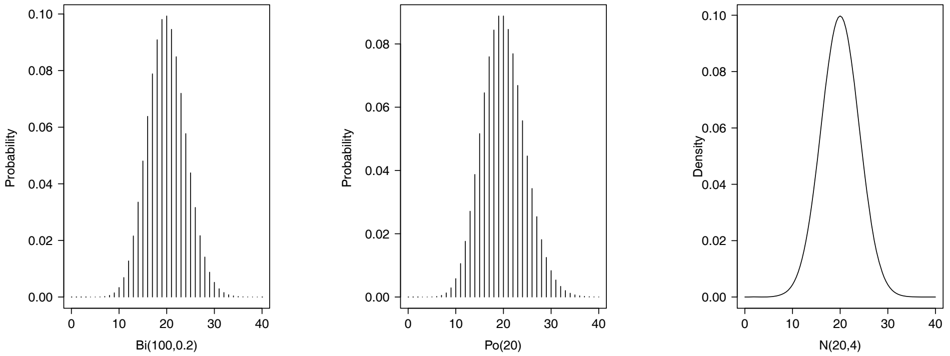


Figure 3: Graphical representation of the probability mass function of the $Bi(100, 0.2)$ distribution and the $Po(20)$ distribution, and probability density function of $N(20, 4)$ distribution.

Table 7: Acceptance percentage of the null hypothesis in the comparison between the Kolmogorov-Smirnov test and the GA-test for various populations. The desired level of significance is 5% and the best result in each case is printed in bold.

	Bi(100,0.2)		Po(20)		N(20,4)	
	GA	K-S	GA	K-S	GA	K-S
Bi(100,0.2)	94.36	**97.56**	**83.46**	96.75	94.94	**86.44**
Po(20)	——	——	94.76	**97.59**	**84.40**	84.95
N(20,4)	——	——	——	——	**95.50**	95.00

A final comparison is carried out with Student's t distributions with several degrees of freedom since these distributions are similar to a standard Normal distribution. The desired level of significance is 0.05, the size of random samples is 100 and the hypothesis test is carried out for 10,000 simulations. The results of the experiment are given in Table 8. Again, the best result in each case is printed in bold, that is, the minimum of the two values when the null hypothesis is false (values outside the diagonal) and the maximum of the two values when the null hypothesis is true (values in the diagonal). It is important to point that the GA-test tends to provide smaller values and therefore tends to accept the null hypothesis less frequently than does the classic test (the classic test therefore tends to be more conservative). As a consequence, the GA-test has a better behaviour when the null hypothesis is false (values outside the diagonal), by differentiating between Student's t distributions with different degrees of freedom better than does the Kolmogorov-Smirnov test, and a worse behaviour (but not much worse) when the null hypothesis is true (values of the diagonal), that is, the Kolmogorov-Smirnov test behaves slightly better under the null hypothesis.

Table 8: Acceptance percentage of the null hypothesis in the comparison between the Kolmogorov-Smirnov test and the GA-test for Student's t distributions. The desired level of significance is 5% and the best result in each case is printed in bold.

	$t(10)$		$t(20)$		$t(30)$		$t(40)$		$t(50)$	
	GA	K-S	GA	K-S	GA	K-S	GA	K-S	GA	K-S
$t(10)$	92.71	**94.56**	**91.03**	94.53	**89.47**	94.55	**88.56**	94.63	**87.78**	94.37
$t(20)$	—	—	94.22	**94.53**	**94.13**	94.85	**93.59**	94.60	**93.97**	94.89
$t(30)$	—	—	—	—	94.53	94.53	**94.33**	94.57	**94.49**	94.84
$t(40)$	—	—	—	—	—	—	94.66	**94.77**	**94.47**	94.57
$t(50)$			—	—	—	—	—	—	**94.66**	94.42

7. Conclusions and Future Work

Several similarity measures between one-dimensional data sets have been developed which can be employed to compare data sets, and a new hypothesis test has been designed. Two comparisons of this test with other classic tests have been made under the null hypothesis that two populations are identical. The main conclusion is that the new test performs reasonably well in comparison with the classic tests considered, and, in certain circumstances, performs even better than said classic tests.

With the distance developed in this paper, various classifications of a data set can be carried out, either by applying the neural network technique, SVM, or via other procedures available.

Although there are other approaches to the choice of the set Q of the percentiles for the $K_{\mathbf{W}}^{Q}$ function from a data set X, such as for example the equal-width interval (Chiu et al. 1991), k-mean clustering (Hartigan 1975), cumulative roots of frequency (González & Gavilan 2001), Ameva (González-Abril, Cuberos, Velasco & Ortega 2009), and the maximum entropy marginal approach (Wong & Chiu 1987), these have not been considered in this paper and will be studied in future papers.

Only the one-dimensional setting is considered in this paper; the possible correlations that can exist between features of multi-dimensional data sets lie outside the scope of this paper and will constitute the focus of study in future work.

Another potential line of research involves the improvement of the design of our hypothesis-testing procedures by using these similarity measures, and the execution of comparisons with other existing methods. For example, the chi-squared test on quantiled bins, or the Wald-Wolfowitz runs test can be tested under the null hypothesis that the two samples come from identical distributions.

Acknowledgements

This research has been partly supported by the Andalusian Regional Ministry of Economy project Simon (TIC-8052).

[Recibido: julio de 2013 — Aceptado: enero de 2014]

References

Anguita, D., Ridella, S. & Sterpi, D. (2004), A new method for multiclass support vector machines, *in* 'Proceedings of the IEEE IJCNN2004', Budapest, Hungary.

Bach, F. R. & Jordan, M. I. (2003), 'Kernel independent component analysis', *Journal of Machine Learning Research* **3**, 1–48.

Bache, K. & Lichman, M. (2013), 'UCI machine learning repository', http://archive.ics.uci.edu/ml, University of California, Irvine, School of Information and Computer Sciences.

Burrell, Q. L. (2005), 'Measuring similarity of concentration between different informetric distributions: Two new approaches', *Journal of the American Society for Information Science and Technology* **56**(7), 704–714.

Chiu, D., Wong, A. & Cheung, B. (1991), Information discovery through hierarchical maximum entropy discretization and synthesis, *in* G. Piatetsky-Shapiro & W. J. Frawley, eds, 'Knowledge Discovery in Databases', MIT Press, pp. 125–140.

Cristianini, N. & Shawe-Taylor, J. (2000), *An introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University press.

González-Abril, L., Cuberos, F. J., Velasco, F. & Ortega, J. A. (2009), 'Ameva: An autonomous discretization algorithm', *Expert Systems with Applications* **36**(3), 5327 – 5332.

González-Abril, L., Velasco, F., Gavilán, J. & Sánchez-Reyes, L. (2010), 'The similarity between the square of the coeficient of variation and the Gini index of a general random variable', *Revista de métodos cuantitativos para la economía y la empresa* **10**, 5–18.

González, L. & Gavilan, J. M. (2001), Una metodología para la construcción de histogramas. Aplicación a los ingresos de los hogares andaluces, *in* 'XIV Reunión ASEPELT-Spain'.

González, L., Velasco, F., Angulo, C., Ortega, J. & Ruiz, F. (2004), 'Sobre núcleos, distancias y similitudes entre intervalos', *Inteligencia Artificial* **8**(23), 113–119.

González, L., Velasco, F. & Gasca, R. (2005), 'A study of the similarities between topics', *Computational Statistics* **20**(3), 465–479.

Hartigan, J. (1975), *Clustering Algorithms*, Wiley, New York.

Hsu, C.-W. & Lin, C.-J. (2002), 'A comparison of methods for multiclass support vector machine', *IEEE Transactions on Neural Networks* **13**(2), 415–425.

Lee, J., Kim, M. & Lee, Y. (1993), 'Information retrieval based on conceptual distance in is-a hierarchies', *Journal of Documentation* **49**(2), 188–207.

Lin, D. (1998), An information-theoretic definition of similarity, *in* 'Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998)', pp. 296–304.

Nielsen, J., Ghugre, N. & Panigrahy, A. (2004), 'Affine and polynomial mutual information coregistration for artifact elimination in diffusion tensor imaging of newborns', *Magnetic Resonance Imaging* **22**, 1319–1323.

Parthasarathy, S. & Ogihara, M. (2000), 'Exploiting dataset similarity for distributed mining', http://ipdps.eece.unm.edu/2000/datamine/18000400.pdf.

Rada, R., Mili, H., Bicknell, E. & Blettner, M. (1989), 'Development and application of a metric on semantic nets', *IEEE Transaction on Systems, Man, and Cybernetics* **19**(1), 17–30.

Salazar, D. A., Vélez, J. I. & Salazar, J. C. (2012), 'Comparison between SVM and logistic regression: Which one is better to discriminate?', *Revista Colombiana de Estadística* **35, 2**, 223–237.

Sheridan, R., Feuston, B., Maiorov, V. & Kearsley, S. (2004), 'Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR', *Journal of Chemical Information and Modeling* **44**, 1912–1928.

Skhölkopf, B. & Smola, A. J. (2002), *Learning with Kernel*, MIT Press.

Vapnik, V. (1998), *Statistical Learning Theory*, John Wiley & Sons, Inc.

Wong, A. & Chiu, D. (1987), 'Synthesizing statistical knowledge from incomplete mixed-mode data', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **9**(6), 796–805.

Appendix. Data Sets of Section 2

The DS_1 and DS_2 data sets are taken from a N(1,1) distribution, the DS_3 data set from a N(0.5,1) distribution, and the DS_4 data set from a N(1.5, 1.25) distribution.

$DS_1 = \{1.47, 0.01, 1.29, -0.27, -0.23, 0.54, -0.20, 3.54, 0.59, -0.03, 2.41, -0.08, 1.48, 1.02, 0.71, 3.40, 0.67, 0.74, 1.64, 2.41, 1.67, 0.74, -0.10, 1.76, 1.82, -1.05, 0.54, 1.20\}$
$DS_2 = \{-0.29, 0.02, 0.84, 1.66, 1.04, 0.69, 2.04, 1.21, 1.71, 1.75, 1.64, 1.10, -1.46, 1.25, -0.10, 1.74, 3.06, -0.53, 0.84, 1.09, 1.26, -0.39, 0.88, 2.15, 1.59, 0.56, 0.37, 3.57\}$
$DS_3 = \{0.22, 0.87, -0.11, 0.29, -0.93, -0.25, 2.05, -0.53, -0.51, 0.80, 0.65, 0.99, 1.28, 0.85, 0.00, -0.28, 0.55, 0.27, -0.68, 1.08, 1.20, 0.44, 0.20, 0.66, 0.29, -0.46, 1.02, 1.99\}$
$DS_4 = \{1.81, -0.41, 1.25, 3.12, 1.91, 1.99, 1.75, 0.93, -0.39, 3.68, -0.69, 1.57, 1.48, 3.59, 0.60, 2.84, 0.37, 1.26, 1.94, -0.19, 1.77, 3.20, 1.11, 4.24, 0.16, 4.48, 0.98, 1.34\}$

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 95 a 110

Locally D-Optimal Designs with Heteroscedasticity: A Comparison between Two Methodologies

Diseños D-óptimos locales con heterocedasticidad: una comparación entre dos metodologías

Jaime Andrés Gaviria[a], Víctor Ignacio López-Ríos[b]

Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Medellín, Colombia

Abstract

The classic theory of optimal experimental designs assumes that the errors of the model are independent and have a normal distribution with constant variance. However, the assumption of homogeneity of variance is not always satisfied. For example when the variability of the response is a function of the mean, it is probably that a heterogeneity model be more adequate than a homogeneous one. To solve this problem there are two methods: The first one consists of incorporating a function which models the error variance in the model, the second one is to apply some of the Box-Cox transformations to both sides on the nonlinear regression model to achieve a homoscedastic model (Carroll & Ruppert 1988, Chapter 4). In both cases it is possible to find the optimal design but the problem becomes more complex because it is necessary to find an expression for the Fisher information matrix of the model. In this paper we present the two mentioned methodologies for the D-optimality criteria and we show a result which is useful to find D-optimal designs for heteroscedastic models when the variance of the response is a function of the mean. Then we apply both methods with an example, where the model is nonlinear and the variance is not constant. Finally we find the D-optimal designs with each methodology, calculate the efficiencies and evaluate the goodness of fit of the obtained designs via simulations.

***Key words*:** D-efficiency, D-optimal design, Box-Cox transformations, Heteroscedasticity.

[a] MSc in Statistics. E-mail: jagaviriab@unal.edu.co
[b] Associate professor. E-mail: vilopez@unal.edu.co

Resumen

La teoría clásica de los diseños experimentales óptimos supone que los errores del modelo son independientes y tienen una distribución normal con varianza constante. Sin embargo, el supuesto de homogeneidad de varianza no siempre se satisface. Por ejemplo, cuando la variabilidad de la respuesta es una función de la media, es probable que un modelo heterocedástico sea más adecuado que uno homogéneo. Para solucionar este problema hay dos métodos: el primero consiste en incorporar una función que modele la varianza del error en el modelo; el segundo consiste en aplicar alguna de las transformaciones de Box-Cox en el modelo de regresión no lineal (Carroll & Ruppert 1988, Capítulo 4). En ambos casos es posible hallar el diseño óptimo, pero el problema se vuelve más complejo porque es necesario encontrar una expresión de la matriz de información de Fisher del modelo. En este artículo se presentan las dos metodologías mencionadas para el criterio D-optimalidad y se muestra un resultado que es útil para encontrar diseños D-óptimos para modelos heterocedásticos cuando la varianza de la respuesta es una función de la media. Luego, se aplican ambos métodos en un ejemplo donde el modelo es no lineal y la varianza no constante. Finalmente se encuentra el diseño D-óptimo con cada metodología, se calculan las eficiencias y se evalúa la bondad del ajuste de los diseños obtenidos a través de simulaciones.

***Palabras clave*:** D-eficiencia, Diseños D-óptimos, heterocedasticidad, transformación de Box-Cox.

1. Introduction

The optimal experimental designs are a tool that allows the researcher to know which factor levels should be experimented in order to obtain a best estimate of the parameters of the model with certain statistical criterion. One of the most popular criteria is the D-optimality which involves finding the design that minimizes the generalized variance of the parameter vector. The design depends on a regression model (1) that relates the response variable Y with the independent variable x

$$Y = \eta(x, \boldsymbol{\beta}) + \epsilon \tag{1}$$

with $\eta(x, \boldsymbol{\beta})$ a linear or nonlinear function of the parameter vector $\boldsymbol{\beta}$ and x.

Besides, if the researcher has the possibility to run N observations of the model (1), then there are the following assumptions:

1. the error components ϵ_i, for $i = 1, 2, \ldots, N$, are independent and
2. have a normal distribution with constant variance σ^2.

For more information about the classic theory of optimal designs see Kiefer (1959), O'Brien & Funk (2003), Atkinson, Donev & Tobias (2007, Chapter 9), López & Ramos (2007). However, in practice there are cases where the homogeneity assumption is not satisfied. For example when the variance of the response is a function of the mean, it can increase or decrease depending of the structure of the variance. The issue of heteroscedasticity in nonlinear regression models is discussed in detail in Seber & Wild (1989, pp. 68-72). Basically there are two methodologies to handle this problem. The first one is to apply some of the Box-Cox transformations to the model (1) with an appropriate λ_1 that stabilizes the variance of the errors. We identify the transformed model like model A:

$$Y_i^{\lambda_1} = \eta(x_i, \boldsymbol{\beta})^{\lambda_1} + \epsilon_i^* \tag{2}$$

where is assumed that the new errors ϵ_i^* have a normal distribution with constant variance.

The second model, which we identify as model B, consists of incorporating the variance structure of the errors in the model as follows:

$$Y_i = \eta(x_i, \boldsymbol{\beta}) + \epsilon_i \tag{3}$$

where the errors ϵ_i are independent $N(0, \sigma^2(\eta(x_i, \boldsymbol{\beta}))^{\lambda_2})$, with λ_2 an adequate power parameter that models the variance of the errors.

As Seber & Wild (1989) emphasize, the difference between models A and B is "that model A transforms so that $y^{(\lambda_1)}$ has a different distribution from y as well as having a homogeneous variance structure, while model B models the variance heterogeneity but leaves the distribution of y unchanged". Also, the authors affirm that model B has often been preferred to model A when the deterministic function is linear, whereas models like A have been preferred in nonlinear models.

Now, in the context of optimal designs when the model has heteroscedasticity, the problem to find D-optimal designs is more complicated than in the homogeneous case, because the D-optimality criterion maximizes the determinant of the Fisher information matrix of the model and the expression of this matrix changes when the variance is not constant. Because the information matrix depends of the model used, the two methodologies mentioned before for handling of heteroscedasticity are traditionally applied in separate ways. For example, Atkinson & Cook (1997) apply some of the Box-Cox transformations that makes the transformed model be linear with a constant variance and then they find local and Bayesian D-optimal designs to several models. On the other hand, in the case of linear models, Atkinson & Cook (1995) find local D-optimal designs for heteroscedastic linear models for various structures of variance, one of them is when the logarithm of variance is a linear function of the independent variable. Other authors have worked with nonlinear models, see for example Dette & Wong (1999).

In this paper we compare the methodologies mentioned above, analyze the structure of the information matrix and we find the D-optimal design for a specific model. Finally we compare the designs obtained through the D- efficiency. This paper is divided in four sections. In section 2 we present a brief summary of both methodologies for the D-optimality criterion and show a result which is useful to find D-optimal designs for heteroscedastic models when the variance of the response is a function of the mean (we omit the proof due to length constraints). In section 3 we illustrate both methods with an example and we compare results using the D-efficiency of each design. Then, we simulate observations of the model for each design and we calculate the relative error and mean square error. Finally, in section 4 we present some conclusions, discussions and suggestions.

2. Methodologies

Starting with a regression model of the form (1) with the usual assumptions, the problem of optimal designs consists to find the levels of the x factor where the researcher should experiment to obtain a best estimate of the parameters of the model under certain statistical criterion. In this paper we focus on the D-optimality criterion, which finds the design that minimizes the generalized variance of the parameter vector (Atkinson et al. 2007, pp. 135). More precisely, a design ξ is defined as a measure of probability with finite support denoted by:

$$\xi = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ w_1 & w_2 & \cdots & w_n \end{bmatrix} \tag{4}$$

where n is the number of support points, $x_1, x_2, \dots, x_n$ are the support points of the design with associated weights $w_i \geq 0$ and such that $\sum_{i=1}^{n} w_i = 1$ (O'Brien & Funk 2003). If the weight w_i is any number between 0 and 1, the design ξ is known as a continuous design. However in practice all designs are exact. This means that the weights w_i are associated with the frequency of the support points (Atkinson et al. 2007, pp. 120).

Now, the main problem of optimal designs is to find a design ξ over a compact region χ, that maximizes a functional of the information matrix $M(\xi)$. This matrix plays an important role in the theory of optimal experimental designs. The structure of this matrix depends on the linear nature of the model and on the assumptions about the errors. When the variance of the errors is constant, this matrix has a known expression, see for example López & Ramos (2007). However, in the case of heteroscedastic models this expression is more complex and depends of the methodology applied. So, in the next two sections we analyze the structure of the Fisher information matrix with each of the two methodologies mentioned before.

2.1. Variance Modelling

When the variance of the error is not constant, one way to solve this problem is to find an adequate function which models the error variance and incorporate it in the regression model. There are many ways to do this, see for example Huet, Bouvier, Poursat & Jolivet (2004, pp. 65) and Seber & Wild (1989, pp. 68-72). One form is when the variance of the response is a power function of the mean:

$$Y_i = \eta(x_i, \boldsymbol{\beta}) + \epsilon_i, \text{ with } var(\epsilon_i) = \sigma^2(\eta(x_i, \boldsymbol{\beta}))^{2\tau} \tag{5}$$

where σ^2 is the constant variance, τ is an unknown parameter and it should be estimated. The model (5) with variance structure is known as the power of the mean variance model (Ritz & Streibig 2008, pp. 74).

Now, some authors have worked the problem to find D-optimal designs modeling the variance. For example Dette & Wong (1999) find D-optimal designs for the Michaelis-Menten model when the variance is a function of the mean and Atkinson & Cook (1995) find D-optimal designs for heteroscedastic linear models. The following result is taken from Downing, Fedorov & Leonov (2001); they show the expression of the information matrix for a more general model than the power of the mean variance model (5):

$$Y = \eta(x, \boldsymbol{\theta}) + \epsilon, \text{ with } var(\epsilon) = S(x, \boldsymbol{\theta}) \tag{6}$$

where $\boldsymbol{\theta}$ is the parameter vector and it can include the parameters of the deterministic function η and those of the function $S(x, \boldsymbol{\theta})$ as a positive function used to model the variance of the error. Observe that the power of the mean variance model (5) is a nested model of the more general model (6). In this case the parameter vector $\boldsymbol{\theta}$ includes all the possible parameters of the model: $\boldsymbol{\beta}, \tau$ and σ^2. So, results about the general model (for instance the next theorem)can be applied in particular for the power of the mean variance model.

Theorem 1. *Information Matrix.*

Let Y with normal distribution, with expected mean $E[Y|x] = \eta(x, \boldsymbol{\theta})$ and variance $Var[Y|x] = S(x, \boldsymbol{\theta})$, where $S(x, \boldsymbol{\theta}) > 0$ is a positive function, $\boldsymbol{\theta}_{q\times q}$ is the parameter vector and χ a compact set. If the N observations $\{y_i, x_i\}_{i=1}^{N}$ are independent, then the Fisher information matrix for the approximate design ξ over the regression design χ is

$$\boldsymbol{M}(\xi, \boldsymbol{\theta})_{q\times q} = \int_{\chi} \boldsymbol{I}(x, \boldsymbol{\theta}) d\xi(x) \tag{7}$$

where

$$I(x, \boldsymbol{\theta})_{q\times q} = \frac{1}{S(x,\boldsymbol{\theta})} \frac{\partial \eta(x,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \eta(x,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} + \frac{1}{2} \frac{1}{S(x,\boldsymbol{\theta})^2} \frac{\partial S(x,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial S(x,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \tag{8}$$

This theorem is the main tool of this methodology, because it allows the researcher many ways of modelling the variance and incorporate it in the model.

Corollary 1. *For the power of the mean variance model given in (1), where the errors are independent and have normal distribution with mean zero and variance $var(\epsilon_i) = \sigma^2(\eta(x_i, \boldsymbol{\beta}))^{2\tau}$ with $\boldsymbol{\beta}, \tau$ and σ^2 parameters, the information matrix is given by*

$$\boldsymbol{M}(\xi, \boldsymbol{\theta}) = \boldsymbol{U}\boldsymbol{W}\boldsymbol{U}^T + \boldsymbol{V}\boldsymbol{W}\boldsymbol{V}^T \tag{9}$$

where

$$\boldsymbol{U}_{(p+2)\times n} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n) \quad \boldsymbol{V}_{(p+2)\times n} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n) \tag{10}$$

$$\boldsymbol{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & & \\ \vdots & & \ddots & \\ 0 & & & w_r \end{pmatrix} \tag{11}$$

and for $i = 1, 2, \ldots, n$:

$$\boldsymbol{u}_i = \left(\frac{1}{\sigma\eta(x_i, \boldsymbol{\beta})^{\tau}} \frac{\partial \eta(x_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}, 0.0 \right)^T_{(p+2)\times 1} \tag{12}$$

$$\boldsymbol{v}_i = \left(\frac{\sqrt{2}\tau}{\eta(x_i, \boldsymbol{\beta})} \frac{\partial \eta(x_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}, \sqrt{2} \log \eta(x_i, \beta), \frac{1}{\sqrt{2}\sigma^2} \right)^T_{(p+2)\times 1} \tag{13}$$

This result is the key at the construction of D-optimal designs and can be implemented computationally to obtain the designs. We will illustrate the use of this corollary with an application in the next section. But before we need the following important result which is one of the equivalence theorems. This theorem allows to verify if the obtained design is in fact the optimal design (Kiefer & Wolfowitz 1960)

Theorem 2. *D-optimality equivalence theorem.*

Let $\boldsymbol{M}(\xi, \boldsymbol{\theta})_{q\times q}$ the information matrix of the design ξ positive, $\Psi(\xi, \boldsymbol{\theta}) = \log |\boldsymbol{M}(\xi, \boldsymbol{\theta})|$ the D-optimality criterion and χ a compact set. Then the design ξ^* is D-optimal if the directional derivative of ϕ in ξ^* on the direction of ξ_x holds

$$\phi(M(\xi^*, \boldsymbol{\theta}), M(\xi_x, \boldsymbol{\theta})) \leq 0 \quad \forall x \in \chi \tag{14}$$

where $\phi(\boldsymbol{M}(\xi^*, \boldsymbol{\theta}), M(\xi_x, \boldsymbol{\theta})) = Tr(\boldsymbol{M}(\xi_x)\boldsymbol{M}^{-1}(\xi^*)) - q$ and ξ_x is the design that puts all probability in x. Also, $\phi(\boldsymbol{M}(\xi^*), \boldsymbol{M}(\xi_x)) = 0$ at the support points of design ξ^*.

This result is useful to verify the D-optimality of a design ξ^*, because one can plot the directional derivative $\phi(M(\xi^*, \theta), M(\xi_x, \theta))$ over $x \in \chi$ and to check that this function at most zero over all experimental region (χ) and also that in the support points of the design, the equality holds.

2.2. Transformation of the Model

The second methodology consists of applying an adequate transformation on the model to obtain constant variance. We focus on the Box-Cox transformations, which are given by Box & Cox (1964).

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{for } \lambda \neq 0 \\ \log y & \text{for } \lambda = 0 \end{cases} \tag{15}$$

The value of the parameter λ usually is unknown, but in some cases it can be assessed depending on the response. For instance, if the response is a volume, the appropriate transformation can be the cube root ($\lambda = 1/3$) and the square root if the response corresponds to count data (Atkinson & Cook 1997).

Now, Atkinson & Cook (1997) find D-optimal designs when a Box-Cox transformation is applied, the resulting model is linear

$$\boldsymbol{Y}^{(\lambda)} = f^T(x)\boldsymbol{\beta} + \epsilon^* \tag{16}$$

and the errors have normal distribution with constant variance $\epsilon^* \sim N(0, \sigma^2)$.

However, as illustrated in the example and since the original model is nonlinear, we must find some appropriate λ such that when we apply the transformation to both sides of the model, the transformed model is linear in the parameters. It is important to observe that in our case, the parameter λ will be known, which is an advantage, because we do not need to estimate this parameter. However, when the parameter λ is unknown, it is possible to find the design, see for example Atkinson (2003) for more details.

Then, the authors show that the information matrix over the design region for the transformed model is (see the details in Atkinson & Cook 1997)

$$\boldsymbol{M}(\xi, \theta) = \int_{\chi} \boldsymbol{I}(\theta) \xi(dx) \tag{17}$$

where the symmetric matrix $I(\theta)$ is given by

$$\begin{aligned} I(\theta) &= -E\left[\frac{\partial^2 \log f(\boldsymbol{Y}_i | x_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right] \\ &= \begin{pmatrix} f f^T & 0 & -\frac{f E(\dot{Y}^{(\lambda)})}{\sigma^2} \\ 0 & \frac{1}{2\sigma^4} & -\frac{E(\epsilon^* \dot{Y}^{(\lambda)})}{\sigma^4} \\ -\frac{f E(\dot{Y}^{(\lambda)})}{\sigma^2} & -\frac{E(\epsilon^* \dot{Y}^{(\lambda)})}{\sigma^4} & \frac{E(\dot{Y}^{(\lambda)})^2 + E(\epsilon^* \ddot{Y}^{(\lambda)})}{\sigma^2} \end{pmatrix} \end{aligned} \tag{18}$$

with $\epsilon^* = Y^{(\lambda)} - f^T(x)\beta, f = f(x)$ and $\dot{Y}^{(\lambda)}, \ddot{Y}^{(\lambda)}$ denote the first and second derivative respect to λ and are given by:

$$\dot{Y}^{(\lambda)} = \frac{Y^\lambda \log Y^\lambda - Y^\lambda + 1}{\lambda^2} \quad \text{and} \tag{19}$$

$$\ddot{Y}^{(\lambda)} = \frac{Y^\lambda (\log Y^\lambda - 1)^2 + Y^\lambda - 2}{\lambda^3} \tag{20}$$

However, these expressions have to be approximated using first-order Taylor approximations, since the expected values can not be calculated exactly. Finally, once the design is found using the above expressions, is necessary verify the D-optimality of the design using a similar result of the equivalence theorem 2.

3. Example

In Section 2 we described the two methodologies commonly used to handle the heteroscedasticity of a model. Now we illustrate these methods with one example.

3.1. PCB Model

The example consists of a study realized in 1972 in Lake Cayuga, New York, where the concentrations of *Polychlorinated biphenyls* (PCB) were made in a group of 28 trout at several ages in years. "The ages of the fish were accurately known because the fish are annually stocked as yearlings and distinctly marked as to year class" (Bates & Watts 1988, pp. 267–268). The data taken from Bates & Watts (1988), are shown in the table 1 and the scatter plot is shown in figure 1.

TABLE 1: Lake Cayuga data.

Age	1.00	1.00	1.00	1.00	2.00	2.00	2.00
Concentration	0.60	1.60	0.50	1.20	2.00	1.30	2.50
Age	3.00	3.00	3.00	4.00	4.00	4.00	5.00
Concentration	2.20	2.40	1.20	3.50	4.10	5.10	5.70
Age	6.00	6.00	6.00	7.00	7.00	7.00	8.00
Concentration	3.40	9.70	8.60	4.00	5.50	10.50	17.50
Age	8.00	8.00	9.00	11.00	12.00	12.00	12.00
Concentration	13.40	4.50	30.40	12.40	13.40	26.20	7.40

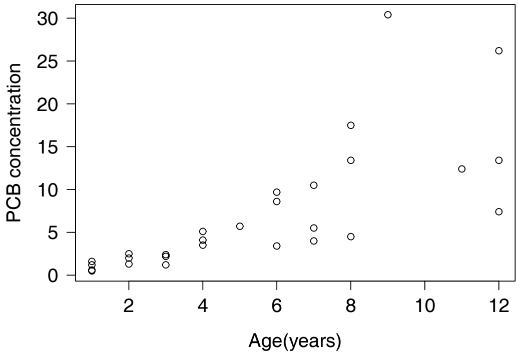


FIGURE 1: Scatter plot of Lake Cayuga data.

The plot of the data shows that the concentration of *Polychlorinated biphenyls* (PCB) increases when the age of the trout does. Also, the relationship between the variables clearly is not linear, so we propose to fit the nonlinear model:

$$Y = \beta_1 e^{\beta_2 x} + \epsilon \tag{21}$$

with β_1, β_2 are unknown parameters to be estimated.

Now, we are going to find the D optimal design for this model with the two methodologies described above. Because our designs are local, we use the data only with the purpose to have a good local value of the parameter vector.

3.1.1. Variance Modelling

First, we apply the methodology consisting on modelling the variance of the errors with an appropriate function. In figure 1, we see that the variability of the concentration increases as a power function of the mean, so we propose to fit the model (21) with variance structure (5)

$$Y = \beta_1 e^{\beta_2 x} + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2(\beta_1 e^{\beta_2 x})^{2\tau}) \tag{22}$$

with τ an unknown parameter to be estimated.

Now, we fit the model (22) in R Development Core Team (2013) and we used the *gnls* function for the generalized nonlinear least squares method. The results of the estimation are showed in table 2.

Table 2: Generalized nonlinear least squares estimation.

Parameter	Estimation
β_1	0.91
β_2	0.31
τ	1.19
σ	0.34

Next, we perform the likelihood ratio test to determine if the model with variance structure (22) is better than the model with constant variance (21). The results of the test are showed in table 3 (the model 1 corresponds to the model with variance structure (22) and the model 2 to the model with constant variance (21)).

Table 3: ANOVA for the likelihood ratio test.

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
(22)	1	4	134.5534	139.8822	-63.27671			
(21)	2	3	178.8002	182.7968	-86.40008	1 vs 2	46.24674	<.0001

The conclusion from this test that is the parameter $\tau \neq 0$, e.g. the model with variance structure (22) is better than the model with constant variance (21) with a signification level of 1%.

3.1.2. *D*-Optimal Design

Now, we find the *D*-optimal design for the model with variance structure (22). Because we work with local designs, we use the estimation of the parameters obtained previously like the local value for θ; that is, we use the local value $\boldsymbol{\theta}_0 = (\beta_1, \beta_2, \tau, \sigma) = (0.91, 0.31, 1.19, 0.34)$. Then we implement the corollary 1 through an algorithm in R Development Core Team (2013) and minimize $-\log(|\boldsymbol{M}(\xi, \boldsymbol{\theta})|)$. In this optimization problem we use the function *nlminb* over the experimental region (χ). The local D-optimal design obtained is shown in table 4 and is denoted by ξ_D. The x_i are the support points of the design and the w_i the weights. As we can see, even though the model with variance structure (22) has four parameters to be estimated, the design consists only of two points, which are the extreme points of the regression range $\chi = [1, 12]$. In this sense, if we could repeat the experiment and our objective are to estimate the parameters with minimum variance, then we measure the *Polychlorinated biphenyls* concentration in trout with ages of one and twelve and with equal number of replicates.

Then we check that the obtained design ξ_D is *D*-optimal. With this in mind, by the *D*-optimality equivalence in theorem 2, we must verify that the directional derivative of Ψ at ξ_D in the direction of the design that puts all mass at x, ξ_x,

Table 4: Local D-optimal design ξ_D to the model (22).

x_i	1.00	12.00
w_i	0.50	0.50

satisfies

$$\phi(\boldsymbol{M}(\xi_D, \boldsymbol{\theta}), \boldsymbol{M}(\xi_x, \boldsymbol{\theta})) = Tr(\boldsymbol{M}(\xi_x)\boldsymbol{M}^{-1}(\xi_D)) - 4 \leq 0 \tag{23}$$

$\forall x \in \chi = [1, 12]$. As we can see in figure 2, this condition holds and the derivative equals zero at the support points, so the design ξ_D is indeed D-optimal.

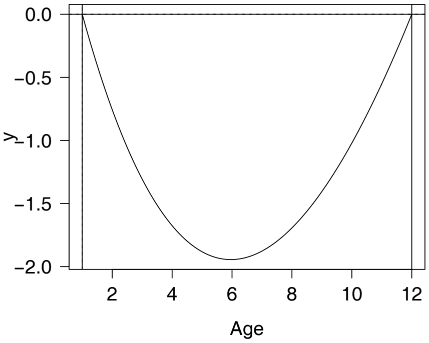


Figure 2: Plot of the directional derivative.

3.1.3. Simulations

We simulate 1,000 times 28 observations of the model with variance structure

$$Y_i = \beta_1 e^{\beta_2 x_i} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2(\beta_1 e^{\beta_2 x_i})^{2\tau}),\ \text{for } i = 1, 2, \dots, 28 \tag{24}$$

taking the values of x_i like the support points of the design ξ_D. Then we simulate the errors $\epsilon_i \sim N(0, \sigma^2(\beta_1 e^{\beta_2})^{2\tau})$ for $i = 1, 2, \dots, 28$; use the estimations obtained in table 2 like the values of the parameters and with the model (24), we calculate the response $y_i's$. Then, with these simulated data, we obtain the estimated parameter $\hat{\theta}$ and calculate the relative and mean square error (RE and MSE respectively). We repeat this process 1,000 times and summarize it in table 5, showing the descriptive measures for both errors. This table shows the mean, median, range and standard deviation for the MSE of each parameter of the model (24) and the relative error in percentage $RE(\theta) \times 100\%$. For the parameter vector θ we propose an overall discrepancy measure, ODM, defined as $ODM(\hat{\theta}) = ||\theta - \hat{\theta}||^2$. From this table, we see that the central tendency measures for the MSE are small as the variability between the simulations. Also, the mean and median for the RE are very close to 10%. In general, all these measures indicate that the local design ξ_D provides good parameter estimates, even though the design only has two experimental points and the model four parameters.

Table 5: Simulations with variance modelling (Std denotes the Standard deviation).

	$MSE(\beta_1)$	$MSE(\beta_2)$	$MSE(\tau)$	$MSE(\sigma)$	$ODM(\theta)$	$RE(\theta)\%$
Mean	9.08e-03	3.13e-04	1.08e-02	5.97e-03	2.61e-02	9.31
Median	4.28e-03	1.50e-04	5.40e-03	2.68e-03	1.86e-02	8.71
Range	9.60e-02	5.78e-03	1.15e-01	1.09e-01	2.01e-01	27.60
Std	1.28e-02	4.61e-04	1.45e-02	8.84e-03	2.56e-02	4.45

3.1.4. Efficiencies

Finally, we show the robustness of the design ξ_D with respect to the choice of the local value θ_0, through the D-efficiency of any design ξ:

$$D_{\text{eff}} = \left(\frac{|\boldsymbol{M}(\xi)|}{|\boldsymbol{M}(\xi_D)|} \right)^{1/p} \tag{25}$$

where p is the number of parameters of the model and $\boldsymbol{M}(\xi)$ denotes the information matrix of the design, where ξ is another design obtained with another local values of parameter vector. With this in mind, we perturb each one of the four parameters of the model (22) in a percentage Δ:

$$\theta_i \pm \Delta \times \theta_i \tag{26}$$

Since the model has four parameters and each one can be perturbed at left, at right or not be perturbed; it is clear that the total number of perturbations is $3^4 = 81$. Then each one of these perturbations will give us a design ξ and with (25) we calculate how far we are of the local D-optimal design. Then for a fixed $\Delta = 0.6$ (we could used another), we obtain 81 designs and for each one we calculate the respective D-efficiency. However, because most of these designs were equal to the two point design ξ_D, we only show in table 9 (see the appendix) the support points, the weights and the D-efficiency of the 36 designs that were different to the optimal. Figure 3 summarizes the results of the efficiencies and shows that the design ξ_D is robust respect the choice of the local value θ_0, because the D-efficiencies are high (at least 0.80).

3.2. Transformation of the Model

Previously we apply the first methodology of variance modelling and find the local D-optimal design. Now we use the second methodology, that consists on applying an adequate transformation on the model. As we described in section 2.2, this transformation should be such that the transformed model is linear and homoscedastic. In this case as the model (1) is exponential, the appropriate Box-Cox transformation consists on applying logarithm to both sides:

$$\log Y = \log \beta_1 + \beta_2 x + \epsilon^* \tag{27}$$

or equivalently in the form:

$$Y^* = \beta_1^* + \beta_2 x + \epsilon^* \tag{28}$$

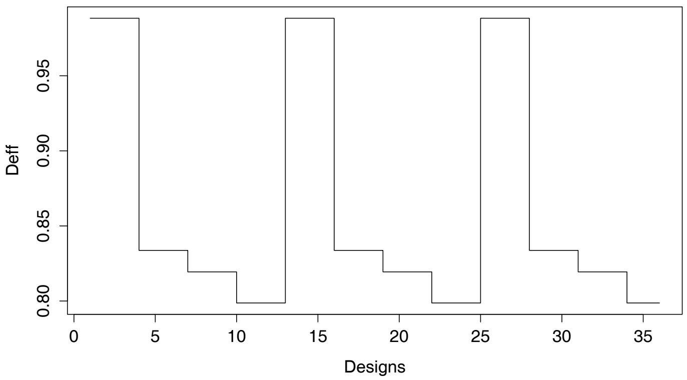


Figure 3: D-efficiencies perturbing θ in 60%.

where $Y^* = \log Y$, $\beta_1^* = \log \beta_1$ and the new errors ϵ^* are normal with constant variance. Then we fitted the linear model (28) and we obtained the estimations $\hat{\boldsymbol{\beta}}^* = (0.03, 0.26)^T$, $\widehat{\sigma} = 0.57$, and then $\hat{\boldsymbol{\beta}} = (e^{0.03}, 0.26)^T = (1.03, 0.26)^T$. Finally, we implemented an algorithm in R Development Core Team (2013) to find the information matrix with $\lambda = 0$ and to obtain the design that minimizes $-\log \boldsymbol{M}(\xi, \boldsymbol{\theta})$ over $\chi = [1, 12]$. The resulting design in table 6, shows that in this case the design is the same obtained with first methodology. However, we have to point out that despite that the resulting design is the same with both methodologies, it is attributed to the fact that with each method we used the best local value for the parameter $\boldsymbol{\theta}$ and as we saw when we calculate the D-efficiencies, the design can have three support points depending on the local value used.

Table 6: D-optimal design to the model (27).

i	1	2
x_i	1.00	12.00
w_i	0.50	0.50

3.2.1. Simulations

Analogously to the first methodology, we simulated 28 observations of the model (28). The results of the 1,000 simulations are summarized in table 7. This is similar to the table 5 and shows the mean, median, range and standard deviation for the MSE of each parameter of the model (27) and relative error in percentage $RE(\theta) \times 100\%$. For the parameter vector $\boldsymbol{\theta}$ we propose a measure defined as $ODM(\hat{\boldsymbol{\theta}}) = ||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}||^2$, which is a kind of square distance between the estimated parameter and the original. The conclusions from these results are similar as the obtained with the first methodology, although when we compare the measures for the relative error (RE), is noteworthy that all the descriptive measures are almost three times the correspondent to the first methodology. But in general, all these measures indicate that the local design ξ_D fits well the model.

Table 7: Simulations for the logarithmic transformation model (Std denotes the Standard deviation).

	$MSE(\beta_1)$	$MSE(\beta_2)$	$MSE(\sigma)$	$ODM(\theta)$	$RE(\theta)\%$
Mean	1.80e-01	9.57e-04	1.27e-02	1.93e-01	34.0
Median	1.39e-01	6.48e-04	8.43e-03	1.53e-01	32.4
Range	1.18e+00	8.65e-03	9.34e-02	1.18e+00	84.0
Std	1.55e-01	1.00e-03	1.36e-02	1.53e-01	13.1

3.2.2. Efficiencies

Finally, we obtain the D-efficiencies following the same procedure described in section 3.1.4. In this case because we perturb three parameters: β_1, β_2 and σ, we only have $3^3 = 27$ combinations (the parameter $\lambda = 0$). But again most of all these designs were equal to the D-optimal design, so we only show in table 8 the six designs that correspond to a perturbation $\Delta = 60\%$ and were different to the optimal. In this table we use the symbols $-$, $+$ or 0 to indicate the specific combinations of the parameters.

For instance, the first design is obtained when we disturb 60% to the left ($-$) the parameters β_1 and σ and we do not perturb (0) the parameter β_2. Then the support points for this design are 1, 6.5 and 12 and the D-efficiency of the design is 0.93. It indicates that if we use this design instead of the unperturbed *D*-optimal design, we would need around 7% more observations to obtain the same efficiency that the *D*-optimal. Even more, it is remarkable that all six designs have exactly 3 support points: The extremes of the interval $[1, 12]$ and the middle point 6.5. The only difference between these designs is the weight (in parentheses with two decimal places) and the *D*-efficiency, that can be 0.89 or 0.93, but in both cases it is high, so we can conclude that the *D*-optimal design is robust respect to the choice of the local value θ_0.

Table 8: Support points, weights and *D*-efficiencies perturbing 60% to left ($-$), right ($+$) or not (0).

Design	β_1	β_2	σ	x_1	x_2	x_3	D_{eff}
1	$-$	0	$-$	1(0.40)	6.5(0.20)	12(0.40)	0.93
2	0	0	$-$	1(0.40)	6.5(0.20)	12(0.40)	0.93
3	$+$	0	$-$	1(0.40)	6.5(0.20)	12(0.40)	0.93
4	$-$	$+$	$-$	1(0.36)	6.5(0.28)	12(0.36)	0.89
5	0	$+$	$-$	1(0.36)	6.5(0.28)	12(0.36)	0.89
6	$+$	$+$	$-$	1(0.36)	6.5(0.28)	12(0.36)	0.89

4. Conclusions

We have presented a brief summary of two methodologies that can be implemented to find *D*-optimal designs when the model under study presents heteroscedasticity. In both cases the main problem is to find an expression for the Fisher information matrix of the model. We have illustrated both methods with

Lake Cayuga data from which clearly do not have constant variance. However, there is an important difference between the methods that applied: The variance modelling methodology has the assumption that the errors of the original model has a normal distribution. However, the second methodology only requires a normal distribution for the transformed model, not the original. This can be an advantage of this methodology compared with variance modelling.

Under both methods, we find the same D-optimal design with two support points and with equal weights. But this fact is attributed only to the local values used in an independent way, that in this case were the estimations of the parameter vector using the data. Because the optimal design is local, we determine the robustness of this design with each methodology by disturbing the parameters of the corresponding model and calculating the D-efficiency of the obtained designs. In both cases, the efficiencies were high indicating that the D-optimal design is a robust design respect the choice of the local value θ_0. Also, with each methodology we simulate $1,000$ observations of the model and calculate some descriptive measures for the relative and mean square errors. The results were similar. The only important difference is that measures for the relative errors of the second methodology were almost three times the correspondent to the first methodology. We cannot conclude which methodology is better because each one has its pros and shortcoming, with the example we obtained similar results.

Finally, we want to point out that we have not study two potential problems: First, the problem of heteroscedasticity for G optimality criterion and second, the problem of nonnormality (for D-optimality or not). Respect to the former, further work includes finding optimal designs for heteroscedastic models with another optimality criteria different to D-optimality. For instance, Wong & Cook (1993) have worked with G-optimal designs with linear models when the variance of the errors is incorporated in the model. With non normality, we did not find too many published papers, so this can be an interesting problem to work. Finally we have found local designs, but other option is to use the Bayesian approach.

Table 9: D-efficiencies, support points and weights with a 60% of perturbation of the parameter vector: disturb to left (−), to right (+) or do not (0).

ξ_i	β_1	β_2	τ	σ	x_1	x_2	x_3	w_1	w_2	w_3	D_{eff}
1	−	0	−	−	1	5.6	12	0.48	0.02	0.5	0.9883
2	0	0	−	−	1	5.6	12	0.48	0.02	0.5	0.9883
3	+	0	−	−	1	5.6	12	0.48	0.02	0.5	0.9883
4	−	+	−	−	1	8.26	12	0.27	0.28	0.45	0.8337
5	0	+	−	−	1	8.26	12	0.27	0.28	0.45	0.8337
6	+	+	−	−	1	8.26	12	0.27	0.28	0.45	0.8337
7	−	0	+	−	1	4.45	12	0.44	0.3	0.26	0.8194
8	0	0	+	−	1	4.45	12	0.44	0.3	0.26	0.8194
9	+	0	+	−	1	4.45	12	0.44	0.3	0.26	0.8194
10	−	+	+	−	1	3.12	12	0.42	0.34	0.24	0.7987
11	0	+	+	−	1	3.12	12	0.42	0.34	0.24	0.7987
12	+	+	+	−	1	3.12	12	0.42	0.34	0.24	0.7987
13	−	0	−	0	1	5.6	12	0.48	0.02	0.5	0.9883
14	0	0	−	0	1	5.6	12	0.48	0.02	0.5	0.9883
15	+	0	−	0	1	5.6	12	0.48	0.02	0.5	0.9883
16	−	+	−	0	1	8.26	12	0.27	0.28	0.45	0.8337
17	0	+	−	0	1	8.26	12	0.27	0.28	0.45	0.8337
18	+	+	−	0	1	8.26	12	0.27	0.28	0.45	0.8337
19	−	0	+	0	1	4.45	12	0.44	0.3	0.26	0.8194
20	0	0	+	0	1	4.45	12	0.44	0.3	0.26	0.8194
21	+	0	+	0	1	4.45	12	0.44	0.3	0.26	0.8194
22	−	+	+	0	1	3.12	12	0.42	0.34	0.24	0.7987
23	0	+	+	0	1	3.12	12	0.42	0.34	0.24	0.7987
24	+	+	+	0	1	3.12	12	0.42	0.34	0.24	0.7987
25	−	0	−	+	1	5.6	12	0.48	0.02	0.5	0.9883
26	0	0	−	+	1	5.6	12	0.48	0.02	0.5	0.9883
27	+	0	−	+	1	5.6	12	0.48	0.02	0.5	0.9883
28	−	+	−	+	1	8.26	12	0.27	0.28	0.45	0.8337
29	0	+	−	+	1	8.26	12	0.27	0.28	0.45	0.8337
30	+	+	−	+	1	8.26	12	0.27	0.28	0.45	0.8337
31	−	0	+	+	1	4.45	12	0.44	0.3	0.26	0.8194
32	0	0	+	+	1	4.45	12	0.44	0.3	0.26	0.8194
33	+	0	+	+	1	4.45	12	0.44	0.3	0.26	0.8194
34	−	+	+	+	1	3.12	12	0.42	0.34	0.24	0.7987
35	0	+	+	+	1	3.12	12	0.42	0.34	0.24	0.7987
36	+	+	+	+	1	3.12	12	0.42	0.34	0.24	0.7987

[Recibido: mayo de 2013 — Aceptado: enero de 2014]

References

Atkinson, A. C. (2003), 'Horwitz's rule, transforming both sides and the design of experiments for mechanistic models', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **52**(3), pp. 261–278.

Atkinson, A. C. & Cook, R. D. (1995), 'D-optimum designs for heteroscedastic linear models', *Journal of the American Statistical Association* **90**(429), 204–212.

Atkinson, A. C. & Cook, R. D. (1997), 'Designing for a response transformation parameter', *Journal of the Royal Statistical Society. Series B (Methodological)* **59**(1), 111–124.

Atkinson, A. C., Donev, A. N. & Tobias, R. D. (2007), *Optimum Experimental Designs with SAS*, Oxford Science Publications, New York.

Bates, D. M. & Watts, D. G. (1988), *Nonlinear Regression Analysis and its Applications*, John Wiley and Sons, New York.

Box, G. E. P. & Cox, D. R. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society. Series B (Methodological)* **26**(2), pp. 211–252.

Carroll, R. & Ruppert, D. (1988), *Transformation and Weighting in Regression, Chapter 4*, Taylor & Francis.

Dette, H. & Wong, K. (1999), 'Optimal Designs When the Variance Is a Function of the Mean', *Biometrics* **55**(3), 925–929.

Downing, D., Fedorov, V. & Leonov, S. (2001), *Extracting Information from the Variance Function: Optimal Design*, Springer, Austria.

Huet, S., Bouvier, A., Poursat, M. & Jolivet, E. (2004), *Statistical Tools for Nonlinear Regression: A Practical Guide With S-PLUS and R Examples* , Springer-Verlag, New York.

Kiefer, J. (1959), 'Optimum experimental designs', *Journal of the Royal Statistical Society. Series B (Methodological)* **21**(2), pp. 272–319.

Kiefer, J. & Wolfowitz, J. (1960), 'The equivalence of two extremum problems', *Canadian Journal of Mathematics* **12**(5), pp. 363–365.

López, V. & Ramos, R. (2007), 'Introducción a los diseños óptimos', *Revista Colombiana de Estadística* **30**(1), 37–51.

O'Brien, T. & Funk, G. M. (2003), 'A gentle introduction to optimal design for regression models', *Journal of the American Statistical Association* **57**(4), 265–267.

R Development Core Team (2013), *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

Ritz, C. & Streibig, J. (2008), *Nonlinear Regression with R*, Springer, New York.

Seber, G. & Wild, C. (1989), *Nonlinear Regression*, John Wiley, New York.

Wong, W. K. & Cook, R. D. (1993), 'Heteroscedastic G-optimal designs', *Journal of the Royal Statistical Society. Series B (Methodological)* **55**(4), pp. 871–880.

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 111 a 125

An Iterative Method for Curve Adjustment Based on Optimization of a Variable and its Application

Un método iterativo para el ajuste de curvas basado en la optimización en una variable y su aplicación al caso lineal en una variable independiente

Rogelio Acosta[1,a], Suitberto Cabrera[2,b], Luis Manuel Vega[1,c], Asela Cabrera[2,d], Nersa Acosta[3,e]

[1]Departamento de Matemática, Universidad de Las Tunas, Las Tunas, Cuba

[2]Departamento de Estadística, Investigación Operativa Aplicadas y Calidad, Universitat Politecnica de Valencia, Valencia, Spain

[3]Universidad de las Ciencias Informáticas, La Habana, Cuba

Abstract

An iterative method for the adjustment of curves is obtained by applying the least squares method reiteratively in functional subclasses, each defined by one parameter, after assigning values to the rest of the parameters which determine a previously determined general functional class. To find the minimum of the sum of the squared deviations, in each subclass, only techniques of optimization are used for real functions of a real variable.The value of the parameter which gives the best approximation in an iteration is substituted in the general functional class, to retake the variable character of the following parameter and repeat the process, getting a succession of functions. In the case of simple linear regression, the convergence of that succession to the least squares line is demonstrated, because the values of the parameters that define each approximation coincide with the values of the parameters obtained when applying the method of Gauss - Seidel to the normal system of equations. This approach contributes to the teaching objective of improving the treatment of the essential ideas of curve adjustment, which is a very important topic in applications, what gives major importance to the optimization of variable functions.

***Key words*:** Curve estimation, Iterative method, Least Squares Method, Linear regression, Teaching materials.

[a] Professor. E-mail: racosta@ult.edu.cu
[b] Professor. E-mail: suicabga@eio.upv.es
[c] Professor. E-mail: racosta@ult.edu.cu
[d] Professor. E-mail: ascabul@eio.upv.es
[e] Professor. E-mail: ndacosta@uci.cu

Resumen

Se obtiene un método iterativo para el ajuste de curvas al aplicar reiteradamente el método de los mínimos cuadrados en subclases funcionales, cada una definida por un parámetro, luego de asignar valores a los restantes parámetros que determinan una clase funcional general, seleccionada previamente. Para hallar el mínimo de la suma de las desviaciones cuadráticas, en cada subclase, solo se utilizan técnicas de optimización para funciones reales de una variable real. El valor del parámetro, que proporciona la mejor aproximación en una iteración, se sustituye en la clase funcional general, para retomar el carácter variable del siguiente parámetro y repetir el proceso, obteniéndose una sucesión de funciones. En el caso de la regresión lineal simple se demuestra la convergencia de esa sucesión a la recta mínimo cuadrática, pues coinciden los valores de los parámetros que definen cada aproximación con los que se obtienen al aplicar el método de Gauss - Seidel al sistema normal de ecuaciones. Este enfoque contribuye al objetivo docente de adelantar el tratamiento de las ideas esenciales del ajuste de curvas, temática muy importante en las aplicaciones, lo que le confiere mayor significación a la optimización de funciones de una variable.

***Palabras clave*:** estimación de curvas, materiales de enseñanza, método de mínimos cuadrados, método iterativo, regresión lineal.

1. Introduction

The method of regression is one of the most important statistical methods for higher education graduates. Its comprehension facilitates obtaining and correctly interpreting the results of different types of models to be applied in their professional careers. Bibliographic research in the scientific literature shows the wide interest and use of regression methods. From such bibliographic analysis, three approaches can be differentiated: The largest one, related to the application of regression methods to different fields and topics of science; see Braga, Silveira, Rodríguez, Henrique de Cerqueira, Aparecido & Barros (2009), Guzmán, Bolivar, Alepuz, González & Martin (2011), Ibarra & Arana (2011) and Santos da Silva, Estraviz, Caixeta & Carolina (2006) a second approach related to the theoretical aspects of the topic; see Núñez, Steyerberg & Núñez (2011), Vega-Vilca & Guzmán (2011), Donal (2001), Ranganatham (2004), Kelley (1999), Schmidt (2005). Lastly, a third group related to the teaching of the method, i.e., how to help students, and professionals in general, in correctly applying regression and interpreting its results see Batanero, Burrill & Reading (2011), Gutiérrez de Ravé, Jiménez-Hornero & Giráldez (2011) and Wei, De Quan & Jian (2001).

Applications of regression methods are found in scientific papers related to agriculture, medicine, environment, economics, sociology and different engineering areas. Using a random sample of one hundred papers published during 2012 and obtained by the authors from the Web of Knowledge, in 32% of them there was a direct application of these methods and in almost half of them (46%) there was a reference to regression.

Such significant use of regression supports its inclusion in the largest part of university curricula. It is generally included in the statistics subject.

In terms of teaching, curve adjustment is generally explained once the methods of optimization of real functions of several real variables are known, along with the solution of linear equation systems. This makes possible the support of procedures that permit to determine the values of the parameters characterizing the functional class of the best adjustment curve sought. Usually, in practice, computer packages are used to determine these parameters.

Taking into consideration the importance of curve adjustment for applications, it is advisable to teach the students these ideas much more before it is usually done in university curricula. How can this purpose be achieved if curve adjustment is preceded by mathematical requirements which seem not to be possible to sever? This paper presents an approach that permits to teach in advance the consideration of these basic ideas of regression in at least one semester, what is justified assuming the following hypotheses:

-The methods of optimization for the real functions of many variables are neglected, what has the immediate implication of not requiring the partial derivation.

-A system of linear equations is not stated, so the corresponding theory is not necessary.

-The reiterative application of the least squares method in functional classes determined by only one parameter, so that in each of them, the corresponding sum of the squared deviations is function of a unique variable. Consequently, optimization techniques for real functions of a real variable are only required.

Though sufficient and varied bibliography about the least squares method is available, it was considered necessary to make explicit some of its basic aspects initially, such as the expression that takes the sum of the squared deviations, as well as the normal system of equations that is formed at stating the necessary conditions for extremes, both in the case of the simple linear regression.

Curve adjustment is, possibly, the most frequently used mathematical resource for solving one of the fundamental problems related to numerous scientific areas: "reconstructing" a function starting from experimental data. Essentially, for the case of one variable functions, this problem may be formulated through the following statement:

"*Given the set of n points $\{(x_1, y_1);\ (x_2, y_2); \dots (x_n, y_n)\}$, where n is a natural number and every two x_k abscissas are different, the goal is determining the $y = f(x)$ function which, within a given prefixed class of functions, best adjusts them*".

2. Materials and Methods

2.1. Curve Adjustment and the Least Squares Method

Generally, the prefixed functional class depends on various parameters, and the purpose of the method used for their estimation is to satisfy some criterion of optimization, which is characteristic of the method; particularly, the objective of the least squares method is to minimize the sum of the squared deviations. Two other alternatives, which are also frequently used are the Maximum Likelihood Method; see Yoshimori & Lahiri (2014), Seo & Lindsay (2013) and Han & Phillips (2013) and for the Bayesian regression method; see Zhao, Valle, Popescu, Zhang & Mallick (2013), Mudgal, Hallmark, Carriquiry & Gkritza (2014) and Choi & Hobert (2013).

In the probably most renowned and significant case of finding the best-adjusting function within the class of linear functions of one independent variable $f(x) = a_1x + a_2$ this problem is solved through the least squares method by determining the values of the a_1 and a_2 parameters (the slope and the intercept with the y-axis, respectively), which provide the minimum value to the sum of the $S(a_1, a_2)$ squared deviations:

$$S(a_1, a_2) = \sum_{k=1}^{n}(f(x_k) - y_k)^2 = \sum_{k=1}^{n}(a_1x_k + a_2 - y_k)^2.$$

Determining the minimum of $S(a_1, a_2)$ requires applying optimization techniques for real functions of two real variables, which initially require the use of the necessary condition on extreme points:

$$\frac{\partial S}{\partial a_1} = 2\sum_{k=1}^{n} x_k(a_1x_k + a_2 - y_k) = 0; \qquad \frac{\partial S}{\partial a_2} = 2\sum_{k=1}^{n}(a_1x_k + a_2 - y_k) = 0 \tag{1}$$

Afterwards, it requires the resolution of the system of two linear equations resulting from it with a_1 and a_2 as unknowns. This system is called the normal equation system, which is expressed as follows:

$$a_1\sum_{k=1}^{n} x_k^2 + a_2\sum_{k=1}^{n} x_k = \sum_{k=1}^{n} x_ky_k; \qquad a_1\sum_{k=1}^{n} x_k + na_2 = \sum_{k=1}^{n} y_k \tag{2}$$

When applying any of the existing techniques for the resolution of system (2), the result is a single solution $a_1 = a_1^{(0)}$, $a_2 = a_2^{(0)}$, given by the expressions:

$$a_1^{(0)} = \frac{n\sum_{k=1}^{n} x_ky_k - \sum_{k=1}^{n} x_k\sum_{k=1}^{n} y_k}{n\sum_{k=1}^{n} x_k^2 - \Big(\sum_{k=1}^{n} x_k\Big)^2}; \qquad a_2^{(0)} = \frac{1}{n}\Big(\sum_{k=1}^{n} y_k - a\sum_{k=1}^{n} x_k\Big) \tag{3}$$

The procedure herein presented is equivalent to applying the Gauss - Seidel Method (McCracken & Dorn 1974) to the normal system of equation (2). This is an iterative method for the resolution of linear equation systems, as it happens with the system in equation (2), or the one resulting from applying the necessary condition in equation (1) to the sum of the squared deviations, when the adjustment takes place in a functional class that is linear with respect to the parameters defining it.

In terms of teaching organization, this approach provides more significance to the optimization methods of real functions of one real variable. At the same time, it permits the introduction of an important application such as curve adjustment, advancing one semester, at least.

2.2. An Iterative Method for the Process of Curve Adjustment

Solving the normal equation system of equation (2) is not possible until a method that permits optimizing a derivable function of two real variables is available. Therefore, a sequence, that only requires applying different variable optimization techniques, one at a time, may be followed. Such method results from realizing the following steps:

1. Prefix the functional class in which the adjustment process will be carried out.

   As it is known, the functional class is characterized by a functional expression involving the independent variable and the p parameters that define it, being p a positive integer. This class of functions is denoted by:

   $$y = f(x, a_1, a_2, \ldots a_p), \tag{4}$$

   where x is the independent variable, and the parameters have been denoted by $a_1, a_2, \ldots, a_p$, for which it is necessary to previously establish an order among them.

2. Keep the variable character of a_1 and assign values to the rest of the parameters.

   The values assigned to the parameters are denoted by $a_2^{(0)}, a_3^{(0)}, \ldots, a_p^{(0)}$, where the sub-index of each identifies the parameter, and the supra - index 0 indicates that it is the initial assignment. These values may be arbitrary or follow a certain criterion, but this is irrelevant to the method being described. Thus, the set of functions is defined as:

   $$y_1 = f_1(x, a_1, a_2^{(0)}, \ldots, a_p^{(0)}) \tag{5}$$

   which is formed by functions of the independent variable x depending on the parameter a_1 that obviously constitutes a subclass of the pre-fixed functional class.

3. Form the sum of the quadratic differences in $y_1 = f_1(x, a_1, a_2^{(0)}, \ldots, a_p^{(0)})$.

   Given the set $\{(x_1, y_1); (x_2, y_2); \ldots; (x_n, y_n)\}$, of n points, the corresponding sum of the quadratic differences to be minimized is formed, which is a function of the a_1 parameter, and is defined by the expression:

   $$S(a_1) = \sum_{k=1}^{n} \Big(f_1(x_k) - y_k\Big)^2 = \sum_{k=1}^{n} \Big(f_1(x_k, a_1, a_2^{(0)}, \ldots, a_p^{(0)}) - y_k\Big)^2$$

4. Apply the necessary extreme condition to $S(a_1)$.

   As $S(a_1)$ is a one variable function, it is enough to state $S'(a_1) = 0$ to thus determine the solution of this equation. This gives the value of parameter a_1, denoted by $a_1^{(1)}$ so that $S(a_1^{(1)})$ is the lowest value of $S(a_1)$. It is important to note that the supra-index 1 in $a_1^{(1)}$ means that this is the first value calculated for a_1.

3. Results and Discussion

The implementation of the previous process guarantees obtaining the function of better adjustment within the $y_1 = f_1(x, a_1, a_2^{(0)}, \dots, a_p^{(0)})$ subclass, in the initially pre-fixed functional class $y = f(x, a_1, a_2, \dots, a_p)$.

In general, it is not expected that the $y_1^{(1)} = f_1(x, a_1^{(1)}, a_2^{(0)}, a_3^{(0)}, \dots, a_p^{(0)})$ function obtained from $y_1 = f_1(x, a_1, a_2^{(0)}, \dots, a_p^{(0)})$ by substituting $a_1^{(1)}$ by a_1, to be a good approximation to the better adjustment within the general prefixed class $y = f(x, a_1, a_2, \dots, a_p)$.

The described process is repeated, leaving the next parameter as arbitrary (in this case a_2) and taking for a_1 the calculated value $a_1^{(1)}$, and for the rest of the parameters the initially assumed values $a_3^{(0)}, \dots, a_p^{(0)}$.

As a result, the value of parameter a_2, denoted by $a_2^{(1)}$, will be obtained, offering the best adjustment function within the subclass:

$$y_2 = f_2(x, a_1^{(1)}, a_2, a_3^{(0)}, \dots, a_p^{(0)})$$

Once the whole set of parameters has been recovered, by proceeding similarly, the following p functions would be obtained:

$$\begin{aligned} y_1^{(1)} &= f_1(x, a_1^{(1)}, a_2^{(0)}, a_3^{(0)}, \dots, a_p^{(0)}) \\ y_2^{(1)} &= f_2(x, a_1^{(1)}, a_2^{(1)}, a_3^{(0)}, \dots, a_p^{(0)}) \\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ y_p^{(1)} &= f_p(x, a_1^{(1)}, a_2^{(1)}, a_3^{(1)}, \dots, a_p^{(1)}) \end{aligned}$$

where each of them is the best adjustment function within the corresponding functional class.

It can be verified that each of these functions is not a worse, but a better approximation than the previous one. Indeed, $y_2^{(1)}$ is reduced to $y_1^{(1)}$taking $a_2^{(1)} = a_2^{(0)}$ for it. So with this value for that parameter, function $y_2^{(1)}$ provides a value for the sum of squared differences that is similar to the one given by $y_1^{(1)}$. This proves that $y_2^{(1)}$ is an approximation not worse than $y_1^{(1)}$. This also holds for the rest of the functions and this step completes the first iteration.

As the values of parameters $a_2^{(1)}, a_3^{(1)}, \dots, a_p^{(1)}$ have a similar purpose to the one followed with numbers $a_2^{(0)}, a_3^{(0)}, \dots, a_p^{(0)}$, the above process may be repeated to find the second iteration, which will be completed once the following new p functions have been determined:

$$\begin{aligned} y_1^{(2)} &= f_1(x, a_1^{(2)}, a_2^{(1)}, a_3^{(1)}, \dots, a_p^{(1)}) \\ y_2^{(2)} &= f_2(x, a_1^{(2)}, a_2^{(2)}, a_3^{(1)}, \dots, a_p^{(1)}) \\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ y_p^{(2)} &= f_p(x, a_1^{(2)}, a_2^{(2)}, a_3^{(2)}, \dots, a_p^{(2)}) \end{aligned}$$

For a third iteration, the calculated values $a_2^{(2)}, a_3^{(2)}, \dots, a_p^{(2)}$ will be used, and the process would be the same successively. A possible stop criterion for the iterative process would be that each of the values determined for the parameters in a given iteration were sufficiently close to the corresponding value in the preceding iteration[1].

Sufficiently close means here, as is common in mathematics, that a certain pre-fixed degree $\epsilon > 0$ of accuracy is fulfilled for all j index ($j = 1, 2, \dots, p$), and inequality $|a_j^{(m)} - a_j^{(m-1)}| < \epsilon$, where m is a natural number indicating the number of order of the iteration.

The objective is proving that this process converges with the function of best adjustment within the pre-fixed functional class $y = f(x, a_1, a_2, \dots, a_p)$, which is the initial step in the iterative method. For the case of pre-fixing the linear functions, the method may be proven to converge.

3.1. Geometrical Interpretation of the Case of Linear Adjustment

If in the initial step of the process the linear function class is pre-fixed in an independent variable $f(x) = ax + b$, where a and b are real numbers, an interesting geometrical interpretation may be given from the described iterative method.

By keeping the notation used in the description of the iterative method, the slope is denoted by a_1 and the intercept with the y axis by a_2, so that this general functional class is then expressed as:

$$y = a_1 x + a_2$$

Prefixing the value of parameter a_2 equal to $a_2^{(0)}$ implies taking the point of coordinates $(0, a_2^{(0)})$ of the y - axis in the system of Cartesian coordinates, and considering the family of all straight lines that pass through such point, except the very axis of the ordinates (y - axis).

This family is formed by the graphics of the $y_1 = a_1 x + a_2^{(0)}$ subclass functions, for all the possible values of the a_1 slope.

Once the best adjustment function $y_1^{(1)} = a_1^{(1)} x + a_2^{(0)}$ in $y_1 = a_1 x + a_2^{(0)}$ is determined, we come back to the arbitrary character of the intercept with the y

[1] This is one of the stop - criteria used in some numeric methods, such as Gauss - Seidel, for the resolution of a linear system.

- axis, taking the new subclass within the $y = f(x) = a_1x + a_2$ class, defined by $y_2 = a_1^{(1)}x + a_2$ to carry out the new adjustment process. As a result, the value of parameter a_2 will be determined as $a_2^{(1)}$, which determines the $y_2^{(1)} = a_1^{(1)}x + a_2^{(1)}$ equation, which is a straight line parallel to the one initially determined, $y_1^{(1)} = a_1^{(1)}x + a_2^{(0)}$.

Therefore, the first iteration concludes, when the following functions are ready:

$$y_1^{(1)} = a_1^{(1)}x + a_2^{(0)}, \qquad y_2^{(1)} = a_1^{(1)}x + a_2^{(1)}$$

Initially, with $y_1^{(1)} = a_1^{(1)}x + a_2^{(0)}$, we determine the angle of inclination of the straight line that passes through the point of coordinates $(0, a_2^{(0)})$ with the positive direction of the axis of abscissas. This is subsequently transferred parallel to itself until it occupies the graphic position corresponding to $y_2^{(1)} = a_1^{(1)}x + a_2^{(1)}$ that passes through the point of coordinates $(0, a_2^{(1)})$. In turn, this is used to implement the second iteration: the new slope $a_1^{(2)}$ and the new intercept $a_2^{(2)}$, and so on.

3.2. One Example of the Application of the Iterative Method

A table with arbitrary or hypothetical data, which determine five points of integer coordinates: A(1, 1), B(2, 3), C(3, 3), D(4, 5) y E(5, 5), is taken.

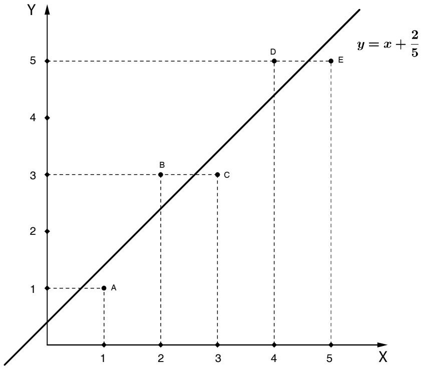


Figure 1: Regression line.

In Figure 1, a regression line $y = x + 0.4$ is represented. It was obtained by the the least squares method, in the general functional class $y = a_1x + a_2$, where the parameters are the slope a_1 and the intercept a_2. The sum of the squared deviations is function of these two parameters, so to obtain $y = x + 0.4$ (it means, $a_1 = 1$ and $a_2 = 0.4$) techniques of optimization for the functions of some variables were required and the exact resolution of the normal system of equations.

In Figure 2, a segment of the first approximation is represented. It is the line that by the origin (of slope 61/55) better adjusts to the five points. It is optimized in the functional class $y = a_1x$, where the parameter is the a_1 slope, what follows

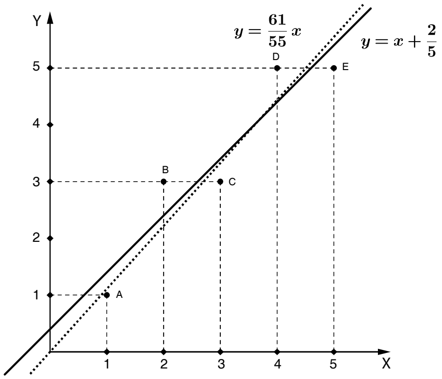


FIGURE 2: Segment of the first approximation.

from assigning, in $y = a_1x + a_2$, the zero value to a_2 parameter. Geometrically, it means that it optimizes in the functional class of all no vertical lines that pass through the origin. The sum of the squared deviations is only function of a_1, so that this optimum (minimum) is determined by techniques of optimization of functions of one variable (it not even requires the ordinary derivative, observing that the sum of the squared deviations is a quadratic function in a_1 variable, whose graphic is a parabola that opens upwards, so that the optimum (minimum) is reached in the abscissa of the vertex (value 61/55). It is maintained with purposes of comparison, the segment of the regression line.

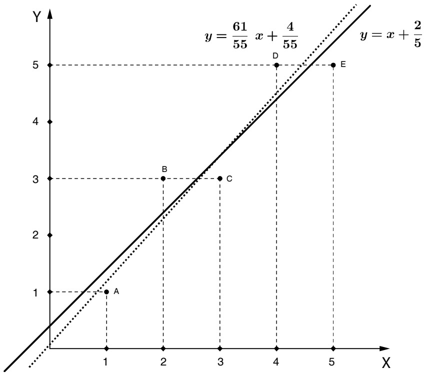


FIGURE 3: Segment of the second approximation.

In Figure 3 a segment of the second approximation, which best adjusts to the five points among all the lines with slope 61/55, is represented. It is optimized in the functional class $y = (61/55)x + a_2$, which is obtained from the functional class $y = a_1x+a_2$ replacing $a_1 = 61/55$ and retaking the variable character of a_2 (notice that for a_2 the value 0 was initially assumed). Geometrically it means that the line of equation $y = (61/55)x$ is paralleled displaced itself up to a position that betters the adjustment (provides the minimal for the sum of the squared deviations which now depends on a_2). The resulting value for the parameter is $a_2 = 4/55$.

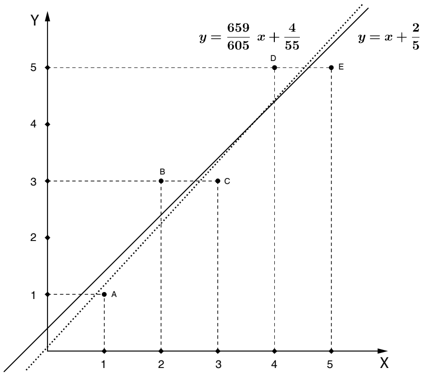


FIGURE 4: Segment of the third approximation.

For a new approximation $a_2 = 4/55$ in the general class $y = a_1x+a_2$ is replaced to optimize in the subclass $y = a_1x + 4/55$, in which the slope a_1 is variable again, so that what is looked for is the line that better adjusts to the data (in the sense of minimizing the corresponding sum of squared deviations) among all those that pass through the axis point (0, 4/55). The result is the new slope value $a_1 = 659/605$, what permits the line with best equation adjustment $y = (659/605)x + 4/55$, a segment of which is represented in Figure 4 together with the one of the least regression line.

The process continues similarly, so that the new adjustment would take place in the functional subclass $y = (659/605)x+a_2$, where the variable character of the second of the parameters is retaken.

3.3. The Gauss - Seidel Method and Convergence in the Case of Linear Adjustment

The issue related to the convergence of the described iterative method has an affirmative answer in the case of linear adjustment, if the set of points fulfills the initially described characteristics; i.e., if within the full set, every pair of points has different abscissas.

As the best adjusting straight line, with the equation $y = a_1^{(0)}x + a_2^{(0)}$, does exist, and the parameters are analytically determined as the only solution by (3) in the standard equation system, an iterative method convergent to the solution of such system would obviously provide, after an adequate number of iterations, an $\alpha \approx a_1^{(0)}$, $\beta \approx a_2^{(0)}$ approximation. This offers the possibility of defining a $y = \alpha x + \beta$ approximation for the best adjustment equation $y = a_1^{(0)}x + a_2^{(0)}$.

One of the simplest iterative methods for the resolution of a linear equation system, easily programmed for its computerized application, is the Gauss - Seidel Method.

For the case of a two - equation system with two unknowns:

$$\begin{aligned} b_{11}a_1 + b_{12}a_2 &= c_1 \\ b_{21}a_1 + b_{22}a_2 &= c_2 \end{aligned}$$

The method is described as follows:

Supposing that in the coefficients matrix, those of the main diagonal are not null, it is possible to find the unknowns a_1 and a_2:

$$a_1 = \frac{1}{b_{11}}(c_1 - b_{12}a_2); \qquad a_2 = \frac{1}{b_{22}}(c_2 - b_{21}a_1) \tag{6}$$

An arbitrary approximation is now defined for the $a_1 = a_1^{(0)}$, $a_2 = a_2^{(0)}$ solution, and it is used to find a new approximation for the a_1 unknown value stemming from the first of the expressions in (6):

$$a_1^{(1)} = \frac{1}{b_{11}}(c_1 - b_{12}a_2^{(0)})$$

The $a_1^{(1)}$ calculated value is substituted in the second of the expressions (6) to determine an approximation to the a_2 unknown value:

$$a_2^{(1)} = \frac{1}{b_{22}}(c_2 - b_{21}a_1^{(1)})$$

At this point the first iteration is fulfilled.

The second iteration is implemented by taking the calculated approximation $a_1^{(1)}$, $a_2^{(1)}$ with the same role played by $a_1 = a_1^{(0)}$, $a_2 = a_2^{(0)}$ in the first iteration.

In this way it is possible to reach the order m iteration defined by the expressions:

$$a_1^{(m)} = \frac{1}{b_{11}}(c_1 - b_{12}a_2^{(m-1)}), \qquad a_2^{(m)} = \frac{1}{b_{22}}(c_2 - b_{21}a_1^{(m-1)}) \tag{7}$$

A sufficient condition for convergence to the solution of the iterations produced through the Gauss - Seidel Method lies in the matrix of the coefficients being diagonally dominant, which means in this case that the $|b_{11}b_{22}| > |b_{21}b_{12}|$ inequality has to be fulfilled (McCracken & Dorn 1974).

If we now define $b_{11} = \sum_{k=1}^{n} x_k^2$; $b_{12} = b_{21} = \sum_{k=1}^{n} x_k$; $a = a_1$, and $b = a_2$ then the standard equation system (2) can be expressed as:

$$\begin{aligned} b_{11}a_1 + b_{12}a_2 &= c_1 \\ b_{21}a_1 + b_{22}a_2 &= c_2 \end{aligned}$$

where $c_1 = \sum_{k=1}^{n} x_k y_k$, $c_2 = \sum_{k=1}^{n} y_k$ so that the expressions in (7) allow determining an approximation to its solution.

It is not difficult to verify that the values given by the expressions (7) for the unique solution of the standard system (2) match in each m iteration, those provided for the a_1 slope and the a_2 intercept by each iteration of the iterative adjustment process here described. Neither is it difficult to prove that the system matrix (2) is diagonally dominant if in the $\{(x_1, y_1); (x_2, y_2); \ldots ; (x_n, y_n)\}$ set of points, where n is a natural number, every two of the x_k abscissas are different, which means that the $n\sum_{k=1}^{n} x_k^2 > \Big(\sum_{k=1}^{n} x_k\Big)^2$ inequality is fulfilled.

Indeed, according to Bronshtein & Semendiaev (1971), in the inequality (which is strict if there are at least two different x_k values):

$$\frac{|x_1 + x_2 + \ldots x_n|}{n} \leq \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}}$$

it would suffice to square both sides to obtain, first:

$$\frac{(x_1 + x_2 + \ldots + x_n)^2}{n^2} \leq \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}$$

Then, multiplying the two sides by n^2 and expressing the sums in a compact form, it results:

$$n\sum_{k=1}^{n} x_k^2 \geq \Big(\sum_{k=1}^{n} x_k\Big)^2$$

As every two of the x_k numbers are supposed to be different, the fulfillment of the $n\sum_{k=1}^{n} x_k^2 > \Big(\sum_{k=1}^{n} x_k\Big)^2$ inequality is finally guaranteed. This proves that the matrix of the standard equation system (2) is diagonally dominant, and in turn implies that the expressions (7), obtained by applying the Gauss - Seidel Method, converge to the unique solution of such system.

At the same time, each iteration of the method was observed to coincide with the parameter values that result, in each step, from the function of better adjustment within the corresponding subclass. Therefore, a conclusion can be advanced so that these functions converge to the least squares straight line of equation $y = a_1^{(0)}x + a_2^{(0)}$.

4. Conclusions

An iterative method has been proposed to obtain an approximation of the best adjustment function to a given set of points, consisting of determining the best adjustment function within a certain subclass of the pre-fixed functional class each time. Each subclass is dependent on a single parameter.

As optimization is used only on one variable, it is not required to explicitly write the standard equation system.

For the case of linear adjustment with one independent variable, the iterative method is revealed to be equivalent to the application of the standard equation system of the Gauss - Seidel Method, which permits to show its convergence to the lowest quadratic straight line of the $y = a_1^{(0)}x + a_2^{(0)}$ equation.

Everything suggests that for other linear functional classes, with respect to the parameters that define them, similar results should be obtained, in the sense of the equivalence between the iterative method and the Gauss - Seidel one. Also, in this way, it may be possible to show that the iterative method is convergent to the best adjustment function obtained when applying the least squares m-ethod.

The proposed approach offers the possibility of focusing the least squares method along with that of curve adjustment, as an application of optimization techniques of the real functions of one real variable, developed during the first semester of higher education diplomas. This would permit speeding the approach of the significant topic of curve adjustment.

Acknowledgments

The authors acknowledge the financial support of AECID (Spain) under Project N° A2/039476/11 - "Institutional Strengthening in the Teaching of Subjects of Statistics, Operative Investigation, Liability and Quality, and Scientific Application in Topics of Regional Interest".

[Recibido: agosto de 2013 — Aceptado: febrero de 2014]

References

Batanero, C., Burrill, G. & Reading, C. (2011), *Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education: A Joint ICMI/IASE Study*, Vol. 18, 2 edn, Springer, New York.

Braga, G., Silveira, C., Rodríguez, V., Henrique de Cerqueira, P., Aparecido, W. & Barros, F. (2009), 'Quantifying herbage mass on rotationally stocked palisadegrass pastures using indirect methods', *Scientia Agricola* **66**(1), 127–131.

Bronshtein, I. & Semendiaev, K. (1971), *Mathematics Manual*, Mir, Moscow.

Choi, H. M. & Hobert, J. P. (2013), 'Analysis of MCMC algorithms for Bayesian linear regression with Laplace errors', *Journal of Multivariate Analysis* **117**(0), 32–40.

Donal, R. J. (2001), 'A taxonomy of global optimization methods bases on response surfaces', *Journal of Global Optimization* **21**, 345–383.

Gutiérrez de Ravé, E., Jiménez-Hornero, F. J. & Giráldez, J. V. (2011), 'A computer application for interpolation algorithms of curves', *Computer Applications in Engineering Education* **19**(1), 40–47.

Guzmán, K. P., Bolivar, I., Alepuz, M. T., González, D. & Martin, M. (2011), 'Impacto en el tiempo asistencial y el estadio tumoral de un programa de diagnóstico y tratamiento rápido del cáncer colorrectal', *Revista Española de Enfermedades Digestivas* **103**(1), 13–19.

Han, C. & Phillips, P. C. (2013), 'First difference maximum likelihood and dynamic panel estimation', *Journal of Econometrics* **175**(1), 35 – 45.

Ibarra, M. & Arana, P. (2011), 'Crecimiento del camarón excavador Parastacus pugnax (Poeppig, 1835) determinado mediante técnica de marcaje', *Latin American Journal of Aquatic Research* **39**(2), 378–384.

Kelley, C. T. (1999), *Iterative Methods for Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, U.S.

McCracken, D. D. & Dorn, W. S. (1974), *Métodos Numéricos y Programación Fortran.*, 1 edn, Editorial Pueblo y Educación, La Habana, Cuba.

Mudgal, A., Hallmark, S., Carriquiry, A. & Gkritza, K. (2014), 'Driving behavior at a roundabout: A hierarchical Bayesian regression analysis', *Transportation Research Part D: Transport and Environment* **26**(0), 20–26.

Núñez, E., Steyerberg, E. W. & Núñez, J. (2011), 'Estrategias para la elaboración de modelos estadísticos de regresión', *Revista Española de Cardiología* **64**(6), 501–507.

Ranganatham, A. (2004), 'The Levenberg - Marquardt algorithm', *Tutorial on LM Algorithm* .

Santos da Silva, L. M., Estraviz, L. C., Caixeta, J. V. & Carolina, S. B. (2006), 'Fitting a Taper function to minimize the sum of absolute deviations', *Scientia Agricola* **63**(5), 460–470.

Schmidt, M. (2005), *Least Squares Optimization with L1-Norm Regularization*, CS542B Project Report.

Seo, B. & Lindsay, B. G. (2013), 'Nearly universal consistency of maximum likelihood in discrete models', *Statistics and Probability Letters* **83**(7), 1699–1702.

Vega-Vilca, J. & Guzmán, J. (2011), 'Regresión PLS y PCA como solución al problema de multicolinealidad en regresión múltiple', *Revista de Matemática: Teoría y Aplicaciones* **18**(1), 09–20.

Wei, W., De Quan, L. & Jian, L. (2001), 'On the system of multiple linear regression of higher education tuition in China', *Journal Advanced Materials Research* **211-212**, 752–755.

Yoshimori, M. & Lahiri, P. (2014), 'A new adjusted maximum likelihood method for the Fay-Herriot small area model', *Journal of Multivariate Analysis* **124**(0), 281 – 294.

Zhao, K., Valle, D., Popescu, S., Zhang, X. & Mallick, B. (2013), 'Hyperspectral remote sensing of plant biochemistry using Bayesian model averaging with variable and band selection', *Remote Sensing of Environment* **132**(0), 102–119.

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 127 a 140

Algorithms to Calculate Exact Inclusion Probabilities for a Non-Rejective Approximate πps Sampling Design

Algoritmos para calcular probabilidades exactas de inclusión para un diseño de muestreo no rechazable πpt

Zaizai Yan[a], Yuxia Xue[b]

Science College, Inner Mongolia University of Technology, Hohhot, P. R. China

Abstract

AP-design, an efficient non-rejective implementation of the πps sampling design, was proposed in the literature as an alternative Poisson sampling scheme. In this paper, we have updated inclusion probabilities formulas in the AP sampling design. The formulas of these inclusion probabilities have been greatly simplified. The proposed results show that the AP design and the algorithms to calculate inclusion probabilities are simple and effective, and the design is possible to be used in practice. Three real examples have also been included to illustrate the performance of these designs.

***Key words*:** AP sampling design, Inclusion probabilities, Poisson sampling.

Resumen

Una implementación del diseño de muestreo πpt, que no es de rechazo, ha sido recientemente propuesta como alternativa al esquema de Poisson. En este trabajo, hemos adaptado las formulas de probabilidades de inclusión en el diseño de muestreo Poisson alternativo (AP por sus siglas en inglés). Estas fórmulas han sido significativamente simplificadas. Los resultados propuestos muestran que el diseño AP y los algoritmos para calcular las probabilidades de inclusión son simples y efectivos, y que el diseño se puede usar en la práctica. Se incluyen tres ejemplos reales para ilustrar el desempeño de la propuesta.

***Palabras clave*:** AP diseño de muestra, probabilidades de inclusión, esquema de Poisson.

[a] Professor. E-mail: zz.yan@163.com
[b] Postgraduate student. E-mail: yuxiaxue_imut@163.com

1. Introduction

Unequal probability sampling is frequently used in surveys in order to increase the efficiency in the estimation of the population characteristics. A sampling design without replacement and with unequal inclusion probabilities which are proportional to a size variable, that is known for all units in the population is usually called a πps sampling design. The πps sampling usually produces more efficient estimates than sampling with equal probabilities. Suppose that the finite population U consists of N units labelled $1, \ldots, N$. An auxiliary variable with value X_i for the unit i is known for all $i = 1, \ldots, N$. Assume that $X_i > 0$, for all i and strict inequality for at least one i. It is required to estimate the total $Y = \sum_i Y_i$ where the sum is over $1, \ldots, N$, given a sample of size n. Let $p_i = nX_i/X, i = 1, \ldots, N$ be the prescribed inclusion probability parameters with $\sum_{i=1}^{N} p_i = n$ with X its corresponding population total. The problem is how to select a sample with fixed size n, so that the probability of each unit i to be included in the sample equals just p_i. Many papers have proposed sampling schemes in which the inclusion probability of unit i is π_i. Some important reference are followings: Sen (1953), Durbin (1967), Brewer (1963), Sampford (1967), Hájek (1964, 1981), Rosén (1997a), Aires (1999), Bondesson & Thorburn (2008), Bondesson & Grafström (2011), Grafströ (2009), Laitila & Olofsson (2011), Olofsson (2011). Most of the schemes with predetermined inclusion probabilities are either difficult to execute or calculate π_{ij}, the second order inclusion probability units i and j, if n is more than 2. Recently, Zaizai, Miaomiao & Yalu (2013) presented a new approximative πps design for fixed sample size n as follows:

1. Draw an initial sample s_0, using Poisson sampling design with probabilities $\{p_i\}_1^N$. The size of the initial sample s_0 is a random variable denoted by n_{s_0}.

2. If $n_{s_0} = n$, then the sampling is finished and the sample $s = s_0$. If $n_{s_0} < n$, then replenish the rest units denoted by s_1, its size $n - n_{s_0}$, by simple random sampling without replacement (SRSWOR) design from $U - s_0$, the final sample $s = s_0 \cup s_1$. If $n_{s_0} > n$, then remove $n_{s_0} - n$ units denoted by s_2, using the SRSWOR-design, from s_0, the final sample $s = s_0 - s_2$. The AP design becomes a non-rejective sampling design. Algorithms for calculating exact first- and second-order inclusion probabilities of the corresponding design are too complex and involve a Jacobi over-relaxation iterative method.

***Note* 1.** We assume that the population is such that $p_i = nX_i/X < 1$, for all i. You need to remove the cases where p_i is larger than 1 and then iterative removing further units if necessary

The purpose of this paper is to simplify calculation of the first-order and second-order inclusion probabilities of the AP design. The analytical expressions of inclusion probabilities for the AP design presented in Section 2 are simpler to operate than the original one.

2. Inclusion Probabilities of AP Design

Now we discuss inclusion probabilities of the AP sampling design. For convenience, we denote the random variable $\sum_{k\in U,k\neq i} I_k$ as $n_{s_0}^{-i}$, the random variable $\sum_{k\in U,k\neq i,k\neq j} I_k$ as $n_{s_0}^{-ij}$, where $I_k = \begin{cases} 1 & \text{if } k \in s_0 \\ 0 & \text{otherwise} \end{cases}$ for all $k \in U$ are indicators for the Poisson sampling. In order to calculate the first and second-order inclusion probabilities of the AP design, we firstly derive the following Proposition and Lemmas. For convenience, the subset $\{1,2,\cdots,i\}$ of U is abbreviated as U_i and $Pr(\sum_{\alpha=1}^{i} I_\alpha = j)$ as P_j^i where $j = 0,1,\ldots,i$; $i = 1,2,\ldots,N$. Then $Pr(n_{s_0} = \nu) = P_\nu^N$, $\nu = 0,1,\ldots,N$.

Proposition 1. *Keep the same assumptions as above and* $q_i = 1 - p_i$. *Then* $P_0^i = \prod_{\alpha=1}^{i} q_\alpha$; $P_k^i = p_i P_{k-1}^{i-1} + q_i P_k^{i-1}$,$1 \leq k \leq i-1$ *and* $P_i^i = \prod_{\alpha=1}^{i} p_\alpha$.

A proof of proposition 1 can be found in Tillé (2006) and Olofsson (2011).

***Note* 2.** Proposition 1 shows that we can calculate $P_0^i, P_1^i, \ldots, P_i^i$ by using $P_0^{i-1}, P_1^{i-1}, \ldots, P_{i-1}^{i-1}$ with initial values $P_0^1 = q_1$ and $P_1^1 = p_1$. By recursive calculation with respect to i, we can finally obtain P_ν^N, $\nu = 0,1,\ldots,N$.

Lemma 1. *Let* $\mu_k = \frac{1}{1-p_k}$, *then*

$$Pr(n_{s_0}^{-k} = \nu) = \mu_k \sum_{j=0}^{\nu} (-1)^{\nu-j} \big(\frac{p_k}{1-p_k}\big)^{\nu-j} P_j^N \tag{1}$$

Lemma 2. *Given the assumptions as in Lemma 1, then*

$$Pr(n_{s_0}^{-kl} = \nu) = \mu_k \mu_l \sum_{j=0}^{\nu} (-1)^{\nu-j} \big(\frac{p_l}{1-p_l}\big)^{\nu-j} \sum_{t=0}^{j} (-1)^{j-t} \big(\frac{p_k}{1-p_k}\big)^{j-t} P_t^N \tag{2}$$

Lemma 1 and Lemma 2 are proved in the appendix. Now we present theorems 1 and 2 which the core results of this paper.

Theorem 1. *Under the AP-design, the algorithms for calculating the first-order inclusion probabilities can be written as*

$$\pi_k = \sum_{\nu=0}^{N-1} C_k(\nu) \mu_k \sum_{j=0}^{\nu} (-1)^{\nu-j} (\frac{p_k}{1-p_k})^{\nu-j} \cdot P_j^N \tag{3}$$

where $C_k(\nu) = \begin{cases} \frac{(N-n)p_k + (n-\nu)}{N-\nu} & \nu = 0,\ldots,n-1, \\ \frac{np_k}{\nu+1} & \nu = n,\ldots,N-1 \end{cases}$ *and* $P_j^N = Pr(n_{s_0} = j)$.

Theorem 2. *Under the AP-design, the analytical formula of the second-order inclusion probabilities is as follows*

$$\pi_{kl} = \sum_{\nu=0}^{N-2} C_{kl}(\nu) \mu_k \mu_l \sum_{j=0}^{\nu} (-1)^{\nu-j} \big(\frac{p_l}{1-p_l}\big)^{\nu-j} \sum_{t=0}^{j} (-1)^{j-t} \big(\frac{p_k}{1-p_k}\big)^{j-t} P_t^N \tag{4}$$

where

$$C_{kl}(\nu)=\begin{cases} q_kq_l\frac{(n-\nu)(n-\nu-1)}{(N-\nu)(N-\nu-1)}+(p_kq_l+p_lq_k)\frac{n-\nu-1}{N-\nu-1}+p_kp_l, & \nu=0,1,\dots,n-2,\\ p_kp_l\frac{n(n-1)}{(\nu+2)(\nu+1)}, & \nu=n-1,\dots,N-2.\end{cases}$$

From Theorem 1 and Theorem 2, we can find that the problem to solve π_k and π_{kl} may be switched into solving a series of $Pr(n_{s_0}=\nu)=P_\nu^N$, $\nu=0,1,\dots,N$. We can recursively calculate P_ν^N by using Proposition 1. Proofs of Theorems 1 and 2 can be found in the appendix.

3. Numerical Examples

The statistical literature contains several proposals for methods generating fixed-size without-replacement πps sampling designs. In practice, πps designs with sample size $n=2$ are widely used and fully studied. Due to the difficulties in the implementation and the complexity in computing of inclusion probabilities, application of πps designs with sample size $n>2$ is relatively less. Instead, approximate πps designs such as the Conditional Poisson design (CP), two-phase πps sampling design (2Pπps), Rosén (1997)'s Pareto design and Zaizai et al. (2013)'s design (AP) have been used. However, there are fast and fairly simple implementations of strict πps designs such as systematic πps sampling. Unfortunately, its variance estimation is cumbersome.

3.1. A Review of some Sampling Designs

Poisson sampling is a method to generate a sample s, which has a random size, from a finite population U consisting of N individuals. Each individual i in the population has a predetermined probability p_i and is included in the sample s. A Poisson sample may be obtained by using N independent Bernoulli trials to determine whether the individual under consideration is to be included in the sample s or not. The first-order inclusion probabilities of the individuals are equal to the target inclusion probabilities under the Poisson sampling design. A major drawback with the Poisson design is the randomness of the sample size which has urged statisticians to develop sampling schemes providing fixed size πps designs.

Conditional Poisson sampling (CP), also called rejective sampling or maximum entropy sampling, was first introduced by Hájek (1964). It is a fixed size sampling design, without replacement, on a finite population, with unequal inclusion probabilities among the units of the population. It was called rejective sampling because Hájek's implementation amounts to drawing samples with the Poisson sampling design which has a random size until the desired size is chosen. In fact, one can also obtain the conditional Poisson design by drawing samples, with replacement, using a multinomial sampling design and rejecting the samples which hold some units of the population more than one.

Laitila & Olofsson (2011) proposed a new method to generate a sample with fixed size and inclusion probabilities proportional to size, viz. the 2Pπps design based on a two-phase approach. Consider a population U of N units. For sample generation, let n be the predetermined sample size and assume target inclusion probabilities, p_k, to be proportional to a size variable, x_k, known for all $k \in U$. The 2Pπps sampling scheme is as follows:

1. Draw a sample, s_0, using a Poisson design with $p_{ak} \propto x_k$ as inclusion probabilities, with expected sample size $E(n_{s_0}) = \sum_U p_{ak} \geq n$.

2. If the size of s_0 is greater than or equal to n, then proceed to step 3 and let $s_a = s_0$. If not, repeat step 1.

3. From the sampled set, s_a, draw a sample s of size n using an SRSWOR design.

It was shown that the first-order inclusion probabilities of the 2Pπps design are asymptotically equal to the target inclusion probabilities. But the 2Pπps design is still a rejective sampling design.

Pareto sampling was introduced by Rosén (1997*a*, 1997*b*). It is a simple method to get a fixed size πps sample though with inclusion probabilities only approximately as desired, which can be described as follows: firstly independent random numbers($U_1, \ldots, U_N$) from $U(0, 1)$ are generated, one value for each population unit ($i = 1, \ldots, N$). Then Pareto distributed ranking variables $Q_i = \frac{U_i(1-U_i)}{p_i(1-p_i)}$, where p_i is the targeted inclusion probability for unit i and $\sum p_i = n$, are calculated. Those n units with the smallest Q-values are selected as a πps sample with fixed size n. Bondesson, Traat & Lundqvist (2006) obtained the formulas of first-order and second-order inclusion probabilities for the Pareto design. The true inclusion probabilities only agree with the target inclusion probabilities approximately.

Zaizai et al. (2013) presented an alternative πps design (AP) as Section 1. The AP design is a non-rejective sampling design.

3.2. Examples

Since the Horvitz-Thompson estimators under the AP design, CP design and (2Pπps) design are unbiased, their precision is measured by the variance. However, the ratio estimators mentioned by Kadilar & Cingi (2004) and the traditional ratio estimator are biased, so their precision is measured by mean square error (MSE). In the following section, the estimators and their variances(or MSEs) under the AP design, CP design, 2Pπps design and SRSWOR are studied using three data sets earlier used in the literature. In this paper the AP design and other designs are applied to three populations in which y-values are known, so these variances or MSEs can be calculated exactly. This is only to show the performance of various designs. In practice the y-values in an interested population will be unknown, the variance or MSE of an estimator cannot be obtained, but can be estimated from a sample. Then, the precision is measured by estimation of variance or MSE. As far as the Horvitz-Thompson estimators under the AP design, CP design and (2Pπps) design, the Yates-Grundy variance estimators can be used as the precision. It is unbiased estimator for the true variance.

Example 1. We have used the data of Kadilar & Cingi (2004) in this section. However, we have considered the data of only Aegean Region of Turkey, as we are interested in unequal probabilities sampling with fixed sample size here. We have applied our proposed method and other unequal probabilities sampling methods, such as the 2Pπps sampling design and the CP sampling design on the data of apple production amount (as interest of variate y) and number of apple trees (as auxiliary variate x) in 105 villages of Aegean Region in 1999 (Source: Institute of Statistics, Republic of Turkey).

For a large size population, we may divide the population into three strata according to size of X_i, and the AP-design can be used to get a sample of fixed size within each stratum independently. Let the population be stratified into 3 strata, where sample sizes and population sizes are $(N_1, n_1) = (41, 8)$,$(N_2, n_2) = (41, 8)$ and $(N_3, n_3) = (23, 4)$ respectively. Finally we use stratification sampling technique to build estimation. The relative differences of the inclusion probabilities for the AP-design ,2Pπps-design and CP-design with respect to target inclusion probabilities can be calculated in each stratum respectively. Then, we can build estimators $\widehat{\overline{Y}}_{HT}^{AP}$, $\widehat{\overline{Y}}_{HT}^{2P\pi ps}$ and $\widehat{\overline{Y}}_{HT}^{CP}$ of population mean $\overline{Y}$ from Table 1, and the variance of $\widehat{\overline{Y}}_{HT}^{AP}$, $\widehat{\overline{Y}}_{HT}^{2P\pi ps}$ and $\widehat{\overline{Y}}_{HT}^{CP}$ are easily computed, respectively. As mentioned previously, it is of interest to compare the efficiency of using alternative sampling schemes, for example, the 2Pπps design, AP design, CP design and SRSWOR design. We conclude that the proposed method is more efficient than the 2Pπps design and SRSWOR design. The empirical comparisons included in Table 1 are of interest. It is noticed that the efficiency of the AP design is almost identical to the 2Pπps design, but it is significantly higher than ratio estimators of the SRSWOR design mentioned by Kadilar & Cingi (2004) (Note: The MSEs here are different from the original literature, because the original literature has 106 datum, one of which is a invalid data and is removed, this article has 105 datum). Although the CP design is more efficient than the AP design, the CP design is not easy to implement. The some important advantages of the proposed sampling design are not only its implementation as non-rejective, but also its inclusion probabilities that can be calculated recursively.

Table 1: The variances of the AP design, $2P\pi ps$ design, CP design with $n = 20$, and MSE of SRSWOR ratio estimators in example 1. Aegean Region data.

Sampling scheme	Method of estimation	Variance (or MSE)
The AP design	$\widehat{\overline{Y}}_{HT}^{AP} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{AP}$	349150
The 2Pπps design	$\widehat{\overline{Y}}_{HT}^{2P\pi ps} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{2P\pi ps}$	375615
The CP design	$\widehat{\overline{Y}}_{HT}^{CP} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{CP}$	188396
SRSWOR	Upadhyaya-Singh 1	2331432
SRSWOR	Upadhyaya-Singh 2	2330455
SRSWOR	Singh-Kakran	2329395
SRSWOR	Sisodia-Dwivedi	2331304
SRSWOR	Traditional	2331436

***Note* 3.** The AP design still is not an exact πps design. The inclusion probabilities will be larger than intended probabilities for small inclusion probabilities and smaller than intended probabilities for large inclusion probabilities. At the extreme case there will be risks of not selecting units which are intended to be taken with probability 1, and of selecting units with intended inclusion probability 0.

Example 2. To analyze the performance of the suggested method in comparison to other methods considered in this paper, a natural population data set from the literature (Singh 1967) is being considered. The descriptions of these populations are given below.

y: Percentage of hives affected by disease.

x: January average temperature.

We shall consider drawing a sample according to the AP design previously developed. The exact and desired first-order inclusion probabilities are listed in Table 2 and the second-order inclusion probabilities are in Table 3. Then, once we get an AP sample, we can build estimator $\widehat{\overline{Y}}_{HT}^{AP}$ of population mean $\overline{Y}$, and the variance of $\widehat{\overline{Y}}_{HT}^{AP}$ is easily computed.

Table 2: The raw data and the first-order inclusion probabilities for the AP design ,the 2Pπps design, the CP design and Pareto design, $N = 10, n = 4$ in example 2. Single data.

Unit i	y	x	p	π_i^{AP}	$\pi_i^{2P\pi ps}$	π_i^{CP}	π_i^{Par}
1	49	35	0.3333333	0.3445468	0.3373678	0.3262696	0.3327040
2	40	35	0.3333333	0.3445468	0.3373678	0.3262696	0.3327040
3	41	38	0.3619048	0.3682212	0.3647676	0.3575523	0.3614987
4	46	40	0.3809524	0.3840479	0.3828163	0.3785839	0.3807203
5	52	40	0.3809524	0.3840479	0.3828163	0.3785839	0.3807203
6	59	42	0.4000000	0.3999062	0.4006775	0.3997285	0.3999585
7	53	44	0.4190476	0.4157930	0.4183399	0.4209603	0.4192101
8	61	46	0.4380952	0.4317052	0.4357925	0.4422518	0.4384713
9	55	50	0.4761905	0.4635925	0.4700272	0.4849000	0.4770065
10	64	50	0.4761905	0.4635925	0.4700272	0.4849000	0.4770065

From Table 4, we see that the proposed method has a smaller variance than the CP design. Although the variance of the 2Pπps design is slightly smaller than proposed method, the AP design is easy to implement and generally applicable. In general, the AP design is extremely efficient and it is significantly higher than ratio estimators of the SRSWOR design mentioned by Kadilar & Cingi (2004).

Example 3. The data we considered here is from 35 Scottish farms in Table 5. Let sample size n be equal to 8. The descriptions of these populations are given below (Asok & Sukhatme 1976, page 916).

y: Acreage under oats in 1957.

x: Recorded acreage of crops and grass for 1947.

The exact first-order and second-order inclusion probabilities for the AP design, 2Pπps design and CP design are calculated. In this example, the efficiencies for the

Table 3: The second-order inclusion probabilities $\pi_{ij}^{AP}$for the AP design, $N = 10, n = 4$ in example 2. Single data.

	Unit j									
	1	2	3	4	5	6	7	8	9	10
Unit i										
1	0.34455	0.09537	0.10268	0.10764	0.10764	0.11267	0.11777	0.12293	0.13347	0.13347
2	0.09537	0.34455	0.10268	0.10764	0.10764	0.11267	0.11777	0.12293	0.13347	0.13347
3	0.10268	0.10268	0.36822	0.11588	0.11588	0.12128	0.12675	0.13230	0.14361	0.14361
4	0.10764	0.10764	0.11588	0.38405	0.12146	0.12711	0.13284	0.13864	0.15047	0.15047
5	0.10764	0.10764	0.11588	0.12146	0.38405	0.12711	0.13284	0.13864	0.15047	0.15047
6	0.11267	0.11267	0.12128	0.12711	0.12711	0.39991	0.13901	0.14506	0.15740	0.15740
7	0.11777	0.11777	0.12675	0.13284	0.13284	0.13901	0.41579	0.15156	0.16442	0.16442
8	0.12293	0.12293	0.13230	0.13864	0.13864	0.14506	0.15156	0.43171	0.17152	0.17152
9	0.13347	0.13347	0.14361	0.15047	0.15047	0.15740	0.16442	0.17152	0.46359	0.18595
10	0.13347	0.13347	0.14361	0.15047	0.15047	0.15740	0.16442	0.17152	0.18595	0.46359

Table 4: The variances of the AP design, 2Pπps design, CP design and Pareto design with $n = 4$ and MSE of SRSWOR ratio estimators in example 2. Single data.

Sampling scheme	Method of estimation	Variance(or MSE)
The AP design	$\widehat{\overline{Y}}_{HT}^{AP} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{AP}$	3.8268
The 2Pπps design	$\widehat{\overline{Y}}_{HT}^{2P\pi ps} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{2P\pi ps}$	3.7047
The CP design	$\widehat{\overline{Y}}_{HT}^{CP} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{CP}$	3.8681
The Pareto design	$\widehat{\overline{Y}}_{HT}^{Par} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{Par}$	3.7334
SRSWOR	Upadhyaya-Singh 1	10.5488
SRSWOR	Upadhyaya-Singh 2	15.6308
SRSWOR	Singh-Kakran	10.9737
SRSWOR	Sisodia-Dwivedi	10.4738
SRSWOR	Traditional	10.5164

AP design, CP design and 2Pπps design are compared. From the results of Table 6, we conclude that the AP design is more efficient than the CP design. Since the CP design and 2Pπ ps design are far more complex than the AP design, the proposed design is significantly better than the CP design and 2Pπ ps design and it is significantly higher than ratio estimators of the SRSWOR design mentioned by Kadilar & Cingi (2004).

A primary purpose of this paper is to extend the theory of finite sampling with unequal probabilities. Although the study variable y of the data presented in Table 5 is often unknown in the real world, they do indicate that substantial reductions in variance can be obtained by using the AP design (Table 1, 4 6). It is the opinion of the authors that the technique suggested in this paper may be an implemented utility in the real world for unknown study variable y. Hence, the proposed method has potential application value.

Table 5: Recorded Acreage of Crops and Grass for 1947 and Acreage Under Oats in 1957 for 35 Farms in Orkney in example 3. Scottish forms data.

Farm No.	x	y	Farm No.	x	y	Farm No.	x	y
1	50	17	13	78	23	25	209	70
2	50	17	14	90	0	26	240	28
3	52	10	15	91	27	27	274	62
4	58	16	16	92	34	28	300	59
5	60	6	17	96	25	29	303	66
6	60	15	18	110	24	30	311	58
7	62	20	19	140	43	31	324	128
8	65	18	20	140	48	32	330	38
9	65	14	21	156	44	33	356	69
10	68	20	22	156	45	34	410	72
11	71	24	23	190	60	35	430	103
12	74	18	24	198	63			

Table 6: The variances of the AP design, 2Pπps design, CP design with $n = 8$ and MSE of SRSWOR ratio estimators in example 3. Scottish forms data.

Sampling scheme	Method of estimation	Variance(or MSE)
The AP design	$\widehat{\overline{Y}}_{HT}^{AP} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{AP}$	15.7658
The 2Pπps design	$\widehat{\overline{Y}}_{HT}^{2P\pi ps} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{2P\pi ps}$	15.3746
The CP design	$\widehat{\overline{Y}}_{HT}^{CP} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{CP}$	16.8456
SRSWOR	Upadhyaya-Singh 1	99.4516
SRSWOR	Upadhyaya-Singh 2	99.5016
SRSWOR	Singh-Kakran	99.2217
SRSWOR	Sisodia-Dwivedi	97.9005
SRSWOR	Traditional	98.5479

4. Conclusions

We have shown that it is feasible to calculate the first-order and second-order inclusion probabilities in the AP design. Expressions for the third-order and fourth-order inclusion probabilities under the AP sampling design can be obtained. The proofs are similar to that of π_k.

This study shows that the AP design possesses approximately the same efficiency with the CP design and 2Pπps design. But the AP design is a non-rejective sampling design and very close to the strict πps design. First and second-order inclusion probabilities can be accurately calculated by using the formula given in this paper. From these numerical illustrations, it is deduced that there is considerable gain in efficiency by using the Horvitz-Thompson estimator under the AP design over the other ratio-type estimators mentioned.

Acknowledgments

The authors are grateful to the Editor-in-Chief and two referees for providing valuable comments on an earlier draft of the paper. The paper is supported by

National Natural Science Foundation of China (11161031,11361036), Natural Science Foundation of Inner Mongolia (2013MS0108) and Doctoral Fund of Ministry of Education of China (20131514110005).

[Recibido: octubre de 2013 — Aceptado: abril de 2014]

References

Aires, N. (1999), 'Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto πps', *Methodology and Computing in Applied Probability Sampling Designs* **1**(4), 457–469.

Asok, C. & Sukhatme, B. V. (1976), 'On sampford's procedure of unequal probability sampling without replacement', *Journal of the American Statistical Association* **71**(365), 912–918.

Bondesson, L. & Grafström, A. (2011), 'An extension of Sampford's method for unequal probability sampling', *Scandinavian Journal of Statistics* **38**(2), 377–392.

Bondesson, L. & Thorburn, L. D. (2008), 'A list sequential sampling method suitable for real-time sampling', *Scandinavian Journal of Statistics* **35**(3), 466–483.

Bondesson, L., Traat, I. & Lundqvist, A. (2006), 'Pareto sampling versus conditional Poisson and Sampford sampling', *Scandinavian Journal of Statistics* **33**(4), 699–720.

Brewer, K. R. W. (1963), 'A model of systematic sampling with unequal probability', *Australian and New Zealand Journal of Statistics* **5**(1), 5–13.

Durbin, J. (1967), 'Design of multi-stage surveys for the estimation of sampling errors', *Journal of the Royal Statistical Society. Series C: Applied Statistics* **16**(2), 152–164.

Grafströ, A. (2009), 'Non-rejective implementations of the Sampford sampling design', *Journal of Statistical Planning and Inference* **139**(6), 2111–2114.

Hájek, J. (1964), 'Asymptotic theory of rejective sampling with varying probabilities from a finite population', *Annals of Mathematical Statistics* **35**(4), 1491–1523.

Hájek, J. (1981), *Sampling from a Finite Population*, Marcel Dekker, New York.

Kadilar, C. & Cingi, H. (2004), 'Ratio estimators in simple random sampling', *Applied Mathematics and Computation* **151**(3), 893–902.

Laitila, T. & Olofsson, J. (2011), 'A two-phase sampling scheme and πps designs', *Journal of Statistical Planning and Inference* **141**(5), 1646–1654.

Olofsson, J. (2011), 'Algorithms to find exact inclusion probabilities for 2pπps sampling designs', *Lithuanian Mathematical Journal* **51**(3), 425–439.

Rosén, B. (1997*a*), 'Asymptotic theory for order sampling', *Journal of Statistical Planning and Inference* **62**(2), 135–158.

Rosén, B. (1997*b*), 'On sampling with probability proportional to size', *Journal of Statistical Planning and Inference* **62**(2), 159–191.

Sampford, M. R. (1967), 'On sampling without replacement with unequal probabilities of selection', *Biometrika* **54**(3-4), 499–513.

Sen, A. R. (1953), 'On the estimate of variance in sampling with varying probabilities', *Journal of the Indian Society of Agricultural Statistics* **5**, 119–127.

Singh, M. P. (1967), 'Multivariate product method of estimation for finite populations', *Journal of the Indian Society of Agricultural Statistics* **31**, 375–378.

Tillé, Y. (2006), *Sampling Algorithms*, Springer, New York.

Zaizai, Y., Miaomiao, L. & Yalu, Y. (2013), 'An efficient non-rejective implementation of the πps sampling designs', *Journal of Applied Statistics* **40**(4), 870–886.

Appendix

The derivation of the recursive formula is stated in this part.

A1. Recursive Formula of the First-Order Inclusion Probabilities

Proof of Lemma 1. We use induction on ν. Let $\nu = 0$, then $P_0^N = Pr(n_{s_0} = 0) = Pr\{\sum_{\alpha=1}^N I_\alpha = 0\} = Pr\{I_k = 0, \sum_{\alpha=1,\alpha\neq k}^N I_\alpha = 0\} = (1-p_k)Pr(n_{s_0}^{-k} = 0)$. Hence $Pr(n_{s_0}^{-k} = 0) = \mu_k P_0^N$, Lemma 1 is true for $\nu = 0$. Assume that equation (1) is true for $\nu = j < N$. Then

$$Pr(n_{s_0}^{-k} = j) = \mu_k \sum_{i=0}^{j} (-1)^{j-i} \big(\frac{p_k}{1-p_k}\big)^{j-i} P_i^N$$

Now, let $\nu = j+1 \leq N$. Then $P_{j+1}^N = Pr(n_{s_0} = j+1) = p_k Pr(n_{s_0}^{-k} = j) + (1-p_k)Pr(n_{s_0}^{-k} = j+1)$. By solving for $Pr(n_{s_0}^{-k} = j+1)$ and substituting in the expression above for $Pr(n_{s_0}^{-k} = j)$, we can get that

$$Pr(n_{s_0}^{-k} = j+1) = \mu_k \sum_{i=0}^{j+1} (-1)^{j+1-i} \big(\frac{p_k}{1-p_k}\big)^{j+1-i} P_i^N$$

□

Proof of Lemma 2. By applying Lemma 1 to the reduced population $U - \{k\}$, we can get that

$$Pr(n_{s_0}^{-kl} = \nu) = \mu_l \sum_{j=0}^{\nu} (-1)^{\nu-j} \big(\frac{p_l}{1-p_l}\big)^{\nu-j} Pr(n_{s_0}^{-k} = j)$$

Again, by substituting the expression for $Pr(n_{s_0}^{-k} = j)$ given by Lemma 1. □

Proof of Theorem 1. Firstly we note that

$$\pi_k = Pr(k \in s) = Pr(k \in s, n_{s_0} < n) + Pr(k \in s, n_{s_0} \geq n) \qquad (A1)$$

The first factor on the right of equation (A1) equals

$$Pr(k \in s, n_{s_0} = 0) + \sum_{\nu=1}^{n-1} [Pr(k \in s, I_k = 1, n_{s_0} = \nu) + Pr(k \in s, I_k = 0, n_{s_0} = \nu)],$$

where $Pr(k \in s, n_{s_0} = 0) = \dfrac{n}{N} Pr(n_{s_0} = 0) = \dfrac{n}{N}(1-p_k) Pr(n_{s_0}^{-k} = 0)$.

When $1 \leq \nu \leq n-1$,

$$\begin{aligned} Pr(k \in s, n_{s_0} = \nu) &= Pr(k \in s, I_k = 1, n_{s_0} = \nu) + Pr(k \in s, I_k = 0, n_{s_0} = \nu) \\ &= p_k \cdot Pr(n_{s_0}^{-k} = \nu - 1) + (1 - p_k) \cdot \frac{n-\nu}{N-\nu} \cdot Pr(n_{s_0}^{-k} = \nu), \end{aligned}$$

where $n_{s_0}^{-k} = \sum_{j \neq k}^{N} I_j$. The last equality follows from the fact that I_k and $n_{s_0}^{-k}$ are independent. After some simple algebraic operation, it follows that

$$\begin{aligned} & Pr(k \in s, n_{s_0} < n) \\ & = \tfrac{n}{N}(1-p_k) Pr(n_{s_0}^{-k} = 0) + \textstyle\sum_{\nu=1}^{n-1} \big[p_k Pr(n_{s_0}^{-k} = \nu - 1) \\ & + (1-p_k) \tfrac{n-\nu}{N-\nu} Pr\big(n_{s_0}^{-k} = \nu\big)\big] \end{aligned} \qquad (A2)$$

With the same notation and technique, we also derive that the second factor on the right of equation (A1) corresponds to

$$Pr(k \in s, n_{s_0} \geq n) = \sum_{\nu=n-1}^{N-1} p_k \cdot \frac{n}{\nu+1} \cdot Pr(n_{s_0}^{-k} = \nu) \qquad (A3)$$

By substituting (A3) and (A2) in the equation (A1) and some algebraic operations, the first-order inclusion probabilities can then be expressed as

$$\pi_k = \sum_{\nu=0}^{n-1} \Big[\frac{(N-n)p_k + (n-\nu)}{N-\nu} \cdot Pr(n_{s_0}^{-k} = \nu) \Big] + \sum_{\nu=n}^{N-1} \frac{np_k}{\nu+1} \cdot Pr(n_{s_0}^{-k} = \nu) \quad (A4)$$

By applying Lemma 1 to $Pr(n_{s_0}^{-k} = \nu)$ of equation (A4), we can get Theorem 1. □

A2. Recursive Formula of the Second-Order Inclusion Probabilities

Proof of Lemma 2. The second-order inclusion probabilities can be written as

$$\pi_{ij} = Pr(i \in s, j \in s, n_{s_0} < n) + Pr(i \in s, j \in s, n_{s_0} \geq n) \qquad (A5)$$

The first expression on the right of equation (A5) equals

$$Pr(i \in s, j \in s, n_{s_0} = 0) + Pr(i \in s, j \in s, n_{s_0} = 1) + \sum_{\nu=2}^{n-1} Pr(i \in s, j \in s, n_{s_0} = \nu),$$

where $Pr(i \in s, j \in s, n_{s_0} = 0) = q_i q_j \frac{n(n-1)}{N(N-1)} Pr\big(n_{s_0}^{-ij} = 0\big)$ and

$$\begin{aligned} & Pr(i \in s, j \in s, n_{s_0} = 1) \\ = \; & Pr(i \in s, j \in s, I_i = 0, I_j = 0, n_{s_0} = 1) \\ & +Pr(i \in s, j \in s, I_i = 1, I_j = 0, n_{s_0} = 1) \\ & +Pr(i \in s, j \in s, I_i = 0, I_j = 1, n_{s_0} = 1) \\ = \; & q_i q_j \frac{(n-1)(n-2)}{(N-1)(N-2)} Pr(n_{s_0}^{-ij} = 1) + (p_i q_j + q_i p_j) \frac{(n-1)}{(N-1)} Pr\big(n_{s_0}^{-ij} = 0\big) \end{aligned}$$

When $2 \leq \nu \leq n-1$,

$$\begin{aligned} & Pr(i \in s, j \in s, n_{s_0} = \nu) \\ = \; & Pr(i \in s, j \in s, I_i = 0, I_j = 0, n_{s_0} = \nu) \\ & +Pr(i \in s, j \in s, I_i = 1, I_j = 0, n_{s_0} = \nu) \\ & +Pr(i \in s, j \in s, I_i = 0, I_j = 1, n_{s_0} = \nu) \\ & +Pr(i \in s, j \in s, I_i = 1, I_j = 1, n_{s_0} = \nu) \\ = \; & q_i q_j \frac{(n-\nu)(n-\nu-1)}{(N-\nu)(N-\nu-1)} P(n_{s_0}^{-ij} = \nu) + (p_i q_j + q_i p_j) \frac{n-\nu}{N-\nu} Pr(n_{s_0}^{-ij} = \nu - 1) \\ & +p_i p_j Pr(n_{s_0}^{-ij} = \nu - 2) \end{aligned}$$

The second factor on the right of equation (A5) corresponds to

$$\sum_{\nu=n}^{N} Pr(i \in s, j \in s, I_i = 1, I_j = 1, n_{s_0} = \nu) = \sum_{\nu=n}^{N} p_i \cdot p_j \cdot \frac{n(n-1)}{\nu(\nu-1)} \cdot Pr(n_{s_0}^{-ij} = \nu - 2)$$

On substituting the expressions above in equation (A5), the π_{ij} becomes

$$\begin{aligned} \pi_{ij} = \sum_{\nu=0}^{n-2} \Big[& (1-p_i)(1-p_j) \frac{(n-\nu)(n-\nu-1)}{(N-\nu)(N-\nu-1)} \\ & +(p_i + p_j - 2p_i p_j) \frac{n-\nu-1}{N-\nu-1} + p_i p_j \Big] Pr\big(n_{s_0}^{-ij} = \nu\big) \end{aligned}$$

$$+\sum_{\nu=n-1}^{N-2} p_i p_j \frac{n(n-1)}{(\nu+2)(\nu+1)} Pr\big(n_{s_0}^{-ij} = \nu\big) \qquad (A6)$$

By using Lemma 2 to $Pr(n_{s_0}^{-ij} = \nu)$ of equation (A6), we may derive Theorem 2. □

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 141 a 158

The Beta-Gompertz Distribution

La distribución Beta-Gompertz

Ali Akbar Jafari[1,a], Saeid Tahmasebi[2,b], Morad Alizadeh[3,c]

[1]Department of Statistics, Yazd University, Yazd, Iran
[2]Department of Statistics, Persian Gulf University, Bushehr, Iran
[3]Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran

Abstract

In this paper, we introduce a new four-parameter generalized version of the Gompertz model which is called Beta-Gompertz (BG) distribution. It includes some well-known lifetime distributions such as Beta-exponential and generalized Gompertz distributions as special sub-models. This new distribution is quite flexible and can be used effectively in modeling survival data and reliability problems. It can have a decreasing, increasing, and bathtub-shaped failure rate function depending on its parameters. Some mathematical properties of the new distribution, such as closed-form expressions for the density, cumulative distribution, hazard rate function, the kth order moment, moment generating function, Shannon entropy, and the quantile measure are provided. We discuss maximum likelihood estimation of the BG parameters from one observed sample and derive the observed Fisher's information matrix. A simulation study is performed in order to investigate the properties of the proposed estimator. At the end, in order to show the BG distribution flexibility, an application using a real data set is presented.

***Key words*:** Beta generator, Gompertz distribution, Maximum likelihood estimation.

Resumen

En este artículo, se introduce una versión generalizada en cuatro parámetros de la distribución de Gompertz denominada como la distribución Beta-Gompertz (BG). Esta incluye algunas distribuciones de duración de vida bien conocidas como la Beta exponencial y distribuciones Gompertz generalizadas como casos especiales. Esta nueva distribución es flexible y puede ser usada de manera efectiva en datos de sobrevida y problemas de confiabilidad. Su función de tasa de falla puede ser decreciente, creciente o en forma de bañera

[a]Professor. E-mail: aajafari@yazd.ac.ir
[b]Professor. E-mail: tahmasebi@pgu.ac.ir
[c]Ph.D Student. E-mail: moradalizadeh78@gmail.com

dependiendo de sus parámetros. Algunas propiedades matemáticas de la distribución como expresiones en forma cerrada para la densidad, función de distribución, función de riesgo, momentos k-ésimos, función generadora de momentos, entropía de Shannon y cuantiles son presentados. Se discute la estimación máximo verosímil de los parámetros desconocidos del nuevo modelo para la muestra completa y se obtiene una expresión para la matriz de información. Con el fin de mostrar la flexibilidad de esta distribución, se presenta una aplicación con datos reales. Al final, un estudio de simulación es desarrollado.

***Palabras clave*:** distribución de Gompertz, estimación máximo verosímil, función Beta.

1. Introduction

The Gompertz (G) distribution is a flexible distribution that can be skewed to the right and to the left. This distribution is a generalization of the exponential (E) distribution and is commonly used in many applied problems, particularly in lifetime data analysis (Johnson, Kotz & Balakrishnan 1995, p. 25). The G distribution is considered for the analysis of survival, in some sciences such as gerontology (Brown & Forbes 1974), computer (Ohishi, Okamura & Dohi 2009), biology (Economos 1982), and marketing science (Bemmaor & Glady 2012). The hazard rate function (hrf) of G distribution is an increasing function and often applied to describe the distribution of adult life spans by actuaries and demographers (Willemse & Koppelaar 2000). The G distribution with parameters $\theta > 0$ and $\gamma > 0$ has the cumulative distribution function (cdf)

$$G(x) = 1 - e^{-\frac{\theta}{\gamma}(e^{\gamma x}-1)}, \quad x \geq 0, \quad \beta > 0, \quad \gamma > 0 \tag{1}$$

and the probability density function (pdf)

$$g(x) = \theta e^{\gamma x} e^{-\frac{\theta}{\gamma}(e^{\gamma x}-1)} \tag{2}$$

This case is denoted by $X \sim \mathrm{G}(\theta, \gamma)$.

Recently, a generalization based on the idea of Gupta & Kundu (1999) was proposed by El-Gohary & Al-Otaibi (2013).

This new distribution is known as generalized Gompertz (GG) distribution which includes the E, generalized exponential (GE), and G distributions (El-Gohary & Al-Otaibi 2013).

In this paper, we introduce a new generalization of G distribution which results of the application of the G distribution to the Beta generator proposed by Eugene, Lee & Famoye (2002), called the Beta-Gompertz (BG) distribution.

Several generalized distributions have been proposed under this methodology: beta-Normal distribution (Eugene et al. 2002), Beta-Gumbel distribution (Nadarajah & Kotz 2004), Beta-Weibull distribution (Famoye, Lee & Olumolade 2005), Beta-exponential (BE) distribution, (Nadarajah & Kotz 2006), Beta-Pareto distribution (Akinsete, Famoye & Lee 2008), Beta-modified Weibull distribution (Silva & Cordeiro 2010), Beta-generalized normal distribution (Cintra & Nascimento 2012). The BG distribution includes some well-known distributions: E distribution, GE distribution (Gupta & Kundu 1999), BE distribution (Nadarajah & Kotz 2006), G distribution, GG distribution (El-Gohary & Al-Otaibi 2013).

This paper is organized as follows: In Section 2, we define the density and failure rate functions and outline some special cases of the BG distribution. In Sections 3 we provide some extensions and properties of the cdf, pdf, kth moment and moment generating function of the BG distribution. Furthermore, in these sections, we derive corresponding expressions for the order statistics, Shannon entropy and quantile measure. In Section 4, we discuss maximum likelihood estimation of the BG parameters from one observed sample and derive the observed Fisher's information matrix.

A simulation study is performed in Section 5. Finally, an application of the BG using a real data set is presented in Section 6.

2. The BG Distribution

In this section, we introduce the four-parameter BG distribution. The idea of this distribution rises from the following general class: If G denotes the cdf of a random variable then a generalized class of distributions can be defined by

$$F(x) = I_{G(x)}(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\int_0^{G(x)} t^{\alpha-1}(1-t)^{\beta-1}dt, \quad \alpha,\beta > 0 \tag{3}$$

where $I_y(\alpha,\beta) = \frac{B_y(\alpha,\beta)}{B(\alpha,\beta)}$ is the incomplete beta function ratio and $B_y(\alpha,\beta) = \int_0^y t^{\alpha-1}(1-t)^{\beta-1}dt$ is the incomplete beta function.

Consider that $g(x) = \frac{dG(x)}{dx}$ is the density of the baseline distribution. Then the probability density function corresponding to (3) can be written in the form

$$f(x) = \frac{g(x)}{B(\alpha,\beta)}[G(x)]^{\alpha-1}[1-G(x)]^{\beta-1} \tag{4}$$

We now introduce the BG distribution by taking $G(x)$ in (3) to the cdf in (1) of the G distribution. Hence, the pdf of BG can be written as

$$f(x) = \frac{\theta e^{\gamma x} e^{-\frac{\beta\theta}{\gamma}(e^{\gamma x}-1)}}{B(\alpha,\beta)}[1-e^{-\frac{\theta}{\gamma}(e^{\gamma x}-1)}]^{\alpha-1} \tag{5}$$

and we use the notation $X \sim BG(\theta,\gamma,\alpha,\beta)$.

Theorem 1. *Let $f(x)$ be the pdf of the BG distribution. The limiting behavior of f for different values of its parameters is given below:*

i. If $\alpha = 1$ then $\lim_{x\to 0^+} f(x) = \theta\beta$

ii. *If* $\alpha > 1$ *then* $\lim_{x \to 0^+} f(x) = 0$.

iii. *If* $0 < \alpha < 1$ *then* $\lim_{x \to 0^+} f(x) = \infty$

iv. $\lim_{x \to \infty} f(x) = 0$

Proof . The proof of parts (i)-(iii) are obvious. For part (iv), we have

$$0 \leq [1 - e^{-\frac{\theta}{\gamma}(e^{\gamma x}-1)}]^{\alpha-1} < 1 \Rightarrow 0 < f(x) < \frac{\theta e^{\gamma x} e^{-\frac{\beta\theta}{\gamma}(e^{\gamma x}-1)}}{B(\alpha, \beta)}$$

It can be easily shown that

$$\lim_{x \to \infty} \theta e^{\gamma x} e^{-\frac{\beta\theta}{\gamma}(e^{\gamma x}-1)} = 0.$$

and the proof is completed. □

The hrf of BG distribution is given by

$$h(x) = \frac{\theta e^{\gamma x} e^{-\frac{\beta\theta}{\gamma}(e^{\gamma x}-1)}}{B(\alpha, \beta) - B_{G(x)}(\alpha, \beta)} [1 - e^{-\frac{\theta}{\gamma}(e^{\gamma x}-1)}]^{\alpha-1} \tag{6}$$

Recently, it is observed (Gupta & Gupta 2007) that the reversed hrf plays an important role in the reliability analysis. The reversed hrf of the $BG(\theta, \gamma, \alpha, \beta)$ is

$$r(x) = \frac{\theta e^{\gamma x} e^{-\frac{\beta\theta}{\gamma}(e^{\gamma x}-1)}}{B_{G(x)}(\alpha, \beta)} [1 - e^{-\frac{\theta}{\gamma}(e^{\gamma x}-1)}]^{\alpha-1} \tag{7}$$

Plots of pdf and hrf function of the BG distribution for different values of its parameters are given in Figure 1 and Figure 2, respectively.

Some well-known distributions are special cases of the BG distribution:

1. If $\alpha = 1$, $\beta = 1$, $\gamma \to 0$, then we get the E distribution.

2. If $\beta = 1$, $\gamma \to 0$, then we get the GE distribution which is introduced by Gupta & Kundu (1999)

3. If $\beta = 1$, then we get the GG distribution which is introduced by El-Gohary & Al-Otaibi (2013).

4. If $\alpha = 1$, $\beta = 1$, then we get the G distribution.

5. If $\gamma \to 0$, then we get the BE which is introduced by Nadarajah & Kotz (2006).

If the random variable X has BG distribution, then it has the following properties:

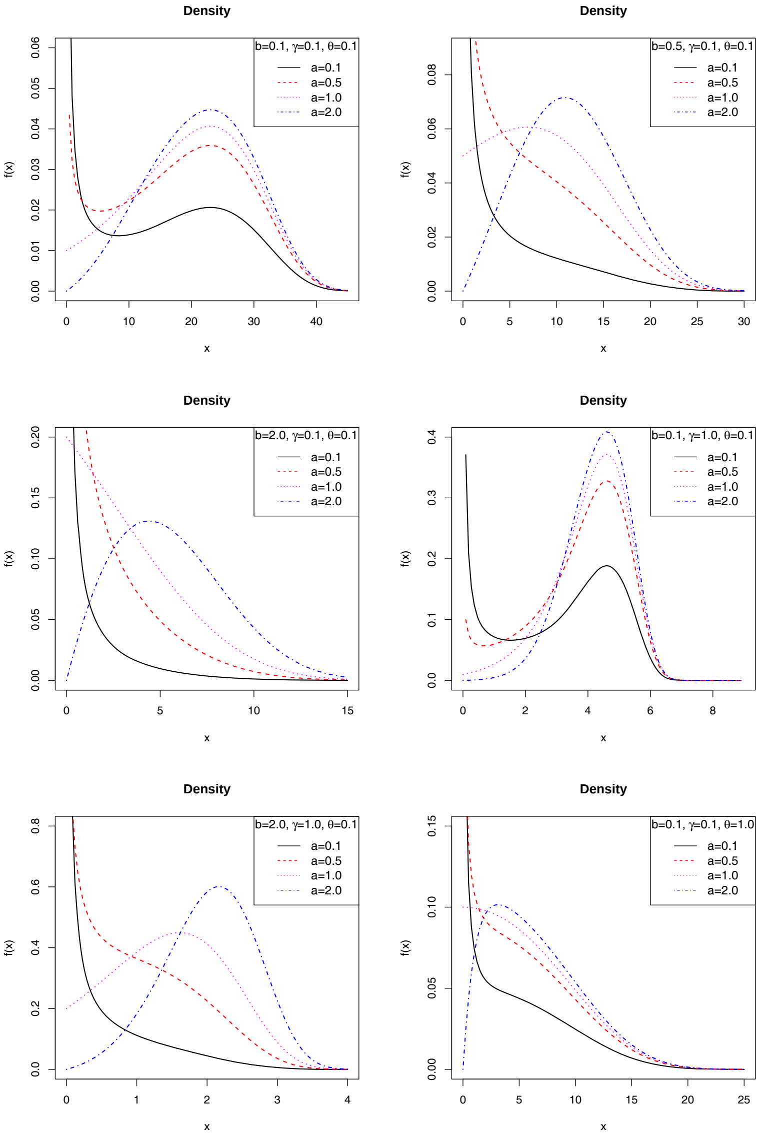


Figure 1: Plots of density functions of BG for different values of parameters.

1. the random variable
$$Y = 1 - e^{-\frac{\theta}{\gamma}(e^{\gamma X}-1)}$$
satisfies the Beta distribution with parameters α and β. Therefore,
$$T = \frac{\theta}{\gamma}(e^{\gamma X} - 1)$$
satisfies the BE distribution with parameters 1, α and β ($BE(1, \alpha, \beta)$).

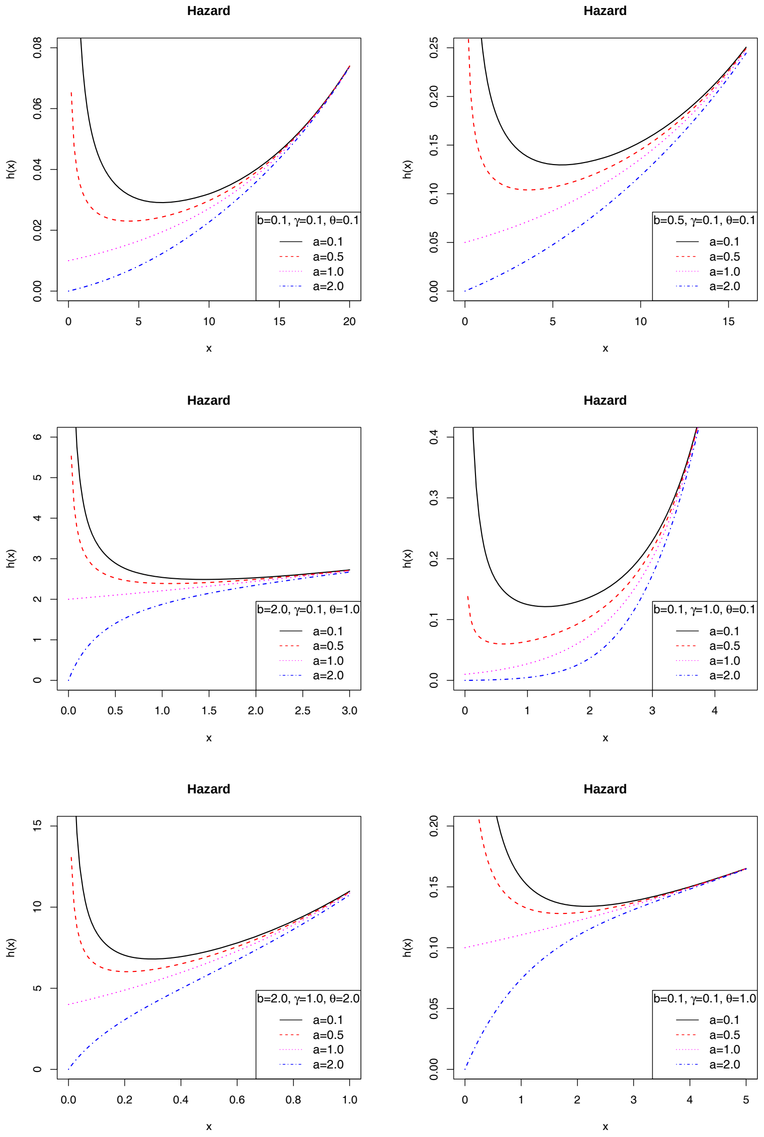


Figure 2: Plots of hrf of BG for different values of parameters.

2. If $\alpha = i$ and $\beta = n - i$, where i and n are positive integer values, then the $f(x)$ is the density function of ith order statistic of G distribution.

3. If V follows Beta distribution with parameters α and β, then

$$X = G^{-1}(V) = \frac{1}{\gamma}\log\left(1 - \frac{\gamma}{\theta}\log(1 - V)\right)$$

follows BG distribution. This result helps in simulating data from the BG distribution.

For checking the consistency of the simulating data set form BG distribution, the histogram for a generated data set with size 100 and the exact BG density with parameters $\theta = 0.1$ and $\gamma = 1.0$, $\alpha = 0.1$, and $\beta = 0.1$, are displayed in Figure 3 (left). Also, the empirical distribution function and the exact distribution function is given in Figure 3 (right).

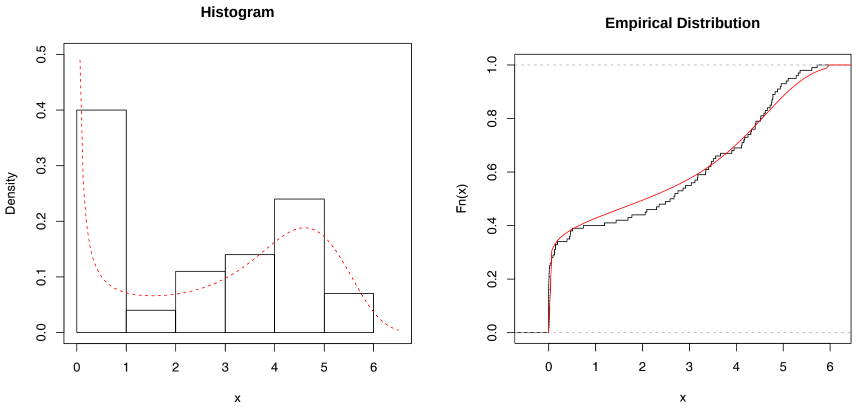


FIGURE 3: The histogram of a generated data set with size 100 and the exact GPS density (left) and the empirical distribution function and exact distribution function (right).

3. Some Extensions and Properties

Here, we present some representations of the cdf, pdf, kth moment and moment generating function of BG distribution. Also, we provide expressions for the order statistics, Shannon entropy and quantile measure of this distribution. The mathematical relation given below will be useful in this section. If β is a positive real non-integer and $|z| < 1$, then (Gradshteyn & Ryzhik 2007, p. 25)

$$(1 - z)^{\beta - 1} = \sum_{j=0}^{\infty} w_j z^j$$

and if β is a positive real integer, then the upper of the this summation stops at $\beta - 1$, where

$$w_j = \frac{(-1)^j \Gamma(\beta)}{\Gamma(\beta - j)\Gamma(j + 1)}$$

Proposition 1. *We can express (3) as a mixture of distribution function of GG distributions as follows:*

$$F(x)=\sum_{j=0}^{\infty}p_j[G(x)]^{\alpha+j}=\sum_{j=0}^{\infty}p_jG_j(x)$$

where $p_j=\frac{(-1)^j\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta-j)\Gamma(j+1)(\alpha+j)}$ *and* $G_j(x)=(G(x))^{\alpha+j}$ *is the distribution function of a random variable which has a GG distribution with parameters* θ, γ, *and* $\alpha+j$. *Also, we can write*

$$\begin{aligned}G(x)^{\alpha+j}&=\sum_{k=0}^{\infty}(-1)^k\binom{\alpha+j}{k}(1-G(x))^k\\&=\sum_{r=0}^{\infty}\sum_{k=r}^{\infty}(-1)^{k+r}\binom{\alpha+j}{k}\binom{k}{r}G(x)^r\end{aligned}\tag{8}$$

and

$$F(x)=\sum_{j=0}^{\infty}\sum_{r=0}^{\infty}\sum_{k=r}^{\infty}p_j(-1)^{k+r}\binom{\alpha+j}{k}\binom{k}{r}G(x)^r=\sum_{r=0}^{\infty}b_rG(x)^r\tag{9}$$

where $b_r=\sum\limits_{j=0}^{\infty}\sum\limits_{k=r}^{\infty}p_j(-1)^{k+r}\binom{\alpha+j}{k}\binom{k}{r}$

Proposition 2. *We can express (5) as a mixture of density functions of a GG distribution as follows:*

$$f(x)=\sum_{j=0}^{\infty}p_j(\alpha+j)g(x)[G(x)]^{\alpha+j-1}=\sum_{j=0}^{\infty}p_jg_j(x)$$

where $g_j(x)$ *is a density function of a random variable with a GG distribution and parameters* θ, γ, *and* $\alpha+j$.

Proposition 3. *The cdf can be expressed in terms of the hypergeometric function and the incomplete beta function ratio (see Cordeiro & Nadarajah 2011) in the following way:*

$$F(x)=\frac{(G(x))^{\alpha}}{\alpha B(\alpha,\beta)}\ {}_2F_1(\alpha,1-\beta;\alpha+1;G(x))$$

where ${}_2F_1(a,b;c;z)=\sum\limits_{k=0}^{\infty}\frac{((a)_k(b)_k)}{((c)_kk!)}z^k$ *and* $(a)_k=a(a+1)\cdots(a+k-1)$

Proposition 4. *The* k*th moment of the BG distribution can be expressed as a mixture of the* k*th moment of GG distributions as follows:*

$$E(X^k)\ =\ \int_0^{\infty}x^k\sum_{j=0}^{\infty}p_j(\alpha+j)g(x)[G(x)]^{\alpha+j-1}=\sum_{j=0}^{\infty}p_jE(X_j^k)\tag{10}$$

where

$$E[X_j^k] = u_{jk}\sum_{i=0}^{\infty}\sum_{r=0}^{\infty}\binom{\alpha+j-1}{i}\frac{(-1)^{i+r}}{\Gamma(r+1)}e^{\frac{\theta}{\gamma}(i+1)}[\frac{\theta}{\gamma}(i+1)]^r[\frac{-1}{\gamma(k+1)}]^{s+1}$$

$u_{jk} = (\alpha+j)\theta\Gamma(k+1)$ *and* $g_j(x)$ *is the density function of a random variable* X_j *which has a GG distribution with parameters* θ*,* γ*, and* $\alpha+j$*.*

Proposition 5. *The moment generating function of the BG distribution can be expressed as a mixture of moment generating function of GG distributions as follows:*

$$M_X(t) \quad = \textstyle\int_0^\infty e^{tx}\sum\limits_{j=0}^{\infty}p_j(\alpha+j)g(x)[G(x)]^{\alpha+j-1} = \sum\limits_{j=0}^{\infty}p_jM_{X_j}(t) \qquad (11)$$

where

$$M_{X_j}(t) = \frac{(\alpha+j)\theta}{\gamma}\sum_{i=0}^{\infty}\sum_{k=0}^{\infty}(-1)^i\binom{\alpha+j-1}{i}\binom{\frac{t}{\gamma}}{k}\frac{\Gamma(k+1)}{[\frac{(i+1)\theta}{\gamma}]^{k+1}}$$

and $g_j(x)$ *is the density function of a random variable* X_j *which has a GG distribution with parameters* θ*,* γ*, and* $\alpha+j$*.*

3.1. Order Statistics

Moments of order statistics play an important role in quality control testing and reliability. For example, if the reliability of an item is high, the duration of an all items fail life test can be too expensive in both time and money.

Therefore, a practitioner needs to predict the failure of future items based on the times of few early failures. These predictions are often based on moments of order statistics.

Let $X_1, X_2, \dots, X_n$ be a random sample of size n from $BG(\theta,\gamma,\alpha,\beta)$. Then the pdf and cdf of the ith order statistic, say $X_{i:n}$, are given by

$$f_{i:n}(x) = \frac{1}{B(i, n-i+1)}\sum_{m=0}^{n-i}(-1)^m\binom{n-i}{m}f(x)F^{i+m-1}(x) \qquad (12)$$

and

$$F_{i:n}(x) = \int_0^x f_{i:n}(t)dt = \frac{1}{B(i, n-i+1)}\sum_{m=0}^{n-i}\frac{(-1)^m}{m+i}\binom{n-i}{m}F^{i+m}(x) \qquad (13)$$

respectively, where $F^{i+m}(x) = (\sum\limits_{r=0}^{\infty}b_rG(x)^r)^{i+m}$. Here and henceforth, we use an equation by Gradshteyn & Ryzhik (2007), page 17, for a power series raised to a positive integer n

$$\left(\sum_{r=0}^{\infty}b_r\,u^r\right)^n = \sum_{r=0}^{\infty}c_{n,r}\,u^r \qquad (14)$$

where the coefficients $c_{n,r}$ (for $r = 1, 2, \ldots$) are easily determined from the recurrence equation

$$c_{n,r} = (r\, b_0)^{-1} \sum_{m=1}^{r} \left[m\,(n+1) - r\right] b_m\, c_{n,r-m}, \tag{15}$$

where $c_{n,0} = b_0^n$. The coefficient $c_{n,r}$ can be calculated from $c_{n,0}, \ldots, c_{n,r-1}$ and hence from the quantities $b_0, \ldots, b_r$.

The equations (12) and (13) can be written as

$$f_{i:n}(x) = \frac{1}{B(i, n-i+1)} \sum_{m=0}^{n-i} \sum_{r=1}^{\infty} \frac{1}{m+i} (-1)^m r c_{i+m,r} g(x) G^{r-1}(x)$$

and

$$F_{i:n}(x) = \frac{1}{B(i, n-i+1)} \sum_{m=0}^{n-i} \sum_{r=0}^{\infty} \frac{1}{m+i} (-1)^m c_{i+m,r} G^{r}(x)$$

Therefore, the sth moments of $X_{i:n}$ follows as

$$\begin{aligned} E[X_{i:n}^s] &= \frac{1}{B(i, n-i+1)} \sum_{m=0}^{n-i} \sum_{r=1}^{\infty} \frac{1}{m+i} (-1)^m r c_{i+m,r} \int_0^{+\infty} t^s g(t) G^{r-1}(t) dt \\ &= \frac{1}{B(i, n-i+1)} \sum_{m=0}^{n-i} \sum_{r=1}^{\infty} \frac{1}{m+i} (-1)^m r c_{i+m,r} \\ &\times \theta \Gamma(s+1) \sum_{i_1=0}^{\infty} \sum_{i_2=0}^{\infty} \binom{r-1}{i_1} \frac{(-1)^{i_1+i_2}}{\Gamma(i_2+1)} e^{\frac{\theta}{\gamma}(i_1+1)} [\frac{\theta(i_1+1)}{\gamma}]^{i_2} [\frac{-1}{\gamma(i_2+1)}]^{s+1} \end{aligned}$$

3.2. Quantile Measure

The quantile function of the BG distribution is given by

$$Q(u) = \frac{1}{\gamma} \log(1 - \frac{\gamma}{\theta} \log(1 - Q_{\alpha,\beta}(u)))$$

where $Q_{\alpha,\beta}(u)$ is the uth quantile of Beta distribution with parameters α and β. The effects of the shape parameters α and β on the skewness and kurtosis can be considered based on quantile measures. The Bowley skewness (Kenney & Keeping 1962) is one of the earliest skewness measures defined by

$$\mathcal{B} = \frac{Q(\frac{3}{4}) + Q(\frac{1}{4}) - 2Q(\frac{1}{2})}{Q(\frac{3}{4}) - Q(\frac{1}{4})}$$

This adds robustness to the measure, since only the middle two quartiles are considered and the other two quartiles are ignored. The Moors kurtosis (Moors 1988) is defined as

$$\mathcal{M} = \frac{Q(\frac{3}{8}) - Q(\frac{1}{8}) + Q(\frac{7}{8}) - Q(\frac{5}{8})}{Q(\frac{6}{8}) - Q(\frac{2}{8})}$$

Clearly, $\mathcal{M} > 0$ and there is a good concordance with the classical kurtosis measures for some distributions. These measures are less sensitive to outliers and they exist even for distributions without moments. For the standard normal distribution, these measures are 0 (Bowley) and 1.2331 (Moors).

In Figures 4 and 5, we plot the measures $\mathcal{B}$ and $\mathcal{M}$ for some parameter values. These plots indicate that both measures $\mathcal{B}$ and $\mathcal{M}$ depend on all shape parameters.

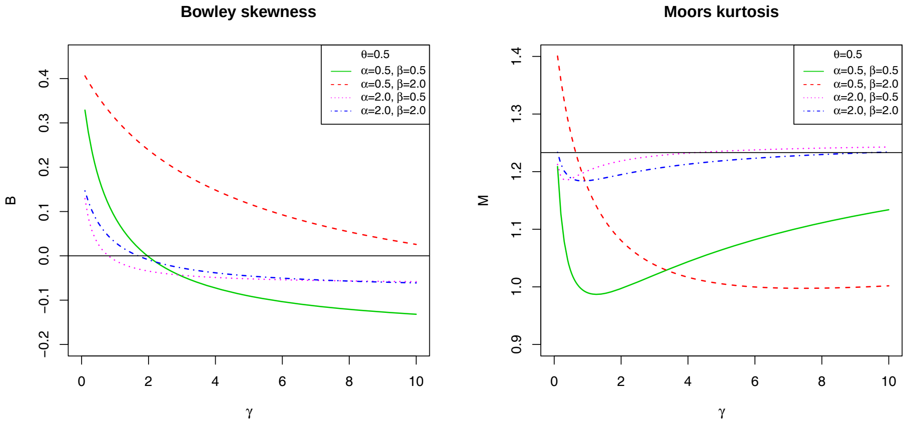


FIGURE 4: The Bowley skewness (left) and Moors kurtosis (right) coefficients for the BG distribution as a function of γ.

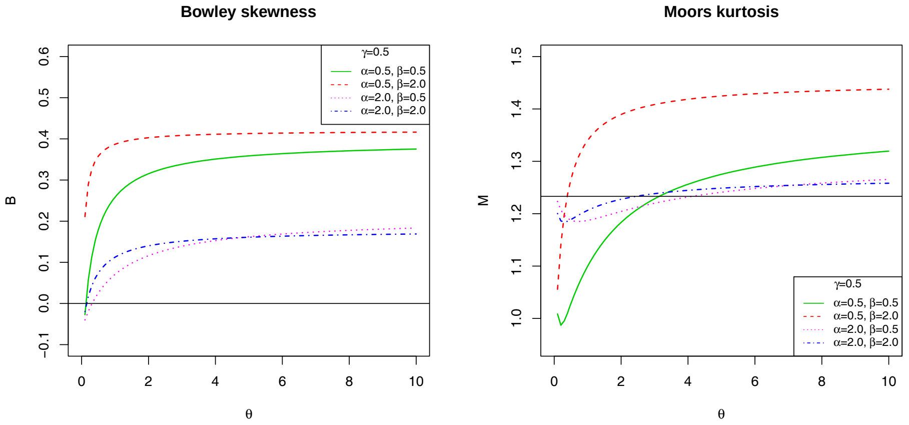


FIGURE 5: The Bowley skewness (left) and Moors kurtosis (right) coefficients for the BG distribution as a function of θ.

3.3. Shannon and Rényi Entropy

If X is a non-negative continuous random variable with pdf $f(x)$, then Shannon's entropy of X is defined by Shannon (1948) as

$$H(f) = E[-\log f(X)] = -\int_0^{+\infty} f(x)\log(f(x))dx$$

and this is usually referred to as the continuous entropy (or differential entropy). An explicit expression of Shannon entropy for BG distribution is obtained as

$$\begin{aligned} H(f) = \log(\frac{B(\alpha,\beta)}{\theta}) - \frac{\theta\beta}{\gamma} - \gamma E(X) \\ + \frac{\theta\beta}{\gamma} M_X(\gamma) + (\alpha-1)[\psi(\alpha+\beta)-\psi(\alpha)] \end{aligned} \tag{16}$$

where $\psi(.)$ is a digamma function.

The Rényi entropy of order λ is defined as

$$H_\lambda(f) = \frac{1}{1-\lambda}\log\int_{-\infty}^{+\infty} f^\lambda(x)dx, \quad \forall\lambda > 0 \ \ (\lambda \neq 1) \tag{17}$$

where

$$H(X) = \lim_{\lambda\to 1} H_\lambda(X) = -\int_{-\infty}^{+\infty} f(x)\log f(x)dx$$

is the Shannon entropy, if both integrals exist. Finally, an explicit expression of Rényi entropy for BG distribution is obtained as

$$\begin{aligned} H_\lambda(f) = -\log(\theta) + \frac{\lambda}{\lambda-1}\log(B(\alpha,\beta)) + \frac{1}{1-\lambda}\Bigg[\log(B(\alpha,(\beta-1)\lambda+1)) \\ + \log\left(\sum_{j=1}^{\infty}\sum_{k=0}^{j}(-1)^k\binom{\lambda-1}{j}\binom{j}{k}\left(\frac{\gamma}{\theta}\right)^j \frac{\Gamma(j+1)}{(j+1)^{k-1+(\beta-1)\lambda}}\right)\Bigg] \end{aligned} \tag{18}$$

4. Estimation and Inference

In this section, we determine the maximum-likelihood estimates (MLEs) of the parameters of the BG distribution from a complete sample. Consider $X_1,\dots,X_n$ is a random sample from BG distribution. The log-likelihood function for the vector of parameters $\Theta = (\theta,\gamma,\alpha,\beta)$ can be written as

$$\begin{aligned} l_n &= l_n(\Theta) \\ &= n\log(\theta) - n\log(B(\alpha,\beta)) + n\gamma\bar{x} - \beta\theta\sum_{i=1}^{n}\log(t_i) \\ &+ (\alpha-1)\sum_{i=1}^{n}\log(1-t_i^\theta) \end{aligned} \tag{19}$$

where $\bar{x} = n^{-1}\sum\limits_{i=1}^{n} x_i$ and $t_i = e^{\frac{-1}{\gamma}(e^{\gamma x_i}-1)}$. The log-likelihood can be maximized either directly or by solving the nonlinear likelihood equations obtained by differentiating (19). The components of the score vector $U(\Theta)$ are given by

$$\begin{aligned}
U_\alpha(\Theta) &= \frac{\partial l_n}{\partial \alpha} = n\psi(\alpha+\beta) - n\psi(\alpha) + \sum_{i=1}^{n} \log(1-t_i^\theta)\\
U_\beta(\Theta) &= \frac{\partial l_n}{\partial \beta} = n\psi(\alpha+\beta) - n\psi(\beta) - \theta\sum_{i=1}^{n} \log(t_i)\\
U_\theta(\Theta) &= \frac{\partial l_n}{\partial \theta} = \frac{n}{\theta} - \beta\sum_{i=1}^{n} \log(t_i) - (\alpha-1)\sum_{i=1}^{n} \frac{t_i^\theta \log(t_i)}{1-t_i^\theta}\\
U_\gamma(\Theta) &= \frac{\partial l_n}{\partial \gamma} = n\bar{x} - \beta\theta\sum_{i=1}^{n} d_i - \theta(\alpha-1)\sum_{i=1}^{n} \frac{d_i t_i^\theta}{1-t_i^\theta}
\end{aligned}$$

where $\psi(.)$ is the digamma function, and

$$d_i = \frac{\partial \log(t_i)}{\partial \gamma} = \frac{1}{\gamma}(-\log(t_i) + \gamma x_i \log(t_i) - x_i)$$

For interval estimation and hypothesis tests on the model parameters, we require the observed information matrix. The 4×4 unit observed information matrix $J = J_n(\Theta)$ is obtained as

$$J = -\begin{bmatrix} J_{\alpha\alpha} & J_{\alpha\beta} & J_{\alpha\theta} & J_{\alpha\gamma}\\ J_{\alpha\beta} & J_{\beta\beta} & J_{\beta\theta} & J_{\beta\gamma}\\ J_{\alpha\theta} & J_{\beta\theta} & J_{\theta\theta} & J_{\theta\gamma}\\ J_{\alpha\gamma} & J_{\beta\gamma} & J_{\theta\gamma} & J_{\gamma\gamma} \end{bmatrix}$$

where the expressions for the elements of J are

$$J_{\alpha\alpha} = \frac{\partial^2 l_n}{\partial \alpha^2} = n\psi'(\alpha+\beta) - n\psi'(\alpha), \qquad J_{\alpha\beta} = \frac{\partial^2 l_n}{\partial\alpha\partial\beta} = \frac{\partial^2 l_n}{\partial\beta\partial\alpha} = n\psi'(\alpha+\beta)$$

$$J_{\alpha\theta} = \frac{\partial^2 l_n}{\partial\alpha\partial\theta} = \frac{\partial^2 l_n}{\partial\theta\partial\alpha} = \sum_{i=1}^{n}\frac{t_i^\theta\log(t_i)}{1-t_i^\theta}, \qquad J_{\alpha\gamma} = \frac{\partial^2 l_n}{\partial\alpha\partial\gamma} = \frac{\partial^2 l_n}{\partial\gamma\partial\alpha} = -\theta\sum_{i=1}^{n}\frac{d_i t_i^\theta}{1-t_i^\theta},$$

$$J_{\beta\beta} = \frac{\partial^2 l_n}{\partial \beta^2} = n\psi'(\alpha+\beta) - n\psi'(\beta), \qquad J_{\beta\theta} = \frac{\partial^2 l_n}{\partial\beta\partial\theta} = \frac{\partial^2 l_n}{\partial\theta\partial\beta} = -\sum_{i=1}^{n}\log(t_i)$$

$$J_{\beta\gamma} = \frac{\partial^2 l_n}{\partial\beta\partial\gamma} = \frac{\partial^2 l_n}{\partial\gamma\partial\beta} = -\theta\sum_{i=1}^{n} d_i, \quad J_{\theta\theta} = \frac{\partial^2 l_n}{\partial\theta^2} = -\frac{n}{\theta^2} + \theta(\alpha-1)\sum_{i=1}^{n}\frac{t_i^\theta(\log(t_i))^2}{(1-t_i^\theta)^2}$$

$$J_{\theta\gamma} = \frac{\partial^2 l_n}{\partial\theta\partial\gamma} = \frac{\partial^2 l_n}{\partial\gamma\partial\theta} = -\beta\sum_{i=1}^{n} d_i - (\alpha-1)\sum_{i=1}^{n}\frac{d_i t_i^\theta}{1-t_i^\theta}\left(\theta\log(t_i) + 1 + \frac{\theta t_i^\theta\log(t_i)}{1-t_i^\theta}\right)$$

$$J_{\gamma\gamma} = \frac{\partial^2 l_n}{\partial\gamma^2} = -\beta\theta\sum_{i=1}^{n} q_i - \theta(\alpha-1)\sum_{i=1}^{n}\frac{t_i^\theta}{1-t_i^\theta}(q_i + \theta d_i^2) - \theta^2(\alpha-1)\sum_{i=1}^{n}\frac{d_i^2 t_i^{2\theta}}{1-t_i^\theta}$$

where $q_i = \frac{\partial d_i}{\partial\gamma} = d_i(x_i - \frac{2}{\gamma}) + \frac{x_i}{\gamma}\log(t_i)$.

5. Simulation Studies

In this section, we performed a simulation study in order to investigate the proposed estimator of parameters based on the proposed MLE method. We generate 10,000 data sets with size n from the BG distribution with parameters a, b, θ, and γ, and compute the MLE's of the parameters. We assess the accuracy of the approximation of the standard error of the MLE's determined through the Fisher information matrix and variance of the estimated parameters. Table 1 show the results for the BG distribution. From these results, we can conclude that:

i. the differences between the average estimates and the true values are almost small,

ii. the MLE's converge to true value in all cases when the sample size increases,

iii. the standard errors of the MLEs decrease when the sample size increases.

From these simulations, we can conclude that estimation of the parameters using the MLE are satisfactory.

6. Application of BG to a Real Data Set

In this section, we perform an application to real data and demonstrate the superiority of BG distribution as compared to some of its sub-models. The data have been obtained from Aarset (1987), and widely reported in some literatures (for example see Silva & Cordeiro 2010). It represents the lifetimes of 50 devices, and also, possess a bathtub-shaped failure rate property. The numerical evaluations were implemented using R software (nlminb function).

Based on some goodness-of-fit measures, the performance of the BG distribution is quantified and compared with others due to five literature distributions: E, GE, BE, G, and GG, distributions. The MLE's of the unknown parameters (standard errors in parentheses) for these distributions are given in Table 2. Also, the values of the log-likelihood functions ($-\log(L)$), the Kolmogorov Smirnov (KS) test statistic with its p-value, the statistics AIC (Akaike Information Criterion), the statistics AICC (Akaike Information Citerion with correction) and BIC (Bayesian Information Criterion) are calculated for the six distributions in order to verify which distribution fits better to these data. All the computations were done using the R software.

The BG distribution yields the highest value of the log-likelihood function and smallest values of the AIC, AICC and BIC statistics. From the values of these statistics, we can conclude that the BG model is better than the other distributions to fit these data. The plots of the densities (together with the data histogram) and cumulative distribution functions (with empirical distribution function) are given in Figure 6. It is evident that the BG model provides a better fit than the other models. In particular, the histogram of data shows that the BG model provides an excellent fit to these data.

Table 1: The simulated MLE's and mean of the standard errors for BG distribution based on information matrix and variance of MLE's.

	parameters				estimations				simulated				Information matrix			
n	α	β	θ	γ	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\theta}$	$\hat{\gamma}$	$s.e.(\hat{\alpha})$	$s.e.(\hat{\beta})$	$s.e.(\hat{\theta})$	$s.e.(\hat{\gamma})$	$s.e.(\hat{\alpha})$	$s.e.(\hat{\beta})$	$s.e.(\hat{\theta})$	$s.e.(\hat{\gamma})$
30	0.5	0.5	0.5	0.5	0.515	0.739	0.555	0.616	0.171	0.646	0.818	0.291	0.111	0.183	0.117	0.072
	0.5	0.5	0.5	1.0	0.516	0.734	0.605	1.164	0.180	0.609	0.800	0.472	0.123	0.168	0.137	0.089
	0.5	0.5	1.0	1.0	0.517	0.858	1.195	1.245	0.166	0.604	0.845	0.583	0.117	0.191	0.154	0.132
	0.5	2.0	0.5	1.0	0.509	1.741	0.902	1.368	0.144	0.824	0.388	0.839	0.111	0.306	0.129	0.222
	0.5	2.0	1.0	1.0	0.497	1.830	1.441	1.478	0.130	0.877	0.575	1.100	0.105	0.354	0.217	0.352
	2.0	2.0	0.5	0.5	2.068	1.916	0.811	0.826	0.972	0.744	0.881	0.544	0.333	0.147	0.217	0.094
50	0.5	0.5	0.5	0.5	0.506	0.612	0.526	0.567	0.125	0.600	0.900	0.214	0.088	0.143	0.111	0.054
	0.5	0.5	0.5	1.0	0.509	0.729	0.556	1.092	0.131	0.633	0.864	0.342	0.085	0.137	0.096	0.077
	0.5	0.5	1.0	1.0	0.512	0.829	1.104	1.149	0.126	0.598	0.941	0.421	0.086	0.147	0.123	0.104
	0.5	2.0	0.5	1.0	0.507	1.896	0.888	1.183	0.111	0.795	0.345	0.631	0.081	0.244	0.082	0.179
	0.5	2.0	1.0	1.0	0.501	1.932	1.202	1.202	0.101	0.856	0.515	0.873	0.080	0.288	0.147	0.282
	2.0	2.0	0.5	0.5	2.201	1.991	0.678	0.659	0.845	0.734	0.856	0.350	0.372	0.114	0.150	0.060
100	0.5	0.5	0.5	0.5	0.498	0.520	0.511	0.532	0.086	0.617	0.948	0.149	0.060	0.101	0.076	0.039
	0.5	0.5	0.5	1.0	0.498	0.531	0.537	1.039	0.090	0.655	0.917	0.241	0.061	0.098	0.073	0.054
	0.5	0.5	1.0	1.0	0.503	0.540	0.984	1.088	0.086	0.634	1.033	0.295	0.059	0.107	0.085	0.076
	0.5	2.0	0.5	1.0	0.504	1.929	0.492	1.062	0.076	0.722	0.349	0.425	0.055	0.172	0.046	0.133
	0.5	2.0	1.0	1.0	0.502	1.917	1.259	1.163	0.072	0.814	0.479	0.608	0.055	0.212	0.084	0.209
	2.0	2.0	0.5	0.5	2.195	2.062	0.550	0.590	0.648	0.738	0.791	0.229	0.261	0.083	0.060	0.044

For this data set, we perform the Likelihood Ratio Test (LRT) for testing the following hypotheses:

1. H_0: E distribution vs. H_1: BG distribution
2. H_0: GE distribution vs. H_1: BG distribution
3. H_0: BE distribution vs. H_1: BG distribution
4. H_0: G distribution vs. H_1: BG distribution, or equivalently H_0: $(\alpha, \beta) = (1, 1)$ vs. H_1: $(\alpha, \beta) \neq (1, 1)$
5. H_0: GG distribution vs. H_1: BG distribution, or equivalently H_0: $\beta = 1$ vs. H_1: $\beta \neq 1$.

Table 2: Parameter estimates (with std.), K-S statistic, *p*-value for K-S, AIC, AICC, BIC, LRT statistic and *p*-value of LRT for the data set.

Distribution	E	GE	BE	G	GG	BG
$\hat{\alpha}$	—	0.9021	0.5236	—	0.2625	0.2158
(std.)	—	(0.1349)	(0.1714)	—	(0.0395)	(0.0392)
$\hat{\beta}$	—	—	0.0847	—	—	0.2467
(std.)	—	—	(0.0828)	—	—	(0.0448)
$\hat{\theta}$	0.0219	0.0212	0.2352	0.0097	0.0001	0.0003
(std.)	(0.0031)	(0.0036)	(0.2111)	(0.0029)	(0.0001)	(0.0001)
$\hat{\gamma}$	—	—	—	0.0203	0.0828	0.0882
(std.)	—	—	—	(0.0058)	(0.0031)	(0.0030)
$-\log(L)$	241.0896	240.3855	238.1201	235.3308	222.2441	220.6714
K-S	0.1911	0.1940	0.1902	0.1696	0.1409	0.1322
p-value (K-S)	0.0519	0.0514	0.0538	0.1123	0.2739	0.3456
AIC	484.1792	484.7710	482.2400	474.6617	450.4881	449.3437
AICC	484.2625	485.0264	482.7617	475.1834	451.0099	450.2326
BIC	486.0912	488.5951	487.9760	482.3977	456.2242	456.9918
LRT	40.8355	39.4273	34.8962	29.3179	3.1444	—
p-value (LRT)	0.0000	0.0000	0.0000	0.0001	0.0762	—

Values of the LRT statistic and its corresponding p-value for each hypotheses are given in Table 2. From these results, we can conclude that the null hypotheses are rejected in all situations, and therefore, the BG distribution is an adequate model.

***Note* 1.** El-Gohary & Al-Otaibi (2013) found the following estimations for the parameters of th GG distribution:

$$\hat{a} = 0.421, \quad \hat{\theta} = 0.00143, \quad \hat{\gamma} = 0.044.$$

Based on these estimations, the log-likelihood function is equal to -224.1274. But we found the following estimations for the parameters of GG distribution:

$$\hat{a} = 0.2625, \quad \hat{\theta} = 0.0001, \quad \hat{\gamma} = 0.0828.$$

Based on these estimations, the log-likelihood function is equal to -222.2441. Therefore, the estimations of El-Gohary & Al-Otaibi (2013) for GG distribution is not the MLE.

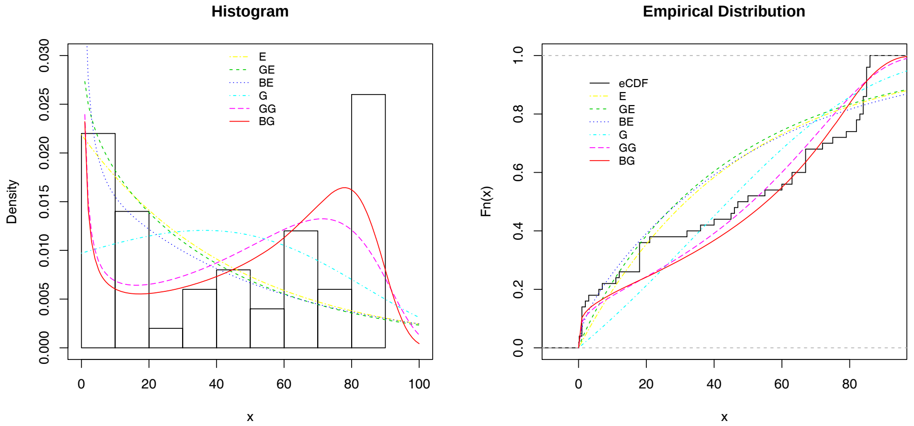


FIGURE 6: Plots (density and distribution) of fitted E, GE, BE, G, GG and BG distributions for the data set.

Acknowledgements

The authors would like to thank the Editor and the anonymous referees for their constructive comments and suggestions that appreciably improved the quality of presentation of this manuscript.

[Recibido: octubre de 2013 — Aceptado: marzo de 2014]

References

Aarset, M. V. (1987), 'How to identify a bathtub hazard rate', *IEEE Transactions on Reliability* **36**(1), 106–108.

Akinsete, A., Famoye, F. & Lee, C. (2008), 'The Beta-Pareto distribution', *Statistics* **42**(6), 547–563.

Bemmaor, A. C. & Glady, N. (2012), 'Modeling purchasing behavior with sudden "death": A flexible customer lifetime model', *Management Science* **58**(5), 1012–1021.

Brown, K. & Forbes, W. (1974), 'A mathematical model of aging processes', *Journal of Gerontology* **29**(1), 46–51.

Cintra, R. J., R. L. C. C. G. M. & Nascimento, A. D. C. (2012), 'Beta generalized normal distribution with an application for SAR image processing', *Statistics: A Journal of Theoretical and Applied Statistics* pp. 1–16. DOI:10.1080/02331888.2012.748776.

Cordeiro, G. M. & Nadarajah, S. (2011), 'Closed-form expressions for moments of a class of Beta generalized distributions', *Brazilian Journal of Probability and Statistics* **25**(1), 14–33.

Economos, A. C. (1982), 'Rate of aging, rate of dying and the mechanism of mortality', *Archives of Gerontology and Geriatrics* **1**(1), 46–51.

El-Gohary, A. & Al-Otaibi, A. N. (2013), 'The generalized Gompertz distribution', *Applied Mathematical Modelling* **37**(1-2), 13–24.

Eugene, N., Lee, C. & Famoye, F. (2002), 'Beta-normal distribution and its applications', *Communications in Statistics - Theory and Methods* **31**(4), 497–512.

Famoye, F., Lee, C. & Olumolade, O. (2005), 'The Beta-Weibull distribution', *Journal of Statistical Theory and Applications* **4**(2), 121–136.

Gradshteyn, I. S. & Ryzhik, I. M. (2007), *Table of Integrals, Series, and Products*, 7 edn, Academic Press, New York.

Gupta, R. C. & Gupta, R. D. (2007), 'Proportional reversed hazard rate model and its applications', *Journal of Statistical Planning and Inference* **137**(11), 3525–3536.

Gupta, R. D. & Kundu, D. (1999), 'Generalized exponential distributions', *Australian & New Zealand Journal of Statistics* **41**(2), 173–188.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995), *Continuous Univariate Distributions*, Vol. 2, 2 edn, John Wiley & Sons, New York.

Kenney, J. F. & Keeping, E. (1962), *Mathematics of Statistics*, Van Nostrand.

Moors, J. J. A. (1988), 'A quantile alternative for kurtosis', *Journal of the Royal Statistical Society. Series D (The Statistician)* **37**(1), 25–32.

Nadarajah, S. & Kotz, S. (2004), 'The Beta Gumbel distribution', *Mathematical Problems in Engineering* **4**, 323–332.

Nadarajah, S. & Kotz, S. (2006), 'The Beta exponential distribution', *Reliability Engineering & System Safety* **91**(6), 689–697.

Ohishi, K., Okamura, H. & Dohi, T. (2009), 'Gompertz software reliability model: Estimation algorithm and empirical validation', *Journal of Systems and Software* **82**(3), 535–543.

Shannon, C. (1948), 'A mathematical theory of communication', *Bell System Technical Journal* **27**, 379–432.

Silva, G. O., O. E. M. & Cordeiro, G. M. (2010), 'The Beta modified Weibull distribution', *Lifetime Data Analysis* **16**(3), 409–430.

Willemse, W. & Koppelaar, H. (2000), 'Knowledge elicitation of Gompertz' law of mortality', *Scandinavian Actuarial Journal* **2**, 168–179.

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 159 a 181

Profile Monitoring for Compositional Data

Monitoreo de perfiles para datos composicionales

Rubén Darío Guevara-González[1,a], José Alberto Vargas-Navas[1,b], Dorian Luis Linero-Segrera[2,c]

[1] Department of Statistics, National University of Colombia, Bogotá, Colombia
[2] Engineering School, National University of Colombia, Bogotá, Colombia

Abstract

In a growing number of quality control applications, the quality of a product or process is best characterized and summarized by a functional relationship between a response variable and one or more explanatory variables. Profile monitoring is used to understand and to check the stability of this relationship over time. In some applications with compositional data, the relationship can be characterized by a Dirichlet regression model. We evaluate five T^2 control charts for monitoring these profiles in Phase I. A real example from production of concrete is given.

***Key words*:** Control chart, Dirichlet distribution, Statistical process control.

Resumen

En un gran número de aplicaciones la calidad de un producto o proceso está mejor representada por una relación funcional entre una variable de respuesta y una o más variables explicatorias. El monitoreo de perfiles permite entender y chequear la estabilidad de esta relación funcional a través del tiempo. En algunas aplicaciones con datos composicionales, la relación puede ser representada por un modelo de regresión Dirichlet. En este artículo nosotros evaluamos cinco cartas de control T^2 para monitorear estos perfiles en Fase I. Un ejemplo real asociado a la producción de concreto es presentado.

***Palabras clave*:** carta de control, control estadístico de procesos, distribución Dirichlet.

[a] Assistant professor. E-mail: rdguevarag@unal.edu.co
[b] Professor. E-mail: javargasn@unal.edu.co
[c] Assistant professor. E-mail: dllineros@unal.edu.co

1. Introduction

In most of the statistical process control (SPC) applications, the quality of a process or product is represented by the distribution of a univariate or multivariate quality characteristic. However, in other applications, process quality is better characterized by a relationship between a response variable and one or more explanatory variables. This relationship is usually known as a profile. In these situations, the focus of the SPC lies on the parameters of the profile monitoring rather than on the monitoring of the univariate or multivariate characteristics. Such profiles can be represented using linear or nonlinear models. Some discussion of the general issues involving profile monitoring can be found in Woodall, Spitzner, Montgomery & Gupta (2004), Woodall (2007), Noorossana, Saghaei & Amiri (2012) and Qiu (2013). Profile practical applications have been reported by many researchers, including Stover & Brill (1998), Kang & Albin (2000), Mahmoud & Woodall (2004), Wang & Tsung (2005) and Kusiak, Zheng & Song (2009). Several control chart approaches for monitoring simple linear profiles have been developed by Kang & Albin (2000), Kim, Mahmoud & Woodall (2003), Zou, Zhang & Wang (2006), Zou, Zhou, Wang & Tsung (2007), Mahmoud, Parker, Woodall & Hawkins (2007), Soleimani, Narvand & Raissi (2013), Zhang, He, Zhang & Woodall (2013), Yeh & Zerehsaz (2013) and Amiri, Zou & Doroudyan (2014). Proposals for monitoring multivariate linear profiles (simple and/or multiple) have been developed by Mahmoud (2008), Noorossana, Eyvazian & Vaghefi (2010), Noorossana, Eyvazian, Amiri & Mahmoud (2010), Eyvazian, Noorossana, Saghaei & Amiri (2011) and Zou, Ning & Tsung (2012).

The linear regression model is commonly used for monitoring profiles. However, it is not appropriate for situations where the response is restricted to the interval $(0, 1)$ since it may yield fitted values in the variable of interest that exceed its lower and upper bounds. Ferrari & Cribari-Neto (2004) proposed a regression model that is tailored for situations where the dependent variable Y is measured continuously on the standard unit interval, i.e. $0 < Y < 1$. The proposed model is based on the assumption that the response is Beta distributed. The Beta distribution is very flexible for modeling proportions since its density can have quite different shapes depending on the values of the two parameters that index the distribution. Vasconcellos & Cribari-Neto (2005) proposed a class of regression models where the response is Beta distributed and the two parameters that index this distribution are related to covariates and regression parameters. However, the proposed regression models are restricted to the univariate case and cannot be applied in many practical situations where data consist of multivariate positive observations summing to one, that is, the study of compositional data, see Aitchison (1986) and Aitchison (2003). Melo, Vasconcellos & Lemonte (2009) proposed a particular structure for compositional data regression, based on the Dirichlet distribution, which is a generalization of the Beta distribution for the simplex sample space. A profile application in a concrete manufacturing plant, which after a preliminary study was found to fit appropriately this structure motivated this paper.

Compositional data are frequently encountered in industries such as the chemical, pharmaceutical, textil, plastic, concrete, steel, asphalt, among other. Several statistical methods for monitoring processes characterized by compositional data have been studied. See for example, Sullivan & Woodall (1996), Boyles (1997), Yang, Cline, Lytton & Little (2004) and Vives-Mestres, Daunis-i Estadella & Martín-Fernández (2013). However, there are not methods for monitoring these processes when the random vectors associated to the compositional data present a functional relationship with a set of explanatory variables.

In this paper, the control charting mechanisms discussed by Williams, Woodall & Birch (2007) and Yeh, Huwang & Li (2009) are extended for monitoring functional relationships in Phase I characterized by a Dirichlet regression model using a regression structure that allows the modeling of relationships between random vectors with Dirichlet distribution and a set of explanatory variables.

The structure of this paper is outlined as follows: In Section 2, we show the Dirichlet regression model for compositional data and the estimation of the model parameters. Five T^2 control charts approaches used for monitoring linear profiles in Phase I with compositional data are presented in Section 3. In Section 4, the performance of the proposed approaches is evaluated through simulation studies. A real example is given in Section 5. In the last section we conclude the paper.

2. Dirichlet Regression

Compositional data are used to indicate how parts contribute to the whole. In most cases they are recorded as closed data, i.e. data summing to a constant, such as 100%. Compositional data occupy a restricted space where variables can vary only from 0 to 100, or any other given constant. Such a restricted space is known formally as a simplex, see Pawlowsky-Glahn & Egozcue (2006).

Let c be a positive number. The p-dimensional closed simplex in $\mathbb{R}^n$ and $(p-1)$-dimensional open simplex in $\mathbb{R}^{p-1}$ are defined by

$$\mathbb{T}_p(c) = \left\{ (y_1, \dots, y_p)^t : y_j > 0, 1 \leq j \leq p, \sum_{j=1}^{p} y_j = c \right\}$$

and

$$\mathbb{V}_{p-1}(c) = \left\{ (y_1, \dots, y_{p-1})^t : y_j > 0, 1 \leq j \leq p-1, \sum_{j=1}^{p-1} y_j < c \right\}$$

respectively, where the superscript t means the function transpose. Furthermore, let $\mathbb{T}_p = \mathbb{T}_p(1)$ and $\mathbb{V}_{p-1} = \mathbb{V}_{p-1}(1)$.

A random vector $\mathbf{Y} = (Y_1, \dots, Y_p)^t \in \mathbb{T}_p$ is said to have a Dirichlet distribution if the density function of $\mathbf{Y_{-p}} = (Y_1, \dots, Y_{p-1})^t$ is

$$f\left(\mathbf{Y_{-p}}|\mathbf{a}\right) = \frac{\Gamma\left(\sum\limits_{j=1}^{p} a_j\right)}{\prod\limits_{j=1}^{p} \Gamma\left(a_j\right)} \prod_{j=1}^{p} y_j^{a_j - 1}, \qquad (y_1, \dots, y_{p-1}) \in \mathbb{V}_{p-1}, \tag{1}$$

where $\mathbf{a} = (a_1, \dots, a_p)^t$ and $a_j > 0$, $j = 1 \dots p$. We will write $\mathbf{Y} \sim Dirichlet_p(a_1, \dots, a_p)$, see Ng, Tian & Tang (2011).

When all $a_j \to 0$, the distribution becomes noninformative. When $p = 2$, the Dirichlet distribution $Dirichlet_2(a_1, a_2)$ reduces to the $Beta(a_1, a_2)$ distribution. The marginal distributions of the components of $\mathbf{Y}$, $Y_j, j = 1, 2, \dots, p$, are distributed as $Beta(a_j, \phi - a_j)$, where $\phi = \sum_{j=1}^{p} a_j$. In this sense, the Dirichlet distribution can be seen as a multivariate extension of the Beta distribution. Therefore, we have

$$E(Y_j) = \frac{a_j}{\phi}, \qquad j = 1, \dots, p \tag{2}$$

$$Var(Y_j) = \frac{a_j(\phi - a_j)}{\phi^2(\phi + 1)}, \qquad j = 1, \dots, p \tag{3}$$

$$Cov(Y_j, Y_l) = -\frac{a_j a_l}{\phi^2(\phi + 1)} < 0, \qquad j \neq l; j, l = 1, \dots, p \tag{4}$$

The Dirichlet distribution is widely used to model data in the form of proportions, where each observation is a vector of positive numbers summing to one. It allows great flexibility of modeling, provided by the appropriate choice of its parameters. See Ng et al. (2011) and Melo et al. (2009).

Gueorguieva, Rosenheck & Zelterman (2008) described a Dirichlet multivariate regression method which is useful for modeling data representing components as a percentage of a total. They described each $\log(a_j)$ as a separate linear function of covariates and regression coefficients. That is, for each component $j = 1, \dots, p$ they used a log-link with

$$\log a_{ij} = \boldsymbol{\beta}_j^t \mathbf{X}_i \tag{5}$$

for covariates X_i recorded on the ith individual $(i = 1, \dots, n)$ and regression coefficients $\boldsymbol{\beta}_j$ to be estimated using maximum likelihood. These estimates are denoted $\widehat{\boldsymbol{\beta}}_j$. The estimates $\widehat{\boldsymbol{a}}_j = \{\widehat{a}_{ij}\}$ of $\boldsymbol{a}_j = \{a_{ij}\}$ are defined by

$$\widehat{a}_{ij} = \exp(\widehat{\boldsymbol{\beta}}_j^t \mathbf{X}_i)$$

Gueorguieva et al. (2008) refer to the $\{\boldsymbol{a}_j\}$ as $meta - parameters$ because they combine the effects of the covariates $\mathbf{X}_i$ using regression parameters $\{\boldsymbol{\beta}_j\}$.

Melo et al. (2009) proposed a generalization of this model. The proposed model is defined by establishing relationships between the parameters that index the Dirichlet distribution and linear predictors on the explanatory variables. They assume a set of independent vector observations $\mathbf{Y}_1, \dots, \mathbf{Y}_n$, where $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ip})$ with $Y_{i1} + \dots + Y_{ip} = 1$, for each i. They suppose that $\mathbf{Y}_i \sim Dirichlet_p(a_{i1}, \dots, a_{ip})$ with

$$a_{ij} = g_j(\beta_{1j} x_{i1} + \dots + \beta_{kj} x_{ik}) \tag{6}$$

where each function $g_j : \mathbb{R} \to (0, \infty)$ is three times differentiable, injective and known, $x_{i1}, \dots, x_{ik}$ are the values corresponding to the ith observation for k explanatory variables and $\beta_{1j}, \dots, \beta_{kj}$ are k unknown parameters corresponding to the jth component. The model, therefore, has kp unknown parameters, which can be estimated through maximum likelihood (See Melo et al. 2009).

The covariates of this regression model affect the vector mean, the variance covariance structure of the distribution of the observations and the higher-order moments. The functions g_j play a similar role to the link functions of generalized linear models, in the sense that they specifically define how the parameters of the distribution of interest are linked to linear combinations of the covariates. The coefficients of this linear combination are unknown. The regression parameters are identifiable if the link functions are injective and the covariates are linearly independent (See Melo et al. 2009).

In the Dirichlet regression model, if $p = 2$ we have the Beta regression model described in Vasconcellos & Cribari-Neto (2005), Gueorguieva et al. (2008) and Melo et al. (2009).

The regression coefficients can be estimated using maximum likelihood. Let $\mathbf{B}$ the $k \times p$ matrix with the β_{hj}'s, $h = 1, 2 \ldots, k$ and $j = 1, 2, \ldots, p$. The log-likelihood function is given by

$$l(\mathbf{B}) = \sum_{i=1}^{n} \left\{ \log[\Gamma(\phi_i)] - \sum_{j=1}^{p} \log[\Gamma(a_{ij})] + \sum_{j=1}^{p} a_{ij} \log(Y_{ij}) \right\} \tag{7}$$

where $\phi_i = a_{i1} + \cdots + a_{ip}$ for each $i = 1, 2, \ldots, n$.

If $\widehat{\mathbf{B}}$ is the maximum likelihood estimator for $\mathbf{B}$, under some regularity conditions, $\sqrt{n}vec(\widehat{\mathbf{B}} - \mathbf{B}) \overset{a}{\sim} N_{kp}(\mathbf{0}, \mathbf{K}(\mathbf{B})^{-1})$, when n is large, with $\overset{a}{\sim}$ denoting asymptotically distributed, N_{kp} representing a kp-variate normal distribution and $\mathbf{K}(\mathbf{B})$ representing the $kp \times kp$ information matrix for the vector version of $\mathbf{B}$ (See Melo et al. 2009). The matrix $\mathbf{K}(\mathbf{B})$ can be obtained as

$$\mathbf{K}(\mathbf{B}) = \left(\mathbf{I}_p \bigotimes \mathbf{X}\right)^t \mathbf{L} \left(\mathbf{I}_p \bigotimes \mathbf{X}\right) \tag{8}$$

where $\bigotimes$ represents the Kronecker product, $\mathbf{L}$ is an $np \times np$ matrix defined in partitioned form as

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_{11} & \cdots & \mathbf{L}_{1p} \\ \vdots & \ddots & \vdots \\ \mathbf{L}_{p1} & \cdots & \mathbf{L}_{pp} \end{pmatrix} \tag{9}$$

where each $\mathbf{L}_{cd}$, with $c = 1, 2, \ldots, p$ and $d = 1, 2, \ldots, p$, is a diagonal matrix having ith element in the diagonal given by

$$l_i^{(cd)} = \begin{cases} -g_c'(\eta_{ic})^2[\psi'(\phi_i) - \psi'(a_{ic})], & c = d \\ -g_c'(\eta_{ic})g_d'(\eta_{id})\psi'(\phi_i), & c \neq d \end{cases} \tag{10}$$

where $\phi_i = a_{i1} + \cdots + a_{ip}$, for each $i = 1, 2, \ldots, n$, $\eta_{ij} = \beta_{1j}X_{i1} + \cdots + \beta_{kj}X_{ik}$ with $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$, g' is the first-order derivative of g with respect to its argument and ψ (digamma function) is the first derivative of the log of the gamma function.

3. Profile Monitoring Control Charts

In this section, we propose and study five Hotelling T^2 control charts for monitoring linear profiles with compositional data using the Dirichlet regression model described in the previous section. The study is limited for Phase I. We consider the same charts analyzed by Yeh et al. (2009) in their study for profile monitoring in binary responses: T^2 based on the sample mean vector and covariance matrix (T^2_{Usual}), T^2 based on the sample average and successive differences estimator (T^2_{SD}) proposed by Sullivan & Woodall (1996), T^2 based on the sample average and intra−profile pooling (T^2_{Int}) Williams et al. (2007), T^2 based on the Minimum Volume Ellipsoid (T^2_{MVE}) and T^2 based on the Minimum Covariance Determinant (T^2_{MCD}) studied by Vargas (2003) and Jensen, Birch & Woodall (2007).

We assume that when the process is in control, the matrix of model parameters is $\mathbf{B}_0$. In Phase I control, m independent samples are taken. In each sample r, $r = 1, \dots, m$, there are a set of n_r independent vector observations $\mathbf{Y}_{1r}, \dots, \mathbf{Y}_{n_r r}$, where $\mathbf{Y}_{ir} = (Y_{ir1}, \dots, Y_{irp})$ with $Y_{ir1} + \cdots + Y_{irp} = 1$, for each $i = 1, \dots, n_r$. We suppose that $\mathbf{Y}_{ir} \sim Dirichlet_p(a_{i1}, \dots, a_{ip})$. We assume that the relationship between the parameters that index the Dirichlet distribution and k explanatory variables $(X_1, \dots, X_k)$ given in equation (6) is $g_j = exp(\cdot)$.

For any given sample r, $r = 1, 2, \dots, m$, $\widehat{\mathbf{B}}_r$ is the maximum likelihood estimator of $\mathbf{B}$. Let $\widehat{\boldsymbol{\beta}}_r = vec(\widehat{\mathbf{B}}_r) = (\widehat{\beta}_{11_r}, \widehat{\beta}_{21_r}, \dots, \widehat{\beta}_{k1_r}, \widehat{\beta}_{12_r}, \widehat{\beta}_{22_r}, \dots, \widehat{\beta}_{k2_r}, \dots, \widehat{\beta}_{1p_r}, \widehat{\beta}_{2p_r}, \dots, \widehat{\beta}_{kp_r})$ $\widehat{\boldsymbol{\beta}}_r$ is a multivariate random vector, where each $\widehat{\beta}_{sj_r}$ represents the estimator of the parameter corresponding to the explanatory variable X_s, $s = 1, \dots, k$, applied on the j components of $\mathbf{Y}_{ir}$.

The Hotelling's T^2 statistic measures the Mahalanobis distance of the corresponding vector from the sample mean vector. The general form of the statistic is

$$T_r^2 = (\widehat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}_0)^t \boldsymbol{\Sigma}_0^{-1} (\widehat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}_0)$$

where $\boldsymbol{\beta}_0$ is the expected value of $\widehat{\boldsymbol{\beta}}_r$ when the process is in control, and $\boldsymbol{\Sigma}_0$ is the in-control covariance matrix of $\widehat{\boldsymbol{\beta}}_r$.

In Phase I control, $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}_0$ both need to be estimated and the performance of the control chart depends on the types of estimates being used. The T^2 statistics for the proposed control charts are calculated by:

$$T^2_{Usual,r} = (\widehat{\boldsymbol{\beta}}_r - \overline{\boldsymbol{\beta}})^t \mathbf{S}^{-1}_{Usual} (\widehat{\boldsymbol{\beta}}_r - \overline{\boldsymbol{\beta}}) \tag{11}$$

where $\overline{\boldsymbol{\beta}} = \frac{1}{m} \sum_{r=1}^m \widehat{\boldsymbol{\beta}}_r$ and $\mathbf{S}_{Usual} = \frac{1}{m-1} \sum_{r=1}^m (\widehat{\boldsymbol{\beta}}_r - \overline{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_r - \overline{\boldsymbol{\beta}})^t$

$$T^2_{SD,r} = (\widehat{\boldsymbol{\beta}}_r - \overline{\boldsymbol{\beta}})^t \mathbf{S}^{-1}_{SD} (\widehat{\boldsymbol{\beta}}_r - \overline{\boldsymbol{\beta}}) \tag{12}$$

where $\mathbf{S}_{SD} = \frac{1}{2(m-1)} \sum_{r=1}^{m-1} (\widehat{\boldsymbol{\beta}}_{r+1} - \widehat{\boldsymbol{\beta}}_r)(\widehat{\boldsymbol{\beta}}_{r+1} - \widehat{\boldsymbol{\beta}}_r)^t$

$$T^2_{Int,r} = (\widehat{\boldsymbol{\beta}}_r - \overline{\boldsymbol{\beta}})^t \mathbf{S}^{-1}_{Int} (\widehat{\boldsymbol{\beta}}_r - \overline{\boldsymbol{\beta}}) \tag{13}$$

where $\mathbf{S}_{Int} = \frac{1}{m}\sum_{r=1}^{m}\widehat{var}(\widehat{\beta}_r)$, which is calculated using the observed information matrix,

$$T^2_{MVE,r} = (\widehat{\boldsymbol{\beta}}_r - \widehat{\boldsymbol{\beta}}_{MVE})^t \mathbf{S}^{-1}_{MVE}(\widehat{\boldsymbol{\beta}}_r - \widehat{\boldsymbol{\beta}}_{MVE}), \tag{14}$$

where $\widehat{\boldsymbol{\beta}}_{MVE}$ and $\mathbf{S}_{MVE}$ are estimates of $\boldsymbol{\beta_0}$ and Σ_0, respectively, based on the MVE method (See Rousseeuw & Van Zomeren 1990), and

$$T^2_{MCD,r} = (\widehat{\boldsymbol{\beta}}_r - \widehat{\boldsymbol{\beta}}_{MCD})^t \mathbf{S}^{-1}_{MCD}(\widehat{\boldsymbol{\beta}}_r - \widehat{\boldsymbol{\beta}}_{MCD}), \tag{15}$$

where $\widehat{\boldsymbol{\beta}}_{MCD}$ and $\mathbf{S}_{MCD}$ are estimates of $\boldsymbol{\beta_0}$ and Σ_0, respectively, based on the MCD method (See Rousseeuw & Van Zomeren 1990).

Although $\widehat{\boldsymbol{\beta}}_r$ is distributed asymptotically normal we do not know its sampling distribution. Therefore, we used simulations to approximate the upper control limit (UCL). For simplicity we consider that the number of components is $p = 2, 3, 4, 5$ and 8. For a chart given we generated m independent samples. For each sample we generated a set of n independent vector observations $\mathbf{Y}_1, \dots, \mathbf{Y}_n$, where $\mathbf{Y}_i \sim Dirichlet_p(a_{i1}, a_{i2}, \dots, a_{ip})$, $i = 1\dots, n$. The parameters of the Dirichlet distribution, a_{ij} with $j = 1, 2, \dots, p$, are described by $a_{ij} = exp(\beta_{0j} + \beta_{1j}X_i)$ with $\beta_{01} = 2$, $\beta_{11} = 3$, $\beta_{02} = 1$, $\beta_{12} = 4$, $\beta_{03} = 3$, $\beta_{13} = -2$, $\beta_{04} = 0$, $\beta_{14} = 2$, $\beta_{05} = -0.1$, $\beta_{15} = 2.5$, $\beta_{06} = 1$, $\beta_{16} = 2$, $\beta_{07} = 3$, $\beta_{17} = 2$, $\beta_{08} = 1$ and $\beta_{18} = 2.5$. The values of the regressor variable (X) can be random but we have assumed that X takes fixed values, $X = 0.1, 0.2, \dots, 0.9$. For the m samples generated we calculate the maximum T^2, denoted by $T^2_{\max}$. This process was then repeated 10,000 times which resulted in 10,000 $T^2_{\max}$ values. The 95th quantile of these $T^2_{\max}$ values was then taken as an estimate of the UCL for that chart.

For each of the proposed charts, we ran the simulations for $m = 30, 60$ and 90 samples with a prespecified type I error probability $\alpha = 0.05$. The UCLs obtained are shown in Table 1. We used the R language to run the simulations, in particular we used the DirichletReg-package written by Maier (2011) to calculate the estimates of $\boldsymbol{\beta}_r$. We also used the functions cov.mve and cov.mcd from the MASS-package to calculate $\widehat{\boldsymbol{\beta}}_{MVE}$ and $\mathbf{S}_{MVE}$ in equation (14), and $\widehat{\boldsymbol{\beta}}_{MCD}$ and $\mathbf{S}_{MCD}$ in equation (15).

If a process is modeled using the multivariate normal regression, the response variables can take any real value, (Noorossana, Eyvazian, Amiri & Mahmoud 2010). However, this assumption is not met here, because the response variables for compositional data are always positive and range only from 0 to 100, or any other constant. Therefore, the use of the multivariate normal regression in this kind of processes can produce invalid results. For more details see Aitchison (2003) and Pawlowsky-Glahn & Egozcue (2006).

4. The Performance Evaluation

In this section we compare the performance of the proposed methods for Phase I, monitoring of compositional data, through linear regression profiles in terms of the overall probability of a signal under step and drift shift and outliers. The

Table 1: Values of simulated UCL for the proposed control charts with $\alpha = 0.05$

Number of components (p)	Total samples (m)	T^2_{Usual}	T^2_{SD}	T^2_{Int}	T^2_{MVE}	T^2_{MCD}
2	30	20.941	24.610	87.627	212.263	246.696
	60	35.680	37.925	117.805	154.041	164.521
	90	47.359	48.993	139.634	163.259	162.749
3	30	19.505	26.127	56.683	161.556	362.290
	60	30.644	34.035	75.480	102.306	124.628
	90	38.699	41.184	87.004	96.878	103.855
4	30	19.905	30.427	49.128	186.489	799.903
	60	28.324	32.960	58.086	87.104	139.786
	90	32.124	35.336	60.897	70.180	81.265
5	30	20.815	37.366	47.412	253.064	1820.518
	60	28.870	35.098	55.731	95.451	204.833
	90	32.973	37.308	60.638	72.088	92.750
8	30	24.550	74.952	54.258	710.379	11972.690
	60	33.487	47.861	61.769	175.620	521.925
	90	38.155	46.609	65.442	103.688	187.424

signal probability is defined as the probability that at least one sample, of a total of m samples, is considered to be out of control. When the process is out of control, a large signal probability indicates better ability of a control chart to detect the out-of-control process. However, when the process is in control, a large signal probability actually works against a control chart since it gives a higher false alarm rate (See Yeh et al. 2009).

We have that $\sqrt{n}vec(\widehat{\mathbf{B}} - \mathbf{B}) \overset{a}{\sim} N_{kp}(\mathbf{0}, \mathbf{K}(\mathbf{B})^{-1})$. Following equations (6), (8), (9) and (10), the information matrix $\mathbf{K}(\mathbf{B})$ depends on the unknown parameters of the regression and of the values assigned to the regressor variable. For simplicity, we consider that $p = 2$. So, for $\boldsymbol{\beta}_0 = c(2, 3, 1, 4)$ and $X = 0.1, 0.2, \ldots, 0.9$ we have that

$$\Sigma_0 = \mathbf{K}(\mathbf{B})^{-1} = \begin{pmatrix} \sigma^2_{\beta_{01}} & \sigma_{\beta_{01}\beta_{11}} & \sigma_{\beta_{01}\beta_{02}} & \sigma_{\beta_{01}\beta_{12}} \\ \sigma_{\beta_{11}\beta_{01}} & \sigma^2_{\beta_{11}} & \sigma_{\beta_{11}\beta_{02}} & \sigma_{\beta_{11}\beta_{12}} \\ \sigma_{\beta_{02}\beta_{01}} & \sigma_{\beta_{02}\beta_{11}} & \sigma^2_{\beta_{02}} & \sigma_{\beta_{02}\beta_{12}} \\ \sigma_{\beta_{12}\beta_{01}} & \sigma_{\beta_{12}\beta_{11}} & \sigma_{\beta_{12}\beta_{02}} & \sigma^2_{\beta_{12}} \end{pmatrix}$$

$$= \begin{pmatrix} 1.0322 & -1.6290 & 0.9807 & -1.5621 \\ -1.6290 & 3.2702 & -1.5615 & 3.1763 \\ 0.9807 & -1.5615 & 1.0041 & -1.5926 \\ -1.5621 & 3.1763 & -1.5926 & 3.2218 \end{pmatrix}$$

Let $\Delta = (\delta_1\sigma_{\beta_{01}}, \delta_2\sigma_{\beta_{11}}, \delta_3\sigma_{\beta_{02}}, \delta_4\sigma_{\beta_{12}})$, where $\delta_j = 0, 1, 2, 3,\ j = 1, 2, 3, 4$. If $\boldsymbol{\beta}_r$ changes from $\boldsymbol{\beta}_0$ to $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + \Delta$, with $\Delta \neq \mathbf{0}$, the process is out-of-control. The level of shifts in $\boldsymbol{\beta}_r$ is described by the non-centrality parameter (ncp). The non-centrality parameter measures the severity of a shift to the out-of-control vector $\boldsymbol{\beta}_1$ from the in-control vector $\boldsymbol{\beta}_0$ and is defined by $ncp = \Delta^t\Sigma_0^{-1}\Delta = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^t\Sigma_0^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$ (See Vargas (2003) and Yeh et al. (2009)).

The out-of-control signal probabilities were calculated based on 5,000 simulations, as the percentage of times the $T^2_{\max}$ exceeds the corresponding UCL.

For step shift or sustained shift, we generate a shift in the vector of parameters of the regressions, $\boldsymbol{\beta}$, from $\boldsymbol{\beta}_0$ to $\boldsymbol{\beta}_1$. The shift starts from the sample l, for $l = [m * k] + 1$, where $[x]$ denotes the largest integer which is less or equal than x and $k = \frac{1}{4}, \frac{1}{2}$ and $\frac{3}{4}$. So, for $k = \frac{1}{2}$ the first half of the samples is in-control, while the second half is in the out-of-control state. The signal probabilities for each control chart, when $m = 30$, were calculated through simulation.

Table 2: Signal probabilities when the intercept and the slope of the profile corresponding to the second component have not changed.

ncp	δ_1	δ_2	δ_3	δ_4	T^2_{Usual}	T^2_{SD}	T^2_{Int}	T^2_{MVE}	T^2_{MCD}
0	0	0	0	0	0.0492	0.0494	0.0464	0.0452	0.0488
175.0568	1	0	0	0	0.0114	0.8052	0.8966	0.0312	0.0346
700.2274	2	0	0	0	0.0134	0.8976	1	0.0264	0.032
1575.512	3	0	0	0	0.0146	0.9078	1	0.0294	0.0328
297.7102	0	1	0	0	0.0086	0.843	0.9912	0.0234	0.031
910.8436	1	1	0	0	0.0102	0.8974	1	0.0284	0.031
1874.091	2	1	0	0	0.0092	0.8996	1	0.0256	0.0344
3187.452	3	1	0	0	0.0096	0.9144	1	0.025	0.0336
1190.841	0	2	0	0	0.009	0.9004	1	0.019	0.028
2242.051	1	2	0	0	0.0072	0.9142	1	0.0268	0.0322
3643.375	2	2	0	0	0.0056	0.9124	1	0.0222	0.029
5394.812	3	2	0	0	0.006	0.9058	1	0.03	0.0284
2679.391	0	3	0	0	0.0096	0.907	1	0.026	0.0328
4168.678	1	3	0	0	0.007	0.9064	1	0.0242	0.0294
6008.079	2	3	0	0	0.0046	0.8998	1	0.0256	0.0322
8197.593	3	3	0	0	0.0072	0.9016	1	0.0268	0.0338
6008.079	2	3	0	0	0.006	0.901	1	0.0262	0.033

Table 2 shows the signal probabilities of the five control charts considered for a step shift occuring in $l = [m/2] + 1$ when the intercept and the slope of the profile corresponding to the second component have not changed. When $ncp = 0$ the signal probabilities for the T^2_{Usual}, T^2_{SD}, T^2_{Int}, T^2_{MVE} and T^2_{MCD} control charts are close to 0.05. For other values of *ncp*, the signal probabilities of the T^2_{Int} control chart are 1 o very near 1, showing an excellent performance to detect step shifts.

Figures 1 to 5 describe the signal probabilities of the T^2_{Usual}, T^2_{SD}, T^2_{Int}, T^2_{MVE} and T^2_{MCD} control charts for a step shift occurring in three scenarios: the last three quarters, the second half, and the last quarter of the 30 samples considered. With exception of the T^2_{Int} control chart, the location of the step shift affects the performance of the T^2 control charts considered. For example, the signal probabilities decrease considerably when the shift stars in the half of the samples. The effect is greater in the T^2_{MVE} and T^2_{MCD} control charts. These charts are more powerful when the shifts occur at $k = \frac{1}{4}$ and $k = \frac{3}{4}$.

For drift shifts, the first sample generated was in control, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0$, and the process parameter vector started to change from the second sample to $\boldsymbol{\beta}_1$, where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + \frac{r-1}{m-1}(\Delta)$, with $r = 2\ldots, m$ and $m = 30$. Figure 6 shows the signal probabilities found by simulation for the 256 possible values of ncp. We observe

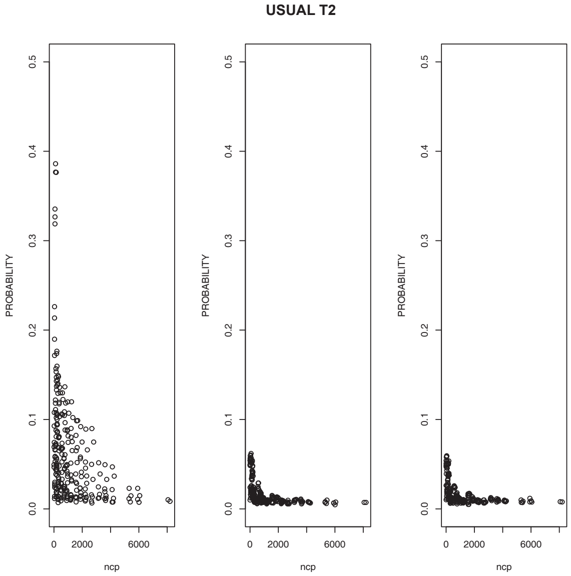


FIGURE 1: Signal probabilities of the T^2_{Usual} control chart for a step shift occurring in $l = [m * k] + 1$, with $k = \frac{1}{4}$ (on the left), $k = \frac{2}{4}$ (on the middle), and $k = \frac{3}{4}$ (on the right), when the number of samples is $m = 30$.

that the T^2_{SD}, T^2_{Int} control charts have a good performance for detecting shifts with trend.

In the scenario considering the presence of outliers, 5 of them were inserted randomly in the m samples, with $m = 30$. They were generated from $\boldsymbol{\beta}_1$, where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + (\Delta)$. Figure 7 presents the signal probabilities calculated by simulations. T^2_{Int}, T^2_{MVE} and T^2_{MCD} control charts have the best performance for detecting outliers.

5. Example of Application

The concrete is a composite material that essentially consists of a mixture of cement, water and aggregates, which is regularly used in infrastructure and buildings construction (Li 2011). The aggregates are rock fragments named coarse aggregate and sand particles called fine aggregate, which be derived from land- or sea-based deposits, from gravel pits or hard-rock quarries, from sand dunes or river courses. The aggregate occupies between 70% and 75% of the concrete volume and affect its strength, durability, workability and cohesiveness. One aspect of interest in the quality of the aggregate is the particle size distribution known as gradation (Alexander & Mindess 2005).

In order to obtain the gradation of the aggregate, a series of standard sieves are nested or stacked, one on top of another, with increasing aperture size from bottom to top, and through which a aggregate sample is passed from top, usually

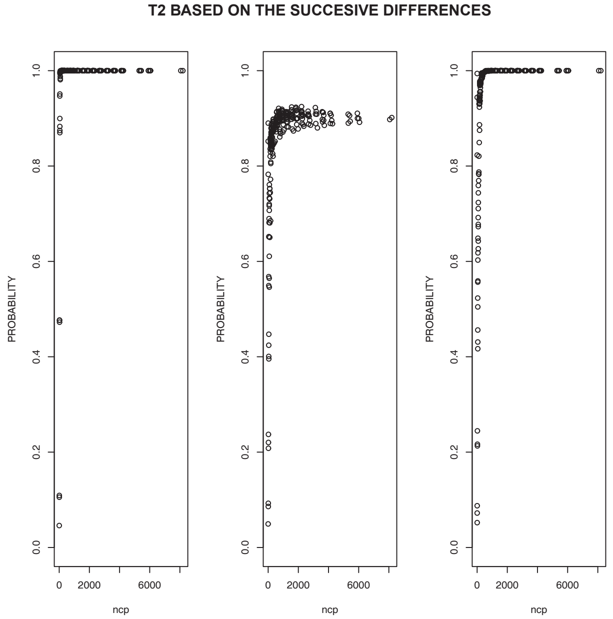


Figure 2: Signal probabilities of the T^2_{SD} control chart for a step shift occurring in $l = [m * k] + 1$, with $k = \frac{1}{4}$ (on the left), $k = \frac{2}{4}$ (on the middle), and $k = \frac{3}{4}$ (on the right), when the number of samples is $m = 30$.

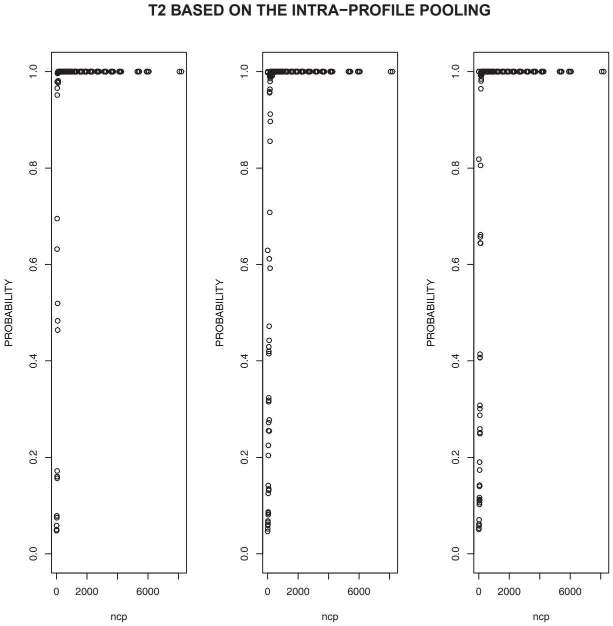


Figure 3: Signal probabilities of the T^2_{Int} control chart for a step shift occurring in $l = [m * k] + 1$, with $k = \frac{1}{4}$ (on the left), $k = \frac{2}{4}$ (on the middle), and $k = \frac{3}{4}$ (on the right), when the number of samples is $m = 30$.

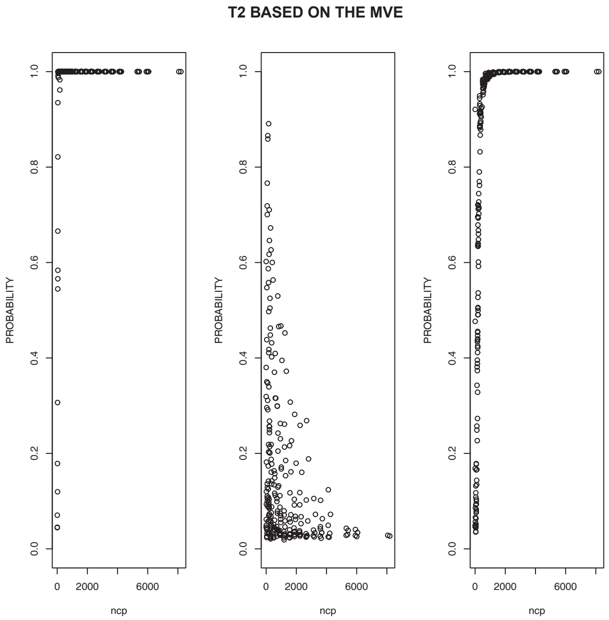


Figure 4: Signal probabilities of the T^2_{MVE} control chart for a step shift occurring in $l = [m * k] + 1$, with $k = \frac{1}{4}$ (on the left), $k = \frac{2}{4}$ (on the middle), and $k = \frac{3}{4}$ (on the right), when the number of samples is $m = 30$.

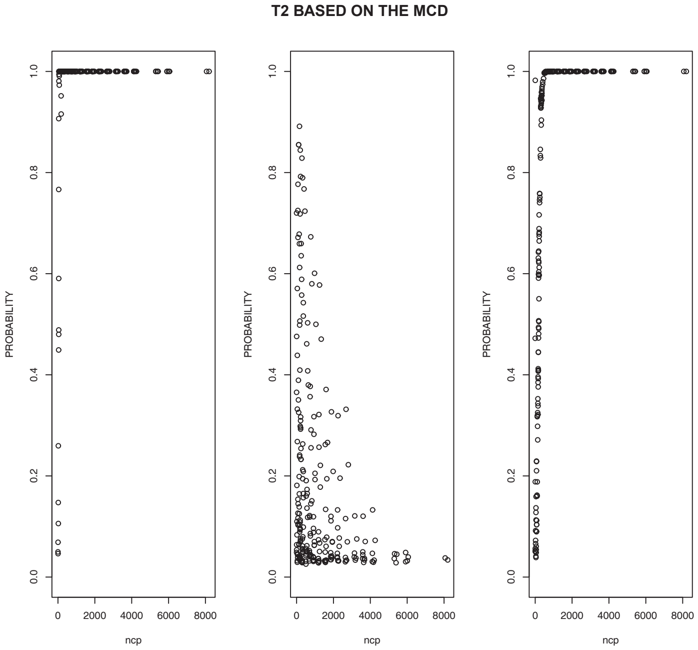


Figure 5: Signal probabilities of the T^2_{MCD} control chart for a step shift occurring in $l = [m * k] + 1$, with $k = \frac{1}{4}$ (on the left), $k = \frac{2}{4}$ (on the middle), and $k = \frac{3}{4}$ (on the right), when the number of samples is $m = 30$.

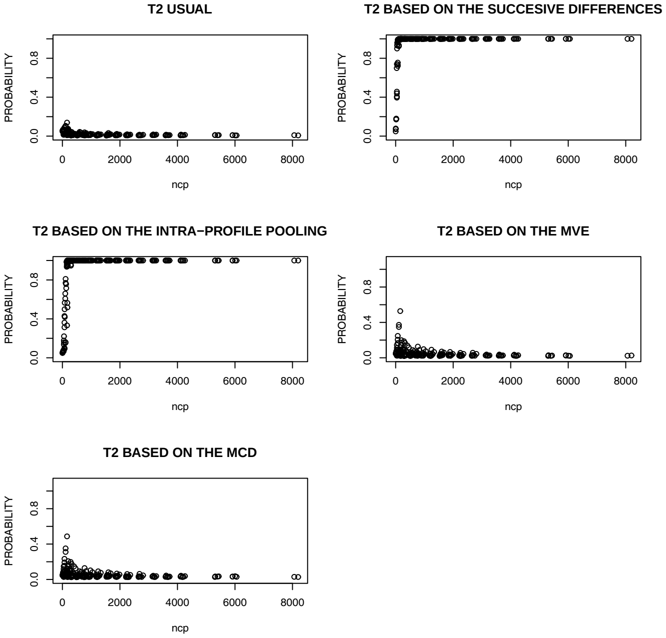


FIGURE 6: Signal probabilities of drift shifts for five control charts: usual, successive differences, intra-profile, MVE and MCD.

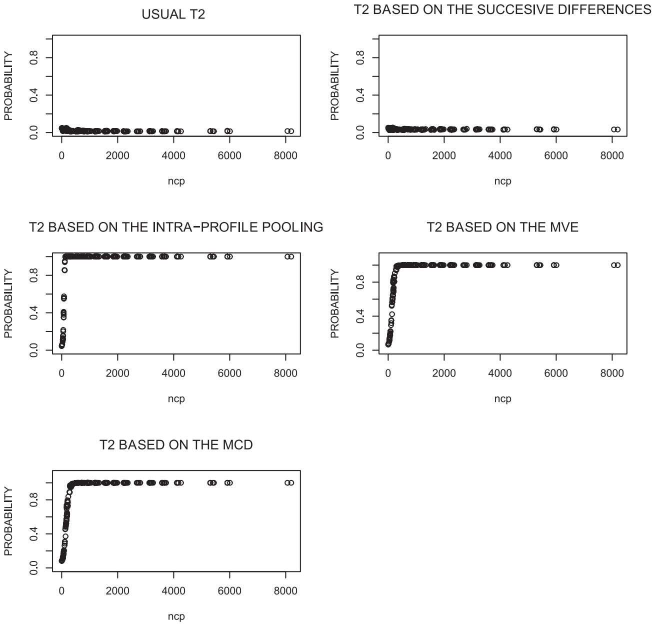


FIGURE 7: Signal probabilities of outliers for five control charts: usual, successive differences, intra-profile, MVE and MCD.

aided by shaking or vibrating the sieves (Alexander & Mindess 2005, Lyons 2008). Figure 8 shows a kind of machine used in the gradation process. The sieves labeled as 200, 100, 50, 30, 16, 8, 4, and P3, have the hole sizes of 0.075 mm, 0.149 mm, 0.297 mm, 0.595 mm, 1.19 mm, 2.38 mm, 4.75 mm and 9.5 mm, respectively. The gradation results are the percent of aggregate retained on each sieve and the fineness modulus, which measures the average particle size. This dimensionless parameter is equal to sum of the percent of aggregate retained on each of sieve divided by 100. A smaller fineness modulus indicates a finer aggregate and a higher value represents a courser aggregate.

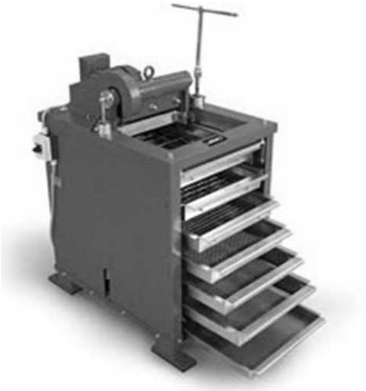


Figure 8: Series of sieves placed one above the other in order of size with the largest sieve at the top.

In this work, 217 aggregate samples from a concrete manufacturing plant were studied. The aggregate samples were daily tested during 31 weeks. The set of daily observations obtained in a week is named weekly sample, therefore, 31 weekly samples were considered. The proportion passing through of each sieve and the fineness modulus were measured in each aggregate sample.

The components $j = 1, 2, \dots, 8$ are defined by the aggregate size retained by each sieve. The proportion passing through the sieve j, $j = 1, \dots, 8$, is the variable Y_j. Each component corresponds to an aggregate with constant size and the proportion of aggregate passing trough them is identified by the vector $\mathbf{Y} = (Y_1, Y_2, \dots, Y_8)$. Figure 9 shows plots of the marginal frequencies of each component for the aggregate samples. We observe that Y_4, Y_5 and Y_7 are skewed, which implies that $\mathbf{Y}$ does not have a multivariate normal distribution.

The weekly sample r, $r = 1, \dots, 31$, contains the daily observations $(x_{ir}, \mathbf{Y}_{ir})$, $i = 1, \dots, n$, with $n = 7$. The vectors $\mathbf{Y}_{1r}, \dots, \mathbf{Y}_{nr}$ are independent and $\mathbf{Y}_{ir} \sim Dirichlet_8(a_{i1}, \dots, a_{i8})$. There is a relationship between the fineness modulus and the proportion of aggregate passing through each sieve. Figures 10 and Figures 11 show these relationships for the components $j = 1, 2, 3, 4, 5, 6, 7, 8$ associated to the sieves 200, 100, 50, 30, 16, 8, 4, and P3, respectively. A likelihood ratio test (LRT) for each sample r, shows that the Dirichlet regression models are significant at the 10%, so the null hypothesis $H_0 : \beta_{11} = \beta_{12} = \cdots = \beta_{18} = 0$ is rejected.

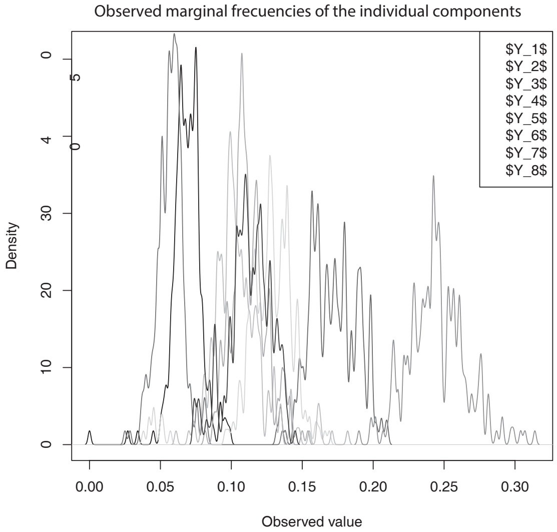


FIGURE 9: Observed marginal frequencies $Y_1, Y_2, \ldots, Y_8$ of the individual components of the aggregate gradings.

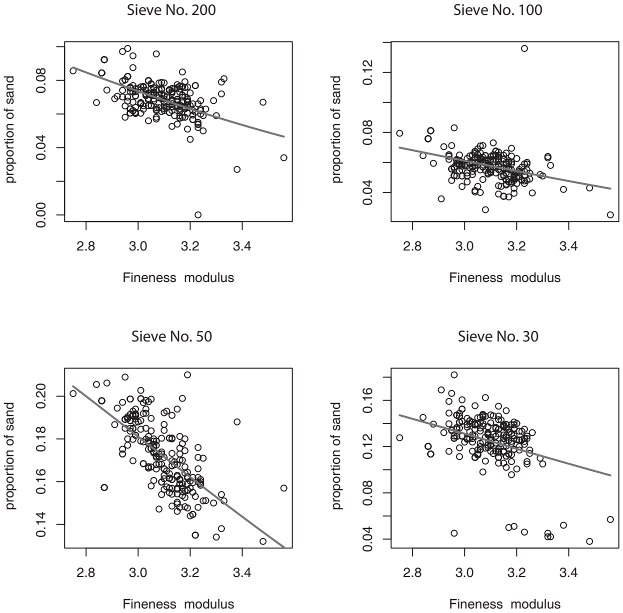


FIGURE 10: Linear relationship between fineness modulus and the proportion of sand passing through the sieves No 200, 100, 50 and 30.

Through simulations we found the upper control limit for each T^2 control chart proposed in the section (3). The T^2 control chart based on successive differences suggests that the process does not present step and drift shifts, but the control chart based on MVE detects the presence of outliers, see Figures 12 and 13.

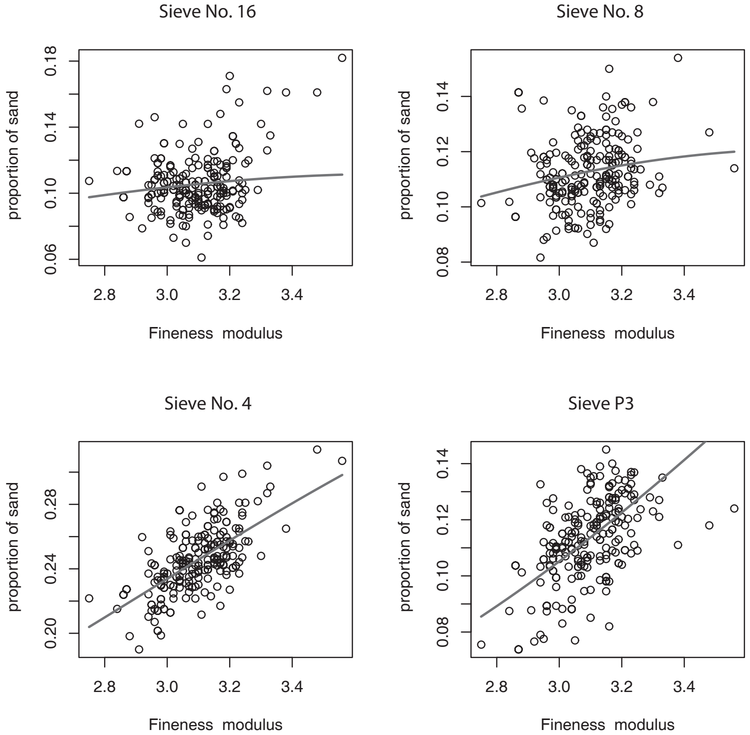


Figure 11: Linear relationship between fineness modulus and the proportion of sand passing through the sieves No 16, 8, 4 and P3.

Although the engineers believed that the process was in-control, the intra-profile control chart shows a lot of points outside the upper control limit, see Figure 15. The usual T^2 control chart detects some of these points, see Figure 16. Figure 17 describes the behaviour of the linear regressions associated with the first sieve from the sample. This graph shows that the profile is not stable. The other sieves have a similar behaviour. As a first result of this application, engineers are reviewing and adjusting the process to ensure that the linear relationship associated with each sieve is in-control.

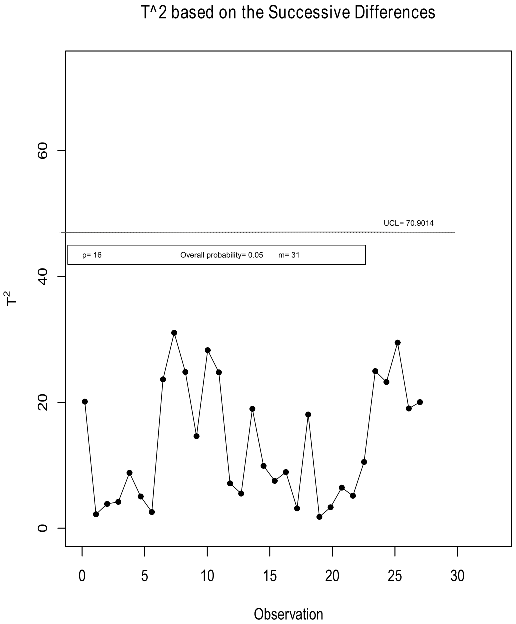


FIGURE 12: T^2 Control chart based on successive differences for the process of grading of sand in a mine of a concrete manufacturing plant.

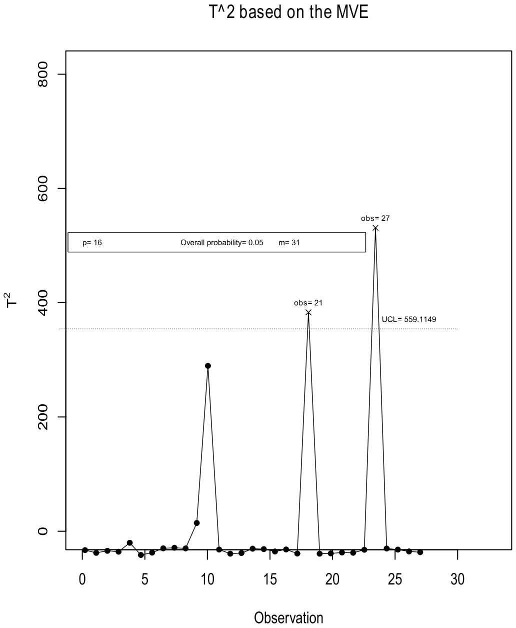


FIGURE 13: T^2 Control chart based on Minimum Volume Ellipsoid (MVE) for the process of grading of sand in a mine of a concrete manufacturing plant.

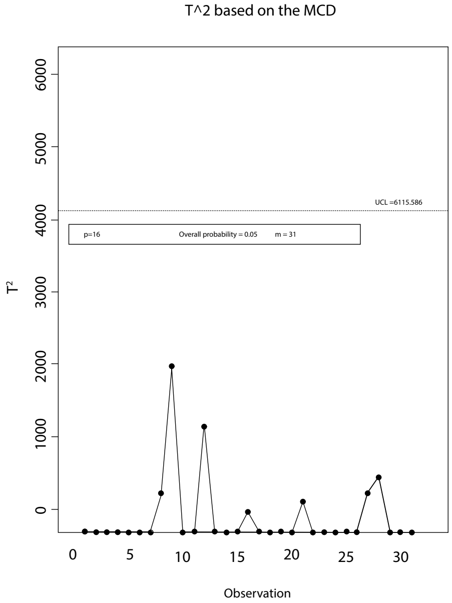


FIGURE 14: T^2_{MCD} for the process of grading of sand in a mine of a concrete manufacturing plant.

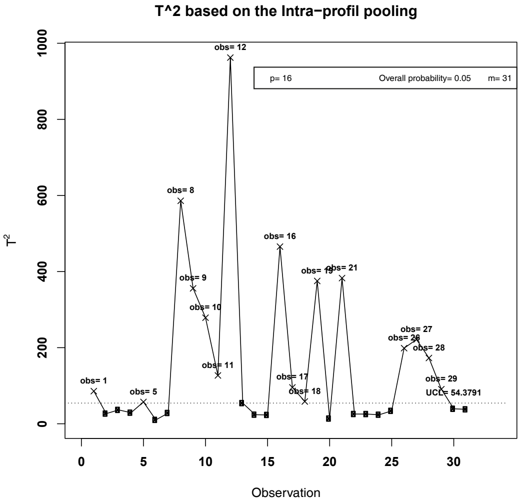


FIGURE 15: T^2 Control chart based on intra-profile pooling for the process of grading of sand in a mine of a concrete manufacturing plant.

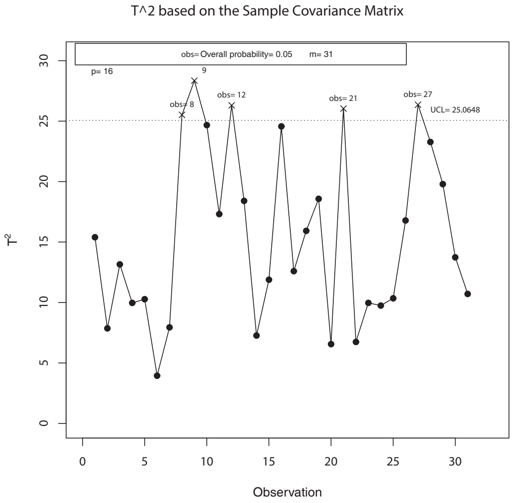


FIGURE 16: Usual T^2 control chart for the process of grading of sand in a mine of a concrete manufacturing plant.

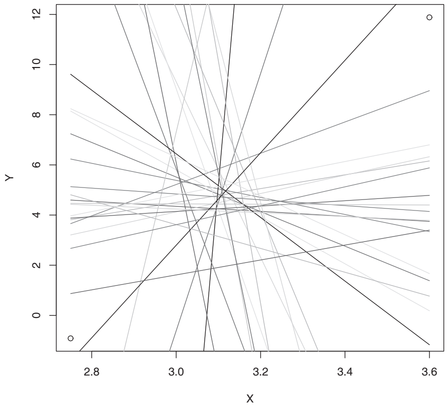


FIGURE 17: Linear regressions for the first component of the sand gradation.

6. Conclusions

In this paper the control charting mechanisms discussed by Williams et al. (2007) and Yeh et al. (2009) have been extended for monitoring compositional data profiles in Phase I processes, whose response variable follows a Dirichlet distribution. This methodology allows us monitoring the linear relationship between the parameters of a Dirichlet distribution and a set of explanatory variables, and assess the stability of the parameters that characterize the studied regression model.

We used five Hotelling's type T^2 control charts and compared their performance for detecting step and drift shifts in the process parameters and outliers in the studied profiles. Simulation procedures suggest that the T^2 control chart called intra-profile pooling is an excellent tool in order to detect outliers in the profiles and step and drift shifts in the parameters of the compositional data profile. The intra-profile pooling T^2 control chart is based on the average of the sample covariance matrices of the estimates of the parameters $\boldsymbol{\beta}$ characterizing the profile. The T^2 control chart based on the vector of successive differences of parameter estimates is a good alternative for detecting step and drift shifts; while the T^2 control charts based on robust estimates for the mean and covariance matrix, minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) methods, are a good option for detecting outliers.

We presented an example of application with real data of the proposed methodology, in order to control the quality of the aggregate gradation in the concrete. The T^2 control chart based on successive differences suggests that the process is in-control and does not present step and drift shifts, the control chart based on MVE detects the presence of some outliers, and the intra-profile and usual T^2 control charts show that the process is out-control. This methodology can be extended to other processes with compositional data.

This paper constitutes an initial solution for monitoring compositional data profiles. It would be worthwhile to study and compare the performance of other control charts like the change point approach. Since the performance of the T^2 control charts deteriorates when number of parameters increases, it is needed more research when the number of components in the response variable increase and/or when the number of covariates increases. Reduction methods for multivariate data or high dimensional methods need also future research.

When the process is out-of-control is important to identify the causes of the anomaly in order to apply appropriate remedial measures. A future work can implement diagnostic aids such as determining the parameters responsible for out-of-control signal.

Finally, some methods for monitoring Dirichlet regression profiles in Phase II can be developed.

Acknowledgments

The authors want to thank Professor William H. Woodall from Virgina Tech, the Editor, and two referees for their helpful comments that have strengthened this paper.

[Recibido: octubre de 2013 — Aceptado: marzo de 2014]

References

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, Chapman & Hall.

Aitchison, J. (2003), *The Statistical Analysis of Compositional Data*, Blackburn Press.

Alexander, M. & Mindess, S. (2005), *Aggregates in Concrete*, Taylor & Francis.

Amiri, A., Zou, C. & Doroudyan, M. H. (2014), 'Monitoring correlated profile and multivariate quality characteristics', *Quality and Reliability Engineering International* **30**(1), 133–142.

Boyles, R. A. (1997), 'Using the chi-square statistic to monitor compositional process data', *Journal of Applied Statistics* **24**(5), 589–602.

Eyvazian, M., Noorossana, R., Saghaei, A. & Amiri, A. (2011), 'Phase II monitoring of multivariate multiple linear regression profiles', *Quality and Reliability Engineering Intenational* **27**(3), 281–296.

Ferrari, S. L. & Cribari-Neto, F. (2004), 'Beta regression for modelling rates and proportions', *Journal of Applied Statistics* **31**(7), 799–815.

Gueorguieva, R., Rosenheck, R. & Zelterman, D. (2008), 'Dirichlet component regression and its applications to psychiatric data', *Computational Statistics and Data Analysis* **52**(12), 5344–5355.

Jensen, W. A., Birch, J. B. & Woodall, W. H. (2007), 'High breakdown estimation methods for phase I multivariate control charts', *Quality and Reliability Engineering International* **23**(5), 615–629.

Kang, L. & Albin, S. L. (2000), 'On-line monitoring when the process yields a linear profile', *Journal of Quality Technology* **32**(4), 418–426.

Kim, K., Mahmoud, M. A. & Woodall, W. H. (2003), 'On the monitoring of linear profiles', *Journal of Quality Technology* **35**(3), 317–328.

Kusiak, A., Zheng, H. & Song, Z. (2009), 'Models for monitoring wind farm power', *Renewable Energy* **34**(3), 583–590.

Li, Z. (2011), *Advanced Concrete Technology*, John Wiley & Sons.

Lyons, A. (2008), *Materials for Architects and Builders*, fourth edn, Elsevier.

Mahmoud, M. A. (2008), 'Phase I analysis of multiple linear regression profiles', *Communications in Statistics-Simulation and Computation* **37**(10), 2106–2130.

Mahmoud, M. A., Parker, P. A., Woodall, W. H. & Hawkins, D. M. (2007), 'A change point method for linear profile data', *Quality and Reliability Engineering International* **23**(2), 247–268.

Mahmoud, M. A. & Woodall, W. H. (2004), 'Phase I analysis of linear profiles with calibration applications', *Technometrics* **46**(4), 380–391.

Maier, M. J. (2011), Dirichletreg: Dirichlet regression in R. R package version 0.3-0.

Melo, T. F., Vasconcellos, K. L. & Lemonte, A. (2009), 'Some restriction tests in a new class of regression models for proportions', *Computational Statistics and Data Analysis* **53**(12), 3972–3979.

Ng, K. W., Tian, G.-L. & Tang, M.-L. (2011), *Dirichlet and related distributions: Theory, methods and applications.*, John Wiley & Sons.

Noorossana, R., Eyvazian, M., Amiri, A. & Mahmoud, M. A. (2010), 'Statistical monitoring of multivariate multiple linear regression profiles in phase I with calibration application', *Quality and Reliability Engineering International* **26**(3), 291–303.

Noorossana, R., Eyvazian, M. & Vaghefi, A. (2010), 'Phase II monitoring of multivariate simple linear profiles', *Computers & Industrial Engineering* **58**(4), 563–570.

Noorossana, R., Saghaei, A. & Amiri, A. (2012), *Statistical Analysis of Profile Monitoring*, John Wiley & Sons.

Pawlowsky-Glahn, V. & Egozcue, J. (2006), *Compositional Data Analysis in the Geosciences: From Theory to Practice*, Geological Society, chapter Compositional data and their analysis: An introduction, pp. 1–10.

Qiu, P. (2013), *Introduction to Statistical Process Control*, CRC Press.

Rousseeuw, P. & Van Zomeren, B. (1990), 'Unmasking multivariate outlier and leverage points', *Journal of the American Statistical Association* **85**(411), 633–639.

Soleimani, P., Narvand, A. & Raissi, S. (2013), 'Online monitoring of autocorrelated linear profiles via mixed model', *International Journal of Manufacturing Technology and Management* **27**(4), 238–250.

Stover, F. & Brill, R. (1998), 'Statistical quality control applied to ion chromatography calibrations', *Journal of Chromatography* **804**(1-2), 37–43.

Sullivan, J. H. & Woodall, W. H. (1996), 'A comparison of multivariate control charts for individual observations', *Journal of Quality Technology* **28**(4), 398–408.

Vargas, J. A. (2003), 'Robust estimation in multivariate control charts for individual observations', *Journal of Quality Technology* **35**(4), 367–376.

Vasconcellos, K. L. & Cribari-Neto, F. (2005), 'Improved maximum likelihood estimation in a new class of Beta regression models', *Brazilian Journal of Probability and Statistics* **19**(1), 13–31.

Vives-Mestres, M., Daunis-i Estadella, J. & Martín-Fernández, J. A. (2013), 'Out of control signals in three part compositional t^2 control chart', *Quality and Reliability Engineering Intenational* **30**(3), 337–346.

Wang, K. & Tsung, F. (2005), 'Using profile monitoring techniques for a data-rich environment with huge sample size', *Quality and Reliability Engineering International* **21**(7), 677–688.

Williams, J. D., Woodall, W. H. & Birch, J. B. (2007), 'Statistical monitoring of nonlinear product and process quality profiles', *Quality and Reliability Engineering International* **23**(8), 925–941.

Woodall, W. H. (2007), 'Current research in profile monitoring', *Producao* **17**(3), 420–425.

Woodall, W. H., Spitzner, D. J., Montgomery, D. C. & Gupta, S. (2004), 'Using control charts to monitor process and product quality profiles', *Journal of Quality Technology* **36**(3), 309–320.

Yang, G., Cline, D. B. H., Lytton, R. L. & Little, D. N. (2004), 'Ternary and multivariate quality control charts of aggregate gradation for hot mix asphalt', *Journal of Materials in Civil Engineering* **16**(1), 28–34.

Yeh, A. B., Huwang, L. & Li, Y.-M. (2009), 'Profile monitoring for a binary response', *IIE Transactions* **41**(11), 931–941.

Yeh, A. & Zerehsaz, Y. (2013), 'Phase I control of simple linear profiles with individual observations', *Quality and Reliability Engineering Intenational* **29**(6), 829–840.

Zhang, Y., He, Z., Zhang, C. & Woodall, W. H. (2013), 'Control charts for monitoring linear profiles with within profile correlation using gaussian process models', *Quality and Reliability Engineering Intenational* . DOI: 10.1002/qre.1502.

Zou, C., Ning, X. & Tsung, F. (2012), 'LASSO-based multivariate linear profile monitoring', *Annals of Operations Research* **192**(1), 3–19.

Zou, C., Zhang, Y. & Wang, Z. (2006), 'A control chart based on a change-point model for monitoring linear profiles', *IIE Transactions* **38**(12), 1093–1103.

Zou, C., Zhou, C., Wang, Z. & Tsung, F. (2007), 'A self-starting control chart for linear profiles', *Journal of Quality Technology* **39**(4), 364–375.

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 183 a 198

Asymmetric Regression Models Bernoulli/Log Proportional-Hazard Distribution

Modelos de regresión asimétrico Bernoulli/distribución Log Hazard proporcional

Guillermo Martínez-Flórez[1,a], Carlos Barrera[2,b]

[1]Departamento de Matemáticas y Estadística, Universidad de Córdoba, Monteríia, Colombia

[2]Facultad de Ciencias Exactas y Aplicadas, Instituto Tecnológico Metropolitano, Medellín, Colombia

Abstract

In this paper we introduce a kind of asymmetric distribution for non-negative data called log-proportional hazard distribution (LPHF). This new distribution is used to study an asymmetrical regression model for data with limited responses (censored) through the mixture of a Bernoulli distribution with logit link and the LPHF distribution. Properties of the LPHF distribution are studied, maximum likelihood parameter estimation and information matrices are addressed. An illustration with real data shows that the model is a new alternative for studies with positive data censored.

***Key words*:** Censoring, Fisher information matrix, Maximum likelihood estimators, Proportional hazard.

Resumen

En este artículo se introduce una forma de distribución asimétrica para datos no-negativos llamada distribución log hazard proporcional (LPHF). Esta nueva distribución es usada para estudiar un modelo de regresión asimétrico para datos con respuestas limitadas (censuradas) a través de mezclas de una distribución Bernoulli con función link logit y la distribución LPHF. Propiedades de la distribución LPHF son estudiadas, se abordan las estimaciones de máxima verosimilitud de los parámetros y las matrices de información. Se presenta una ilustración con datos reales, donde se muestra que el modelo propuesto es una nueva alternativa para estudios con datos positivos censurados.

***Palabras clave*:** censura, estimadores de máxima verosimilitud, hazard proporcional, matriz de información de Fisher.

[a]Professor. E-mail: gmartinez@correo.unicordoba.edu.co
[b]Associate professor. E-mail: carlosbarrera@itm.edu.co

1. Introduction

The fundamental law of geochemistry enunciated by Ahrens (1954), "the concentration of a chemical element in a rock is distributed log-normal", is an application of the log-normal distribution. This distribution is also widely used to model different types of information, including income in the economy and lifetime distributions from materials, among others.

In many of these situations, both the kurtosis and the asymmetry of the distribution are above or below the expected for the log-normal model, reason why it is necessary to think in a more flexible model that achieves such deviation in modeling positive data.

In the case of positive data, Azzalini, dal Cappello & Kotz (2003), Mateu-Figueras & Pawlosky-Glanh (2003) and Mateu-Figueras, Pawlosky-Glanh & Barcelo-Vidal (2004) introduce the univariate distribution log-skew-normal (LSN), which contains as special case, the log-normal model.

Its density function is given by:

$$\phi_{LSN}(y;\xi,\eta,\lambda)=\frac{2}{\eta y}\phi\left(\frac{\log(y)-\xi}{\eta}\right)\Phi\left(\lambda\frac{\log(y)-\xi}{\eta}\right),\qquad y\in\mathbb{R}^+$$

where $\xi\in\mathbb{R}$, is a location parameter, $\eta\in\mathbb{R}^+$, is a scale parameter, λ is an asymmetry parameter, $\phi(\cdot)$ is the density function of a standard normal distribution and $\Phi(\cdot)$ is the respective cumulative distribution function. Notice that if $\lambda=0$ then the ordinary log-normal distribution follows as it is the case with the ordinary skew-normal model. Also, the information matrix is singular, thus regularity conditions are no longer satisfied. One consequence of this fact is that likelihood ratio statistics is no longer distributed according to the central chi-square distribution (Arellano-Valle & Azzalini 2008).

Moreover, in many cases the asymmetrical positive random variable in the study is limited, and in turn this is explained by a set of auxiliary covariates $X_1,\ X_2,\ ...,X_p$, thus extensions to the censured case with covariates should be addressed. The study of random variables with limited responders with covariates was presented by Tobin (1958) who studied the model popularly known as Tobit.

This model has been extensively studied in the case of normally distributed errors and is defined by considering that the observed random variable $y_i=\max\{y_i^*,0\}$ with $y_i^*=\mathbf{x}_i'\boldsymbol{\beta}+\epsilon_i,\ i=1,2,\ldots,n$; where the error term $\epsilon_i\sim N(0,\sigma^2)$, $i=1,\ldots,n$, $\mathbf{x}_i$ is a $p\times 1$ vector of known independent variables and $\boldsymbol{\beta}$ is a $p\times 1$ unknown parameter vector.

Although the Tobit model is an alternative for censoring data; in some situations the proportion of censored data cannot be well explained by the normal model since the tail of this distribution is more or less heavier than the proportion of censored data.

For instance, Moulton & Halsey (1995) show an application with 330 children in Haiti during 1987-1990 (*see* Job, Halsey, Boulos, Holt, Farrell, Albrecht, Brutus, Adrien, Andre, Chan, Kissinger, Boulos & the CiteSoleil/JHU 1991) which examines the; Immunogenecity of children before the implementation of a vaccine, here the number of observations below the detection limit was 86 observations (26.1%), which exceeds the expected value with a normal model, whereby the proportion of censure cannot be explained by the Tobit model. These authors use asymmetric models such as log-normal model.

In this sense, other works have been published, for instance, other distributions: such as the log-skew-normal has been implemented by Chai & Bailey (2008) and recently, Martínez-Florez, Bolfarine & Gómez (2013) implemented the model log-alpha-power-normal.

The model proposed by Moulton & Halsey (1995) is a generalization of the Cragg (1971) model, that in the classical literature is known as *the two-part model,* which is an alternative to Tobit when the data rate below or above the threshold is quite different from the probability of the tail obtained with the normal model.

The probability density function of y_i under Cragg (1971) model can be expressed as

$$g(y_i) = p_i I_i + (1 - p_i) f(y_i)(1 - I_i)$$

where p_i is the probability determining the relative contribution made by the point distribution to the overall mixture distribution, f is a density function with positive support, and $I_i = 0$ if $y_i > T$ and $I_i = 1$ if $y_i \leq T$.

Given the nature of the random variables involved in the Cragg (1971) model, different processes determine the respective components of the model.

A positive response necessarily comes from f, on the other hand, a T value comes from the point mass distribution. This model, however, does not consider the situation of a lower limit and that part of the observations may be below this lower limit.

If allowed to some limiting responses are the result of interval censored to f, we have the generalization of the two-part model exposed by Moulton & Halsey (1995). This means that an observed T value can be either a realization from the point-mass distribution or a partial observation from f with critical value not precisely known but lying somewhat in $(0, T)$ for a small pre-specified constant T. Formally,

$$g(y_i) = [p_i + (1 - p_i)F(T)]I_i + (1 - p_i) f(y_i)(1 - I_i)$$

where F is the cumulative distribution corresponding to f. If we vary the basic density f and the link function corresponding to p_i, we can generate a large family of mixed models. Models such as probit/trucated-normal, logit/lognormal, logit/log-gamma, probit/log-skew-normal and logit/log-alpha-power normal have been considered in practical applications in biology, economy, agricultural and so on (Chai & Bailey 2008, Martínez-Florez, Bolfarine & Gómez 2013). Notice that for $p_i = 0$, $i = 1, \ldots, n$, Moulton & Halsey (1995) model reduces to the Tobit model (Tobin 1958).

This is an extension of the log-normal distribution allowing for one extra parameter which will be presented in the next section.

2. Proportional Hazard Distribution

Recently, Martínez-Florez, Moreno-Arenas & Vergara-Cardozo (2013) introduced a new asymmetric model which is called proportional hazard model, this model is defined as follows:

Let F be a continuous cumulative distribution function with probability density function f, and hazard function $h = f/(1-F)$. We say that Z has a proportional hazard distribution associated with F, f and the parameter $\alpha > 0$ if its probability density function is

$$\varphi_F(z;\alpha) = \alpha f(z)\{1 - F(z)\}^{\alpha-1}, \quad z \in \mathbb{R},$$

where α is a positive real number. We use the notation $Z \sim PHF(\alpha)$. The distribution function of the PHF model is given by

$$\mathbb{F}(z) = 1 - \{1 - F(z)\}^{\alpha},\ z \in \mathbb{R}.$$

This is why this type of distribution can also be regarded as an exponentiated distribution or a fractional order statistic distribution, widely studied in the literature.

If Z is a random variable from a standard $PHF(\alpha)$ distribution then the location-scale extension of Z is obtained from the transformation $X = \xi + \eta Z$, where $\xi \in R$ and $\eta \in R^+$, is a scale parameter.

In the particular case where $F = \Phi(\cdot)$, we have the family of distributions called proportional hazard normal (PHN) and denoted $PHN(\xi, \eta, \alpha)$.

In Martínez-Florez, Moreno-Arenas & Vergara-Cardozo (2013), we can see the behavior of the $PHN(0, 1, \alpha)$ density and the model hazard function for some values of the α parameter.

2.1. Log Proportional-Hazard Distribution

Let Y be a random variable with support in R^+, we say that Y follows a univariate log-proportional-hazard distribution with parameter α, if the transformed variable $X = \log(Y) \sim PHF(\alpha)$. We denote $Y \sim LPHF(\alpha)$.

Then, the pdf for the random variable Y can be written as

$$\varphi_{LF}(y;\alpha) = \frac{\alpha}{y} f(\log(y))\left\{1 - F(\log(y))\right\}^{\alpha-1}, \qquad y \in R^+$$

where F is an absolutely continuous distribution function with density function $f = dF$. This model is called standard log proportional-hazard distribution.

Let $X \sim PHF(\xi, \eta, \alpha)$, where $\xi \in \mathbb{R}$ is a location parameter and $\eta \in \mathbb{R}^+$ is a scale parameter. Hence, the transformation $X = \ln(Y)$ leads to the location-scale log proportional-hazard model, with pdf given by

$$\varphi_{LF}(y;\xi,\eta,\alpha) = \frac{\alpha}{\eta y} f\left(\frac{\log(y) - \xi}{\eta}\right)\left\{1 - F\left(\frac{\log(y) - \xi}{\eta}\right)\right\}^{\alpha-1}, \qquad y \in \mathbb{R}^+$$

We use the notation $Y \sim LPHF(\xi, \eta, \alpha)$, so that $LPHN(\alpha) = LPHN(0, 1, \alpha)$.

Its cumulative distribution function can be written as

$$\mathcal{F}_F(y;\alpha) = 1 - \{1 - F(\log(y))\}^{\alpha}, \qquad y \in \mathbb{R}^+. \tag{1}$$

According to (1), the inversion method can be used for generating from a random variable with distribution $LPHF(\xi, \eta, \alpha)$. That is, if $U \sim U(0, 1)$, then, random variable $Y = e^{\xi + \eta F^{-1}(1-(1-U)^{1/\alpha})}$ is distributed according to the LPHF distribution with vector of parameters $\boldsymbol{\theta} = (\xi, \eta, \alpha)'$.

In the special case where $f = \phi(\cdot)$ and $F = \Phi(\cdot)$, the density and distribution functions of the standard normal distribution, respectively, we have the standard log proportional-hazard-normal distribution.

We will denote this extension by using the notation $Y \sim LPHN(\xi, \eta, \alpha)$.

Figure 1 shows the pdf's for the LPHN distribution for α equals 0.75, 1, 2 and 3. Is clearly seen that the shape of the distribution is affected when changes the value of α. For the log-normal case, when $\alpha = 1$, the kurtosis is smaller than when $\alpha = 2$ and, similarly, for the log-skew case, when $\alpha = 3$. Furthermore, when $\alpha = 0.75$ the kurtosis for the log-normal is greater. Asymmetry is always positive and also controlled by parameter α. Hence, α controls asymmetry as well as kurtosis for the LPHN distribution.

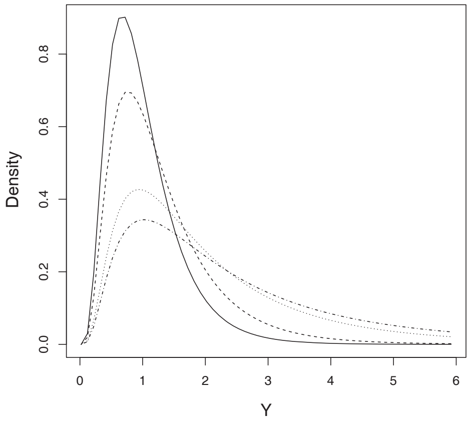


Figure 1: Plots of pdf $\varphi_{L\Phi}(y; 0.5, 0.75, \alpha)$, for α equals to 3 (solid line), 2 (dashed line), 1 (dotted line) and 0.75 (dashed and dotted line).

The r-th moment for the random variable $Y \sim LPHN$ is calculated numerically. Using the results of the central moments $\acute{\mu}_r$, the coefficients of variation, asymmetry and kurtosis are obtained.

Figure 2 shows the behavior of the mean and the coefficients of asymmetry and kurtosis of the LPHN model.

The survival and hazard functions for the LPHN model are, respectively, given by

$$S(t) = \{1 - \Phi(\log(t))\}^{\alpha} \;\text{ and }\; r(t) = \frac{\alpha}{t}\frac{\phi(\log(t))}{1 - \Phi(\log(t))} = \alpha r_{LN}(t)$$

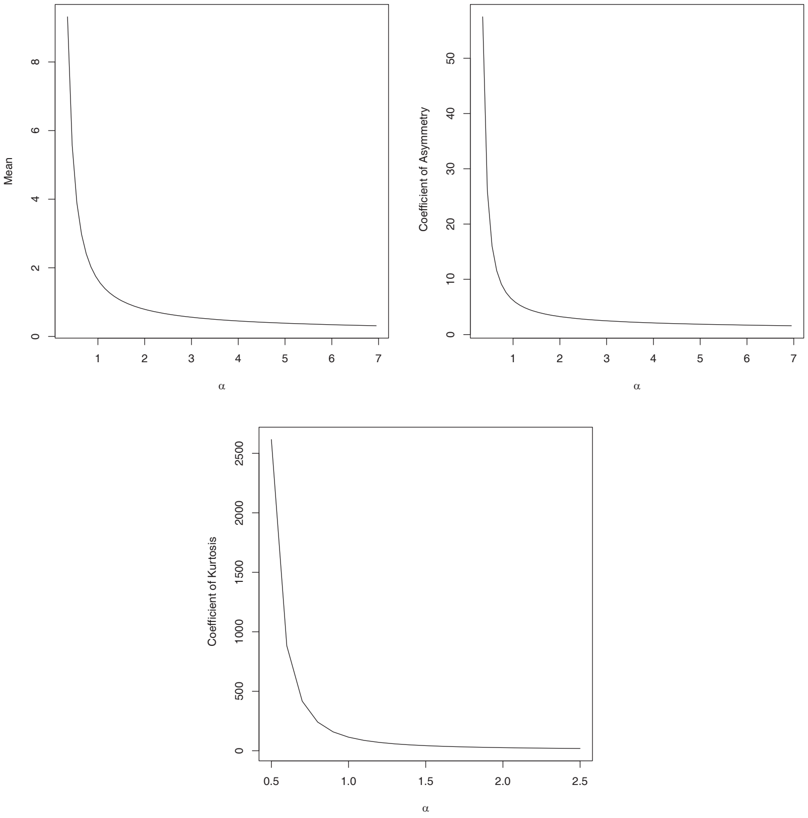


FIGURE 2: Behavior of some characteristic values of the LPHN model. (a) mean, (b) asymmetry coefficient and (c) coefficient of kurtosis.

where $r_{LN}(\cdot)$ is the hazard function of the log-normal distribution. Then, the hazard index T is proportional to the hazard index of the log-normal distribution.

2.2. Inference for Log Proportional-Hazard-Normal Model

For a random sample of size n, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ with $Y_i \sim LPHN(\xi, \eta, \alpha)$, the log-likelihood function of $\boldsymbol{\theta} = (\xi, \eta, \alpha)'$ given $\mathbf{Y}$ is

$$\ell(\boldsymbol{\theta}; \mathbf{Y}) = n\log(\alpha) - n\log(\eta) - \sum_{i=1}^{n}\log(y) + \sum_{i=1}^{n}\log(\phi(z_i)) + (\alpha-1)\sum_{i=1}^{n}\log(1-\Phi(z_i)),$$

where $z_i = \frac{\log(y_i)-\xi}{\eta}$. The corresponding score equations are similar to the obtained in Martínez-Florez, Moreno-Arenas & Vergara-Cardozo (2013), we only need to consider the change in the log-likelihood function and obtain the MLE estimators using numerical methods.

The observed information matrix for location-scale PHN follows from minus the second derivatives of the log-likelihood function. This result is similar to that obtained by Martínez-Florez, Moreno-Arenas & Vergara-Cardozo (2013) but with minor changes due to the difference in the log-likelihood function.

2.2.1. Expected Information Matrix for the Location-Scale PHN

Considering $a_{kj} = \mathbb{E}\{z_i^k w_i^j\}$, where $w_i = \frac{\phi(z_i)}{1-\Phi(z_i)}$, the expected information matrix entries are:

$$I_{\xi\xi} = \frac{1}{\eta^2}\left[1 + (\alpha - 1)(a_{02} - a_{11})\right] \qquad I_{\eta\xi} = \frac{2}{\eta^2}a_{10} + \frac{\alpha - 1}{\eta^2}\left[a_{01} - a_{02} + a_{12}\right]$$

$$I_{\eta\eta} = -\frac{1}{\eta^2} + \frac{3}{\eta^2}a_{20} + \frac{\alpha - 1}{\eta^2}\left[a_{22} + 2a_{11} - a_{31}\right]$$

$$I_{\alpha\xi} = -\frac{1}{\eta}a_{01} \qquad I_{\alpha\eta} = -\frac{1}{\eta}a_{11} \qquad I_{\alpha\alpha} = \frac{1}{\alpha^2}$$

The expected values of the above variables are generally calculated using numerical integration. When $\alpha = 1$, $\varphi_{L\Phi}(x;\xi,\eta,1) = \frac{1}{\eta y}\phi\left(\frac{\log(y)-\xi}{\eta}\right)$, the location-scale log-normal density function. Thus, the information matrix becomes

$$I(\boldsymbol{\theta}) = \begin{pmatrix} 1/\eta^2 & 0 & -a_{01}/\eta \\ 0 & 2/\eta^2 & -a_{11}/\eta \\ -a_{01}/\eta & -a_{11}/\eta & 1 \end{pmatrix}$$

Numerical integration shows that the determinant is $|I(\boldsymbol{\theta})| = \frac{1}{\eta^4}[2 - a_{11}^2 - 2a_{01}^2] \neq 0$, so in the case of a log-normal distribution the model's information matrix is nonsingular. The upper left 2×2 submatrix is the log-normal distribution's information matrix.

For large n and under regularity conditions we have

$$\hat{\boldsymbol{\theta}} \xrightarrow{A} N_3(\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1})$$

and the conclusion follows that $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically approaches the normal distribution with $I(\boldsymbol{\theta})^{-1}$ as covariance matrix, for large samples.

This result shows that the information matrix for the LPHN model is nonsingular and therefore the inference for large samples can be made, contrary to the log-skew-normal model, whose information matrix is singular when $\lambda = 0$, that consequently resulting likelihood ratio statistic is not distributed as a chi-square.

Note that as in the LPHN model, the information matrix of the log-skew-normal model has the same structure or shape that the location-scale skew-normal model, $SN(\xi, \eta, \lambda)$, where now $Z = (\log(y) - \xi)/\eta$. Is well known and was demonstrated by Azzalini (1985), that the information matrix of the skew-normal model is singular when its parameter of asymmetry $\lambda = 0$.

3. Asymmetric Regression Model Logit/LPHN

We now extend the LPHN model to the case of random variables with a limit of detection and the presence of covariates. Specifically, we consider the case of models with limited response and excess zeros in the response variable. Considering extensions of the generalized the two-part model of Moulton & Halsey (1995) to the situations logit/log proportional hazard-normal model, jointly with covariates at each step of the model. Initially, we develop the case of censored random variables LPHN. Thus, calling p_0 the proportion of observations at or below threshold point T, the censored model $LPHN(\xi,\eta,\alpha)$ is represented by the probability density function

$$g(y_i)=\begin{cases} p_{0i}+(1-p_{0i})\left[1-\left\{1-\Phi\left(\frac{\log(T)-\xi}{\eta}\right)\right\}^{\alpha}\right], & \text{if } y_i\leq T\\ (1-p_{0i})\frac{\alpha}{\eta y}\phi\left(\frac{\log(y_i)-\xi}{\eta}\right)\left\{1-\Phi\left(\frac{\log(y_i)-\xi}{\eta}\right)\right\}^{\alpha-1}, & \text{if } y_i>T\end{cases}$$

Now we extend this model to the case of presence of covariates in limited response and when the response is not limited.

The above model can be extended to the situation where only a proportion $100p_0\%$ of censored observations comes from the censored LPHN, with the remaining $100(1-p_0)\%$ of the observations coming from the population of low responders, located below or at the point T.

Modeling this mixture as the outcome of a Bernoulli random variable D with

$$pr(D=1)=1-p_0$$

while for $D=0$, $Y\leq T$ with probability one. The contribution of y_i to the likelihood conditioning on $D=1$ when Y is assumed to follow a LPHN model can be written as

$$\left[1-(1-p_0)\left\{1-\Phi\left(\frac{\log(T)-\xi}{\eta}\right)\right\}^{\alpha}\right]^{I_i}$$
$$\left[\frac{(1-p_0)\alpha}{\eta y_i}\phi\left(\frac{\log(y_i)-\xi}{\eta}\right)\left\{1-\Phi\left(\frac{\log(y_i)-\xi}{\eta}\right)\right\}^{\alpha-1}\right]^{1-I_i}$$

Then, assuming that the response $y_i=T$ is explained by the set of explanatory variables $X_{11},X_{12},\dots,X_{1p}$, then we model this mixture as the outcome of a Bernoulli random variable with logit link function with

$$p_{0i}=prob(y_i=T)=\frac{\exp\left(x'_{(1)i}\beta_{(1)}\right)}{1+\exp\left(x'_{(1)i}\beta_{(1)}\right)}$$

and

$$1-p_{0i}=\frac{1}{1+\exp\left(x'_{(1)i}\beta_{(1)}\right)}$$

where $x_{(1)i} = (1, x_{1i1}, \ldots, x_{1ip})'$, is a covariate vector of dimension $p+1$ associated with the parameter vectors $\boldsymbol{\beta}_{(1)} = (\beta_{10}, \beta_{11}, \ldots, \beta_{1p})'$.

Taking into account the LPHN model, we have a covariate vector $x_{(2)} = (1, X_{21}, X_{22}, \ldots, X_{2r})'$ of dimension r, possibly different from $x_{(1)}$ and parameter vector $\boldsymbol{\beta}_{(2)} = (\beta_{20}, \beta_{21}, \ldots, \beta_{2r})'$, for which

$$\log(y_i) \sim PHN(x'_{(2)i}\boldsymbol{\beta}_{(2)}, \eta, \alpha), \quad y_i > T$$

where $x_{(2)i} = (1, x_{2i1}, \ldots, x_{2ir})'$.

This mixture of distributions we will call "linear logistic regression model" with proportional hazard-normal distribution and will be denoted by

$$RLLPHN(\beta_{(1)}, \beta_{(2)}, \eta, \alpha)$$

The logarithm of the likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\beta}'_{(1)}, \boldsymbol{\beta}'_{(2)}, \eta, \alpha)'$ given $X_{(1)}$, $X_{(2)}$ and Y, is given by

$$\begin{aligned}
\ell(\boldsymbol{\theta}) = \sum_i I_i \log\left[1 + \exp(x'_{(1)i}\boldsymbol{\beta}_{(1)})\left[1 - \{1 - \Phi(z_{Ti})\}^{\alpha}\right]\right] \\
- \sum_{i=1}^{n} \log\left[1 + \exp(x'_{(1)i}\boldsymbol{\beta}_{(1)})\right] \\
+ \sum_i (1 - I_i)\left\{\log(\alpha) - \log(\eta y_i) + x'_{(1)i}\boldsymbol{\beta}_{(1)} + \log(\phi(z_i)) + (\alpha - 1)\log(1 - \Phi(z_i))\right\}
\end{aligned}$$

where $z_{Ti} = \frac{\log(T) - x'_{(2)i}\boldsymbol{\beta}_{(2)}}{\eta}$ and $z_i = \frac{\log(y_i) - x'_{(2)i}\boldsymbol{\beta}_{(2)}}{\eta}$.

We denote by $\sum_0$ the sum over censored observations and $\sum_1$ the sum over noncensored observations. The score function corresponding to the log-likelihood function is given by (for $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, r$)

$$\begin{aligned}
U(\beta_{(1)j}) = \sum_0 \frac{x_{1ij}\exp(x'_{(1)i}\beta_{(1)})\left[1 - \{1 - \Phi(z_{T_i})\}^{\alpha}\right]}{1 + \exp(x'_{(1)i}\beta_{(1)})\left[1 - \{1 - \Phi(z_{T_i})\}^{\alpha}\right]} \\
- \sum_{i=1}^{n} \frac{x_{1ij}\exp(x'_{(1)i}\beta_{(1)})}{1 + \exp(x'_{(1)i}\beta_{(1)})} + \sum_1 x_{1ij}
\end{aligned}$$

$$
\begin{aligned}
U(\beta_{(2)k}) &= -\sum_{0} \frac{x_{2ik}\exp(x'_{(1)i}\beta_{(1)})\varphi_{L\Phi}(T, x'_{(2)i}\beta_{(2)}, \eta, \alpha)}{1+\exp(x'_{(1)i}\beta_{(1)})\left[1-\{1-\Phi(z_{T_i})\}^{\alpha}\right]}\\
&\quad -\frac{1}{\eta}\sum_{1} x_{2ik}\left[-z_i-(\alpha-1)\frac{\phi\left(z_i\right)}{1-\Phi\left(z_i\right)}\right]\\
U(\eta) &= -\sum_{0} \frac{z_{T_i}\exp(x'_{(1)i}\beta_{(1)})\varphi_{L\Phi}(T, x'_{(2)i}\beta_{(2)}, \eta, \alpha)}{1+\exp(x'_{(1)i}\beta_{(1)})\left[1-\{1-\Phi(z_{T_i})\}^{\alpha}\right]}\\
&\quad -\frac{1}{\eta}\sum_{1}\left[1-z_i^2-(\alpha-1)z_i\frac{\phi\left(z_i\right)}{1-\Phi\left(z_i\right)}\right]\\
U(\alpha) &= -\sum_{0} \frac{\exp(x'_{(1)i}\beta_{(1)})\left\{1-\Phi\left(z_{T_i}\right)\right\}^{\alpha}\log\left(1-\Phi(z_{T_i})\right)}{1+\exp(x'_{(1)i}\beta_{(1)})\left[1-\{1-\Phi(z_{T_i})\}^{\alpha}\right]}\\
&\quad +\sum_{1}\left[\frac{1}{\alpha}+\log\left(1-\Phi(z_i)\right)\right]
\end{aligned}
$$

The system of equations obtained by equating the score to zero has no solution in closed form, and tends to be solved using iterative numerical methods.

The resulting equations require numerical procedures such as the Newton-Raphson or quasi-Newton method. These optimization algorithms can be found in the packages *maxLik* or *optimx* of the R software.

The observed information matrix is given by $J(\boldsymbol{\theta}) = -H(\boldsymbol{\theta}) = -\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$, where $H(\boldsymbol{\theta})$ is the hessian matrix, which is obtained in the Appendix for the vector of parameters $\boldsymbol{\theta}$. In addition can be obtained information matrix defined as less n^{-1} times the expected value of the observed information matrix.

4. Numerical Illustration

The application of the logit/LPHN model, is carried out using the data described by Moulton & Halsey (1995) in a study of measles vaccines conducted in Haiti during 1987-1990. The detection limit was 0.1 international units (UI), or $\log(0.1) = -2.306$ in the natural log-scale. The codification for the covariates involved in the study were $X_1 = EZ$ (vaccine type; 0:Schwarz , 1:Edmonston-Zagreb); $X_2 = HI$ (vaccine dose; 0:medium, 1:high) and $X_3 = FEM$ (gender; 0:male, 1:female).

Such as Moulton & Halsey (1995), the aim in the present analysis is to study the immunogenicity differential between boys and girls using the logit/ log-proportional-hazard-normal (logit/LPHN) model.

4.1. Models

A variety of models can be adjusted given the covariates in the study. We adjust some of these models were carefully chosen from the cases studied by Moulton & Halsey (1995).

Model 1: Covariates and censored data in limited response, without censored data and covariates in the point-mass distribution located at zero;

Model 2: Censored data and covariates in limited response, without covariates in the point-mass distribution located at zero;

Model 3: Censored data, covariates in limited response and in the point-mass distribution located at zero;

Model 4: Censored data, covariates in limited response and in the point-mass distribution located at zero, a particular model.

The summary statistics we have $\overline{\log(y)} = -0.1793$, $s^2 = 1.1055$, $\sqrt{b_1} = 0.7521$ and $b_2 = 2.6286$ where the quantities $\sqrt{b_1}$ and b_2 correspond to the sample coefficients of asymmetry and kurtosis for values above 0.1. The high asymmetry degree indicated by the sample coefficient of asymmetry ($\sqrt{b_1}$) reveals that it seems worthwhile trying to fit an asymmetric model for this data set.

Moulton & Halsey (1995), and Moulton & Halsey (1996) modeled this data using the hybrids logit/log-normal (logit/LN) and logit/log-gamma (logit/LGM) models.

As a first attempt, we fitted the ordinary Tobit model with covariates (model 1), which resulted in a poor fit to the data set under study. Here $\hat{\beta}_{(2)0} = 0.565$, $\hat{\beta}_{(2)1} = 0.248$, $\hat{\beta}_{(2)2} = -0.191$ and $\hat{\beta}_{(2)3} = 0.262$, and $AIC = 1291.81$.

We adjust the mixtures logit/LN and logit/LGM, for 1-4 models, finding in both cases the model 4 presents the best fit. The estimates for these models are given in the Table 1. Note that δ is the shape parameter of the LGM model.

Table 1: Parameter estimation (standard error) and model fitting for one and two components hybrid Bernoulli/log-distributions.

		Bernoulli component			Log-distributions components		
density	AIC	INT	EZ	HI	δ	INT	FEM
LN	986.19	0.652	0.808	0.422		-0.401	0.264
		(0.220)	(0.304)	(0.288)		(0.112)	(0.155)
LGM	1022.43	0.572	0.656	0.374	-2.833	-1.179	0.053
		(0.201)	(0.261)	(0.255)	(0.510)	(0.088)	(0.056)

Hence, there is a clear indication that the conditions under which the Tobit model is adequate, are not satisfied for the measles vaccine data set.

Estimates (MLEs) for the model parameters 1-4, were obtained, and the results are shown in Tables 2.

To compare model fit, we computed the AIC criterion (Akaike 1974).

Table 2: Parameter estimation (standard error) and model fitting for one and two components hybrid logit/LPHN.

		Bernoulli component				Log-distributions components				
Model	AIC	INT	EZ	HI	FEM	INT	EZ	HI	FEM	α
(1)	1022.95					2.418	0.256	0.081	0.180	6.669
						(0.966)	(0.192)	(0.191)	(0.191)	(2.661)
(2)	992.17	1.051				-1.014	-0.162	-0.012	0.271	0.391
		(0.138)				(0.379)	(0.148)	(0.148)	(0.149)	(0.229)
(3)	979.98	0.875	1.027	0.385	-0.612	-1.514	-0.241	-0.104	0.142	0.074
		(0.258)	(0.330)	(0.273)	(0.291)	(0.480)	(0.178)	(0.153)	(0.143)	(0.152)
(4)	975.44	0.488	0.911	0.368		-1.609			0.286	0.152
		(0.203)	(0.275)	(0.262)		(0.002)			(0.060)	(0.010)

We started by fitting the censored LPHN model with covariates (Model 1). It is also fitted by the Bernoulli/LPHN model with covariates and logit link (models 2-4), for which the results are presented in the Table 2. According to the criterion AIC, the best fit clearly is presented by the hybrid logit/LPHN model.

In the case of the Bernoulli/LPHN model, we found that of all hybrid models fitted, the best is the Model 4.

In the continuous component we has that $E(Y) \neq X_{(2)}\beta_{(2)}$ since $E(\epsilon) \neq 0$. In order to have $E(Y) = X_{(2)}\beta_{(2)}$ we must correct the intercept taking $\beta^*_{(2)0} = \beta_{(0)} + E(\epsilon)$, where $\epsilon \sim LPN(0, \eta, \alpha)$, That is, the corrected estimator for the intercept of the regression model corresponding to the continuous part. Therefore, for model 4, we found that $\hat{\beta}^*_{(2)0} = -0.333$.

Here, covariates EZ and HI entered only in the Bernoulli component, and covariate FEM is the only associated with the LPHN component. Based on the Model 4, for those observations above the detection limit, the girls had $\exp(0.286) = 1.331$, and hence greater measles antibody concentration than boys.

As mentioned at the beginning of this illustration, the goal was to show that the model censored logit/LPHN was a good alternative to adjust the data set vaccine now we are going to show that this model is indeed different from the model censored logit/LN, so, we test the hypothesis

$$H_0: \ \alpha = 1 \ \text{ versus } \ H_1: \ \alpha \neq 1$$

Using the likelihood ratio statistics, we have that

$$-2\log(\Lambda) = -2(-511.18 + 480.72) = 60.91$$

which is greater than the 5% critical chi-square value 3.84, then we conclude that the logit/LPHN model fits the data better than the logit/LN model.

As a proof of good fit of the proposed model, we can confirm that the proportion of observations below the detection limit is 26.1% and the estimated proportion from model 2 with the hybrid model logit/LPHN is 25.90%.

Finally, in order to check the fit of the model estimates, we make the QQplot of the standardized residuals or scaled residuals of the continuous part, $e_i = (\log(y_i) -$

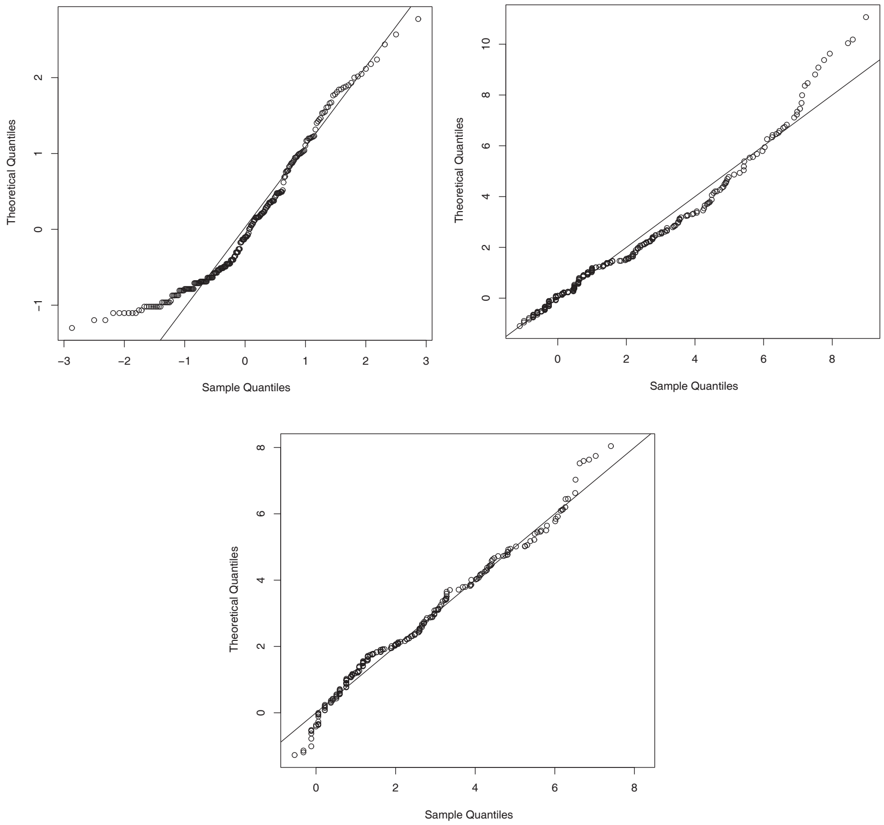


FIGURE 3: QQ-plot of the scaled residuals e_i, from the fit of Model 4. (a) log-normal (b)log-gamma and (c) log-proportional hazard-normal.

$x'_{(2)i}\hat{\beta}_{(2)})/\hat{\eta}$ based on the model and to the LN, LGM and LPHN distributions. Figure 3 presents QQplots for the scaled residuals.

Here, we can see that for vaccine data, the model LPHN fits better than the LN and LGM models, and thus, the mixed model logit/LPHN may be a new option to adjust censored data with covariates.

5. Conclusions

We proposed a new distribution that is used to study an asymmetrical regression model for data with limited responses through the mixture of a Bernoulli distribution with logit link and the LPHF distribution. Additionally, we made an illustration with real data and showed that the proposed model is an alternative for censored positive data.

[Recibido: noviembre de 2013 — Aceptado: abril de 2014]

References

Ahrens, L. H. (1954), *Quantitative Spectrochemical Analysis of Silicates*, London, Pergamon Press.

Akaike, H. (1974), 'A new look at statistical model identification', *IEEE Transaction on Automatic Control* **AU-19**, 716–722.

Arellano-Valle, R. B. & Azzalini, A. (2008), 'The centred parametrization for the multivariate skew-normal distribution', *Journal of Multivariate Analysis* **99**, 1362–1382.

Azzalini, A. (1985), 'A class of distributions which includes the normal ones', *Scandinavian Journal of Statistics* **12**, 171–178.

Azzalini, A., dal Cappello, T. & Kotz, S. (2003), 'Log-skew-normal and log-skew-t distributions as models for family income data', *Journal of Income Distribution* **11**, 12–20.

Chai, H. & Bailey, K. (2008), 'Use of log-normal distribution in analysis of continuous data with a discrete component at zero', *Statistics in Medicine* **27**, 3643–3655.

Cragg, J. (1971), 'Some statistical models for limited dependent variables with application to the demand for durable goods', *Econometrica* **39**, 829–844.

Job, J., Halsey, N., Boulos, R., Holt, E., Farrell, D., Albrecht, P., Brutus, J., Adrien, M., Andre, J., Chan, E., Kissinger, P., Boulos, C. & the CiteSoleil/JHU, P. T. (1991), 'Successful immunization of infants at 6 months of age with high dose edmonston-zagreb measles vaccine', *Pediatric Infectious Diseases Journal* **10**, 303–311.

Martínez-Florez, G., Bolfarine, H. & Gómez, H. W. (2013), 'Asymmetric regression models with limited responses with an application to antibody response to vaccine', *Biometrical Journal* **55**, 156–172.

Martínez-Florez, G., Moreno-Arenas, G. & Vergara-Cardozo, S. (2013), 'Properties and inference for proportional hazard models', *Revista Colombiana de Estadística* **36**(1), 95–114.

Mateu-Figueras, G. & Pawlosky-Glanh (2003), Una alternativa a la distribución log-normal, *in* 'Actas del XXVII Congreso Nacional de Estadística e Investigación Operativa (SEIO)', Sociedade de Estadítica e Investigación Operativa, España, pp. 1849–1858.

Mateu-Figueras, G., Pawlosky-Glanh, . & Barcelo-Vidal, C. (2004), The natural law in geochemistry: Lognormal or log skew-normal?, *in* '32th International Geological Congress', International Union of Soil Sciences, Florence, Italy, pp. 1849–1858.

Moulton, L. & Halsey, N. (1995), 'A mixture model with detection limits for regression analyses of antibody response to vaccine', *Biometrics* **51**, 1570–1578.

Moulton, L. & Halsey, N. (1996), 'A mixed Gamma model for regression analyses of quantitative assay data', *Vaccine* **14**, 1154–1158.

Tobin, J. (1958), 'Estimation of relationships for limited dependent variables', *Econometrica* **26**, 24–36.

Appendix

In this appendix we present the Hessian matrix for the logit/LPHN model. Its elements are given by

$$\begin{aligned}U(\beta_{(1)j}\beta_{(1)r}) = \sum_{0} x_{1ij}x_{1ir}&\left[\frac{\exp(x'_{(1)i}\beta_{(1)})[1-\{1-\Phi(z_{T_i})\}^{\alpha}]}{\{1+\exp(x'_{(1)i}\beta_{(1)})[1-\{1-\Phi(z_{T_i})\}^{\alpha}]\}^2}\right]\\ &-\sum_{i=1}^{n}\frac{x_{1ij}x_{1ir}\exp(x'_{(1)i}\beta_{(1)})}{[1+\exp(x'_{(1)i}\beta_{(1)})]^2},\end{aligned}$$

$$U(\beta_{(2)k}\beta_{(1)j}) = \frac{-\alpha}{\eta}\sum_{0}\frac{x_{2ik}x_{1ij}\phi(z_{T_i})\exp(x'_{(1)i}\beta_{(1)})\{1-\Phi(z_{T_i})\}^{\alpha-1}}{\{1+\exp(x'_{(1)i}\beta_{(1)})[1-\{1-\Phi(z_{T_i})\}^{\alpha}]\}^2}$$

$$U(\beta_{(1)j}\eta) = \frac{-\alpha}{\eta}\sum_{0}\frac{x_{1ij}z_{T_i}\phi(z_{T_i})\exp(x'_{(1)i}\beta_{(1)})\{1-\Phi(z_{T_i})\}^{\alpha-1}}{\{1+\exp(x'_{(1)i}\beta_{(1)})[1-\{1-\Phi(z_{T_i})\}^{\alpha}]\}^2},$$

$$U(\beta_{(1)j}\alpha) = -\sum_{0}\frac{x_{ij}\exp(x'_{(1)i}\beta_{(1)})\{1-\Phi(z_{T_i})\}^{\alpha}\log(1-\Phi(z_{T_i}))}{[1+\exp(x'_{(1)i}\beta_{(1)})[1-\{1-\Phi(z_{T_i})\}^{\alpha}]]^2},$$

$$\begin{aligned}U(\beta_{(2)k}\beta_{(2)s}) =&\frac{-\alpha}{\eta^2}\sum_{0}x_{2ik}x_{2is}\left\{[z_{T_i}+(\alpha-1)M_i]\,A_i+A_i^2\right\}+\\ &\frac{1}{\eta^2}\sum_{1}x_{2ik}x_{2is}\left\{-1+(\alpha-1)z_iM_i-(\alpha-1)M_i^2\right\},\end{aligned}$$

$$\begin{aligned}U(\beta_{(2)k}\eta) =&\frac{\alpha}{\eta^2}\sum_{0}\left\{\left[x_{2ik}-x_{2ik}z_{T_i}^2-(\alpha-1)x_{2ik}z_{T_i}M_i\right]A_i-\alpha x_{2ik}z_{T_i}A_i^2\right\}\\ &+\frac{1}{\eta^2}\sum_{1}\left\{x_{2ik}\left[\frac{-2z_i}{\eta}-(1-z_i^2)(\alpha-1)M_i-z_i(\alpha-1)M_i^2\right]\right\},\end{aligned}$$

$$\begin{aligned}
U(\beta_{(2)k}\alpha) =& \frac{-1}{\eta}\sum_{0}\Bigg\{[1+\alpha\log(1-\Phi(z_{T_i}))]x_{2ik}A_i\\
&+\frac{\alpha x_{2ik}[1-\Phi(z_{T_i})]\log(1-\Phi(z_{T_i}))}{\phi(z_{T_i})}A_i^2\Bigg\}+\frac{1}{\eta}\sum_{1}x_{2ik}M_i,
\end{aligned}$$

$$\begin{aligned}
U(\eta\eta) =& \frac{\alpha}{\eta^2}\sum_{0}\left\{\left[2z_{T_i}-z_{T_i}^3+(\alpha-1)z_{T_i}^2M_i\right]A_i+\alpha z_{T_i}^2A_i^2\right\}+\\
&\frac{1}{\eta^2}\sum_{1}\left\{1-z_i^2-(\alpha-1)z_iM_i\left[2-\frac{z_i^2-\phi(z_i)z_i}{\phi(z_i)}M_i\right]\right\},
\end{aligned}$$

$$\begin{aligned}
U(\eta\alpha) =& \frac{-1}{\eta}\sum_{0}\Bigg\{[z_{T_i}+\alpha z_{T_i}\log(1-\Phi(z_{T_i}))]A_i\\
&+\left[\frac{\alpha z_{T_i}(1-\Phi(z_{T_i}))\log(1-\Phi(z_{T_i}))}{z_{T_i}\phi(z_{T_i})}\right]A_i^2\Bigg\}+\frac{1}{\eta}\sum_{1}z_iM_i,
\end{aligned}$$

$$\begin{aligned}
U(\alpha\alpha) =& -\sum_{0}\Bigg\{\frac{\{1-\Phi(z_{T_i})\}\log^2(1-\Phi(z_{T_i}))}{\phi(z_{T_i})}A_i\\
&+\left[\frac{\{1-\Phi(z_{T_i})\}\log(1-\Phi(z_{T_i}))}{\Phi(z_{T_i})}A_i\right]^2\Bigg\}-\sum_{1}\frac{1}{\alpha^2}
\end{aligned}$$

where

$$A_i=\frac{\phi(z_i)\exp(x'_{(1)i}\beta_{(1)})\{1-\Phi(z_{T_i})\}^{\alpha-1}}{1+\exp(x'_{(1)i}\beta_{(1)})[1-\{1-\Phi(z_{T_i})\}^{\alpha}]}\quad\text{and}\quad M_i=\frac{\phi(z_{T_i})}{1-\Phi(z_{T_i})}.$$

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 199 a 211

A New Difference-Cum-Exponential Type Estimator of Finite Population Mean in Simple Random Sampling

Un nuevo estimador tipo diferencia-cum-exponencial de la media de una población finita en muestras aleatorias simple

Javid Shabbir[1,a], Abdul Haq[1,b], Sat Gupta[2,c]

[1] Department of Statistics, Quaid-I-Azam University, Islamabad, Pakistan

[2] Department of Mathematics and Statistics, The University of North Carolina at Greensboro, Greensboro, USA

Abstract

Auxiliary information is frequently used to improve the accuracy of the estimators when estimating the unknown population parameters. In this paper, we propose a new difference-cum-exponential type estimator for the finite population mean using auxiliary information in simple random sampling. The expressions for the bias and mean squared error of the proposed estimator are obtained under first order of approximation. It is shown theoretically, that the proposed estimator is always more efficient than the sample mean, ratio, product, regression and several other existing estimators considered here. An empirical study using 10 data sets is also conducted to validate the theoretical findings.

***Key words*:** Ratio estimator, Auxiliary Variable, Exponential type estimator, Bias, MSE, Efficiency.

Resumen

Información auxiliar se utiliza con frecuencia para mejorar la precisión de los estimadores al estimar los parámetros poblacionales desconocidos. En este trabajo, se propone un nuevo tipo de diferencia-cum-exponencial estimador de la población finita implicar el uso de información auxiliar en muestreo aleatorio simple. Las expresiones para el sesgo y el error cuadrático medio del estimador propuesto se obtienen en primer orden de aproximación. Se muestra teóricamente, que el estimador propuesto es siempre más eficiente que la media de la muestra, la relación de, producto, regresión y varios otros

[a] Professor. E-mail: javidshabbir@gmail.com
[b] Lecturer. E-mail: aaabdulhaq@gmail.com
[c] Professor. E-mail: sngupta@uncg.edu

estimadores existentes considerados aquí. Un estudio empírico utilizando 10 conjuntos de datos también se lleva a cabo para validar los resultados teóricos.

***Palabras clave*:** estimador de razón, variables auxiliares, estimador tipo exponencial, sesgo, error cuadrático medio.

1. Introduction

In sample surveys, auxiliary information can be used either at the design stage or at the estimation stage or at both stages to increase precision of the estimators of population parameters. The ratio, product and regression methods of estimation are commonly used in this context. Recently many research articles have appeared where authors have tried to modify existing estimators or construct new hybrid type estimators. Some contribution in this area are due to Bahl & Tuteja (1991), Singh, Chauhan & Sawan (2008), Singh, Chauhan, Sawan & Smarandache (2009), Yadav & Kadilar (2013), Haq & Shabbir (2013), Singh, Sharma & Tailor (2014) and Grover & Kaur, (2011, 2014).

Consider a finite population $U = \{U_1, U_2, \ldots, U_N\}$. We draw a sample of size n from this population using simple random sampling without replacement scheme. Let y and x respectively be the study and the auxiliary variables and y_i and x_i, respectively be the observations on the ith unit. Let $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ be the sample means and $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$, be the corresponding population means. We assume that the mean of the auxiliary variable $(\bar{X})$ is known. Let $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ be the sample variances and $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$and $S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$, be the corresponding population variances. Let ρ_{yx} be the correlation coefficient between y and x. Finally let $C_y = \frac{S_y}{\bar{Y}}$ and $C_x = \frac{S_x}{\bar{X}}$ respectively be the coefficients of variation for y and x.

In order to obtain the bias and mean squared error (MSE) for the proposed estimator and existing estimators considered here, we define the following relative error terms: Let $\delta_0 = \frac{\bar{y}-\bar{Y}}{\bar{Y}}$ and $\delta_1 = \frac{\bar{x}-\bar{X}}{\bar{X}}$, such that $E(\delta_i) = 0$ for $(i = 0, 1)$, $E(\delta_0^2) = \lambda C_y^2$, $E(\delta_1^2) = \lambda C_x^2$ and $E(\delta_0\delta_1) = \lambda \rho_{yx} C_y C_x$, where $\lambda = \left(\frac{1}{n} - \frac{1}{N}\right)$.

In this paper, our objective is to propose an improved estimator of the finite population mean using information on a single auxiliary variable in simple random sampling. Expressions for the bias and mean squared error (MSE) of the proposed estimator are derived under first order of approximation. Based on both theoretical and numerical comparisons, we show that the proposed estimator outperforms several existing estimators. The outline of the paper is as follows: in Section 2, we consider several estimators of the finite population mean that are available in literature. The proposed estimators are given in Section 3 along with the corresponding bias and MSE expressions. In Section 4, we provide theoretical comparisons to evaluate the performances of the proposed and existing estimators. An empirical study is conducted in Section 5, and some concluding remarks are given in Section 6.

2. Some Existing Estimators

In this section, we consider several estimators of finite population mean.

2.1. Sample Mean Estimator

The variance of the sample mean $\bar{y}$, the usual unbiased estimator, is given by

$$Var(\bar{y}) = \lambda \bar{Y}^2 C_y^2 \tag{1}$$

2.2. Traditional Ratio and Product Estimators

Using information on the auxiliary variable, Cochran (1940) suggested a ratio estimator $\hat{\bar{Y}}_R$ for estimating $\bar{Y}$. It is given by

$$\hat{\bar{Y}}_R = \bar{y}(\frac{\bar{X}}{\bar{x}}) \tag{2}$$

The MSE of $\hat{\bar{Y}}_R$, to first order of approximation, is given by

$$MSE(\hat{\bar{Y}}_R) \approx \lambda \bar{Y}^2 (C_y^2 + C_x^2 - 2\rho_{yx} C_y C_x) \tag{3}$$

On similar lines, Murthy (1964) suggested a product estimator ($\hat{\bar{Y}}_P$), given by

$$\hat{\bar{Y}}_P = \bar{y}\left(\frac{\bar{x}}{\bar{X}}\right) \tag{4}$$

The MSE of $\hat{\bar{Y}}_P$, to first order of approximation, is given by

$$MSE(\hat{\bar{Y}}_P) \approx \lambda \bar{Y}^2 (C_y^2 + C_x^2 + 2\rho_{yx} C_y C_x) \tag{5}$$

The ratio and product estimators are widely used when the correlation coefficient between the study and the auxiliary variable is positive and negative, respectively. Both of the estimators, $\hat{\bar{Y}}_R$ and $\hat{\bar{Y}}_P$, show better performances in comparison with $\bar{y}$ when $\rho_{yx} > \frac{C_x}{2C_y}$ and $\rho_{yx} < -\frac{C_x}{2C_y}$, respectively.

2.3. Regression Estimator

The usual regression estimator $\hat{\bar{Y}}_{Reg}$ of $\bar{Y}$, is given by

$$\hat{\bar{Y}}_{Reg} = \bar{y} + b(\bar{X} - \bar{x}) \tag{6}$$

where b is the usual slope estimator of the population regression coefficient β (Cochran 1977). The estimator $\hat{\bar{Y}}_{Reg}$ is biased, but the bias approaches zero as the sample size n increases.

Asymptotic variance of $\hat{\bar{Y}}_{Reg}$, is given by

$$Var(\hat{\bar{Y}}_{Reg}) = \lambda \bar{Y}^2 C_y^2 (1 - \rho_{yx}^2) \tag{7}$$

The regression estimator $\hat{\bar{Y}}_{Reg}$ performs better than the usual mean estimator $\bar{y}$, ratio estimator $\hat{\bar{Y}}_R$ and product estimator $\hat{\bar{Y}}_P$ when $\lambda \bar{Y}^2 \rho_{yx}^2 C_y^2 > 0$, $\lambda \bar{Y}^2 (C_x - \rho_{yx} C_y)^2 > 0$ and $\lambda \bar{Y}^2 (C_x + \rho_{yx} C_y)^2 > 0$, respectively.

2.4. Bahl & Tuteja (1991) Estimators

Bahl & Tuteja (1991) suggested ratio-and product type estimators of $\bar{Y}$, given respectively by

$$\hat{\bar{Y}}_{BT,R} = \bar{y} \exp\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right) \tag{8}$$

and

$$\hat{\bar{Y}}_{BT,P} = \bar{y} \exp\left(\frac{\bar{x} - \bar{X}}{\bar{x} + \bar{X}}\right) \tag{9}$$

The MSEs of $\hat{\bar{Y}}_{BT,R}$ and $\hat{\bar{Y}}_{BT,P}$, to first order of approximation, are given by

$$MSE(\hat{\bar{Y}}_{BT,R}) \approx (1/4) \lambda \bar{Y}^2 (4C_y^2 + C_x^2 - 4\rho_{xy} C_y C_x) \tag{10}$$

and

$$MSE(\hat{\bar{Y}}_{BT,P}) \approx (1/4) \lambda \bar{Y}^2 (4C_y^2 + C_x^2 + 4\rho_{xy} C_y C_x) \tag{11}$$

2.5. Singh et al. (2008) Estimator

Following Bahl & Tuteja (1991), Singh et al. (2008) suggested a ratio-product exponential type estimator $\hat{\bar{Y}}_{S,RP}$ of $\bar{Y}$, given by

$$\hat{\bar{Y}}_{S,RP} = \bar{y}[\alpha \exp(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}) + (1 - \alpha) \exp(\frac{\bar{x} - \bar{X}}{\bar{x} + \bar{X}})] \tag{12}$$

where α is an arbitrary constant.

The minimum MSE of $\bar{Y}_{S,RP}$, up to first order of approximation, at optimum value of α, i.e., $\alpha_{(opt)} = \frac{1}{2} + \frac{\rho_{yx} C_y}{C_x}$, is given by

$$MSE_{\min}(\hat{\bar{Y}}_{S,RP}) \approx \lambda \bar{Y}^2 (1 - \rho_{yx}^2) C_y^2 = Var(\hat{\bar{Y}}_{Reg}) \tag{13}$$

The minimum MSE of $\hat{\bar{Y}}_{S,RP}$ is exactly equal to variance of the linear regression estimator $(\hat{\bar{Y}}_{Reg})$.

2.6. Rao (1991) Estimator

Rao (1991) suggested a regression-type estimator of $\bar{Y}$, given by

$$\hat{\bar{Y}}_{R,Reg} = k_1\bar{y} + k_2(\bar{X} - \bar{x}) \tag{14}$$

where k_1 and k_2 are suitably chosen constants.

The minimum MSE of $\bar{Y}_{R,Reg}$, upto first order of approximation, at optimum values of k_1 and k_2, i.e., $k_{1(opt)} = \frac{1}{1+\lambda(1-\rho_{yx}^2)C_y^2}$ and $k_{2(opt)} = -\frac{\bar{Y}\rho_{yx}C_y}{\bar{X}C_x[-1+\lambda(-1+\rho_{yx}^2)C_y^2]}$, is given by

$$MSE_{\min}(\hat{\bar{Y}}_{R,Reg}) \approx \bar{Y}^2\left\{1 + \frac{1}{-1+\lambda(-1+\rho_{yx}^2)C_y^2}\right\} \tag{15}$$

2.7. Grover & Kaur (2011) Estimator

Following Rao (1991) and Bahl & Tuteja (1991), Grover & Kaur (2011) suggested an exponential type estimator of $\bar{Y}$, given by

$$\hat{\bar{Y}}_{GK} = [d_1\bar{y} + d_2(\bar{X} - \bar{x})]\exp(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}) \tag{16}$$

where d_1 and d_2 are suitably chosen constants.

The minimum MSE of $\hat{\bar{Y}}_{GK}$, up to first order of approximation, at optimum values of d_1 and d_2 i.e., $d_{1(opt)} = \frac{-8+\lambda C_x^2}{8\left\{-1+\lambda(-1+\rho_{yx}^2)C_y^2\right\}}$ and

$$d_{2(opt)} = \frac{\bar{Y}[-8\rho_{yx}C_y + C_x\left\{4 - \lambda C_x^2 + \lambda\rho_{yx}C_yC_x + 4\lambda(-1+\rho_{yx}^2)C_y^2\right\}}{8\bar{X}C_x\left\{-1+\lambda(-1+\rho_{yx}^2)C_y^2\right\}}$$

is given by

$$MSE_{\min}(\hat{\bar{Y}}_{GK}) \approx \frac{\lambda\bar{Y}^2[\lambda C_x^4 - 16(-1+\rho_{yx}^2)(-4+\lambda C_x^2)C_y^2]}{64[-1+\lambda(-1+\rho_{yx}^2)C_y^2]} \tag{17}$$

Grover & Kaur (2011) derived the result

$$MSE_{min}(\hat{\bar{Y}}_{GK}) \approx Var(\hat{\bar{Y}}_{Reg}) - \frac{\lambda^2\bar{Y}^2\left\{C_x^2 + 8(1-\rho_{yx}^2)C_y^2\right\}^2}{64\left\{1+\lambda(1-\rho_{yx}^2)C_y^2\right\}} \tag{18}$$

Equation (18) shows that $\hat{\bar{Y}}_{GK}$ is more efficient than the linear regression estimator $\hat{\bar{Y}}_{Reg}$.

Since regression estimator $\hat{\bar{Y}}_{Reg}$ is always better than $\bar{y}$, $\hat{\bar{Y}}_R$, $\hat{\bar{Y}}_P$, $\hat{\bar{Y}}_{BT,R}$, $\hat{\bar{Y}}_{BT,P}$, it can be argued that $\hat{\bar{Y}}_{GK}$ is also always better than these estimators.

3. Proposed Estimator

In this section, an improved difference-cum-exponential type estimator of the finite population mean $\bar{Y}$ using a single auxiliary variable is proposed. Expressions for the bias and MSE of the proposed estimator are obtained upto first order of approximation.

The conventional difference estimator $(\hat{\bar{Y}}_D)$ of $\bar{Y}$, is given by

$$\hat{\bar{Y}}_D = \bar{y} + w_1(\bar{X} - \bar{x}) \tag{19}$$

where w_1 is a constant.

From (8), (12), and (14), a difference-cum-exponential type estimator $(\hat{\bar{Y}}^*_D)$ of $\bar{Y}$may be given by

$$\hat{\bar{Y}}^*_D = \left[\hat{\bar{Y}}^*_{S,RP} + w_1(\bar{X} - \bar{x})\right] \exp\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right) \tag{20}$$

where $\hat{\bar{Y}}^*_{S,RP} = \frac{\bar{y}}{2}\left[\exp\left(\frac{\bar{X}-\bar{x}}{\bar{x}+\bar{X}}\right) + \exp\left(\frac{\bar{x}-\bar{X}}{\bar{x}+\bar{X}}\right)\right]$ is the average of exponential ratio and exponential product estimators $\hat{\bar{Y}}_{BT,R}$ and $\hat{\bar{Y}}_{BT,P}$ respectively.

Following Searls (1964) and Bahl & Tuteja (1991), Yadav & Kadilar (2013) suggested the following estimator for $\bar{Y}$:

$$\hat{\bar{Y}}_{YK} = w_2\, \bar{y} \exp\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right) \tag{21}$$

where w_2 is a suitably chosen constant.

By combining the ideas in (20) and (21), a modified difference-cum-exponential type estimator of $\bar{Y}$, is given by

$$\hat{\bar{Y}}^*_P = [\hat{\bar{Y}}^*_{S,RP} + w_1(\bar{X} - \bar{x}) + w_2\bar{y}] \exp\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right) \tag{22}$$

where w_1 and w_2 are unknown constants to be determined later.

Rewriting $\hat{\bar{Y}}^*_P$ as

$$\hat{\bar{Y}}^*_P = \left[\frac{\bar{y}}{2}\left\{\exp(\frac{\bar{X} - \bar{x}}{\bar{x} + \bar{X}}) + \exp\left(\frac{\bar{x} - \bar{X}}{\bar{x} + \bar{X}}\right)\right\} + w_1(\bar{X} - \bar{x}) + w_2\bar{y}\right] \exp\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right)$$

Solving $\hat{\bar{Y}}^*_P$ in terms of $\delta_i(i = 0, 1)$, to first order of approximation, we can write

$$\begin{aligned}\hat{\bar{Y}}^*_P - \bar{Y} \approx{}& \bar{Y}w_2 + \bar{Y}\delta_o - \frac{1}{2}\bar{Y}\delta_1 - \bar{X}\delta_1 w_1 + \bar{Y}\delta_o w_2 - \frac{1}{2}\bar{Y}\delta_1 w_2\\ & - \frac{1}{2}\bar{Y}\delta_o\delta_1 + \frac{1}{2}\bar{Y}w_1^2 + \frac{1}{2}\bar{X}\delta_1^2 w_1 - \frac{1}{2}\bar{Y}\delta_o\delta_1 w_2 + \frac{3}{8}\bar{Y}\delta_1^2 w_2\end{aligned} \tag{23}$$

Taking expectation on both sides of (23), we get the bias of $\hat{\bar{Y}}^*_P$, given by

$$Bias(\hat{\bar{Y}}^*_P) \approx \frac{1}{8}[8\bar{Y}w_2 + \lambda C_x^2\left\{4\bar{X}w_1 + \bar{Y}(4 + 3w_2)\right\} - 4\bar{Y}\lambda C_Y C_x(1 + w_2)\rho_{yx}] \tag{24}$$

Squaring both sides of (23) and using first order of approximation, we get

$$\begin{aligned}(\hat{\bar{Y}}_P^* - \bar{Y})^2 \approx\ & \bar{Y}^2 w_2^2 + \bar{Y}^2\delta_o^2 - \bar{Y}^2\delta_o\delta_1 \\ & + \frac{1}{4}\bar{Y}^2\delta_1^2 - 2\bar{X}\bar{Y}\delta_o\delta_1 w_1 + \bar{X}\bar{Y}\delta_1^2 w_1 + \bar{X}^2\delta_1^2 w_1^2 + 2\bar{Y}^2\delta_o^2 w_2 \\ & - 3\bar{Y}^2\delta_o\delta_1 w_2 + \frac{3}{2}\bar{Y}^2\delta_1^2 w_2 - 2\bar{X}\bar{Y}\delta_o\delta_1 w_1 w_2 + 2\bar{X}\bar{Y}\delta_1^2 w_1 w_2 \\ & + \bar{Y}^2\delta_o^2 w_2^2 - 2\bar{Y}^2\delta_o\delta_1 w_2^2 + \bar{Y}^2\delta_1^2 w_2^2\end{aligned} \tag{25}$$

Taking expectation on both sides of (25), the MSE of $\hat{\bar{Y}}_P^*$, to first order of approximation, is given by

$$\begin{aligned}MSE(\hat{\bar{Y}}_P^*) \approx\ & \frac{1}{4}\lambda C_x^2\left\{(\bar{Y} + 2\bar{X}w_1)^2 + 2\bar{Y}(3\bar{Y} + 4\bar{X}w_1)w_2 + 4\bar{Y}^2 w_2^2\right\} \\ & + \bar{Y}^2\left\{w_2^2 + \lambda C_Y^2(1 + w_2)^2\right\} \\ & - \bar{Y}\lambda\rho_{yx}C_yC_x(1 + w_2)(\bar{Y} + 2\bar{X}w_1 + 2\bar{Y}w_2)\end{aligned} \tag{26}$$

Partially differentiating (26) with respect to w_1 and w_2, we get

$$\frac{\partial MSE(\hat{\bar{Y}}_P^*)}{\partial w_1} = \bar{X}\lambda C_x\left\{-2\bar{Y}\rho_{yx}C_y(1 + w_2) + C_x(\bar{Y} + 2\bar{X}w_1 + 2\bar{Y}w_2)\right\}$$

$$\begin{aligned}\frac{\partial MSE(\hat{\bar{Y}}_P^*)}{\partial w_2} =\ & \frac{1}{2}\bar{Y}[4\bar{Y}\left\{w_2 + \lambda C_y^2(1 + w_2)\right\} - 2\lambda\rho_{yx}C_yC_x\left\{2\bar{X}w_1 + \bar{Y}(3 + 4w_2)\right\} \\ & + \lambda C_x^2\left\{4\bar{X}w_1 + \bar{Y}(3 + 4w_2)\right\}]\end{aligned}$$

Setting $\frac{\partial MSE(\hat{\bar{Y}}_P^*)}{\partial w_2}$=0 for $i = 0, 1$, the optimum values of w_1 and w_2 are given by

$$w_{1(opt)} = \frac{\bar{Y}[-4\rho_{yx}C_y + C_x\left\{2 - \lambda C_x^2 + \lambda\rho_{yx}C_yC_x + 2\lambda(-1 + \rho_{yx}^2)C_y^2\right\}]}{4\bar{X}C_x\left\{-1 + \lambda(-1 + \rho_{yx}^2)C_y^2\right\}}$$

and $w_{2(opt)} = \frac{\lambda(C_x^2 - 4(-1+\rho_{yx}^2)C_y^2)}{4\left\{-1+\lambda(-1+\rho_{yx}^2)C_y^2\right\}}$, respectively.

Substituting the optimum values of w_1 and w_2 in (26), we can obtain the minimum MSE of $\hat{\bar{Y}}_P^*$, as given by

$$MSE_{\min}(\hat{\bar{Y}}_P^*) \approx \frac{\lambda\bar{Y}^2\left\{\lambda C_x^4 - 8(-1 + \rho_{yx}^2)(-2 + \lambda C_x^2)C_y^2\right\}}{16\left\{-1 + \lambda(-1 + \rho_{yx}^2)C_y^2\right\}} \tag{27}$$

After some simplifications, (27) can be written as

$$MSE_{\min}(\hat{\bar{Y}}_P^*) \approx MSE(\hat{\bar{Y}}_{Reg}) - (T_1 + T_2) \tag{28}$$

where $T_1 = \frac{\lambda^2\bar{Y}^2\left\{C_x^2+8(1-\rho_{yx}^2)C_y^2\right\}^2}{64\left\{1+\lambda(1-\rho_{yx}^2)C_y^2\right\}}$ and $T_2 = \frac{\lambda^2\bar{Y}^2C_x^2\left\{3C_x^2+16(1-\rho_{yx}^2)C_y^2\right\}}{64\left\{1+\lambda(1-\rho_{yx}^2)C_y^2\right\}}$

Note that both quantities, T_1 and T_2, are always positive.

4. Efficiency Comparisons

In this section, we compare the proposed estimator with the existing estimators considered in Section 2 and derive the following observations:

Observation (i): By (1) and (28)

$$Var(\bar{y}) - MSE_{\min}(\hat{\bar{Y}}_P^*) = \lambda \bar{Y}^2 \rho_{yx}^2 C_y^2 + T_1 + T_2 > 0$$

Observation (ii): By (3) and (28)

$$MSE(\hat{\bar{Y}}_R) - MSE_{\min}(\hat{\bar{Y}}_P^*) = \lambda \bar{Y}^2 (C_x - \rho_{yx} C_y)^2 + T_1 + T_2 > 0$$

Observation (iii): By (5), and (28)

$$MSE(\hat{\bar{Y}}_P) - MSE_{\min}(\hat{\bar{Y}}_P^*) = \lambda \bar{Y}^2 (C_x + \rho_{yx} C_y)^2 + T_1 + T_2 > 0$$

Observation (iV): By (7), (13) and (28)

$$MSE(\hat{\bar{Y}}_{Reg}) - MSE_{\min}(\hat{\bar{Y}}_P^*) = MSE(\hat{\bar{Y}}_{S,RP}) - MSE_{\min}(\hat{\bar{Y}}_P^*) = T_1 + T_2 > 0$$

Observation (V): By (10) and (28)

$$MSE(\hat{\bar{Y}}_{BT,R}) - MSE_{\min}(\hat{\bar{Y}}_P^*) = \frac{1}{4} \lambda \bar{Y}^2 (C_x - 2\rho_{yx} C_y)^2 + T_1 + T_2 > 0$$

Observation (Vi): By (11) and (28)

$$MSE(\hat{\bar{Y}}_{BT,P}) - MSE_{\min}(\hat{\bar{Y}}_P^*) = \frac{1}{4} \lambda \bar{Y}^2 (C_x + 2\rho_{yx} C_y)^2 + T_1 + T_2 > 0$$

Observation (Vii): By (15) and (28)

$$MSE(\hat{\bar{Y}}_{R,Reg}) - MSE_{\min}(\hat{\bar{Y}}_P^*) = \frac{\lambda^2 \bar{Y}^2 C_x^2 \left\{ C_x^2 + 16(1 - \rho_{yx}^2) C_y^2 \right\}}{64 \left\{ 1 + \lambda (1 - \rho_{yx}^2) C_y^2 \right\}} + T_2 > 0$$

Observation (Viii): By (18) and (28)

$$MSE(\hat{\bar{Y}}_{GK}) - MSE_{\min}(\hat{\bar{Y}}_P^*) = T_2 > 0$$

In the light of the eight observations made above, we can argue that the proposed estimator performs better than all of the estimators considered here.

5. Empirical Study

In this section, we consider 10 real data sets to numerically evaluate the performances of the proposed and the existing estimators considered here.

Population 1: [Source: Cochran (1977), pp. 196] Let y be the peach production in bushels in an orchard and x be the number of peach trees in the orchard

in North Carolina in June 1946. The summary statistics for this data set are:
$N = 256$, $n = 100$, $\bar{Y} = 56.47$, $\bar{X} = 44.45$, $C_y = 1.42$, $C_x = 1.40$, $\rho_{yx} = 0.887$.

Population 2: [Source: Murthy (1977), pp. 228] Let y be the output and x be the number of workers. The summary statistics for this data set are:
$N = 80$, $n = 10$, $\bar{Y} = 51.8264$, $\bar{X} = 2.8513$, $C_y = 0.3542$, $C_x = 0.9484$, $\rho_{yx} = 0.915$.

Population 3: [Source: Das (1988)] Let y be the number of agricultural laborers for 1971 and x be the number of agricultural laborers for 1961. The summary statistics for this data set are:
$N = 278$, $n = 25$, $\bar{Y} = 39.068$, $\bar{X} = 25.111$, $C_y = 1.4451$, $C_x = 1.6198$, $\rho_{yx} = 0.7213$.

Population 4: [Source: Steel, Torrie & Dickey (1960), pp. 282] Let y be the log of lef burn in sacs and x be the chlorine percentage. The summary statistics for this data set are:
$N = 30$, $n = 6$, $\bar{Y} = 0.6860$, $\bar{X} = 0.8077$, $C_y = 0.7001$, $C_x = 0.7493$, $\rho_{yx} = -0.4996$.

Population 5: [Source: Maddala (1977), pp. 282] Let y be the consumption per capita and x be the deflated prices of veal. The summary statistics for this data set are:
$N = 16$, $n = 4$, $\bar{Y} = 7.6375$, $\bar{X} = 75.4343$, $C_y = 0.2278$, $C_x = 0.0986$, $\rho_{yx} = -0.6823$.

Population 6: [Source: Kalidar & Cingi (2007)] Let y be the level of apple production (in 100 tones) and x be the number of apple trees in 104 villages in the East Anatolia Region in 1999. The summary statistics for this data set are:
$N = 104$, $n = 20$, $\bar{Y} = 6.254$, $\bar{X} = 13931.683$, $C_y = 1.866$, $C_x = 1.653$, $\rho_{yx} = 0.865$.

Population 7: [Source: Kalidar & Cingi (2005)] Let y be the apple production amount in 1999 and x be the number of apple trees in 1999 in Black sea region of Turkey. The summary statistics for this data set are:
$N = 204$, $n = 50$, $\bar{Y} = 966$, $\bar{X} = 26441$, $C_y = 2.4739$, $C_x = 1.7171$, $\rho_{yx} = 0.71$.

Population 8: [Source: Cochran (1977)] Let y be the number of 'placebo' children and x be the number of paralytic polio cases in the placebo group. The summary statistics for this data set are:
$N = 34$, $n = 10$, $\bar{Y} = 4.92$, $\bar{X} = 2.59$, $C_y = 1.01232$, $C_x = 1.07201$, $\rho_{yx} = 0.6837$.

Population 9: [Source: Srivnstava, Srivastava & Khare (1989)] Let y be the measurement of weight children and x be the mid-arm circumference of children. The summary statistics for this data set are:
$N = 55$, $n = 30$, $\bar{Y} = 17.08$, $\bar{X} = 16.92$, $C_y = 0.12688$, $C_x = 0.07$, $\rho_{yx} = 0.54$.

Population 10: [Source: Sukhatme & Chand (1977)] Let y be the apple trees of bearing age in 1964 and x be the bushels harvested in 1964. The summary statistics for this data set are:
$N = 200$, $n = 20$, $\bar{Y} = 1031.82$, $\bar{X} = 2934.58$, $C_y = 1.59775$, $C_x = 2.00625$, $\rho_{yx} = 0.93$.

In Table 1, the MSE values and percent relative efficiencies (PREs) of all the estimators considered here are reported based on Populations 1-10.

We observe from Table 1 that:

1. The ratio estimator ($\hat{\bar{Y}}_R$) performs better than $\bar{y}$ in Populations 1, 3, 6-10 because the condition $\rho_{yx} > \frac{C_x}{2C_y}$ is satisfied. In other Populations 2, 4 and 5, its performance is poor.

2. The product estimator ($\hat{\bar{Y}}_P$) performs better than $\bar{y}$ in Population 5 because the condition $\rho_{yx} < -\frac{C_x}{2C_y}$ is satisfied.

3. The exponential ratio estimator ($\hat{\bar{Y}}_{BT,R}$) performs better than $\bar{y}$ in Populations 1-3, 6-10 because the condition $\rho_{yx} > \frac{C_x}{4C_y}$ is satisfied.

4. The exponential product estimator ($\hat{\bar{Y}}_{BT,P}$) performs better than $\bar{y}$ in Populations 4 and 5 because the condition $\rho_{yx} < -\frac{C_x}{4C_y}$ is satisfied.

5. It is also observed that, regardless of positive or negative correlation between the study and the auxiliary variable, the estimators, $\hat{\bar{Y}}_{Reg}$, $\hat{\bar{Y}}_{R,Reg}$, $\hat{\bar{Y}}_{GK}$ and $\hat{\bar{Y}}^*_P$, always perform better than the unbiased sample mean, ratio and product estimators considered here in all populations. Among all competitive estimators, the proposed estimator ($\hat{\bar{Y}}^*_P$) is preferable.

6. Conclusion

In this paper, we have suggested an improved difference-cum-exponential type estimator of the finite population mean in simple random sampling using information on a single auxiliary variable. Expressions for the bias and MSE of the proposed estimator are obtained under first order of approximation. Based on both the theoretical and numerical comparisons, we showed that the proposed estimator always performs better than the sample mean estimator, traditional ratio and product estimators, linear regression estimator, Bahl & Tuteja (1991) estimators, Rao (1991) estimator, and Grover & Kaur (2011) estimator. Hence, we recommend the use of the proposed estimator for a more efficient estimation of the finite population mean in simple random sampling.

Acknowledgments

The authors are thankful to the Editor-in-Chief and two anonymous referees for their valuable comments and suggestions that led to an improved version of the article.

TABLE 1: MSE values and PREs of different estimators with respect to $\bar{y}$.

		Estimators								
Population		$\bar{y}$	$\hat{\bar{Y}}_R$	$\hat{\bar{Y}}_P$	$\hat{\bar{Y}}_{Reg}, \hat{\bar{Y}}_{S,RP}$	$\hat{\bar{Y}}_{BT,R}$	$\hat{\bar{Y}}_{BT,P}$	$\hat{\bar{Y}}_{R,Reg}$	$\hat{\bar{Y}}_{GK}$	$\hat{\bar{Y}}^*_P$
1	MSE	39.1829	8.7384	145.8014	8.355	14.4389	82.9704	8.3332	8.3012	8.2551
	PRE	100	448.3998	26.8742	468.975	271.3702	47.2252	470.2037	472.0147	474.6535
2	MSE	29.4854	96.4018	385.3576	4.7995	10.0951	154.5729	4.7909	4.4372	3.5644
	PRE	100	30.586	7.6514	614.345	292.0779	19.0754	615.4427	664.5098	827.2156
3	MSE	116.031	74.1901	449.4337	55.6631	58.6653	246.2871	53.7046	52.2123	50.3002
	PRE	100	156.3967	25.8171	208.4522	197.7846	47.1121	216.0543	222.2292	230.6769
4	MSE	0.0308	0.0989	0.0331	0.0231	0.056	0.0231	0.022	0.0216	0.021
	PRE	100	31.1051	92.9307	133.2623	54.9124	133.0384	139.7975	142.7234	146.3193
5	MSE	0.5676	1.0091	0.3387	0.3033	0.7618	0.4265	0.3018	0.3016	0.3015
	PRE	100	56.2431	167.5887	187.1024	74.5067	133.0649	188.0754	188.163	188.2545
6	MSE	5.4999	1.3871	18.2446	1.3847	2.3645	10.7933	1.3374	1.2933	1.2349
	PRE	100	396.4953	30.1454	397.18	232.6006	50.9568	411.2418	425.2586	445.3895
7	MSE	86226.1674	42781.393	212750.8436	42759.5564	54118.7925	139103.5178	40886.0545	40403.4109	39865.5124
	PRE	100	201.5506	40.5292	201.6536	159.3276	61.9871	210.8938	213.4131	216.2926
8	MSE	1.7511	1.1791	6.2503	0.9325	0.9742	3.5097	0.8979	0.8773	0.8519
	PRE	100	148.5052	28.0157	187.7743	179.747	49.8911	195.0081	199.5885	205.5391
9	MSE	0.0712	0.0504	0.1352	0.0504	0.0554	0.0978	0.0504	0.0504	0.0504
	PRE	100	141.1359	52.6256	141.1632	128.5059	72.7795	141.1876	141.1903	141.1931
10	MSE	122303.2646	29494.9337	600785.7109	16523.171	27689.8347	313335.2233	16270.6541	14996.4826	12647.4936
	PRE	100	414.6585	20.3572	740.1925	441.6901	39.0327	751.6801	815.5463	967.0158

[Recibido: octubre de 2013 — Aceptado: marzo de 2014]

References

Bahl, S. & Tuteja, R. (1991), 'Ratio and product type exponential estimators', *Journal of Information and Optimization Sciences* **12**(1), 159–164.

Cochran, W. G. (1940), 'The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce', *The Journal of Agricultural Science* **30**(02), 262–275.

Cochran, W. G. (1977), *Sampling Techniques*, 3 edn, John Wiley and Sons, New York.

Das, A. (1988), Contribution to the Theory of Sampling Strategies Based on Auxiliary Information, Ph.D. thesis, Bidhan Chandra Krishi Viswavidyalay, Nadia West Bengal, India.

Grover, L. K. & Kaur, P. (2011), 'An improved estimator of the finite population mean in simple random sampling', *Model Assisted Statistics and Applications* **6**(1), 47–55.

Grover, L. K. & Kaur, P. (2014), 'A generalized class of ratio type exponential estimators of population mean under linear transformation of auxiliary variable', *Communications in Statistics-Simulation and Computation* **43**(7), 1552–1574.

Haq, A. & Shabbir, J. (2013), 'Improved family of ratio estimators in simple and stratified random sampling', *Communications in Statistics-Theory and Methods* **42**(5), 782–799.

Kalidar, C. & Cingi, H. (2005), 'A new estimator using two auxiliary variables', *Applied Mathematics and Computation* **162**(2), 901–908.

Kalidar, C. & Cingi, H. (2007), 'Improvement in variance estimation in simple random sampling', *Communication in Statistics-Theory and Methods* **36**, 2075–2081.

Maddala, G. S. (1977), *Econometrics*, Economics handbook series, McGraw Hills Publication Company, New York.

Murthy, M. (1964), 'Product method of estimation', *Sankhya A* **26**(1), 69–74.

Murthy, M. N. (1977), *Sampling: Theory and Methods*, Statistical Pub. Society.

Rao, T. (1991), 'On certain methods of improving ratio and regression estimators', *Communications in Statistics-Theory and Methods* **20**(10), 3325–3340.

Searls, D. T. (1964), 'The utilization of a known coefficient of variation in the estimation procedure', *Journal of the American Statistical Association* **59**(308), 1225–1226.

Singh, H. P., Sharma, B. & Tailor, R. (2014), 'Hartley-Ross type estimators for population mean using known parameters of auxiliary variate', *Communications in Statistics-Theory and Methods* **43**(3), 547–565.

Singh, R., Chauhan, P. & Sawan, N. (2008), 'On linear combination of ratio and product type exponential estimator for estimating the finite population mean', *Statistics in Transition* **9**(1), 105–115.

Singh, R., Chauhan, P., Sawan, N. & Smarandache, F. (2009), 'Improvement in estimating the population mean using exponential estimator in simple random sampling', *Bulletin of Statistics and Economics* **3**(13), 13–18.

Srivnstava, R. S., Srivastava, S. & Khare, B. (1989), 'Chain ratio type estimator for ratio of two population means using auxiliary characters', *Communications in Statistics-Theory and Methods* **18**(10), 3917–3926.

Steel, R., Torrie, J. & Dickey, D. (1960), *Principles and Procedures of Statistics*, McGraw-Hill Companies, Michigan.

Sukhatme, B. & Chand, L. (1977), Multivariate ratio-type estimators, *in* 'Proceedings of the Social Statistics Section', American Statistical Association, Michigan, pp. 927–931.

Yadav, S. K. & Kadilar, C. (2013), 'Improved exponential type ratio estimator of population variance', *Revista Colombiana de Estadística* **36**(1), 145–152.

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 213 a 224

Generalized Exponential Type Estimator for Population Variance in Survey Sampling

Estimadores tipo exponencial generalizado para la varianza poblacional en muestreo de encuestas

Amber Asghar[1,a], Aamir Sanaullah[2,b], Muhammad Hanif[2,c]

[1]Department of Mathematics & Statistics, Virtual University of Pakistan, Lahore, Pakistan

[2]Department of Statistics, NCBA & E, Lahore, Pakistan

Abstract

In this paper, generalized exponential-type estimator has been proposed for estimating the population variance using mean auxiliary variable in single-phase sampling. Some special cases of the proposed generalized estimator have also been discussed. The expressions for the mean square error and bias of the proposed generalized estimator have been derived. The proposed generalized estimator has been compared theoretically with the usual unbiased estimator, usual ratio and product, exponential-type ratio and product, and generalized exponential-type ratio estimators and the conditions under which the proposed estimators are better than some existing estimators have also been given. An empirical study has also been carried out to demonstrate the efficiencies of the proposed estimators.

***Key words*:** Auxiliary variable, Single-phase sampling, Mean square error, Bias.

Resumen

En este artículo, de tipo exponencial generalizado ha sido propuesto con el fin de estimar la varianza poblacional a través de una variables auxiliar en muestreo en dos fases. Algunos casos especiales del estimador medio y el sesgo del estimador generalizado propuesto son derivados. El estimador es comprado teóricamente con otros disponibles en la literatura y las condiciones bajos los cuales éste es mejor. Un estudio empírico es llevado a cabo para comprar la eficiencia de los estimadores propuestos.

***Palabras clave*:** Información auxiliar, muestras en dos fases, error cuadrático medio, sesgo.

[a]Lecturer. E-mail: zukhruf10@gmail.com
[b]Lecturer. E-mail: chaamirsanaullah@yahoo.com
[c]Associate professor. E-mail: drmianhanif@gmail.com

1. Introduction

In survey sampling, the utilization of auxiliary information is frequently acknowledged to higher the accuracy of the estimation of population characteristics. Laplace (1820) utilized the auxiliary information to estimate the total number of inhabitants in France. Cochran (1940) prescribed the utilization of auxiliary information as a classical ratio estimator. Recently, Dash & Mishra (2011) prescribed the few estimators with the utilization of auxiliary variables. Bahl & Tuteja (1991) proposed the exponential estimator under simple random sampling without replacement for the population mean. Singh & Vishwakarma (2007), Singh, Chauhan, Sawan & Smarandache (2011), Noor-ul Amin & Hanif (2012),Singh & Choudhary (2012), Sanaullah, Khan, Ali & Singh (2012), Solanki & Singh (2013*b*) and Sharma, Verma, Sanaullah & Singh (2013) suggested exponential estimators in single and two-phase sampling for population mean.

Estimating the finite population variance has great significance in various fields such as in matters of health, variations in body temperature, pulse beat and blood pressure are the basic guides to diagnosis where prescribed treatment is designed to control their variation. Therefore, the problem of estimating population variance has been earlier taken up by various authors. Gupta & Shabbir (2008) suggested the variance estimation in simple random sampling by using auxiliary variables. Singh & Solanki (2009, 2010) proposed the estimator for population variance by using auxiliary information in the presences of random non-response. Subramani & Kumarapandiyan (2012) proposed the variance estimation using quartiles and their functions of an auxiliary variable. Solanki & Singh (2013*b*) suggested the improved estimation of population mean using population proportion of an auxiliary character. Singh & Solanki (2013) introduced the new procedure for population variance by using auxiliary variable in simple random sampling. Solanki & Singh (2013*a*) and Singh & Solanki (2013) also developed the improved classes of estimators for population variance. Singh et al. (2011), and Yadav & Kadilar (2013) proposed the exponential estimators for the population variance in single and two-phase sampling using auxiliary variables.

In this paper the motivation is to look up some exponential-type estimators for estimating the population variance using the population mean of an auxiliary variable. Further, it is proposed a generalized form of exponential-type estimators. The remaining part of the study is organized as follows: The Section 2 introduced the notations and some existing estimators of population variance in brief. In Section 3, the proposed estimator has been introduced, Section 4 is about the efficiency comparison of the proposed estimators with some available estimators, section 5 and 6 is about numerical comparison and conclusions respectively.

2. Notations and some Existing Estimators

Let $(x_i, y_i), i = 1, 2, \ldots, n$ be the n pairs of sample observations for the auxiliary and study variables respectively from a finite population of size N under simple random sampling without replacement (SRSWOR). Let S_y^2 and s_y^2 are variances respectively for population and sample of the study variable say y . Let $\bar{X}$ and $\bar{x}$ are means respectively for the population and sample mean of the auxiliary variable say x. To obtain the bias and mean square error under simple random sampling without replacement, let us define

$$\left.\begin{array}{c} e_0 = \dfrac{s_y^2 - S_y^2}{S_y^2}, \quad e_1 = \dfrac{\bar{x} - \bar{X}}{\bar{X}} \\ s_y^2 = S_y^2(1+e_0), \quad \bar{x} = \bar{X}(1+e_1) \end{array}\right\} \tag{1}$$

where, e_i is the sampling error, Further, we may assume that

$$E(e_0) = E(e_1) = 0 \tag{2}$$

When single auxiliary mean information is known, after solving the expectations, the following expression is obtained as

$$\left.\begin{array}{l} E(e_0^2) = \dfrac{\delta_{40}}{n}, \quad E(e_1^2) = \dfrac{C_x^2}{n}, \quad E(e_0 e_1) = \dfrac{\delta_{21} C_x}{n} \\ \text{where} \\ \delta_{pq} = \dfrac{\mu_{pq}}{\mu_{20}^{p/2} \mu_{02}^{q/2}}, \quad \text{and} \quad \mu_{pq} = \dfrac{1}{N} \sum (y_i - \bar{Y})^p (x_i - \bar{X})^q \end{array}\right\} \tag{3}$$

(p, q) be the non-negative integer and μ_{02}, μ_{20} are the second order moments and δ_{pq} is the moment's ratio and $C_x = \dfrac{S_x}{\bar{X}}$ is the coefficient of variation for auxiliary variable X . The unbiased estimator for population variance

$$S_y^2 = \frac{1}{N-1} \sum_i^N (Y_i - \bar{Y})^2$$

is defined as

$$t_0 = s_y^2 \tag{4}$$

and its variance is

$$var(t_0) = \frac{s_y^4}{n} [\delta_{40} - 1] \tag{5}$$

Isaki (1983) proposed a ratio estimator for population variance in single-phase sampling as

$$t_1 = s_y^2 \frac{S_x^2}{s_x^2} \tag{6}$$

The bias and the mean square error (MSE) of the estimator in (6), up to first order-approximation respectively are

$$Bias(t_1) = \frac{S_y^2}{n} [\delta_{04} - \delta_{22}] \tag{7}$$

$$MSE(t_1) \approx \frac{S_y^4}{n} [\delta_{40} + \delta_{04} - 2\delta_{22}] \tag{8}$$

Singh et al. (2011) suggested ratio-type exponential estimator for population variance in single-phase sampling as

$$t_2 = s_y^2 \exp\left[\frac{S_x^2 - s_x^2}{S_x^2 + s_x^2}\right] \tag{9}$$

The bias and *MSE*, up to first order-approximation is

$$Bias(t_2) = \frac{S_y^2}{n}\left[\frac{\delta_{04}}{8} - \frac{\delta_{22}}{2} + \frac{3}{8}\right] \tag{10}$$

$$MSE(t_2) \approx \frac{S_y^4}{n}\left[\delta_{40} + \frac{\delta_{04}}{4} - \delta_{22} - \frac{1}{4}\right] \tag{11}$$

Singh et al. (2011) proposed exponential product type estimator for population variance in single-phase sampling as

$$t_3 = s_y^2 \exp\left[\frac{s_x^2 - S_x^2}{s_x^2 + S_x^2}\right] \tag{12}$$

The bias and *MSE*, up to first order-approximation is

$$Bias(t_3) = \frac{S_y^2}{n}\left[\frac{\delta_{04}}{8} + \frac{\delta_{22}}{2} - \frac{5}{8}\right] \tag{13}$$

$$MSE(t_3) \approx \frac{S_y^4}{n}\left[\delta_{40} + \frac{\delta_{04}}{4} + \delta_{22} - \frac{9}{4}\right] \tag{14}$$

Yadav & Kadilar (2013) proposed the exponential estimators for the population variance in single-phase sampling as

$$t_4 = s_y^2 \exp\left[\frac{S_x^2 - s_x^2}{S_x^2 + (\alpha - 1)s_x^2}\right] \tag{15}$$

The bias and *MSE*, up to first order-approximation is

$$Bias(t_4) = \frac{S_y^2}{n}\left[\frac{\delta_{04} - 1}{2\alpha^2}(2\alpha(1 - \lambda) - 1)\right] \tag{16}$$

$$MSE(t_4) \approx \frac{S_y^4}{n}\left[(\delta_{40} - 1) + \frac{(\delta_{04} - 1)}{\alpha^2}(1 - 2\alpha\lambda)\right] \tag{17}$$

where, $\lambda = \frac{\delta_{22}-1}{\delta_{04}-1}$ and $\alpha = \frac{1}{\lambda}$.

3. Proposed Generalized Exponential Estimator

Following Bahl & Tuteja (1991), new exponential ratio-type and product-type estimators for population variance are as

$$t_5 = s_y^2 \exp\left[\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right] \tag{18}$$

$$t_6 = s_y^2 \exp\left[\frac{\bar{x} - \bar{X}}{\bar{x} + \bar{X}}\right] \tag{19}$$

Equations (18) and (19) lead to the generalized form as

$$t_{EG} = \lambda\ s_y^2\ \exp\left[\alpha\left(1 - \frac{a\bar{x}}{\bar{X} + (a-1)\bar{x}}\right)\right] = \lambda\ s_y^2\ \exp\left[\alpha\left(\frac{\bar{X} - \bar{x}}{\bar{X} + (a-1)\bar{x}}\right)\right] \tag{20}$$

where the three different real constants are $0 < \lambda \leq 1$, and $-\infty < \alpha < \infty$ and $a > 0$. It is observed that for different values of λ, α and a in (20), we may get various exponential ratio-type and product-type estimators as new family of t_{EG} i.e. $G = 0, 1, 2, 3, 4, 5$. From this family, some examples of exponential ratio-type estimators may be given as follows: It is noted that, for $\lambda = 1, \alpha = 0$ and $a = a_0, t_{EG}$ in (20) is reduced to

$$t_{E0} = s_y^2 \exp(0) = s_y^2 \tag{21}$$

which is an unbiased employing no auxiliary information.

For $\lambda = 1, \alpha = 0$ and $a = 0, t_{EG}$ in (20) is reduced to

$$t_{E1} = s_y^2 \exp(1) \tag{22}$$

For $\lambda = 1, \alpha = 1$ and $a = 2, t_{EG}$ in (20) is reduced to

$$t_{E2} = s_y^2 \exp\left[\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right] = t_5 \tag{23}$$

For $\lambda = 1, \alpha = 1$ and $a = 1, t_{EG}$ in (20) is reduced to

$$t_{E3} = s_y^2 \exp\left[\frac{\bar{X} - \bar{x}}{\bar{X}}\right] \tag{24}$$

Some example for exponential product-type estimators may be given as follows:

For $\lambda = 1, \alpha = -1$ and $a = 2, t_{EG}$ in (20) is reduced to

$$t_{E4} = s_y^2 \exp\left[-\left\{\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right\}\right] = t_6 \tag{25}$$

For $\lambda = 1, \alpha = -1$ and $a = 1, t_{EG}$ in (20) is reduced to

$$t_{E5} = s_y^2 \exp\left[-\left\{\frac{\bar{X} - \bar{x}}{\bar{X}}\right\}\right] \tag{26}$$

3.1. The Bias and Mean Square Error of Proposed Estimator

In order to obtain the bias and *MSE*, (20) may be expressed in the form of e's by using (1), (2) and (3) as

$$t_{EG} = \lambda\ S_y^2(1 + e_0)\ \exp\left[\alpha \frac{-e_1}{1 + (a-1)(1 + e_1)}\right] \tag{27}$$

Further, it is assumed that the contribution of terms involving powers in e_0 and e_1 higher than two is negligible

$$t_{EG} \approx \lambda\ S_y^2\ \left[1 + e_0 - \frac{\alpha e_1}{a} + \frac{\alpha^2 e_1^2}{2a^2} - \frac{\alpha e_0 e_1}{a}\right] \tag{28}$$

In order to obtain the bias, subtract S_y^2 both sides and taking expectation of (28), after some simplification, we may get the bias as

$$Bias(t_{EG}) \approx \frac{S_y^2}{n}\ \left[\lambda\left\{1 + \frac{\alpha^2}{2a^2}C_x^2 - \frac{\alpha}{a}\delta_{21}C_x\right\}\right] - S_y^2 \tag{29}$$

Expanding the exponentials and ignoring higher order terms in e_0 and e_1 , we may have on simplification

$$t_{EG} - S_y^2 \approx \lambda\ s_y^2\ \left[\left\{1 + e_0 - \frac{\alpha e_1}{a}\right\} - 1\right] \tag{30}$$

Squaring both sides and taking the expectation we may get the MSE of (t_{EG}) from as (30)

$$MSE(t_{EG}) \approx \frac{S_y^4}{n}\ \left[\lambda^2\left\{1 + (\delta_{40} - 1) - 2\frac{\alpha}{a}\delta_{21}C_x + \frac{\alpha^2}{a^2}C_x^2\right\} + (1 - 2\lambda)\right] \tag{31}$$

or

$$MSE(t_{EG}) \approx \frac{S_y^4}{n}\ \left[\lambda^2\left\{1 + (\delta_{40} - 1) - 2\omega\delta_{21}C_x + \omega^2 C_x^2\right\} + (1 - 2\lambda)\right] \tag{32}$$

where, $\omega = \frac{\alpha}{a}$, The MSE (t_{EG}) is minimized for the optimal values of λ and ω as, $\omega = \delta_{21}(C_x)^{-1}$ and $\lambda = (\delta_{40} - \delta_{21}^2)^{-1}$. The minimum MSE (t_{EG}) is obtained as

$$MSE_{\min}(t_{EG}) \approx \frac{S_y^4}{n}\ \left[1 - \frac{1}{\delta_{40} - \delta_{21}^2}\right] \tag{33}$$

On substituting the optimal values of $\lambda = (\delta_{40} - \delta_{21}^2)^{-1}$, α and a into (20), we may get the asymptotically optimal estimator as

$$t_{asym} = \frac{s_y^2}{\delta_{40} - \delta_{21}^2}\exp\left[\frac{\delta_{21}(\bar{X} - \bar{x})}{\bar{X} + (C_x - 1)\bar{x}}\right] \tag{34}$$

The values of λ, α and a can be obtained in prior from the previous surveys, for case in point, see Murthy (1967), Ahmed, Raman & Hossain (2000), Singh & Vishwakarma (2008), Singh & Karpe (2010) and Yadav & Kadilar (2013).

In some situations, for the practitioner it is not possible to presume the values of λ, α and a by employ all the resources, it is worth sensible to replace λ, α and a in (20) by their consistent estimates as

$$\hat{\omega} = \hat{\delta_{21}}(\hat{C_x})^{-1} \text{ and } \hat{\lambda} = (\hat{\delta_{40}} - \hat{\delta_{21}^2})^{-1} \tag{35}$$

$\hat{\delta_{21}}$, and $\hat{C}$ respectively are the consistent estimates of δ_{21}, and C_x.

As a result, the estimator in (34) may be obtained as

$$\hat{t}_{asym} = \frac{s_y^2}{\hat{\delta_{40}} - \hat{\delta_{21}^2}} \exp\left[\frac{\hat{\delta_{21}}(\bar{X} - \bar{x})}{\bar{X} + (\hat{C}_x - 1)\bar{x}}\right] \tag{36}$$

Similarly the *MSE* (t_{EG}) in (33) may be given as,

$$MSE_{\min}(\hat{t}_{asym}) \approx \frac{s_y^4}{n}\left[1 - \frac{1}{\hat{\delta_{40}} - \hat{\delta_{21}^2}}\right] \tag{37}$$

Thus, the estimator $\hat{t}_{asym}$, given in (36), is to be used in practice. The bias and *MSE* expression for the new family of t_{EG}, can be obtained by putting different values of λ, α and a in (29) and (31) as

$$Bias(t_{E2}) \approx \frac{S_y^2}{n}\left[\frac{1}{8}C_x^2 - \frac{1}{2}\delta_{21}C_x\right] \tag{38}$$

$$Bias(t_{E3}) \approx \frac{S_y^2}{n}\left[\frac{1}{2}C_x^2 - \delta_{21}C_x\right] \tag{39}$$

$$Bias(t_{E4}) \approx \frac{S_y^2}{n}\left[\frac{1}{8}C_x^2 + \frac{1}{2}\delta_{21}C_x\right] \tag{40}$$

$$Bias(t_{E5}) \approx \frac{S_y^2}{n}\left[\frac{1}{2}C_x^2 + \delta_{21}C_x\right] \tag{41}$$

$$MSE(t_{E2}) \approx \frac{S_y^4}{n}\left[(\delta_{40} - 1) - \delta_{21}C_x + \frac{1}{4}C_x^2\right] \tag{42}$$

$$MSE(t_{E3}) \approx \frac{S_y^4}{n}\left[(\delta_{40} - 1) - 2\delta_{21}C_x + C_x^2\right] \tag{43}$$

$$MSE(t_{E4}) \approx \frac{S_y^4}{n}\left[(\delta_{40} - 1) + \delta_{21}C_x + \frac{1}{4}C_x^2\right] \tag{44}$$

$$MSE(t_{E5}) \approx \frac{S_y^4}{n}\left[(\delta_{40} - 1) + 2\delta_{21}C_x + C_x^2\right] \tag{45}$$

4. Efficiency Comparision of Proposed Estimators with some Available Estimators

The efficiency comparisons have been made with the sample variance (t_0), Isaki (1983) ratio estimator (t_1), Singh et al. (2011) ratio (t_2), and product (t_3), estimators and Yadav & Kadilar (2013) ratio (t_4), estimator using (5),(8),(11),(14) and (17) respectively with the proposed generalized estimator and class of proposed estimators.

$MSE\ (t_{EG}) < Var\ (t_0)$

$$\left\langle if\ \frac{\delta_{40}+\frac{1}{f}}{2} > 1 \right\rangle \tag{46}$$

$MSE\ (t_{EG}) < MSE\ (t_1)$

$$\left\langle if\ \delta_{40}+\delta_{04}-2\delta_{22}+\frac{1}{f} > 1 \right\rangle \tag{47}$$

$MSE\ (t_{EG}) < MSE\ (t_2)$

$$\left\langle if\ \delta_{40}+\frac{\delta_{04}}{4}-\delta_{22}-\frac{1}{4}+\frac{1}{f} > 1 \right\rangle \tag{48}$$

$MSE\ (t_{EG}) < MSE\ (t_3)$

$$\left\langle if\ \delta_{40}+\frac{\delta_{04}}{4}+\delta_{22}-\frac{9}{4}+\frac{1}{f} > 1 \right\rangle \tag{49}$$

$MSE\ (t_{EG}) < MSE\ (t_4)$

$$\left\langle if\ \frac{f[(d-\delta_{40})-(\delta_{22}-1)^2]}{(d-f-\delta_{40}\delta_{21}^2)} > 1 \right\rangle \tag{50}$$

$MSE\ (t_{E2}) < Var\ (t_0)$

$$\left\langle if\ \frac{4\ \delta_{21}}{C_x} > 1 \right\rangle \tag{51}$$

$MSE\ (t_{E2}) < MSE\ (t_1)$

$$\left\langle if\ \frac{4(\delta_{40}-2\delta_{22}+\delta_{21}C_x+1)}{C_x^2} > 1 \right\rangle \tag{52}$$

$MSE\ (t_{E2}) < MSE\ (t_2)$

$$\left\langle if\ \frac{(\delta_{40}-4\delta_{22}+4\delta_{21}C_x+3)}{C_x^2} > 1 \right\rangle \tag{53}$$

$MSE\ (t_{E3}) < Var\ (t_0)$

$$\left\langle if\ \frac{2\ \delta_{21}}{C_x} > 1 \right\rangle \tag{54}$$

$MSE\ (t_{E3}) < MSE\ (t_1)$

$$\left\langle if\ \frac{(\delta_{04}-2\delta_{22}+2\delta_{21}C_x+1)}{C_x^2} > 1 \right\rangle \tag{55}$$

$MSE\ (t_{E3}) < MSE\ (t_2)$

$$\left\langle if\ \frac{(\frac{\delta_{04}}{4} - \delta_{22} + 2\delta_{21}C_x + \frac{3}{4})}{C_x^2} > 1 \right\rangle \tag{56}$$

$MSE\ (t_{E4}) < Var\ (t_0)$

$$\left\langle if\ -\ \frac{4\ \delta_{21}}{C_x} > 1 \right\rangle \tag{57}$$

$MSE\ (t_{E4}) < MSE\ (t_3)$

$$\left\langle if\ \frac{(\delta_{04} - 4\delta_{22} - 4\delta_{21}C_x - 1)}{C_x^2} > 1 \right\rangle \tag{58}$$

$MSE\ (t_{E5}) < Var\ (t_0)$

$$\left\langle if\ -\ \frac{2\ \delta_{21}}{C_x} > 1 \right\rangle \tag{59}$$

$MSE\ (t_{E5}) < MSE\ (t_3)$

$$\left\langle if\ \frac{(\frac{\delta_{04}}{4} - \delta_{22} - 2\delta_{21}C_x - \frac{1}{4})}{C_x^2} > 1 \right\rangle \tag{60}$$

where $f = \delta_{40} - \delta_{21}^2$ and $d = \delta_{40}\delta_{04} - \delta_{04} + 1$.

When the above conditions are satisfied the proposed estimators are more efficient than t_0, t_1, t_2, t_3 and t_4.

5. Numerical Comparison

In order to examine the performance of the proposed estimator, we have taken two real populations. The Source, description and parameters for two populations are given in Table 1 and Table 2

Table 1: Source and Description of Population 1 & 2.

Population	Source	Y	X
1	Murthy (1967, pg. 226)	output	number of workers
2	Gujarati (2004, pg. 433)	average (miles per gallon)	top speed(miles per hour)

The comparison of the proposed estimator has been made with the unbiased estimator of population variance, the usual ratio estimator due to Isaki (1983), Singh et al. (2011) exponential ratio and product estimators and Yadav & Kadilar (2013) generalized exponential-type estimator. Table 3 shows the results of Percentage Relative Efficiency (PRE) for Ratio and Product type estimators. These estimators are compared with respect to sample variance.

Table 2: Parameters of Populations.

Parameter	1	2
N	25	81
n	25	21
$\bar{Y}$	33.8465	2137.086
$\bar{X}$	283.875	112.4568
C_y	0.3520	0.1248
C_x	0.7460	0.4831
ρ_{yx}	0.9136	-0.691135
δ_{40}	2.2667	3.59
δ_{21}	0.5475	0.05137
δ_{04}	3.65	6.820
δ_{22}	2.3377	2.110

where ρ_{yx} is the correlation between the study and auxiliary variable.

Table 3: Percent Relative Efficiencies (PREs) for Ratio and Product type estimators with respect to sample variance (t_0).

Estimator	Population 1	Population 2
$t_0 = s_y^2$	100	100
t_1	102.05	*
t_2	214.15	*
t_3	*	86.349
t_4	214.440	108.915
t_{E2}	127.04	*
t_{E3}	125.898	*
t_{E4}	*	96.895
t_{E5}	*	90.145
t_{EG}	**257.371**	**359.123**

'*' shows the data is not applicable

6. Conclusions

Table 3 shows that the proposed generalized exponential-type estimator (t_{EG}) is more efficient than the usual unbiased estimator (t_0), Isaki (1983) ratio estimator, Singh et al. (2011) exponential ratio and product estimators and Yadav & Kadilar (2013) generalized exponential-type estimator. Further, it is observed that the class of exponential-type ratio estimators t_{E2}, and t_{E3}, are more efficient than the usual unbiased estimator and Isaki (1983) ratio estimator. Furthermore, it is observed that the class of exponential-type product estimators t_{E4} and t_{E5}, are more efficient than Singh et al. (2011) exponential product estimator.

Acknowledgment

The authors are indebted to two anonymous referees and the Editor for their productive comments and suggestions, which led to improve the presentation of this manuscript.

[Recibido: noviembre de 2013 — Aceptado: abril de 2014]

References

Ahmed, M. S., Raman, M. S. & Hossain, M. I. (2000), 'Some competitive estimators of finite population variance multivariate auxiliary information', *Information and Management Sciences* **11**(1), 49–54.

Bahl, S. & Tuteja, R. K. (1991), 'Ratio and product type exponential estimator', *Information and Optimization Sciences* **12**, 159–163.

Cochran, W. G. (1940), 'The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce', *The Journal of Agricultural Science* **30**, 262–275.

Dash, P. R. & Mishra, G. (2011), 'An improved class of estimators in two-phase sampling using two auxiliary variables', *Communications in Statistics-Theory and Methods* **40**, 4347–4352.

Gujarati, D. (2004), *Basic Econometrics*, 4 edn, The McGraw-Hill Companies.

Gupta, S. & Shabbir, J. (2008), 'Variance estimation in simple random sampling using auxiliary information', *Hacettepe Journal of Mathematics and Statistics* **37**, 57–67.

Isaki, C. (1983), 'Variance estimation using auxiliary information', *Journal of the American Statistical Association* **78**, 117–123.

Laplace, P. S. (1820), *A Philosophical Essay on Probabilities*, English Translation, Dover.

Murthy, M. (1967), *Sampling Theory and Methods*, Calcutta Statistical Publishing Society, Kolkatta, India.

Noor-ul Amin, M. & Hanif, M. (2012), 'Some exponential estimators in survey sampling', *Pakistan Journal of Statistics* **28**(3), 367–374.

Sanaullah, A., Khan, H., Ali, A. & Singh, R. (2012), 'Improved ratio-type estimators in survey sampling', *Journal of Reliability and Statistical Studies* **5**(2), 119–132.

Sharma, P., Verma, H. K., Sanaullah, A. & Singh, R. (2013), 'Some exponential ratio- product type estimators using information on auxiliary attributes under second order approximation', *International Journal of Statistics and Economics* **12**(3), 58–66.

Singh, B. K. & Choudhary, S. (2012), 'Exponential chain ratio and product type estimators for finite population mean under double sampling scheme', *Journal of Science Frontier Research in Mathematics and Design Sciences* **12**(6), 0975–5896.

Singh, H. P. & Karpe, N. (2010), 'Estimation of mean, ratio and product using auxiliary information in the presence of measurement errors in sample surveys', *Journal of Statistical Theory and Practice* **4**(1), 111–136.

Singh, H. P. & Solanki, R. S. (2009), 'Estimation of finite population variance using auxiliary information in presence of random non-response', *Gujarat Statistical Review* **1**, 37–637.

Singh, H. P. & Solanki, R. S. (2010), 'Estimation of finite population variance using auxiliary information in presence of random non-response', *Gujarat Statistical Review* **2**, 46–58.

Singh, H. P. & Solanki, R. S. (2013), 'A new procedure for variance estimation in simple random sampling using auxiliary information', *Statistical Papers* **54**(2), 479–497.

Singh, H. P. & Vishwakarma, G. (2008), 'Some families of estimators of variance of stratified random sample mean using auxiliary information', *Journal of Statistical Theory and Practice* **2**(1), 21–43.

Singh, H. P. & Vishwakarma, K. (2007), 'Modified exponential ratio and product estimators for finite population mean in double sampling', *Australian Journal of Statistics* **36**, 217–225.

Singh, R. S., Chauhan, P., Sawan, N. & Smarandache, F. (2011), 'Improved exponential estimator for population variance using two auxiliary variables', *Italian Journal of Pure and Applied Mathematics* **28**, 101–108.

Solanki, R. S. & Singh, H. P. (2013*a*), 'An improved class of estimators for the population variance', *Model Assisted Statistics and Applications* **8**(3), 229–238.

Solanki, R. S. & Singh, H. P. (2013*b*), 'Improved estimation of population mean using population proportion of an auxiliary character', *Chilean Journal of Statistics* **4**(1), 3–17.

Subramani, J. & Kumarapandiyan, G. (2012), 'Variance estimation using quartiles and their functions of an auxiliary variable', *International Journal of Statistics and Applications* **2**(5), 67–72.

Yadav, S. K. & Kadilar, C. (2013), 'Improved exponential type ratio estimator of population variance', *Revista Colombiana de Estadística* **36**(1), 145–152.

Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 225 a 245

The Poisson-Lomax Distribution

Distribución Poisson-Lomax

Bander Al-Zahrani[a], Hanaa Sagor[b]

Department of Statistics, King Abdulaziz University, Jeddah, Saudi Arabia

Abstract

In this paper we propose a new three-parameter lifetime distribution with upside-down bathtub shaped failure rate. The distribution is a compound distribution of the zero-truncated Poisson and the Lomax distributions (PLD). The density function, shape of the hazard rate function, a general expansion for moments, the density of the rth order statistic, and the mean and median deviations of the PLD are derived and studied in detail. The maximum likelihood estimators of the unknown parameters are obtained. The asymptotic confidence intervals for the parameters are also obtained based on asymptotic variance-covariance matrix. Finally, a real data set is analyzed to show the potential of the new proposed distribution.

***Key words*:** Asymptotic variance-covariance matrix, Compounding, Lifetime distributions, Lomax distribution, Poisson distribution, Maximum likelihood estimation.

Resumen

En este artículo se propone una nueva distribución de sobrevida de tres parámetros con tasa fallo en forma de bañera. La distribución es una mezcla de la Poisson truncada y la distribución Lomax. La función de densidad, la función de riesgo, una expansión general de los momentos, la densidad del r-ésimo estadístico de orden, y la media así como su desviación estándar son derivadas y estudiadas en detalle. Los estimadores de máximo verosímiles de los parámetros desconocidos son obtenidos. Los intervalos de confianza asintóticas se obtienen según la matriz de varianzas y covarianzas asintótica. Finalmente, un conjunto de datos reales es analizado para construir el potencial de la nueva distribución propuesta.

***Palabras clave*:** mezclas, distribuciones de sobrevida, distribució n Lomax, distribución Poisson, estomación máximo-verosímil.

[a] Professor. E-mail: bmalzahrani@kau.edu.sa

[b] Ph.D student. E-mail: hsagor123@gmail.com

1. Introduction

Marshall & Olkin (1997) introduced an effective technique to add a new parameter to a family of distributions. A great deal of papers have appeared in the literature used this technique to propose new distributions. In their paper, Marshall & Olkin (1997) generalized the exponential and Weibull distributions. Alice & Jose (2003) followed the same approach and introduced Marshall-Olkin extended semi-Pareto model and studied its geometric extreme stability. Ghitany, Al-Hussaini & Al-Jarallah (2005) studied the Marshall-Olkin Weibull distribution and established its properties in the presence of censored data. Marshall-Olkin extended Lomax distribution was introduced by Ghitany, Al-Awadhi & Alkhalfan (2007). Compounding Poisson and exponential distributions have been considered by many authors; e.g. Kus (2007) proposed the Poisson-exponential lifetime distribution with a decreasing failure rate function. Al-Awadhi & Ghitany (2001) used the Lomax distribution as a mixing distribution for the Poisson parameter and obtained the discrete Poisson-Lomax distribution. Cancho, Louzada-Neto & Barriga (2011) introduced another modification of the Poisson-exponential distribution.

Let $Y_1, Y_2, \dots, Y_Z$ be independent and identically distributed random variables each has a density function f, and let Z be a discrete random variable having a zero-truncated Poisson distribution with probability mass function

$$P_Z(z) \equiv P_Z(z, \lambda) = \frac{e^{-\lambda}\lambda^z}{z!(1-e^{-\lambda})}, \quad z \in \{1, 2, \dots\},\ \lambda > 0. \tag{1}$$

Suppose that X is a random variable representing the lifetime of a parallel-system of Z components, i.e. $X = \max\{Y_1, Y_2, \dots, Y_z\}$, and Y's and Z are independent. The conditional distribution function of $X|Z$ has the probability density function (pdf)

$$f_{X|Z}(x|z) = zf(x)[F(x)]^{z-1}. \tag{2}$$

where $F(x)$ is the cumulative distribution function (cdf) corresponding to $f(x)$.

A compound probability function (pdf) of $f_{X|Z}(x|z)$ and $P_Z(z)$, where X is a continuous random variable (r.v.) and Z a discrete r.v. is defined by

$$g_X(x) = \sum_{z=1}^{\infty} f_{X|Z}(x|z) P_Z(z). \tag{3}$$

Substitution of (1) and (2) in (3) then yields

$$\begin{aligned} g_X(x) &= \sum_{z=1}^{\infty} zf(x)[F(x)]^{z-1} \left(\frac{\lambda^z e^{-\lambda}}{z!(1-e^{-\lambda})}\right) \\ &= \frac{\lambda f(x) e^{-\lambda(1-F(x))}}{(1-e^{-\lambda})}, \quad x > 0,\ \lambda > 0. \end{aligned}$$

The reliability and the hazard rate functions of X are, respectively, given by

$$\bar{G}(x,\lambda) = \frac{1-e^{-\lambda\bar{F}(x)}}{(1-e^{-\lambda})}, \quad x>0, \tag{4}$$

$$h_G(x,\lambda) = \frac{\lambda f(x)e^{-\lambda\bar{F}(x)}}{1-e^{-\lambda\bar{F}(x)}} = \frac{\lambda f(x)}{e^{\lambda\bar{F}(x)}-1}. \tag{5}$$

In this paper we propose a new lifetime distribution by compounding Poisson and Lomax distributions. As we have mentioned in the previous chapters, the Lomax distribution with two parameters is a special case of the generalized Pareto distribution, and ti is also known as the Pareto of the second type. A random variable X is said to have the Lomax distribution, abbreviated as $X \sim \mathrm{LD}(\alpha,\beta)$, if it has the pdf

$$f_{LD}(x;\alpha,\beta) = \alpha\beta\left(1+\beta x\right)^{-(\alpha+1)}, \quad x>0,\ \alpha,\beta>0. \tag{6}$$

Here α and β are the shape and scale parameters, respectively. Analogous tu above, the survival and hazard functions associated with (6) are given by

$$\bar{F}_{LD}(x;\alpha,\beta) = \left(1+\beta x\right)^{-\alpha}, \quad x>0, \tag{7}$$

$$h_{LD}(x;\alpha,\beta) = \frac{\alpha\beta}{1+\beta x}, \quad x>0. \tag{8}$$

The rest of the paper is organized as follows. In Section 2, we give explicit forms and interpretation for the distribution function and the probability density function. In Section 3, we discuss the distributional properties of the proposed distribution. Section 4 discusses the estimation problem using the maximum likelihood estimation method. In Section 5, an illustrative example, model selections, goodness-of-fit tests for the distribution with estimated parameters are all presented. Finally, we conclude in Section 6.

2. Model Formulation

Substitution of (7) in (4) yields the following reliability function:

$$\bar{G}(x;\alpha,\beta,\lambda) = \frac{1-e^{-\lambda(1+\beta x)^{-\alpha}}}{(1-e^{-\lambda})}, \quad x>0,\alpha,\beta,\lambda>0. \tag{9}$$

The pdf associated with (9) is expressed in a closed form and is given by

$$g(x;\alpha,\beta,\lambda) = \frac{\alpha\beta\lambda\left(1+\beta x\right)^{-(\alpha+1)}e^{-\lambda(1+\beta x)^{-\alpha}}}{(1-e^{-\lambda})}, \quad x>0,\alpha,\beta,\lambda>0. \tag{10}$$

The density function given by (10) can be interpreted as a compound of the zero-truncated Poisson distribution and the Lomax distribution. Suppose that $X = \max\{Y_1, Y_2, \ldots, Y_z\}$, and each Y is distributed according to the Lomax distribtion.

The variable Z has zero-truncated Poisson distribution and the variables Y's and Z are independent. Then the conditional distribution function of $X|Z$ has the pdf

$$f_{X|Z}(x|z;\alpha,\beta)=z\alpha\beta(1+\beta x)^{-(\alpha+1)}[1-(1+\beta x)^{-\alpha}]^{z-1}. \tag{11}$$

The joint distribution of the random variables X and Z, denoted by $f_{X,Z}(x,z)$, is given by

$$f_{X,Z}(x,z)=\frac{z}{z!(1-e^{-\lambda})}\ \alpha\beta(1+\beta x)^{-(\alpha+1)}[1-(1+\beta x)^{-\alpha}]^{z-1}e^{-\lambda}\lambda^{z}, \tag{12}$$

the marginal pdf of X is as follows.

$$\begin{aligned} f_X(x;\alpha,\beta,\lambda) &= \frac{\alpha\beta\lambda e^{-\lambda}(1+\beta x)^{-(\alpha+1)}}{(1-e^{-\lambda})}\sum_{z=1}^{\infty}\frac{[(1-(1+\beta x)^{-\alpha})\lambda]^{z-1}}{(z-1)!} \\ &= \frac{\alpha\beta\lambda e^{-\lambda}(1+\beta x)^{-(\alpha+1)}e^{\lambda(1-(1+\beta x)^{-\alpha})}}{(1-e^{-\lambda})} \\ &= \frac{\alpha\beta\lambda(1+\beta x)^{-(\alpha+1)}e^{-\lambda(1+\beta x)^{-\alpha}}}{(1-e^{-\lambda})}, \end{aligned}$$

which is the distribution with the pdf given by (10). The distribution of X may be referred to as the Poisson-Lomax distribution. Symbolically it is abbreviated by $X\sim PLD(\alpha,\beta,\lambda)$ to indicate that the random variable X has the Poisson-Lomax distribution with parameters α, β and λ.

3. Distributional Properties

In this section, we study the distributional properties of the PLD. In particular, if $X\sim PLD(\alpha,\beta,\lambda)$ then the shapes of the density function, the shapes of the hazard function, moments, the density of the rth order statistics, and the mean and median deviations of the PLD are derived and studied in detail.

3.1. Shapes of pdf

The limit of the Poisson-Lomax density as $x\to\infty$ is 0 and the limit as $x\to 0$ is $\alpha\beta\lambda/(e^{\lambda}-1)$. The following theorem gives simple conditions under which the pdf is decreasing or unimodal.

Theorem 1. *The pdf, $g(x)$, of $X\sim PLD(\alpha,\beta,\lambda)$ is decreasing (unimodal) if the function $\xi(x)\geq 0$ (<0) where $\xi(x)=\alpha(1-\lambda(1+\beta x)^{-\alpha})+1$, independent of β.*

***Proof*.** The first derivative of $g(x)$ is given by

$$g'(x)=-\frac{\alpha\beta^2\lambda}{1-e^{-\lambda}}\ (1+\beta x)^{-(\alpha+2)}\ e^{-\lambda(1+\beta x)^{-\alpha}}\xi((1+\beta x)^{-\alpha}),$$

where $\xi(y)=\alpha(1-\lambda y)+1$, and $y=(1+\beta x)^{-\alpha}<1$. Then we have the following:

(i) If $\xi(1) = \alpha(1-\lambda)+1 > 0$, then $\xi(y) > 0$ for all $y < 1$, and hence, $g'(x) \leq 0$ for all $x > 0$, i.e. the function $g(x)$ is decreasing.

(ii) If $\xi(1) < 0$, then $\xi(y)$ has a unique zero at $y_\xi = \frac{\alpha+1}{\alpha\lambda} < 1$. Since $y = (1+\beta x)^{-\alpha}$ is one to one transformation, it follows that $g(x)$ has also a unique critical point at $x_g = \frac{1}{\beta}(y_\xi^{-1/\alpha} - 1)$.

Finally, since $g(0) = \alpha\beta\lambda/(e^\lambda - 1)$ and $g(\infty) = 0$ then x_g must be a point of absolute maximum for $g(x)$. ■

***Note* 1.** It should be noted that:

(i) When $\lambda \in (0,1]$, $g(x)$ is decreasing in $x > 0$ for all values of $\alpha, \beta > 0$.

(ii) When $\lambda > 1$, $g(x)$ may still exhibit a decreasing behavior, depending on the values of α, λ such that $\alpha(1-\lambda)+1 > 0$.

(iii) The mode of the Poisson-Lomax distribution is given by

$$Mode(x) = \begin{cases} 0, & \text{if} \quad \alpha(1-\lambda)+1 \geq 0, \\ \frac{1}{\beta}\left[\left(\frac{\alpha\lambda}{\alpha+1}\right)^{1/\alpha} - 1\right] & \text{otherwsie.} \end{cases} \tag{13}$$

Figure 1 shows the pdf curves for the $PLD(\alpha,\beta,\lambda)$ for selected values of the parameters α, β and λ. From the curves, it is quite evident that the PLD is positively skewed distribution. It becomes highly positively skewed for large values of the involved parameters.

3.2. Hazard Rate Function

The hazard rate function (hrf) of a random variable X is defined by $h(x) = f(x)/\bar{F}(x)$, where $\bar{F} = 1 - F$. The hazard function of $X \sim PLD(\alpha,\beta,\lambda)$ is given by

$$h(x) = \frac{\alpha\beta\lambda(1+\beta x)^{-(\alpha+1)}}{e^{\lambda(1+\beta x)^{-\alpha}} - 1}, \quad x > 0. \tag{14}$$

The following theorem gives simple conditions under which the hrf, given in (14), is decreasing or unimodal.

Theorem 2. *The hrf, $h(x)$, of $X \sim PLD(\alpha,\beta,\lambda)$ is decreasing (unimodal) if $\eta(x) \geq 0 (< 0)$ where $\eta(x) = -(\alpha+1) + (\alpha+1-\alpha\lambda(1+\beta x)^{-\alpha})\ e^{\lambda(1+\beta x)^{-\alpha}}$, independent of β.*

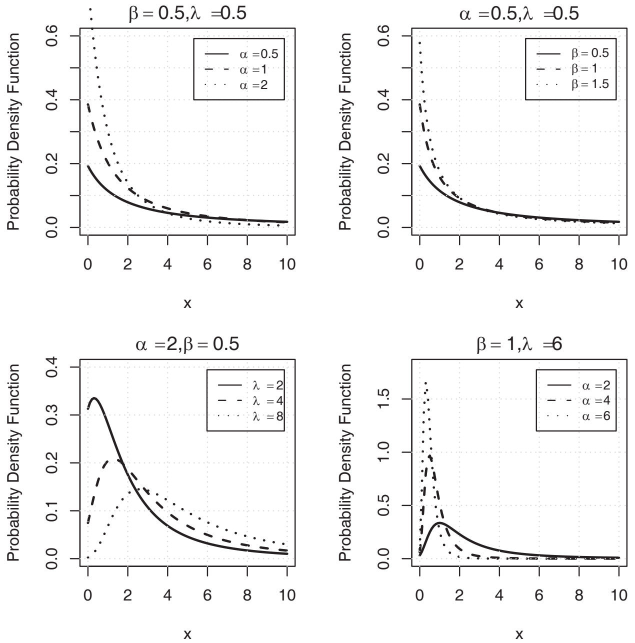


Figure 1: Plot of the probability density function for different values of the parameters α, β and λ.

***Proof*.** The first derivative of $h(x)$ with respect to x is given by

$$\begin{aligned}
h'(x) &= \frac{-\alpha\beta^2\lambda(1+\beta x)^{-(\alpha+2)}}{(e^{\lambda(1+\beta x)^{-\alpha}}-1)^2}\left[(\alpha+1)\ (e^{\lambda(1+\beta x)^{-\alpha}}-1)-\alpha\lambda(1+\beta x)^{-\alpha}e^{\lambda(1+\beta x)^{-\alpha}}\right] \\
&= \frac{-\alpha\beta^2\lambda(1+\beta x)^{-(\alpha+2)}}{(e^{\lambda(1+\beta x)^{-\alpha}}-1)^2}\left[(\alpha+1-\alpha\lambda(1+\beta x)^{-\alpha})\ e^{\lambda(1+\beta x)^{-\alpha}}-(\alpha+1)\right] \\
&= \frac{-\alpha\beta^2\lambda(1+\beta x)^{-(\alpha+2)}}{(e^{\lambda(1+\beta x)^{-\alpha}}-1)^2}\ \eta((1+\beta x)^{-\alpha}),
\end{aligned}$$

where $\eta(y) = -(\alpha+1)+(\alpha+1-\alpha\lambda y)e^{\lambda y}$, and $y = (1+\beta x)^{-\alpha} < 1$. The remaining of the proof is similar to that of Theorem 1. ■

***Note* 2.** The following should be noted.

(i) For $\lambda \in (0, 1]$, $h(x)$ is decreasing in $x > 0$ for all values of $\alpha, \beta > 0$.

(ii) For $\lambda > 1$, $h(x)$ may still exhibit a decreasing behavior, depending on the values of α and λ such that $(1 + (1 - \lambda)\alpha)e^{\lambda} - (\alpha + 1) \geq 0$.

(iii) Since $(1 + (1 - \lambda)\alpha)e^{\lambda} - (\alpha + 1) \geq 0$ implies that $\alpha(1 - \lambda) + 1 \geq 0$, then a decreasing hrf implies decreasing pdf. The converse is not necessarily true, e.g. $\alpha = 2$, $\lambda = 2$ implies decreasing pdf but unimodal hrf.

(iv) Since $(1+(1-\lambda)\alpha)e^{\lambda}-(\alpha+1)<0$ implies that $\alpha(1-\lambda)+1<0$, then a unimodal pdf implies unimodal hrf. The converse is not necessarily true, e.g., $\alpha=2$, $\lambda=2$ implies unimodal hrf but decreasing pdf.

Figure 2 shows the hrf curves for the $PLD(\alpha,\beta,\lambda)$ for selected values of the parameters α,β and λ.

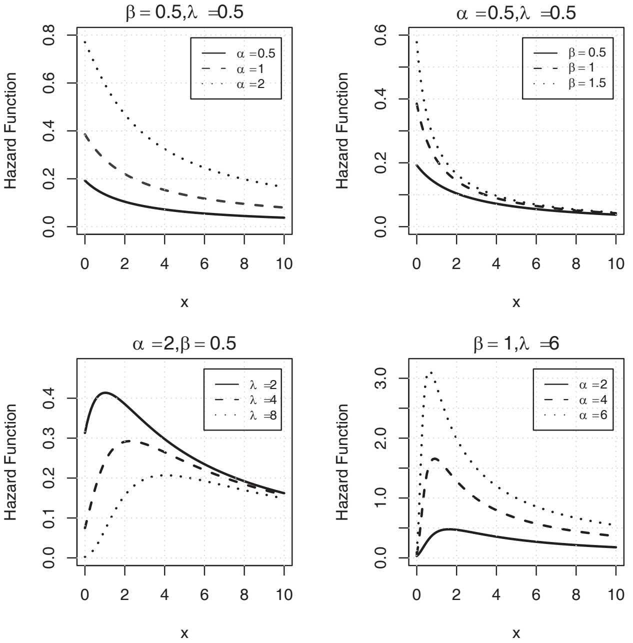


FIGURE 2: Plot of the hazard function for different values of the parameters α,β and λ.

3.3. Moments

We present an infinite sum representation for the rth moment, $\mu'_r=\mathrm{E}[X^r]$, and consequently the first four moments and variance for the PLD.

Theorem 3. *The rth moment about the origin of a random variable X, where $X\sim PLD(\alpha,\beta,\lambda)$, and $\alpha,\beta,\lambda>0$, is given by the following:*

$$\mu'_r=\mathrm{E}[X^r]=\frac{\alpha}{\beta^r(1-e^{-\lambda})}\sum_{n=0}^{\infty}\sum_{j=0}^{r}\binom{r}{j}\frac{\lambda^{n+1}(-1)^{n+r-j+1}}{(j-\alpha(n+1))n!},\quad r=1,2,\dots \tag{15}$$

***Proof*.** The rth moment of X can be determined by direct integration using the pdf, i.e. $\mu'_r=\int x^r f(x)dx$. We use the Maclaurin expansion of $e^x=\sum_{n=0}^{\infty}x^n/n!$, for all x. We also use the series representation

$$(1-w)^k=\sum_{j=0}^{k}\binom{k}{j}(-1)^j\ w^j,\ \text{ where } k \text{ is a positive integer.}$$

Therefore, after some transformations and integrations we have

$$\mathrm{E}\left[X^r\right] = \int_0^{\infty} x^r \ \frac{\alpha\beta\lambda\left(1+\beta x\right)^{-(\alpha+1)} \ e^{-\lambda(1+\beta x)^{-\alpha}}}{\left(1-e^{-\lambda}\right)} \ dx.$$

Setting $y = 1+\beta x$, $dx = dy/\beta$ yields

$$\begin{aligned}
\mathrm{E}\left[X^r\right] &= \frac{\alpha\lambda}{\beta^r(1-e^{-\lambda})}\int_1^{\infty}(y-1)^r y^{-(\alpha+1)} \ e^{-\lambda y^{-\alpha}} dy \\
&= \frac{\alpha\lambda}{\beta^r(1-e^{-\lambda})}\int_1^{\infty}\left\{\sum_{j=0}^{r}\binom{r}{j}y^{j-\alpha-1}(-1)^{r-j}\sum_{n=0}^{\infty}\frac{(-\lambda y^{-\alpha})^n}{n!}\right\}dy \\
&= \frac{\alpha\lambda}{\beta^r(1-e^{-\lambda})}\int_1^{\infty}\sum_{n=0}^{\infty}\sum_{j=0}^{r}\binom{r}{j}\frac{\lambda^{n+1}(-1)^{n+r-j}y^{j-\alpha(n+1)-1}}{n!} \ dy \\
&= \frac{\alpha}{\beta^r(1-e^{-\lambda})}\sum_{n=0}^{\infty}\sum_{j=0}^{r}\binom{r}{j}\frac{\lambda^{n+1}(-1)^{n+r-j+1}}{(j-\alpha(n+1))n!}.
\end{aligned}$$

This completes the proof of the theorem. ■

An alternative representation formula for (15) can readily be found by expanding and substituting in the binomial expansion.

$$\mu_r' = \frac{r!}{\beta^r(1-e^{-\lambda})}\sum_{k=1}^{\infty}\frac{(-1)^{k+r-1}\lambda^k}{k!(1-k\alpha)\cdots(r-k\alpha)}, \quad \alpha \neq \frac{i}{k}, \ i = 1,2,\cdots \tag{16}$$

One may use this representation to obtain the mean and the variance of X.

Corollary 1. *Let $X \sim PLD(\alpha,\beta,\lambda)$, where $\alpha,\beta,\lambda > 0$. Then the first four moments of X are given, respectively, as follows:*

$$\left.\begin{aligned}
\mu &= \mathrm{E}\left[X\right] = \tfrac{1}{\beta(1-e^{-\lambda})}\textstyle\sum_{k=1}^{\infty}\tfrac{(-1)^k\lambda^k}{k!(1-k\alpha)}, \\
\mu_2' &= \mathrm{E}\left[X^2\right] = \tfrac{2}{\beta^2(1-e^{-\lambda})}\textstyle\sum_{k=1}^{\infty}\tfrac{(-1)^{k+1}\lambda^k}{k!(1-k\alpha)(2-k\alpha)}, \\
\mu_3' &= \mathrm{E}\left[X^3\right] = \tfrac{6}{\beta^3(1-e^{-\lambda})}\textstyle\sum_{k=1}^{\infty}\tfrac{(-1)^{k+2}\lambda^k}{k!(1-k\alpha)(2-k\alpha)(3-k\alpha)}, \\
\mu_4' &= \mathrm{E}\left[X^4\right] = \tfrac{24}{\beta^4(1-e^{-\lambda})}\textstyle\sum_{k=1}^{\infty}\tfrac{(-1)^{k+3}\lambda^k}{k!(1-k\alpha)(2-k\alpha)(3-k\alpha)(4-k\alpha)}.
\end{aligned}\right\} \tag{17}$$

***Proof*.** Applying relations (15) or (16) for $r = 1,2,3$ and $r = 4$ yields the desired results. ■

Based on the results given in relations (17), the variance of X, denoted by $\sigma^2 = \mu_2' - \mu^2$ is given by

$$\sigma^2 = \frac{2}{\beta^2(1-e^{-\lambda})}\sum_{k=1}^{\infty}\frac{(-1)^{k+1}\lambda^k}{k!(1-k\alpha)(2-k\alpha)} - \left[\frac{1}{\beta(1-e^{-\lambda})}\sum_{k=1}^{\infty}\frac{(-1)^k\lambda^k}{k!(1-k\alpha)}\right]^2$$

It can be noticed from Table 1 that both the mean and the variance of the PL distribution are decreasing functions of α and β but they are increasing in λ. Table 2 shows the skewness and kurtosis of the PLD for various selected values of the parameters α, β and λ. The skewness is free of parameter β. Both the skewness and kurtosis are decreasing functions of α and both are increasing of λ.

Table 1: Mean and variance of PLD for various values of α, β and λ.

		$\beta = 0.5$		$\beta = 1.0$		$\beta = 2.0$	
λ	α	μ	σ^2	μ	σ^2	μ	σ^2
0.5	4.0	0.1184	1.6233	0.0592	0.4058	0.0296	0.1014
	4.5	0.1013	1.1089	0.0506	0.2772	0.0253	0.0693
	5.0	0.0885	0.8062	0.0442	0.2015	0.0221	0.0503
	5.5	0.0785	0.6128	0.0392	0.1532	0.0196	0.0383
	6.0	0.0706	0.4816	0.0353	0.1204	0.0176	0.0301
1.5	4.0	0.5890	1.9955	0.2945	0.4988	0.1472	0.1247
	4.5	0.5018	1.3402	0.2509	0.3350	0.1254	0.0837
	5.0	0.4369	0.9618	0.2184	0.2404	0.1092	0.0601
	5.5	0.3869	0.7237	0.1934	0.1809	0.0967	0.0452
	6.0	0.3471	0.5641	0.1735	0.1410	0.0867	0.0352
2.0	4.0	0.8104	2.0752	0.4052	0.5188	0.2026	0.1297
	4.5	0.6892	1.377	0.3446	0.3442	0.1723	0.0860
	5.0	0.5993	0.9791	0.2996	0.2447	0.1498	0.0611
	5.5	0.5301	0.7313	0.2650	0.1828	0.1325	0.0457
	6.0	0.4752	0.5668	0.2376	0.1417	0.1188	0.0354
4.0	4	1.4409	2.3195	0.7204	0.5798	0.3602	0.1449
	4.5	1.2179	1.4705	0.6089	0.3676	0.3044	0.0919
	5	1.0542	1.0089	0.5271	0.2522	0.2635	0.0630
	5.5	0.9289	0.7322	0.4644	0.1830	0.2322	0.0457
	6	0.8301	0.5542	0.4150	0.1385	0.2075	0.0346

3.4. L-moments

Suppose that a random sample $X_1, X_2, \dots, X_n$ is collected from $X \sim PLD(\theta)$, where $\theta = (\alpha, \beta, \lambda)$. In what follows, we derive a general representation for the L-moments of X.

The rth population L-moments is given by

$$\begin{aligned} E[X_{r:n}] &= \int_0^\infty x f(X_{r:n})\, dx \\ &= \int_0^\infty x \sum_{i=0}^{r-1} \sum_{j=0}^{n-r+i} \binom{r-1}{i} \binom{n-r+i}{j} (-1)^{i+j} \left\{ \frac{n! \alpha\beta\lambda}{(r-1)!(n-r)!} \right. \\ &\quad \left. \times \frac{(1+\beta x)^{-(\alpha+1)} e^{-\lambda(1+\beta x)^{-\alpha}(j+1)}}{(1-e^{-\lambda})^{n-r+i+1}} \right\} dx. \end{aligned}$$

Table 2: Skewness and kurtosis of PLD for various values of α, β and λ.

		$\beta = 0.5$		$\beta = 1.0$		$\beta = 2.0$	
λ	α	γ_3	γ_4	γ_3	γ_4	γ_3	γ_4
0.5	4.5	3.6525	65.367	3.6525	16.3418	3.6525	4.0854
	5.0	3.1739	24.114	3.1739	6.0285	3.1739	1.5071
	5.5	2.8845	12.696	2.8845	3.1741	2.8845	0.7935
	6.0	2.6904	7.8535	2.6904	1.9633	2.6904	0.4908
	6.5	2.5510	5.3396	2.5510	1.3349	2.5510	0.3337
1.5	4.5	3.0490	75.423	3.049	18.855	3.0490	4.7139
	5.0	2.5371	26.405	2.5371	6.6014	2.5371	1.6503
	5.5	2.2239	13.345	2.2239	3.3362	2.2239	0.8340
	6.0	2.0116	7.9879	2.0116	1.9969	2.0116	0.4992
	6.5	1.8579	5.2867	1.8579	1.3216	1.8579	0.3304
2.0	4.5	3.0915	84.916	3.0915	21.229	3.0915	5.3072
	5.0	2.5372	29.211	2.5372	7.3029	2.5372	1.8257
	5.5	2.1963	14.561	2.1963	3.6404	2.1963	0.9101
	6.0	1.9641	8.6212	1.9641	2.1553	1.9641	0.5388
	6.5	1.7952	5.6554	1.7952	1.4138	1.7952	0.3534
4.0	4.5	3.8191	128.068	3.8191	32.017	3.8191	8.0042
	5.0	3.1425	42.525	3.1425	10.631	3.1425	2.6578
	5.5	2.7233	20.595	2.7233	5.1489	2.7233	1.2872
	6.0	2.4357	11.905	2.4357	2.9764	2.4357	0.7441
	6.5	2.2251	7.6554	2.2251	1.9138	2.2251	0.4784

Let $y = (1 + \beta x)$ so $x = (y - 1)/\beta$ and $dx = (1/\beta)dy$. After some transformation, we arrive to the formula:

$$\mathrm{E}\left[X_{r:n}\right] = \frac{1}{\beta} \sum_{m=0}^{\infty} \sum_{i=0}^{r-1} \sum_{j=0}^{n-r+i} \frac{(j+1)^m (-\lambda)^{m+1}\ A_{ij}}{(m+1)!(1-\alpha(m+1))}, \tag{18}$$

where A_{ij} is

$$A_{ij} = \frac{n!(-1)^{i+j}}{(r-1)!(n-r)!\left(1-e^{-\lambda}\right)^{n-r+i+1}} \binom{r-1}{i} \binom{n-r+i}{j}.$$

One readily can use the relation (18) to obtain the first L-moments of $X_{r:n}$. For example, we take $r = n = 1$ to obtain $\lambda_1 = \mathrm{E}\left[X_{1:1}\right]$ which is the mean of the random variable X.

$$\lambda_1 = E[X_{1:1}] = \frac{1}{\beta(1-e^{-\lambda})} \sum_{m=0}^{\infty} \frac{(-\lambda)^{m+1}}{(m+1)!(1-\alpha(m+1))},$$

This result is consistent with that obtained in relation (17). The other two L-moments, λ_2 and λ_3, are respectively given by

$$\begin{aligned} \lambda_2 \quad &= \tfrac{1}{\beta} \Bigg[\textstyle\sum_{m=0}^{\infty} \sum_{i=0}^{1} \sum_{j=0}^{i} \binom{1}{i} \binom{i}{j} \frac{(j+1)^m (-1)^{i+j+m+1} \lambda^{m+1}}{(m+1)!(1-\alpha(m+1))(1-e^{-\lambda})^{i+1}} \\ &\qquad - \textstyle\sum_{m=0}^{\infty} \sum_{j=0}^{1} \binom{1}{j} \frac{(j+1)^m (-1)^{j+m+1} \lambda^{m+1}}{(m+1)!(1-\alpha(m+1))(1-e^{-\lambda})^2} \Bigg] \end{aligned}$$

and

$$\begin{aligned}\lambda_3 &= \frac{1}{\beta}\Bigg[\textstyle\sum_{m=0}^{\infty}\sum_{i=0}^{2}\sum_{j=0}^{i}\binom{2}{i}\binom{i}{j}\frac{(j+1)^m(-1)^{i+j+m+1}\lambda^{m+1}}{(m+1)!(1-\alpha(m+1))(1-e^{-\lambda})^{i+1}} \\ &\quad -2\textstyle\sum_{m=0}^{\infty}\sum_{i=0}^{1}\sum_{j=0}^{i+1}\binom{1}{i}\binom{i+1}{j}\frac{2(j+1)^m(-1)^{i+j+m+1}\lambda^{m+1}}{(m+1)!(1-\alpha(m+1))(1-e^{-\lambda})^{i+2}} \\ &\quad +\textstyle\sum_{m=0}^{\infty}\sum_{j=0}^{2}\binom{2}{j}\frac{2(j+1)^m(-1)^{j+m+1}\lambda^{m+1}}{(m+1)!(1-\alpha(m+1))(1-e^{-\lambda})^{3}}\Bigg]\end{aligned}$$

The method of L-moments consists of equating the first L-moments of a population, λ_1, λ_2 and λ_3, to the corresponding L-moments of a sample, l_1, l_2 and l_3, thus getting a number of equations that are needed to be solved, numerically, in terms of the unknown parameters, θ.

3.5. Order Statistics

Let $X_1, X_2, \dots, X_n$ be a random sample of size n from the PL distribution in (10) and let $X_{1:n}, \dots, X_{n:n}$ denote the corresponding order statistics. Then, the pdf of $X_{r:n}$, $1 \leq r \leq n$, is given by (see, David & Nagaraja 2003, Arnold, Balakrishnan & Nagaraja 1992)

$$g_{(r)}(x) = C_{r,n} g(x)[G(x)]^{r-1}[1-G(x)]^{n-r}, \quad 0 < x < \infty, \tag{19}$$

where $C_{r,n} = [B(r, n-r+1)]^{-1}$, with $B(a,b)$ being the complete beta function.

Theorem 4. *Let $G(x)$ and $g(x)$ be the cdf and pdf of a Poisson-Lomax distribution for a random variable X. The density of the rth order statistic, say $g_{(r)}(x)$ is given by*

$$g_{(r)}(x) = \alpha\beta\lambda C_{r,n}\sum_{i=0}^{r-1}\sum_{j=0}^{n-r+i}\binom{r-1}{i}\binom{n-r+i}{j} \frac{(-1)^{i+j}(1+\beta x)^{-(\alpha+1)}\ e^{-\lambda(1+\beta x)^{-\alpha}(j+1)}}{(1-e^{-\lambda})^{n-r+i+1}} \tag{20}$$

Proof. First it should be noted that (19) can be written as follows:

$$g_{(r)}(x) = C_{r,n}\sum_{i=0}^{r-1}\binom{r-1}{i}(-1)^i g(x)[\bar{G}(x)]^{n-r+i} \tag{21}$$

then the proof follows by replacing the reliability, $\bar{G}(x)$, and the pdf, $g(x)$, of $X \sim PLD(\alpha, \beta, \lambda)$ which are obtained from (9) and (10), respectively, and substituting them into relation (21), and expanding the term $(1-e^{-\lambda(1+\beta x)^{-\alpha}})^{n-r+i}$ using the binomial expansion. ■

3.6. Quantile Function

Let X denote a random variable with the probability density function given by (10). The quantile function, denoted by $Q(u)$, is

$$Q(u) = \inf\{x \in R : F(x) \geq u\}, \quad \text{where } 0 < u < 1$$

By inverting the distribution function, $F = 1 - \bar{F}$, we can write the following:

$$Q(u) = \frac{1}{\beta}\left[\left(\frac{-\ln(u(1-e^{-\lambda}) + e^{-\lambda})}{\lambda}\right)^{-1/\alpha} - 1\right] \tag{22}$$

The first quartile, the median and the third quartile can be obtained simply by applying (22). The quartiles; Q_1 first quartile, Q_2 second quartile, or the median, and Q_3 third quartile are obtained in Table 3.

Table 3: The quartile values of the PLD for different values of α, β and λ.

		$\beta = 0.5$			$\beta = 1.0$			$\beta = 2.0$		
λ	α	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
0.5	4.0	0.1870	0.4583	0.9647	0.0935	0.2291	0.4824	0.0467	0.1146	0.2412
	4.5	0.1654	0.4025	0.8379	0.0827	0.2013	0.4189	0.0413	0.1006	0.2095
	5.0	0.1482	0.3589	0.7403	0.0741	0.1794	0.3701	0.0371	0.0897	0.1851
	5.5	0.1343	0.3238	0.6629	0.0672	0.1619	0.3315	0.0336	0.0809	0.1657
	6.0	0.1228	0.2949	0.6002	0.0614	0.1474	0.3001	0.0307	0.0737	0.1500
1.5	4.0	0.2893	0.6431	1.2469	0.1446	0.3216	0.6234	0.0723	0.1608	0.3117
	4.5	0.2552	0.5625	1.0767	0.1276	0.2813	0.5384	0.0638	0.1406	0.2692
	5.0	0.2282	0.4998	0.9470	0.1141	0.2499	0.4735	0.0571	0.1249	0.2368
	5.5	0.2065	0.4496	0.8450	0.1032	0.2248	0.4225	0.0516	0.1124	0.2112
	6.0	0.1885	0.4086	0.7626	0.0942	0.2043	0.3813	0.0471	0.1021	0.1907
2.0	4.0	0.3521	0.7418	1.3856	0.1760	0.3709	0.6928	0.0880	0.1855	0.3464
	4.5	0.3101	0.6474	1.1933	0.1550	0.3237	0.5966	0.0775	0.1618	0.2983
	5.0	0.2770	0.5742	1.0473	0.1385	0.2871	0.5237	0.0693	0.1435	0.2618
	5.5	0.2503	0.5158	0.9328	0.1252	0.2579	0.4664	0.0626	0.1289	0.2332
	6.0	0.2283	0.4681	0.8408	0.1142	0.2341	0.4204	0.0571	0.1170	0.2102
4.0	4.0	0.6324	1.1205	1.8827	0.3162	0.5602	0.9414	0.1581	0.2801	0.4707
	4.5	0.5533	0.9700	1.6068	0.2766	0.4850	0.8034	0.1383	0.2425	0.4017
	5.0	0.4917	0.8548	1.4003	0.2458	0.4274	0.7001	0.1229	0.2137	0.3501
	5.5	0.4424	0.7640	1.2401	0.2212	0.3820	0.6201	0.1106	0.1910	0.3100
	6.0	0.4020	0.6900	1.1125	0.2010	0.3452	0.5562	0.1005	0.1726	0.2781

3.7. Mean Deviations

The mean deviation about the mean and the mean deviation about the median are, respectively, defined by

$$\delta_1(\mu) = 2\mu F(\mu) - 2\mu + 2\int_{\mu}^{\infty} z f(z) dz \tag{23}$$

$$\delta_2(M) = 2MF(M) - M - \mu + 2\int_{M}^{\infty} z f(z) dz \tag{24}$$

Theorem 5. *Let X be a random variable distributed according to the PL distribution. Then the mean deviation about the mean, δ_1, and the mean deviation about the median, δ_2, are given as follows:*

$$\begin{aligned}\delta_1(\mu) = \frac{2}{1-e^{-\lambda}}\Bigg\{&\mu\left(e^{-\lambda(1+\beta\mu)^{-\alpha}}-1\right)-\frac{\alpha}{\beta}\sum_{n=0}^{\infty}\frac{\lambda^{n+1}(-1)^n}{n!}\\ &\times\left(\frac{(1+\beta\mu)^{1-\alpha(n+1)}}{1-\alpha(n+1)}+\frac{(1+\beta\mu)^{-\alpha(n+1)}}{\alpha(n+1)}\right)\Bigg\}\end{aligned} \tag{25}$$

and

$$\begin{aligned}\delta_2(M) = \frac{1}{1-e^{-\lambda}}\Bigg\{&M\left(2e^{-\lambda(1+\beta M)^{-\alpha}}-e^{-\alpha}-1\right)\\ &+\frac{1}{\beta}\sum_{n=0}^{\infty}\frac{(-1)^n\lambda^{n+1}}{(n+1)!(1-(n+1)\alpha)}\\ &-\frac{2\alpha}{\beta}\sum_{n=0}^{\infty}\frac{\lambda^{n+1}(-1)^n}{n!}\Bigg(\frac{(1+\beta M)^{1-\alpha(n+1)}}{1-\alpha(n+1)}\\ &+\frac{(1+\beta M)^{-\alpha(n+1)}}{\alpha(n+1)}\Bigg)\Bigg\}\end{aligned} \tag{26}$$

***Proof*.** The proof follows by plugging the density function of the PLD into equation (23) and working out the integration I, where

$$I=\int_{\mu}^{\infty}xg(x)dx=\frac{\alpha\beta\lambda}{1-e^{-\lambda}}\int_{\mu}^{\infty}x(1+\beta x)^{-(\alpha+1)}e^{-\lambda(1+\beta x)^{-\alpha}}dx$$

Setting $y=1+\beta x$, so $dy=\beta dx$ and using the expansion $e^x=\sum_{n=0}^{\infty}x^n/n!$, yields

$$I=\frac{-\alpha}{\beta(1-e^{-\lambda})}\sum_{n=0}^{\infty}\frac{\lambda^{n+1}(-1)^n}{n!}\left(\frac{(1+\beta\mu)^{1-\alpha(n+1)}}{1-\alpha(n+1)}+\frac{(1+\beta\mu)^{-\alpha(n+1)}}{\alpha(n+1)}\right)$$

Substituting I into relation (23) and manipulating the other terms gives directly the desired result. Similarly, the measure $\delta_2(M)$ can be obtained. ■

4. Estimation

In this section we consider maximum likelihood estimation (MLE) to estimate the involved parameters. Asymptotic distribution of $\hat{\boldsymbol{\theta}}=(\hat{\alpha},\hat{\beta},\hat{\lambda})$ are obtained using the elements of the inverse Fisher information matrix.

4.1. Maximum Likelihood Estimation

The idea behind the maximum likelihood parameter estimation is to determine the parameters that maximize the probability (likelihood) of the sample data. For this purpose, let $X_1, X_2, \dots, X_n$ is be random sample from $X \sim PLD(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\alpha, \beta, \lambda)$. Then the likelihood function of the observed sample is given by

$$\begin{aligned} L(\boldsymbol{\theta}; x) &= \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \frac{\lambda\alpha\beta(1+\beta x_i)^{-(\alpha+1)}\ e^{-\lambda(1+\beta x_i)^{-\alpha}}}{(1-e^{-\lambda})} \\ &= \frac{(\lambda\alpha\beta)^n}{(1-e^{-\lambda})^n} \prod_{i=1}^{n}(1+\beta x_i)^{-(\alpha+1)}\ e^{-\lambda\sum_{i=1}^{n}(1+\beta x_i)^{-\alpha}} \end{aligned} \tag{27}$$

The log-likelihood function is given by

$$\begin{aligned} \ell(x; \alpha, \beta, \lambda) &= n\ln(\alpha) + n\ln(\beta) + n\ln(\lambda) - (\alpha+1)\sum_{i=1}^{n}\ln(1+\beta x_i) \\ &\quad -\lambda\sum_{i=1}^{n}(1+\beta x_i)^{-\alpha} - n\ln(1-e^{-\lambda}) \end{aligned} \tag{28}$$

The MLEs of α, β and λ say $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\lambda}$, respectively, can be worked out by the solutions of the system of equations obtained by letting the first partial derivatives of the total log-likelihood equal to zero with respect to $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\lambda}$. Therefore, the system of equations is as follows:

$$\begin{aligned} \frac{\partial\ell}{\partial\alpha} &= \frac{n}{\alpha} - \sum_{i=1}^{n}\ln(1+\beta x_i) + \lambda\sum_{i=1}^{n}(1+\beta x_i)^{-\alpha}\ln(1+\beta x_i) = 0 \\ \frac{\partial\ell}{\partial\beta} &= \frac{n}{\beta} - (\alpha+1)\sum_{i=1}^{n}\frac{x_i}{1+\beta x_i} + \alpha\lambda\sum_{i=1}^{n}x_i(1+\beta x_i)^{-(\alpha+1)} = 0 \\ \frac{\partial\ell}{\partial\lambda} &= \frac{n}{\lambda} - \sum_{i=1}^{n}(1+\beta x_i)^{-\alpha} - \frac{n}{(e^{\lambda}-1)} = 0 \end{aligned}$$

For simplicity, we define A_i to be as $A_i = 1 + \beta x_i$. Thus, we have

$$\hat{\alpha} = n\left[\sum_{i=1}^{n}\ln(A_i)\ (1-\lambda A_i^{-\alpha})\right]^{-1} \tag{29}$$

$$\hat{\beta} = n\left[\sum_{i=1}^{n}\frac{x_i}{A_i}\ (\alpha+1-\alpha\lambda A_i^{-\alpha})\right]^{-1} \tag{30}$$

$$\hat{\lambda} = n\left[\sum_{i=1}^{n}A_i^{-\alpha} + \frac{n}{e^{\lambda}-1}\right]^{-1} \tag{31}$$

The solutions of nonlinear equations (29), (30) and (31) are complicated to obtain, therefore an iterative procedure is applied to solve these equations numerically.

4.2. Asymptotic Distribution

We obtain the asymptotic distribution of $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta}, \hat{\lambda})$. The asymptotic variances of MLEs are given by the elements of the inverse of the Fisher information matrix. The Fisher information matrix of $\boldsymbol{\theta}$, denoted by $\boldsymbol{J}(\boldsymbol{\theta}) = \boldsymbol{E}(\boldsymbol{I}, \boldsymbol{\theta})$, where $\boldsymbol{I}_{ij}$, $i, j = 1, 2, 3$ is the observed information matrix. The second partial derivatives of the maximum likelihood function are given as the following:

$$\begin{aligned}
I_{11} &= -\frac{n}{\alpha^2} - \lambda \sum_{i=1}^{n} (1+\beta x_i)^{-\alpha} [\ln(1+\beta x_i)]^2 \\
&= -\frac{n}{\alpha^2} - \lambda \sum_{i=1}^{n} A_i^{-\alpha} [\ln(A_i)]^2 \\
I_{12} = I_{21} &= \sum_{i=1}^{n} \frac{-x_i}{(1+\beta x_i)} + \lambda \sum_{i=1}^{n} x_i (1+\beta x_i)^{-(\alpha+1)} \left[1 - \alpha \ln(1+\beta x_i)\right] \\
&= \sum_{i=1}^{n} \left[\frac{x_i}{A_i} \left(-1 - \lambda\alpha A_i^{-\alpha} \ln(A_i) + \lambda A_i^{-\alpha}\right) \right] \\
I_{13} = I_{31} &= \sum_{i=1}^{n} (1+\beta x_i)^{-\alpha} \ln(1+\beta x_i) = \sum_{i=1}^{n} A_i^{-\alpha} \ln(A_i) \\
I_{22} &= -\frac{n}{\beta^2} + (\alpha+1) \sum_{i=1}^{n} \frac{x_i^2}{(1+\beta x_i)^2} - \lambda\alpha(\alpha+1) \sum_{i=1}^{n} \frac{x_i^2 (1+\beta x_i)^{-\alpha}}{(1+\beta x_i)^2} \\
&= -\frac{n}{\beta^2} + (\alpha+1) \sum_{i=1}^{n} \left(\frac{x_i}{A_i}\right)^2 \left(1 - \lambda\alpha A_i^{-\alpha}\right) \\
I_{23} = I_{32} &= \alpha \sum_{i=1}^{n} x_i (1+\beta x_i)^{-(\alpha+1)} = \alpha \sum_{i=1}^{n} x_i A_i^{-(\alpha+1)} \\
I_{33} &= -\frac{n}{\lambda^2} + \frac{n e^{\lambda}}{(e^{\lambda}-1)^2}
\end{aligned}$$

The exact mathematical expressions for $\boldsymbol{J}(\boldsymbol{\theta}) = \boldsymbol{E}(\boldsymbol{I}, \boldsymbol{\theta})$ are complicated to obtain. Therefore, the observed Fisher information matrix can be used instead of the Fisher information matrix. The variance-covariance matrix may be approximated as $\boldsymbol{V}_{ij} = \boldsymbol{I}_{ij}^{-1}$. The asymptotic distribution of the maximum likelihood can be written as follows (see Miller 1981).

$$\left[(\hat{\alpha} - \alpha), (\hat{\beta} - \beta), (\hat{\lambda} - \lambda)\right] \sim N_3\left(\boldsymbol{0}, \boldsymbol{V}\right) \tag{32}$$

Since $\boldsymbol{V}$ involves the parameters α, β and λ, we replace the parameters by the corresponding MLEs in order to obtain an estimate of $\boldsymbol{V}$, which is denoted by $\hat{\boldsymbol{V}}$. By using (32), approximate $100(1-\vartheta)\%$ confidence intervals for α, β and λ are determined, respectively, as

$$\hat{\alpha} \pm Z_{\vartheta/2} \sqrt{\hat{\boldsymbol{V}}_{11}}, \quad \hat{\beta} \pm Z_{\vartheta/2} \sqrt{\hat{\boldsymbol{V}}_{22}}, \quad \hat{\lambda} \pm Z_{\vartheta/2} \sqrt{\hat{\boldsymbol{V}}_{33}},$$

where Z_{ϑ} is the upper 100ϑ-th percentile of the standard normal distribution.

In the order to numerically illustrate the estimation of the involved parameters, we have simulated the ML estimators for different sample sizes. The calculation of the estimation is based on $10,000$ simulated samples from the standard PLD. Table 4 shows the MLEs, mean squared errors (MSE) and 95% confidence limits (LCL & UCL) for the parameters α, β, and λ. The true values of the parameters used for simulation were $\alpha = 1, \beta = 1$, and $\lambda = 2$. It is observed that when the sample size n increases, the MLE of α and λ decrease to approach the true one while the MLEs of the parameters β increase.

Table 4: Simulation study: parameter values used for simulation (TRUE) $\alpha = 1, \beta = 1, \lambda = 2$, MLEs, mean squared errors (MSE) and 95% confidence limits (LCL & UCL) for the parameters.

				95% Confi. Limits	
Parameters	n	Estimates	MSE	LCL	UCL
α	20	1.10868	0.05159	-2.00901	4.22637
	30	1.08199	0.03129	-1.12927	3.29326
	40	1.06866	0.02202	-0.62073	2.75807
	50	1.06119	0.01762	-1.43696	3.55935
	60	1.05224	0.01431	0.10648	1.99800
	70	1.04646	0.01203	0.15111	1.94181
	80	1.04378	0.01034	0.01529	2.07227
	90	1.03915	0.00871	0.18454	1.89376
	100	1.03811	0.00791	0.21745	1.85878
	200	1.02512	0.00375	0.30619	1.74405
β	20	0.94360	0.05699	0.52240	1.36480
	30	0.94997	0.03854	0.60608	1.29387
	40	0.95472	0.03019	0.65637	1.25308
	50	0.96011	0.02421	0.69225	1.22797
	60	0.96078	0.02043	0.71629	1.20527
	70	0.96329	0.01748	0.73662	1.18997
	80	0.96387	0.01600	0.75180	1.17594
	90	0.96371	0.01401	0.76389	1.16353
	100	0.97031	0.01216	0.77951	1.16110
	200	0.97528	0.00683	0.83990	1.11065
λ	20	2.07641	0.05612	0.38236	3.77045
	30	2.05300	0.03373	0.67893	3.42706
	40	2.03975	0.02294	0.85301	3.22649
	50	2.03150	0.01773	0.97162	3.09137
	60	2.02744	0.01478	1.06066	2.99422
	70	2.02349	0.01221	1.12896	2.91801
	80	2.02025	0.01077	1.18387	2.85662
	90	2.01885	0.00929	1.23050	2.80719
	100	2.01723	0.00845	1.26951	2.76495
	200	2.00944	0.00388	1.48125	2.53762

5. Application

We have considered a dataset corresponding to remission times (in months) of a random sample of 128 bladder cancer patients given in Lee & Wang (2003). The data are given as follows: 0.08, 2.09, 3.48, 4.87, 6.94 , 8.66, 13.11, 23.63, 0.20, 2.23, 3.52, 4.98, 6.97, 9.02, 13.29, 0.40, 2.26, 3.57, 5.06, 7.09, 9.22, 13.80, 25.74, 0.50, 2.46 , 3.64, 5.09, 7.26, 9.47, 14.24, 25.82, 0.51, 2.54, 3.70, 5.17, 7.28, 9.74, 14.76, 26.31, 0.81, 2.62, 3.82, 5.32, 7.32, 10.06, 14.77, 32.15, 2.64, 3.88, 5.32, 7.39, 10.34, 14.83, 34.26, 0.90, 2.69, 4.18, 5.34, 7.59, 10.66, 15.96, 36.66, 1.05, 2.69, 4.23, 5.41, 7.62, 10.75, 16.62, 43.01, 1.19, 2.75, 4.26, 5.41, 7.63, 17.12, 46.12, 1.26, 2.83, 4.33, 5.49, 7.66, 11.25, 17.14, 79.05, 1.35, 2.87, 5.62, 7.87, 11.64, 17.36, 1.40, 3.02, 4.34, 5.71, 7.93, 11.79, 18.10, 1.46, 4.40, 5.85, 8.26, 11.98, 19.13, 1.76, 3.25, 4.50, 6.25, 8.37, 12.02, 2.02, 3.31, 4.51, 6.54, 8.53, 12.03, 20.28, 2.02, 3.36, 6.76, 12.07, 21.73, 2.07, 3.36, 6.93, 8.65, 12.63, 22.69. We have fitted the Poisson-Lomax distribution to the dataset using MLE, and compared the proposed PLD with Lomax, extended Lomax and Lomax-Logarithmic distributions.

The model selection is carried out using the AIC (Akaike information criterion), the BIC (Bayesian information criterion), the CAIC (consistent Akaike information criteria) and the HQIC (Hannan-Quinn information criterion).

$$\left.\begin{aligned}
&\text{AIC} = -2l(\hat{\theta}) + 2q,\\
&\text{BIC} = -2l(\hat{\theta}) + q\log(n),\\
&\text{HQIC} = -2l(\hat{\theta}) + 2q\log(\log(n)),\\
&\text{CAIC} = -2l(\hat{\theta}) + \tfrac{2qn}{n-q-1}
\end{aligned}\right\} \tag{33}$$

where $l(\hat{\theta})$ denotes the log-likelihood function evaluated at the maximum likelihood estimates, q is the number of parameters, and n is the sample size. Here we let $\boldsymbol{\theta}$ denote the parameters, i.e., $\boldsymbol{\theta} = (\alpha, \beta, \lambda)$. An iterative procedure is applied to solve equations (29), (30) and (31) and consequently obtain $\hat{\boldsymbol{\theta}} = (\hat{\alpha} = 2.8737, \hat{\beta} = 8.2711, \hat{p} = 3.3515)$. At these values we calculate the log-likelihood function given by (28) and apply relation (33). The model with minimum AIC (or BIC, CAIC and HQIC) value is chosen as the best model to fit the data. From Table 5, we conclude that the PLD is best comparable to the Lomax, extended Lomax and Lomax-Logarithmic models.

Table 5: MLEs (standard errors in parentheses) and the measures AIC, BIC, HQIC and CAIC.

	Estimates				Measures			
Models	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\lambda}$	AIC	BIC	HQIC	CAIC
Lomax	13.9384	121.0222			831.67	837.37	833.98	831.76
	(15.3837)	(142.6940)						
MOEL	23.7437	2.0487	2.2818		825.08	833.64	828.56	825.27
	(35.8106)	(2.5891)	(0.5551)					
PLD	2.8737	8.2711		3.3515	824.77	833.33	828.25	824.96
	(0.8869)	(4.8795)		(1.0302)				

For an ordered random sample, $X_1, X_2, \dots, X_n$, from $PLD(\alpha, \beta, \lambda)$, where the parameters α, β and λ are unknown, the Kolmogorov-Smirnov D_n, Cramér-von Mises W_n^2, Anderson and Darling A_n^2, Watson U_n^2 and Liao-Shimokawa L_n^2 tests statistics are given as follows: (For details see e.g. Al-Zahrani (2012) and references therein).

$$
\begin{aligned}
D_n &= \max_i \left[\frac{i}{n} - G_{PL}(x_i, \hat{\alpha}, \hat{\beta}, \hat{\lambda}), G_{PL}(x_i, \hat{\alpha}, \hat{\beta}, \hat{\lambda}) - \frac{i-1}{n}\right] \\
W_n^2 &= \frac{1}{12n} + \sum_{i=1}^{n} \left[G_{PL}(x_i, \hat{\alpha}, \hat{\beta}, \hat{\lambda}) - \frac{2i-1}{2n}\right]^2 \\
A_n^2 &= -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log(G_{PL}(x_i, \hat{\alpha}, \hat{\beta}, \hat{\lambda})) + \log(1 - G_{PL}(x_i, \hat{\alpha}, \hat{\beta}, \hat{\lambda}))\right]^2 \\
U_n^2 &= W_n^2 + \sum_{i=1}^{n}\left[\frac{G_{PL}(x_i, \hat{\alpha}, \hat{\beta}, \hat{\lambda})}{n} - \frac{1}{2}\right]^2 \\
L_n &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\max_i \left[\frac{i}{n} - G_{PL}(x_i, \hat{\alpha}, \hat{\beta}, \hat{\lambda}), G_{PL}(x_i, \hat{\alpha}, \hat{\beta}, \hat{\lambda}) - \frac{i-1}{n}\right]}{\sqrt{G_{PL}(x_i, \hat{\alpha}, \hat{\beta}, \hat{\lambda})[1 - G_{PL}(x_i, \hat{\alpha}, \hat{\beta}, \hat{\lambda})]}}.
\end{aligned}
$$

Table 6 indicates that the test statistics D_n, W_n^2, A_n^2, U_n^2 and L_n have the smallest values for the data set under PLD model with regard to the other models. The proposed model offers an attractive alternative to the Lomax, Lomax-Logarithmic and extended Lomax models. Figure 3 displays the empirical and fitted densities for the data. Estimated survivals for data are shown in Figure 4. The Poisson-Lomax distribution approximately provides an adequate fit for the data. The quantile-quantile or Q-Q plot is used to check the validity of the distributional assumption for the data. Figure 5 shows that the data seems to follow a PLD reasonably well, except some points on extreme.

Table 6: Goodness-of-fit tests.

	Statistics				
Distribution	D_n	W_n^2	A_n^2	U_n^2	L_n
Lomax	0.0967	0.2126	1.3768	31.7017	1.0594
MOEL	0.0302	0.0151	0.0926	31.5177	0.3728
LLD	0.0821	0.1274	0.8739	31.6200	0.8491
PLD	0.0281	0.0134	0.0835	31.5164	0.3567

6. Concluding Remarks

In this paper we have proposed a new distribution, referred to as the PLD. A mathematical treatment of the proposed distribution including explicit formulas for the density and hazard functions, moments, order statistics, and mean and median deviations have been provided. The estimation of the parameters has been approached by maximum likelihood. Also, the asymptotic variance-covariance matrix of the estimates has been obtained. Finally, a real data set was analyzed to show the potential of the proposed PLD. The result indicates that the PLD may be used for a wider range of statistical applications. Further study can be conducted on the proposed distribution. Here, we mention some of possible directions which are still open for further works. The problem of parameter estimation can be studied using e.g. Bayesian approach and making future prediction. The parameters of the proposed distribution can be estimated based on censored data. Some recurrence relations can be established for the single moments and product moments of order statistics.

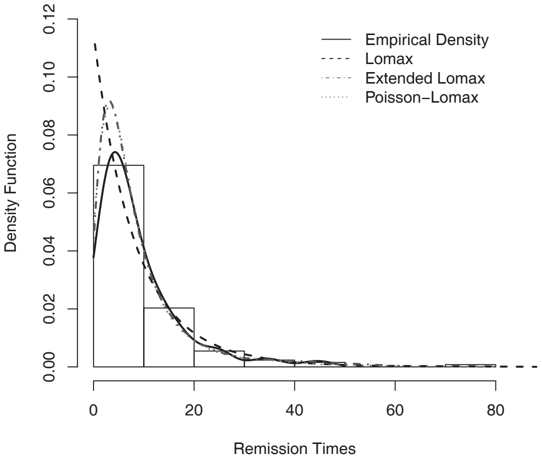


FIGURE 3: Estimated densities for bladder cancer data.

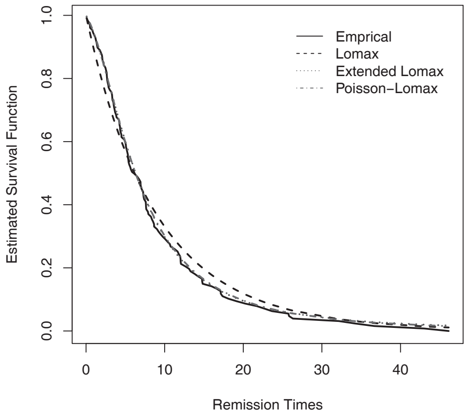


FIGURE 4: Estimated survivals for bladder cancer data.

Acknowledgments

The authors are grateful to the Editor and the anonymous referees for their valuable comments and suggestions that improved the presentation of the paper. This study is a part of the Master Thesis of the second named author whose work was supervised by the first named author.

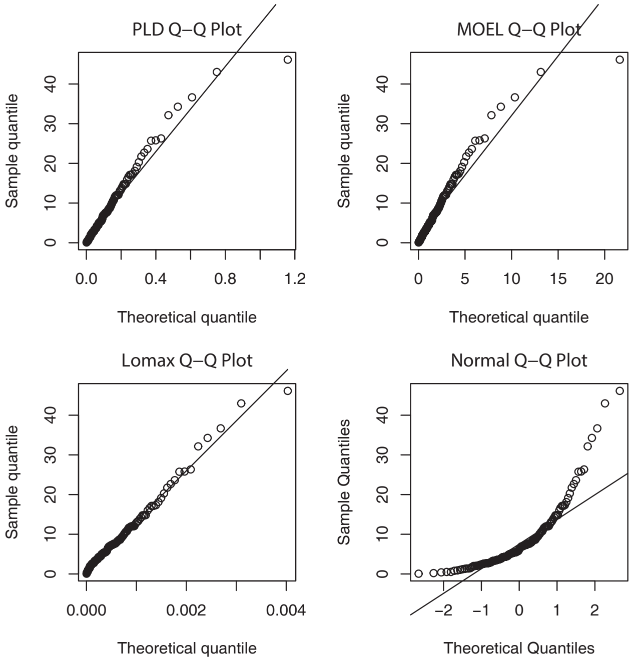


Figure 5: The Q-Q plot for bladder cancer data.

[Recibido: noviembre de 2013 — Aceptado: abril de 2014]

References

Al-Awadhi, S. A. & Ghitany, M. E. (2001), 'Statistical properties of Poisson-Lomax distribution and its application to repeated accidents data', *Journal of Applied Statistical Sciences* **10**(4), 365–372.

Al-Zahrani, B. (2012), 'Goodness-of-fit for the Topp-Leone distribution with unknown parameters', *Applied Mathematical Sciences* **6**(128), 6355–6363.

Alice, T. & Jose, K. K. (2003), 'Marshall-Olkin Pareto processes', *Far East Journal of Theoretical Statistics* **2**(9), 117–132.

Arnold, B. C., Balakrishnan, N. & Nagaraja, H. H. N. (1992), *A First Course in Order Statistics*, John Wiley & Sons, New York.

Cancho, V. G., Louzada-Neto, F. & Barriga, G. D. (2011), 'The Poisson-exponential lifetime distribution', *Computational Statistics & Data Analysis* **55**(1), 677–686.

David, H. & Nagaraja, H. N. (2003), *Order Statistics*, John Wiley & Sons, Hoboken, New Jersey.

Ghitany, M. E., Al-Awadhi, F. A. & Alkhalfan, L. A. (2007), 'Marshall-Olkin extended Lomax distribution and its application to censored data', *Communications in Statistics-Theory and Methods* **36**(10), 1855–1866.

Ghitany, M. E., Al-Hussaini, E. K. & Al-Jarallah, R. A. (2005), 'Marshall-Olkin extended Weibull distribution and its application to censored data', *Journal of Applied Statistics* **32**(10), 1025–1034.

Kus, C. (2007), 'A new lifetime distribution', *Computational Statistics & Data Analysis* **51**(9), 4497–4509.

Lee, E. T. & Wang, J. W. (2003), *Statistical Methods for Survival Data Analysis*, 3 edn, John Wiley, New York.

Marshall, A. W. & Olkin, I. (1997), 'A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families', *Biometrika* **84**(3), 641–652.

Miller, J. R. (1981), *Survival Analysis*, John Wiley, New York.

Information for Authors

The Colombian Journal of Statistics publishes original articles of theoretical, methodological and educational kind in any branch of Statistics. Purely theoretical papers should include illustration of the techniques presented with real data or at least simulation experiments in order to verify the usefulness of the contents presented. Informative articles of high quality methodologies or statistical techniques applied in different fields of knowledge are also considered. Only articles in English language are considered for publication.

The Editorial Committee assumes that the works submitted for evaluation have not been previously published and are not being given simultaneously for publication elsewhere, and will not be without prior consent of the Committee, unless, as a result of the assessment, decides not publish in the journal. It is further assumed that when the authors deliver a document for publication in the Colombian Journal of Statistics, they know the above conditions and agree with them `http://www.estadistica.unal.edu.co/revista`.

Material

All submitted articles to Colombian Journal of Statistics must be presented in pdf format, with text, figures and tables in black and white. All authors must submit a blinded and an unblinded version. Also, sign a letter by each author, where they all accept the submitting rules of Colombian Journal of Statistics (`http://www.ciencias.unal.edu.co/publicaciones/estadistica/rce/CartaArticulos.txt`). Once an article is accepted for publication, authors must send to the editorial board the following files: sources in LaTeX and figures in pdf, tiff, gif, png eps format, all of them in black and white.

To ease the preparation of authors' material we recommend the use of Miktex and the *revcoles* template, available in our website.

Every article must include:

- Title, both in English and Spanish
- Each author's full names, email address, as well as affiliation
- Abstract both in English and Spanish (*Resumen*). The abstract should not be longer than 200 words. The first 100 words should succinctly describe the paper's motivation and contribution.
- Three (3) to six (6) keywords using the Current Index to Statistics (CIS), both in English and Spanish
- If the article is based on a undergraduate thesis, master thesis or a doctoral dissertation, it should be included as a reference
- If the article is based on a research project, it must include the title of the project and the grant that supported it.
- All references must be cited in the text

References and footnotes

If two or more works by the same author or team of authors have the same publication date, list them by order of appearance in the text and distinguish them by lowercase “a”, “b”, and so on, after the date: “(1970a)”. The Harvard [4] package does it automatically.

Works accepted for publication but not off press are listed as “in press” instead of the anticipated date of publication; this may be changed on page proofs if the work comes off press by that time.

We recommend minimizing the use of footnotes and discourage the use of those that cite another footnote.

Figures and Tables

Both figures and tables should be numbered consecutively with arabic numerals, and clearly titled and labeled, and the identifying number must be cited in the text. Once the article is accepted the tables should be designed so they fit the printing area of Colombian Journal of Statistics. That includes the contents of tables, their length, number of representative digits, tiles, subtitles, labels and footnotes.

Figures must be visually clear and capable of withstanding reduction. All elements of figures and tables such as bars, segments, words, symbols and numbers must be printed in black.

Mathematical Material

Numbered mathematical expressions should be typed and centered on a separate line and identified by consecutive arabic numerals in parentheses placed flush with the right margin. Short expressions requiring only one line should remain in the text unless there is need to refer to them elsewhere by number. Lengthy equations should be handled by the use of definitions or broken to conform to the column format.

Keep in mind that space is placed around all operation symbols and before and after function words such as log, sin, y ln [unless they precede or follow a parentheses, e.g., $\log(x + y)$].].

Editorial Style

In addition to content, manuscripts are judged on their clarity. Consequently, well-written and well-structured papers that will be of interest to a wide segment of the readership are preferred.

Although the production office does not undertake major revision or rewriting of manuscripts, it is our policy to copyedit all manuscripts accepted for publication in accordance with the accepted rules of correct grammar, usage, spelling, and punctuation. In addition, deleting redundant words and phrases and punctuation.

[4] http://tug.ctan.org/tex-archive/macros/latex/contrib/harvard

Avoid common problems of style:

1. Use quotation marks only when a standard term is used in a nonstandard way and to indicate the beginning and ending of a direct quotation.

2. Hyphens are used when two or more adjectives or an adjective and a noun together modify another noun; for example, *goodness-of-fit* test is the equivalent of *test for goodness of fit.* Most words with prefixes such as sub and non are not hyphenated, for example, *subtable*, *nonnormal.*

3. Italics are used to introduce important terms, when appropriate; they are to be used sparingly to indicate emphasis.

4. Abbreviations and acronyms should be minimized; those that are used are spelled out on their first appearances in the manuscript with the shortened form given in parentheses, for example, *best linear unbiased estimate* (BLUE).

5. Numbers under 10 are spelled out when they are not part of an equation or an expression containing symbols.

6. The sign % is always used when giving a specific percentage, for example, 23 %, not 23 percent. Otherwise use the word *percent.*

Manuscript Length

There is no maximum length for manuscripts, but it is much more difficult and time-consuming to get reviews for long manuscripts. An efficient writing style with selective use of tables and figures is appreciated. Most manuscripts accepted for publication have fewer than 30 double-spaced pages, including text, figures, tables, and references.

Data

Whenever a dataset is used, its source should be fully documented. When it is not practical to include the whole of a dataset in the paper, the paper should state how the complete dataset was obtained or built.

Results Based on Computation

Papers reporting results based on computation should provide enough information so that readers can evaluate the quality of the results. Such information includes estimated accuracy of results, as well as descriptions of pseudorandom-number generators, numerical algorithms, computers, programming languages, and major software components that were used.

Appendices

Lengthy technical portions of a manuscript should appear in a separate appendix to the manuscript.

Review Process

All articles are first reviewed by the Editorial Committee and afterwards assigned to specialized referees, which follows a double-blind process. The authors do not know the identities of the reviewers, and the referees do not know the names of the authors. Yet the editor is not blinded. Authors are solely responsible for removing clues about their identities, including references to unpublished work and sources of funding, from their manuscripts and supporting documents. The Editorial Committee decides whether to accept or deny articles, based on the review process.

Legal Responsibility

The authors assume full responsibility for the use of material with registered intellectual property such as figures, tables, photographs, etc.

The Colombian Journal of Statistics thanks the following people for their collaboration in Volume 36th (2013).

Alaa H Abdel Hamid, Ph.D.
Ana Ferreira, Ph.D.
Andre Khuri, Ph.D.
Andrea Cardini, Ph.D.
Andrew Gelman, Ph.D.
Arijit Chaudhuri, Ph.D.
Barry Arnold, Ph.D.
Barry Goodwin, Ph.D.
Biswabrata Pradhan, Ph.D.
Brij Behare Khare, Ph.D.
Broderick O. Oluyede, Ph.D.
Bruno Apolloni, Ph.D.
Celia María Fernandes, Ph.D.
Cem Kadilar, Ph.D.
Chang-lin Mei, Ph.D.
Ching-Wen Chen, Ph.D.
Choongrak Kim, Ph.D.
Chris Bambey Guure, Ph.D.
Claudia Neves, Ph.D.
David Cox, Ph.D.
Diana Bilkova, Ph.D.
Elizabeth Hashimoto, Ph.D.
Eloisa Diaz-Frances, Ph.D.
Enrico Colossimo, Ph.D.
Ewa Gudowska, Ph.D.
Federico Márquez, Ph.D.
Francisco Torres Avilés, Ph.D.
G. Kumar Pandiyan, Ph.D.
Gilles Bourgault, Ph.D.
Gwo Dong Lin, Ph.D.
H. David Sheets, Ph.D.
Héctor Nuñez, Ph.D.
Héctor Zapata, Ph.D.
Hernan Tejeda, Ph.D.
Hongtu Zhu, Ph.D.
Housila P. Singh, Ph.D.
Huaihou Chen, Ph.D.
Hugo Salinas, Ph.D.
Ivana Mala, Ph.D.
Johannes Bausch, Ph.D.
John Hinde, Ph.D.
Jordan Stoyanov, Ph.D.
José Alfredo Jiménez, Ph.D.
José Casals, Ph.D.
Kalliopi Mylona, Ph.D.
Karla Muñoz, Ph.D.
Kiyoshi Taniguchi, Ph.D.
Kris Boudt, Ph.D.
Loane Muni Toke, Ph.D.
Luis Guillermo Díaz, M.Sc.
Luis Pericchi, Ph.D.
Luz Mery González, Ph.D.
Mahmoud Abd-Elfatahh, Ph.D.
Margaret A. Oliver, Ph.D.
Maria Castro, Ph.D.
Markus Pauly, Ph.D.
Mauricio Sadinle, Ph.D. (c)
Michael Smithson, Ph.D.
Miriam Zelditch, Ph.D.
Niel Hens, Ph.D.
Osvaldo Venegas, Ph.D.
Paulo Ramos, Ph.D.
Peter Kim, Ph.D.
Pierre Duchesne, Ph.D.
Raghaw Raman Sinha, Ph.D.
Rajesh Singh, Ph.D.
Rajesh Singh, Ph.D.
Raúl Alberto Pérez, Ph.D.
Reinaldo Arellano-Valle, Ph.D.
Reza Drikvandi, Ph.D.
Reza Modarres, Ph.D.
Rosemary Bailey, Ph.D.
Roxy Peck, Ph.D.
Rubén Guevara, Ph.D.
Sanjoy Sinha, Ph.D.
Santiago Medina Hurtado, Ph.D.
Serge Provost, Ph.D.
Silvia Ferrari, Ph.D.
Simon Price, Ph.D.
Simos Meintanis, Ph.D.
Svetlana Borovkova, Ph.D.
Yen Chang Chang, Ph.D.
Eugen Ursu, Ph.D.
Ying Yuan, Ph.D.
Zaizai Yan, Ph.D.
Zhimin Hong, Ph.D.
Zhu Wang, Ph.D.