CHILEAN JOURNAL OF **STATISTICS** Volume 11 Number 1 - April 2020

www.soche.cl/chjs

Volume 11 Number 1 - April 2020

CHILEAN JOURNAL OF **STATISTICS**

CHILEAN JOURNAL OF STATISTICS

Edited by Víctor Leiva and Carolina Marchant

Volume 11 Number 1
April 2020
ISSN: 0718-7912 (print)
ISSN: 0718-7920 (online)

Published by the
Chilean Statistical Society
SOCHE
SOCIEDAD CHILENA DE ESTADÍSTICA

Aims

The Chilean Journal of Statistics (ChJS) is an official publication of the Chilean Statistical Society (`www.soche.cl`). The ChJS takes the place of *Revista de la Sociedad Chilena de Estadística*, which was published from 1984 to 2000.

The ChJS is an international scientific forum strongly committed to gender equality, open access of publications and data, and the new era of information. The ChJS covers a broad range of topics in statistics, data science, data mining, artificial intelligence, and big data, including research, survey and teaching articles, reviews, and material for statistical discussion. In particular, the ChJS considers timely articles organized into the following sections: Theory and methods, computation, simulation, applications and case studies, education and teaching, development, evaluation, review, and validation of statistical software and algorithms, review articles, letters to the editors.

The ChJS editorial board plans to publish one volume per year, with two issues in each volume. On some occasions, certain events or topics may be published in one or more special issues prepared by a guest editor.

Chilean Journal of Statistics

Volume 11, Number 1
April 2020

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

Contents

Chilean Journal of Statistics
Vol. 11, No. 1, April 2020, 1–2

Eleventh Volume – First Number
Editorial Paper

Starting a new decade of the Chilean Journal of Statistics in COVID-19 pandemic times with new editors-in-chief

Welcome to the first issue of the eleventh volume of the Chilean Journal of Statistics (ChJS). Today, April 29, 2020, the ChJS celebrates eleven years of life in a historic period marked by the uncertainty generated by the global pandemic due to COVID-19. Pandemics like this have been in history, but it is the first time that we have lived in a status of global quarantine. In these pandemic times, much of humanity is in isolation and social distancing. We will certainly overcome this situation and one of the keys to do this is science and the generation of accurate knowledge.

For this volume, the ChJS would be nothing without the valuable contributions of renowned international researchers who have honored us by publishing their interesting works in our journal; all of these papers are available for free at http://chjs.mat.utfsm.cl/issues.html. We also thank all the anonymous reviewers who have contributed to keeping the top quality standards of the ChJS.

Although the ChJS is published by the Chilean Statistical Society (www.soche.cl) and belongs to the Chilean statistical community, our journal can be recognized as an international publication since its editorial board is composed of colleagues from practically the five continents. Our current Editorial Board, presented at http://chjs.mat.utfsm.cl/board.html, is a mixture of experienced editors and talented young researchers, the latter mainly from Chile and Brazil, who with great interest and enthusiasm have honored us by accepting to be part of the ChJS. They are having their first editorial experiences, although they all have extensive experience as researchers as well as reviewers for prestigious international journals.

We would also like to thank the members of the Directory of the Chilean Statistical Society (https://soche.cl/quienes-somos) headed by its President, Dr. Jorge Figueroa and Directors Danilo Alvares, Eduardo Alarcón, Carolina Marchant, Yolanda Gómez, Soledad Estrella, Tarik Faouzi and Guido del Pino for ratifying the former Editor-in-Chief of the ChJS and also naming the new Editor-in-Chief. The current Editors-in-Chief will do our best to bring ChJS to the highest standards of professionalism, fairness and quality that all scientific journals must strive for.

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

After this presentation note, the first issue of the eleventh volume of the ChJS comprises four papers authored by researchers from Brazil, Colombia, India, Kenya, Mexico, and US. Our first paper is authored by Luz Milena Zea Fernández and Thiago A.N. de Andrade. The authors derived new unconditioned and log-linear regression models based on the erf-G family of distributions. The second paper is authored by Thodur Parthasarathy Sripriya, Mamandur Rangaswamy Srinivasan, and Meenakshisundaram Subbiah, who presented a methodology for detecting outliers in $I \times J$ contingency tables through the level of susceptibility, a useful methodology for categorical data. In the third paper, Adolphus Wagala introduced a likelihood ratio test for correlated paired multivariate samples. The fourth paper is authored by Josmar Mazucheli, Sudeep R. Bapat, and André Felipe B. Menezes, who developed a new one-parameter unit-Lindley distribution and its application.

Finally, we would like the Chilean statistical community, as well as the international statistical community, our prestigious Editorial Board and past authors to champion ChJS as an emerging international journal and to encourage others to submit new works to the ChJS. Currently, we are indexed by several international systems, including the Institute for Scientific Information (ISI) Web of Science in the Emerging Sources Citation Index. The ChJS faces important challenges for the near future, such as reaching the Science Citation Index and looking for partnerships with prestigious publishers, societies and associations. However, just as with statistics itself, our success will depend on a team effort. Each one of us is important in meeting these challenges. We need you all.

Carolina Marchant and Víctor Leiva
Editors-in-Chief
Chilean Journal of Statistics

Chilean Journal of Statistics
Vol. 11, No. 1, April 2020, 3–23

Statistical Modeling
Research Paper

The erf-G family: new unconditioned and log-linear regression models

Luz Milena Zea Fernández[1] and Thiago A. N. de Andrade[2,*]

[1]Department of Statistics, Federal University of Rio Grande do Norte, Natal, Brazil,
[2]Department of Statistics, Federal University of Pernambuco, Recife, Brazil,

(Received: 08 March 2019 · Accepted in final form: 04 June 2019)

Abstract

In this paper, we propose a new generator of distributions called the erf-G family. Our proposal provides special distributions without adding complexity to parametric spaces of resulting models. We also furnish empirical evidence that the proposed family may solve issues of flat or quasi-red likelihoods in some baselines. In particular, we detail six special models from the erf-G family. We also derive a new log-linear regression model considering a kind of censoring. We discuss censored and uncensored maximum likelihood estimation methods for the proposed models. In order to study asymptotic properties of considered estimators, we carry out a Monte Carlo simulation study. Finally, using applications to real data we illustrate that proposed models may outperform classic lifetime models.

Keywords: Error function · Flat likelihood · Generalized distributions · Log-linear regression models.

Mathematics Subject Classification: Primary 60E10 · Secondary 60E05.

1. Introduction

From both theoretical and applied perspectives, the proposal of new probability distributions is crucial to describe natural phenomena. There are several ways to extend well-known distributions. One of the most popular ways is to consider distribution generators. Some of them are: Marshall-Olkin (Marshall and Olkin, 1997), beta (Eugene et al., 2002), gamma (Zografos and Balakrishnan, 2009), (Ristic and Balakrishnan, 2012) and (Nadarajah et al., 2015), Kumaraswamy (Cordeiro and de Castro, 2011), exponentiated generalized (Cordeiro et al., 2013), red odd exponentiated half-logistic (Afify et al., 2017) classes of models, among others.

Several generators (beyond of these referred above) have provided models more flexible than classic ones, used widely in applications into the lifetime context. However, from a literature review, such generators have the disadvantage of adding complexity to

*Corresponding author. Email: thiagoan.andrade@gmail.com

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

the parametric space of resulting models. In this paper, we use the error function (erf) as a way to outperform this issue. The erf (also known as Gauss error function) is an important special function, that appear often as solutions from several mathematical and physical problems. Its applications include probability theory, statistics, mass and momentum transfer, branches of mathematical physics, partial differential equations describing diffusion process, among others. For more details, we refer to Chevillard (2012).

We propose and study the erf-G family in details. Some of erf-G special cases are introduced and discussed. We derive explicit expressions for some of its mathematical properties and also propose a log-linear regression (llr) model with log-erf$G$ response variables. A discussion about estimation and hypothesis inference is furnished for both proposed unconditioned and llr models. Simulations results and two applications to real data indicates that our proposals may outperform well-defined lifetime models. We also highlight that our study of the erf-G model has very clear and forceful motivations: (i) it does not impose more complex parametric spaces to resulting models; (ii) it may provide concavity to distributions with flat or quasi-flat likelihoods (details are explored in Section 3); and (iii) it can generate bathtub failure rate functions. For the reasons listed above, we strongly believe it is important to study in detail the erf-G distribution. We hope that this new distribution is part of the arsenal of applied researchers and will be used in many practical situations.

This paper is organized as follows. In Section 2, we define some erf-G special models. Inferential tools, including: (i) linear representations for the erf-G probability density function (PDF) and cumulative distribution function (CDF), (ii) estimation and hypotheses inference procedures and (iii) regression models, are provided in Section 3. Mathematical properties of the new family are presented in Section 4. Simulations and applications to real data are provided in Section 5. In Section 6, main conclusions are listed.

2. Genesis of the new model and some of its special models

In this Section, we present the design of the new model and some of its many special models.

2.1 General context

First we consider the traditional error function given by

$$\operatorname{erf}(z) = \frac{1}{\sqrt{\pi}} \int_{-z}^{z} \exp(-t^2)\,\mathrm{d}t = \frac{2}{\sqrt{\pi}} \int_{0}^{z} \exp(-t^2)\,\mathrm{d}t, \quad z \in \mathrm{R}. \tag{1}$$

From now on, we advocate that replacing z in (1) by $G(x)/[1-G(x)]$ for $x \in \mathcal{X} \subset \mathrm{R}$ collapses a new and efficient generator of distributions. Let $G(x)$ be a cumulative distribution function (CDF). The following operator may be considered as the CDF of a potential family of models:

$$F(x) = \operatorname{erf}\left[\frac{G(x)}{1-G(x)}\right], \, x \in \mathcal{X}. \tag{2}$$

We denote this case as the erf-G family. A stochastic conception of this class which may furnish insight about the relation between new erf-G models and their respective baselines (with CDF G) is given by the following theorem.

Theorem *Let $Z > 0$ be a random variable with CDF given by $F_Z(z) = \text{erf}(z)\,\text{I}_{(0,\infty)}(z)$. Thus, $X = G^{-1}\left[Z\,(1+Z)^{-1}\right]$ is a stochastic transformation having CDF*

$$F_X(x) = \text{erf}\left[\frac{G(x)}{1-G(x)}\right],$$

where $G(x)$ represents the CDF of a baseline distribution.

The proof of this theorem holds from the basic probability manipulations. It reveals that distributions into the new family can understood as a quantile of $Y \sim G$ associated with a mapping $\mathcal{X} \to (0,1)$.

Now, let $X \sim \text{erf-}G(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \text{R}^p$, where $\boldsymbol{\Theta}$ represents the parametric space. The PDF of X and hazard rate function (HRF) are given respectively by

$$f(x) = \frac{2g(x;\boldsymbol{\theta})\exp\left[-\left(\frac{G(x;\boldsymbol{\theta})}{1-G(x;\boldsymbol{\theta})}\right)^2\right]}{\sqrt{\pi}(1-G(x;\boldsymbol{\theta}))^2}, \quad x \in \text{R}. \tag{3}$$

and

$$h(x) = \frac{2g(x;\boldsymbol{\theta})\exp\left[-\left(\frac{G(x;\boldsymbol{\theta})}{1-G(x;\boldsymbol{\theta})}\right)^2\right]}{\sqrt{\pi}(1-G(x;\boldsymbol{\theta}))^2\left\{1-\text{erf}\left[\frac{G(x;\boldsymbol{\theta})}{1-G(x;\boldsymbol{\theta})}\right]\right\}}, \quad x \in \text{R}.$$

2.2 Some special models

The erf-G model is completely new. There is, therefore, a great variety of new distributions, based on (2), that can be explored by statisticians and applied researchers. In what follows, we discuss some special models.

2.2.1 The erf-Gumbel model

The Gumbel distribution is a statistical model defined in real support widely used in engineering problems (de Andrade et al., 2015). Its CDF is given by $G(x;\mu,\sigma) = \exp\{-\exp[-(x-\mu)/\sigma]\}$, where $-\infty < \mu < \infty$ and $\sigma > 0$ are the location and scale parameters, respectively. Applying its CDF and PDF in (2) and (3), we obtain the erf-Gumbel (erfGum) model, having CDF and PDF given by

$$F(x) = \text{erf}\left\{\frac{1}{\exp[z_1(x)]-1}\right\}, \quad x \in \text{R},$$

and

$$f(x) = \frac{2z_1(x)\exp\left\{-z_1(x)-\left[\frac{1}{\exp[z_1(x)]-1}\right]^2\right\}}{\sqrt{\pi}\sigma\{1-\exp[z_1(x)]\}^2}, \quad x \in \text{R},$$

respectively, where $z_1(x) = \exp\left[-(x-\mu)/\sigma\right]$. Figure 1 presents erfGum PDF curves for some selected parameters. The Gumbel distribution is asymmetric. As we can see in the Figure 1, the erfGum model can accommodate asymmetric shapes.

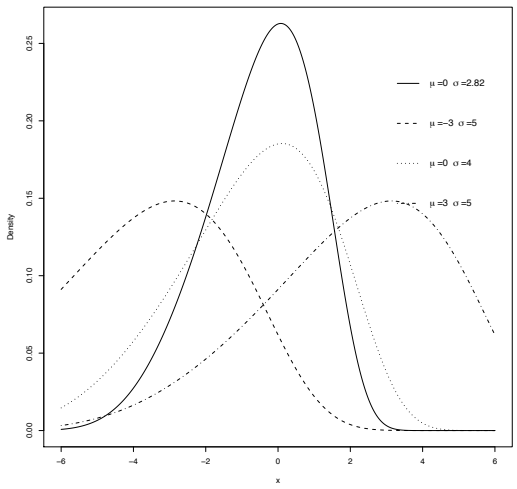


Figure 1. The PDF of the erfGumbel model for some σ and μ parameter values.

2.2.2 The erf-normal model

Let ϕ and Φ be the PDF and CDF of the standard normal model, respectively. Evaluating these equation in (2) and (3), we obtain the erf-normal (erfN) model, with CDF and PDF expressed by

$$F(x) = \text{erf}\left[\frac{\Phi(z_2(x))}{\Phi(-z_2(x))}\right], \quad x \in \text{R},$$

and

$$f(x) = \frac{\sqrt{2}\exp\left\{-z_2(x)^2/2 - [\Phi(z_2(x))/\Phi(-z_2(x))]^2\right\}}{\pi[\Phi(-z_2(x))]^2}, \quad x \in \text{R},$$

where $z_2(x) = (x-\mu)/\sigma$. Plots for the erfN PDF at selected parameter values are displayed in Figure 2. Based on Figure 2, likewise that the erfGum, the erfN distribution may present asymmetrical behaviour in contrast with its baseline.

2.2.3 The erf-gamma model

As third special model, applying gamma model (having shape α and scale β) CDF and PDF in (2) and (3), we get the erf-gamma (erfΓ) model with CDF and PDF expressed as

$$F(x) = \text{erf}\left[\frac{\gamma(\alpha,\beta x)}{\Gamma(\alpha,\beta x)}\right], \quad x > 0,$$

where $\Gamma(s,x) = \int_x^\infty \text{t}^{\text{s}-1}\exp(-\text{t})\,\text{dt}$ and $\gamma(s,x) = \int_0^x \text{t}^{\text{s}-1}\exp(-\text{t})\,\text{dt}$ are the upper and lower incomplete gamma functions, and

$$f(x) = \frac{2\beta^\alpha\Gamma(\alpha)x^{\alpha-1}\exp\left[-\left(\frac{\gamma(\alpha,\beta x)}{\Gamma(\alpha,\beta x)}\right)^2 - \beta x\right]}{\sqrt{\pi}\,[\Gamma(\alpha,\beta x)]^2}, \quad x > 0,$$

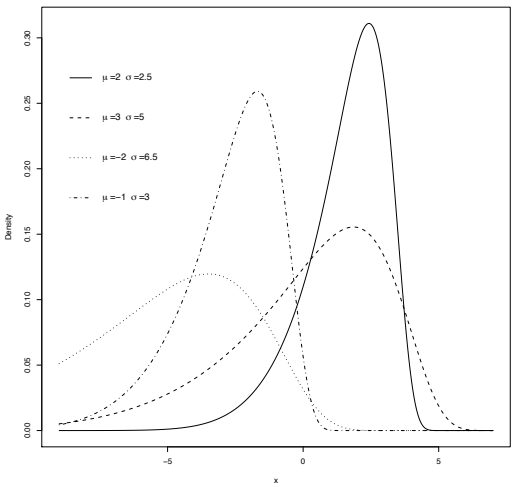
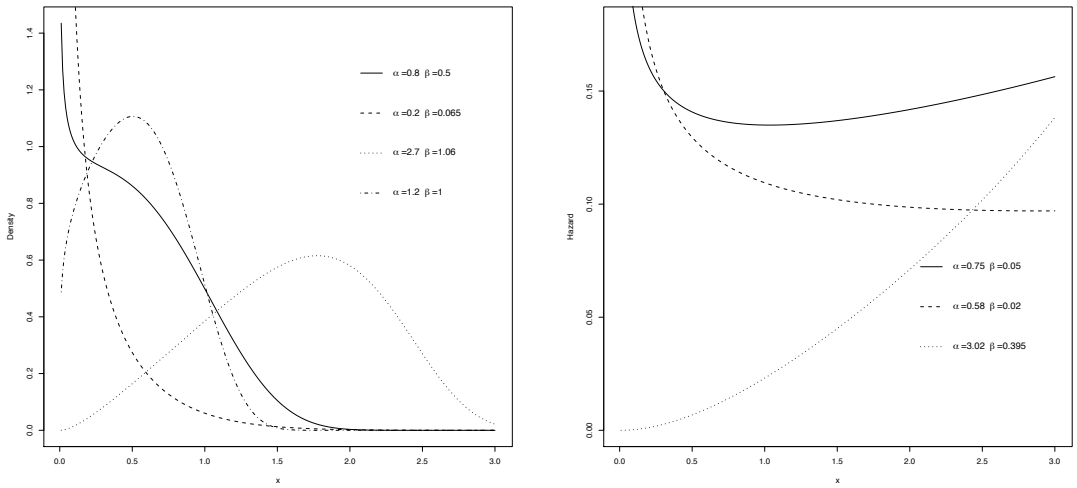


Figure 2. The PDF of the erfN model for some σ and μ parameter values.

where Γ represents the gamma function. The HRF of the erfΓ model is defined by

$$h(x) = \frac{2\beta^{\alpha}\Gamma(\alpha)x^{\alpha-1}\exp\left[-\left(\dfrac{\gamma(\alpha,\beta x)}{\Gamma(\alpha,\beta x)}\right)^2-\beta x\right]}{\sqrt{\pi}\left[\Gamma(\alpha,\beta x)\right]^2\left\{1-\operatorname{erf}\left(\dfrac{\gamma(\alpha,\beta x)}{\Gamma(\alpha,\beta x)}\right)\right\}}, \quad x>0.$$

Plots of the erfΓ PDF and HRF for selected parameter values are presented in Figure 3. At least, the associated HRF can assume bathtub, increasing and decreasing shapes. In contrast with the gamma model, which assumes only monotone HRF shapes.



(a) The PDF of the erfΓ model (b) The HRF of the erfΓ model

Figure 3. The PDF and HRF of the erfΓ model for some α and β parameter values.

2.2.4 The erf-Weibull model

The Weibull distribution can be considered as a standard model for lifetime data and, therefore, is interesting to study a special model generated from it. From evaluating Weibull

CDF and PDF in (2) and (3), we obtain the erf-Weibull (erfW) model, characterized by CDF and PDF given by

$$F(x) = \operatorname{erf}\left[\exp\left(\alpha x^{\beta}\right) - 1\right], \quad x > 0,$$

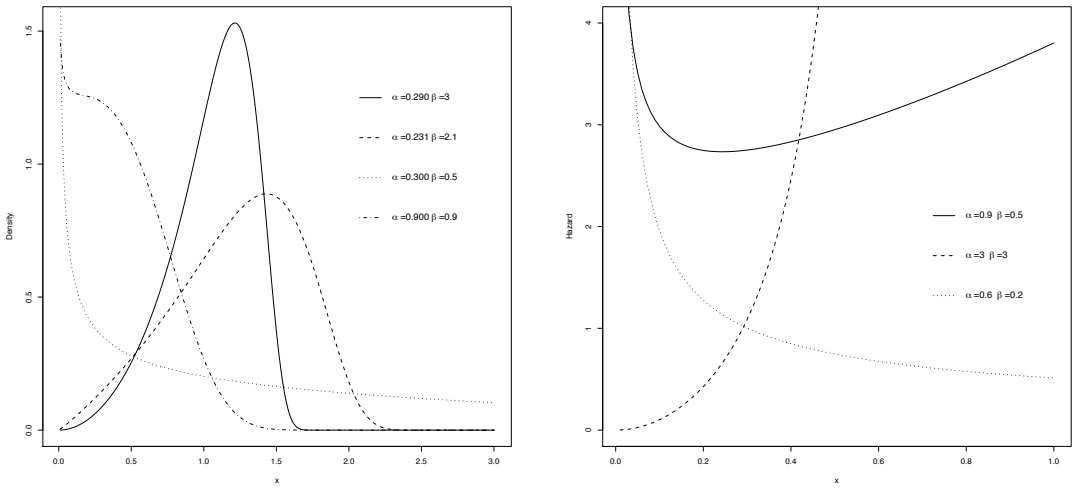
and

$$f(x) = 2\pi^{-1/2}\,\alpha\,\beta\,x^{\beta-1}\exp\left[\alpha x^{\beta} - \left(\exp(\alpha x^{\beta}) - 1\right)^2\right], \quad x > 0.$$

The erfW hazard can be expressed as

$$h(x) = \frac{2\,\alpha\,\beta\,x^{\beta-1}\exp\left[\alpha x^{\beta} - \left(\exp(\alpha x^{\beta}) - 1\right)^2\right]}{\sqrt{\pi}\,\{1 - \operatorname{erf}\left[\exp(\alpha x^{\beta}) - 1\right]\}}, \quad x > 0.$$

Plots of the erfW PDF for selected parameter values are displayed in Figure 4. This figure provides possible shapes of the erfW HRF, which includes the bathtub shape. It represents a gain on the Weibull model, which has constant and monotone shapes.



(a) The PDF of the erfW model (b) The HRF of the erfW model

Figure 4. The PDF and HRF of the erfW model for some α and β parameter values.

2.2.5 The erf-log-logistic distribution

In the survival analysis context, the log-logistic distribution is one of the possible choices when you want to model data with a unimodal failure rate. For $x > 0$, the CDF of the log-logistic model is given by $G(x;\alpha,\,\beta) = 1 - \left[1 + \left(\frac{x}{\alpha}\right)^{\beta}\right]^{-1}$, where $\alpha > 0$ and $\beta > 0$ are shape parameters. Thus, the CDF and PDF regard to the erf-log-logistic (erfLL) distribution are given by

$$F(x) = \operatorname{erf}\left[\left(\frac{x}{\alpha}\right)^{\beta}\right], \quad x > 0,$$

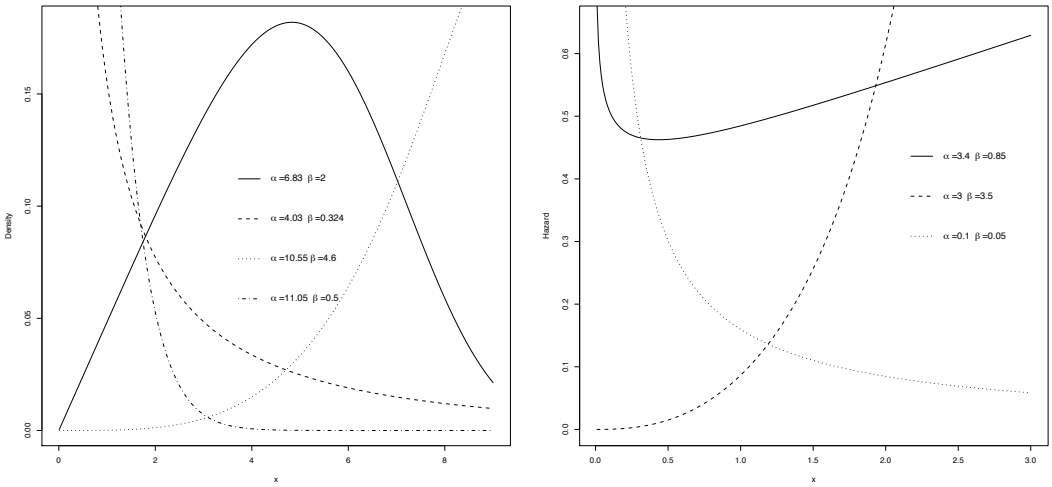
and

$$f(x) = \frac{2\,\beta\,x^{\beta-1}}{\sqrt{\pi}\,\alpha^{\beta}}\,\exp\left[-\left(\frac{x}{\alpha}\right)^{2\beta}\right], \quad x > 0.$$

The HRF of the erfLL distribution is easily defined as

$$h(x) = \frac{2\,\beta\,x^{\beta-1}\exp\left[-\left(\frac{x}{\alpha}\right)^{2\beta}\right]}{\sqrt{\pi}\,\alpha^{\beta}\left\{1-\text{erf}\left[\left(\frac{x}{\alpha}\right)^{\beta}\right]\right\}}, \quad x > 0.$$

Plots of the erfLL PDF for selected parameter values are displayed in Figure 5. Figure 5 also provides some possible shapes of the erfLL hazard function for appropriate parameter values, including bathtub, increasing and decreasing shapes. These plots indicate that the erfLL model is fairly flexible and can be used to fit several types of positive data.



(a) The PDF of the erfLL model (b) The HRF of the erfLL model

Figure 5. The PDF and HRF of the erfLL model for some α and β parameter values.

2.2.6 The erf-Frechet distribution

The CDF of the Frechet model is given by $G(x;\delta,\lambda) = \exp(-\delta^{\lambda}x^{-\lambda})$ for $x > 0$ and $\delta, \lambda > 0$. An important generalization based on this distribution was proposed by da Silva et al. (2013). Considering $G(x)$ as the Frechet CDF in equations (2) and (3), we get the erf-Frechet (erfF) model with CDF and PDF expressed as

$$F(x) = \text{erf}\left[\left(\exp(\delta^{\lambda}x^{-\lambda}) - 1\right)^{-1}\right]$$

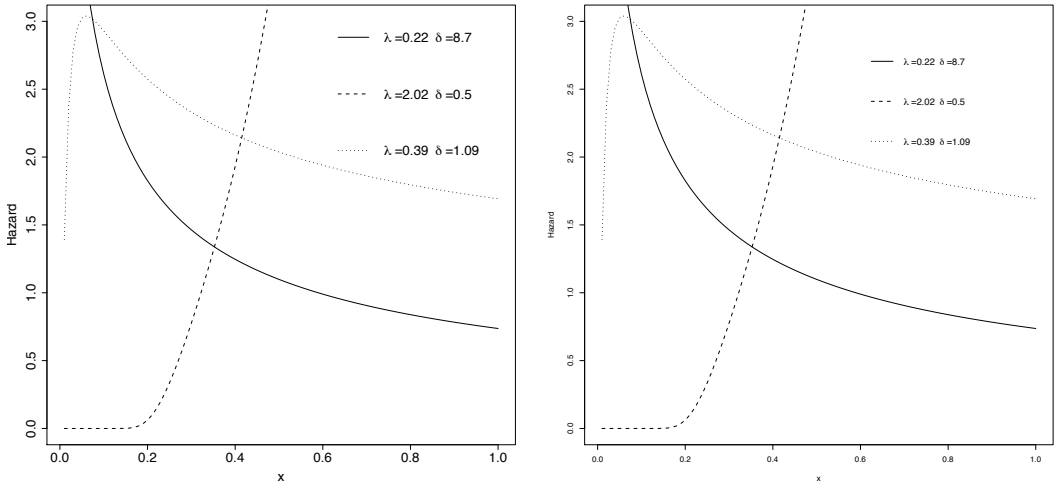
and

$$f(x) = \frac{2\lambda\,\delta^{\lambda}\,x^{-\lambda-1}\exp\left\{-\delta^{\lambda}x^{-\lambda} - \left[\exp(\delta^{\lambda}x^{-\lambda}) - 1\right]^{-2}\right\}}{\sqrt{\pi}\left[1-\exp(-\delta^{\lambda}x^{-\lambda})\right]^{2}}. \tag{4}$$

The risk function associated appears as

$$h(x) = \frac{2\lambda\,\delta^{\lambda}\,x^{-\lambda-1}\exp\left\{-\delta^{\lambda}x^{-\lambda} - \left[\exp(\delta^{\lambda}x^{-\lambda}) - 1\right]^{-2}\right\}}{\sqrt{\pi}\left[1-\exp(-\delta^{\lambda}x^{-\lambda})\right]^{2}\left\{1 - \operatorname{erf}\left[(\exp(\delta^{\lambda}x^{-\lambda}) - 1)^{-1}\right]\right\}}.$$

Some plots for the erfF PDF and HRF are provide in Figure 6. The erfF HRF covers the inverted bathtub shape in contrast with the Frechet HRF, that assumes only monotone behavior.



(a) The PDF of the erfF model (b) The HRF of the erfF model

Figure 6. The PDF and HRF of the erfF model for some λ and δ parameter values.

3. Miscellaneous

In this Section, we provide a complete background for inferential processes.

3.1 A linear expansion

General expressions for the PDF and CDF functions are highly appreciated by applied researchers, as they allow approximate results to be obtained when analytical solutions are not available. Here, we refer to some works that consider these expansions: Cordeiro et al. (2015), Leao et al. (2013), de Andrade et al. (2016) and Afify et al. (2017). This section aims to provide expansions for (2) and (3) in order to determine representations for some erf-G mathematical properties, which do not present closed-forms. First, consider the Maclaurin expansion for the erf function given by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\sum_{k=0}^{\infty}\frac{(-1)^{k}x^{2k+1}}{k!(2k+1)}. \tag{5}$$

By applying (5) in (2), one has that

$$F(x) = \frac{2}{\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k \left[\frac{G(x)}{1-G(x)}\right]^{2k+1}}{k!(2k+1)}. \tag{6}$$

From the Taylor expansion, we have

$$\frac{x}{1-x} = \sum_{i=1}^{\infty} x^i \quad \text{for} \quad |x| < 1, \tag{7}$$

(7) applied in (6) collapses

$$F(x) = \frac{2}{\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(2k+1)} \left[\sum_{i=1}^{\infty} G(x)^i\right]^{2k+1}. \tag{8}$$

Setting ℓ as a positive integer number, we have

$$\left(\sum_{k=0}^{\infty} a_k x^k\right)^{\ell} = \sum_{m=0}^{\infty} c_{\ell,m} x^m, \tag{9}$$

where

$$c_{\ell,_0} = a_0^{\ell}, \quad c_{\ell,m} = \frac{1}{m\, a_0} \sum_{j=1}^{m} (j\ell - m + j) a_j c_{\ell,m-j}, \quad m \geq 1.$$

From (9) in (8), we get

$$F(x) = \frac{2}{\sqrt{\pi}} \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \frac{(-1)^k d_{2k+1,m}}{k!(2k+1)} G(x)^{m+2k+1} = \sum_{k,m=0}^{\infty} b_{k,m}\, G(x)^{m+2k+1}, \tag{10}$$

where $d_{2k+1,0} = 1, d_{2k+1,m} = \frac{1}{m} \sum_{j=1}^{m} \big[2j(k+1) - m\big] d_{2k+1,m-j}, m \geq 1$ and

$$b_{k,m} = \frac{2\,(-1)^k\, d_{2k+1,m}}{\sqrt{\pi}\, k!\, (2k+1)}.$$

By applying the derivate with respect to x in (10), erf-G PDF can be express as

$$f(x) = \sum_{k,m=0}^{\infty} b_{k,m}\, (m+2k+1)\, g(x)\, G(x)^{m+2k} = \sum_{k,m=0}^{\infty} a_{k,m}\, g(x)\, G(x)^{m+2k}, \tag{11}$$

where $a_{k,m} = b_{k,m}\,(m+2k+1)$. Equations (10) and (11) indicate that erf-G random variables can be represented as a linear combination of exp-G distributions (discussed in detailed by Tahir and Nadarajah (2015)) having additional parameter $m+1$.

3.2 Maximum likelihood estimation

Let $x_1, \ldots, x_n$ be a n-points observed sample obtained from $X \sim \text{erf}G(\boldsymbol{\theta})$. The log-likelihood function for the vector of parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathrm{R}^p$ is expressed as

$$\ell(\boldsymbol{\theta}) = n \log\left(\frac{2}{\sqrt{\pi}}\right) + \sum_{i=1}^{n} \log\left[g(x_i|\boldsymbol{\theta})\right] - 2\sum_{i=1}^{n} \log\left[1 - G(x_i|\boldsymbol{\theta})\right] - \sum_{i=1}^{n} \frac{G(x_i|\boldsymbol{\theta})^2}{[1 - G(x_i|\boldsymbol{\theta})]^2}, \tag{12}$$

In this case, the jth element of the score vector, $\mathbf{U}(\boldsymbol{\theta}) = [U_1(\boldsymbol{\theta}), \ldots, U_p(\boldsymbol{\theta})]^\top = \left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_p}\right]^\top$, is given by

$$\begin{aligned} U_j(\boldsymbol{\theta}) = &\sum_{i=1}^{n} \frac{\dot{g}(x_i|\boldsymbol{\theta})}{g(x_i|\boldsymbol{\theta})} + 2\sum_{i=1}^{n} \frac{\dot{G}(x_i|\boldsymbol{\theta})}{[1 - G(x_i|\boldsymbol{\theta})]} - 2\sum_{i=1}^{n} \frac{G(x_i|\boldsymbol{\theta})\dot{G}(x_i|\boldsymbol{\theta})[1 - G(x_i|\boldsymbol{\theta})]^2}{[1 - G(x_i|\boldsymbol{\theta})]^4} \\ &- 2\sum_{i=1}^{n} \frac{G(x_i|\boldsymbol{\theta})^2\dot{G}(x_i|\boldsymbol{\theta})[1 - G(x_i|\boldsymbol{\theta})]}{[1 - G(x_i|\boldsymbol{\theta})]^4}, \end{aligned}$$

where $\dot{g}(x_i|\boldsymbol{\theta}) = \partial g(x_i;\boldsymbol{\theta})/\partial\theta_j$ and $\dot{G}(x_i|\boldsymbol{\theta}) = \partial G(x_i;\boldsymbol{\theta})/\partial\theta_j$. Thus, the maximum likelihood estimator (ML estimator) are given by

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\theta\in\Theta}\{\ell(\boldsymbol{\theta})\}$$

or, equivalently, $\hat{\boldsymbol{\theta}}$ is a root of the non-linear equations system defined by $\mathbf{U}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$.

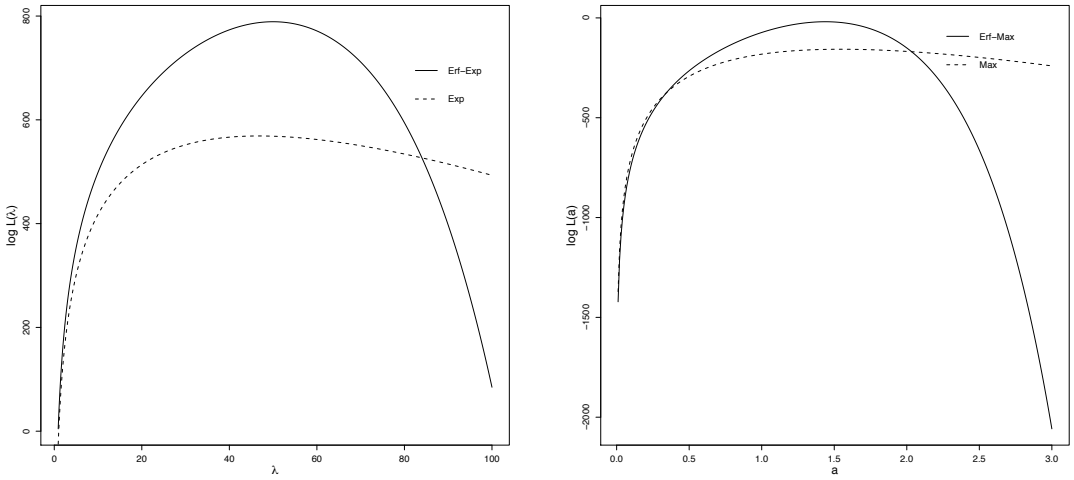
To illustrate as the erf-G model can modify geometrically a G distribution log-likelihood, we compare two pairs of distributions: (exponential (Exp), erf-exponential (erfExp)) and (Maxwell (Max), erf-Maxwell (erfMax)). The erfExp log-likelihood function is given by

$$\ell(\lambda) = n \log\left(\frac{2\lambda}{\sqrt{\pi}}\right) + \lambda\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}(1 - \mathrm{e}^{\lambda x_i}).$$

From Figure 7, it is noticeable that the erf-G structure may provide concavity to distributions with flat or quasi-flat likelihoods. It advocates in favor of the proposed family. Among other advantages, a greater concavity of likelihood provides better quality in the estimation process. In the next section, we illustrate that the maximum likelihood estimates (ML estimates) based on (12) may be more accurate than those obtained from the corresponding baseline.

3.3 The log-erf-Frechet regression model

In several applications, lifetimes are related to exatory variables. Regression models are sought for this end. Let T be a random variable with PDF (4), then $Y = \log(T)$ has the log-erf-Frechet (lerfF) distribution, denoted as $Y \sim$ lerfF. Taking the parametrization $\delta = \exp(\mu)$ and $\lambda = 1/\sigma$, the PDF of Y can be written as



(a) The log-likelihood function for the Exp and erfExp distributions

(b) The log-likelihood function for the Max and erfMax distribution

Figure 7. The log-likelihood function for the Exp, erfExp, Max and erfMax distributions.

$$f(y,\mu,\sigma) = \frac{2}{\sqrt{\pi}\sigma}\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\exp\left\{\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}\left(\exp\left\{\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}-1\right)^{-2} \times \exp\left[-\left(\exp\left\{\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}-1\right)^{-2}\right], \tag{13}$$

for $-\infty < y < \infty$, $-\infty < \mu < \infty$ and $\sigma > 0$. Now, if $T \sim \text{erfF}(\delta, \lambda)$, then $Y = \log(T) \sim \text{lerfF}(\mu, \sigma)$ with CDF

$$F_Y(y) = \text{erf}\left[\left(\exp\left\{\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}\right)^{-1}\right],$$

and survival function (sf) given by

$$S(y;\mu,\sigma) = 1 - \text{erf}\left[\left(\exp\left\{\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}\right)^{-1}\right]. \tag{14}$$

Now, we are in position of defining the standardized random variable $Z = (Y-\mu)/\sigma$ with PDF

$$\pi(z) = \frac{2}{\sqrt{\pi}}\exp(-z)\exp\big[\exp(-z)\big]\big\{\exp[\exp(-z)]-1\big\}^{-2}\exp\big(-\{\exp[\exp(-z)]-1\}^{-2}\big). \tag{15}$$

Considering the substitution $u = \big\{\exp[\exp(-z)]-1\big\}^{-1}$, the r-th moment of Z is given by

$$\text{E}(Z^r) = \frac{2}{\sqrt{\pi}}\int_0^\infty \big\{-\log[\log(u^{-1}+1)]\big\}^r \exp(-u^2)\text{d}u.$$

Using the `Mathematica` software, it is possible to verify that the second ordinary moment of Z is finite:

$$\mathrm{E}(Z^2) = \frac{2}{\sqrt{\pi}} \int_0^\infty \left\{ -\log[\log(u^{-1}+1)] \right\}^2 \exp(-u^2)\mathrm{d}u = 0.321075 < \infty.$$

Let $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^\top$ be the exatory variable vector associated with the ith response variable Y_i for $i = 1, \ldots, n$.

Consider the sample $(Y_1, \boldsymbol{x_1}), \ldots, (Y_n, \boldsymbol{x_n})$ of n independent variables, where each random response is defined by $Y_i = \min\{\log(T_i), \log(c_i)\}$ and $\log(T_i)$ and $\log(c_i)$ are the log −lifetime and log −censoring, respectively. We consider non-informative censorship such that the lifetimes and censorship times are independent.

The linear regression model for the lerfF response variable, Y_i, is given by

$$Y_i = \boldsymbol{x_i}^\top \boldsymbol{\beta} + \sigma Z_i, \quad i = 1, 2, \ldots, n. \tag{16}$$

where Z_i is a random variable with PDF (15), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and $\sigma > 0$ are unknown parameters, and $\boldsymbol{x_i}$ is the ith explanatory random variables vector.

In this case, the location of $(Y_1, \ldots, Y_n)^\top$ is $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ such that $\mu_i = \boldsymbol{x_i}^\top \boldsymbol{\beta}$ or, in matrix terms, $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ with model matrix $\boldsymbol{X} = (x_1, \ldots, x_n)^\top$.

Let F and C be the sets of individuals for which y_i is the log −lifetime or log −censoring, respectively.

The total log-likelihood function for the parameters $\boldsymbol{\theta} = (\sigma, \boldsymbol{\beta}^\top)^\top$ of model (16) has the form

$$\ell(\boldsymbol{\theta}) = \sum_{i \in F} \ell_i(\boldsymbol{\theta}) + \sum_{i \in C} \ell_i^{(c)}(\boldsymbol{\theta}),$$

where $\ell_i(\boldsymbol{\theta}) = \log[f(y_i)]$, $\ell_i^{(c)}(\boldsymbol{\theta}) = \log[S(y_i)]$, $f(y_i)$ and $S(y_i)$ are given in equations (13) and (14). Then, the log −likelihood function reduces to

$$\begin{aligned}
\ell(\boldsymbol{\theta}) = {} & q\left(\log(2) - \frac{\log(\pi)}{2} - \log(\sigma)\right) + \sum_{i \in F} \left\{ -\left(\frac{y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta}}{\sigma}\right) + \exp\left[-\left(\frac{y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta}}{\sigma}\right)\right] \right. \\
& -2\log\left(\exp\left\{\exp\left[-\left(\frac{y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta}}{\sigma}\right)\right]\right\} - 1\right) - \left(\exp\left\{\exp\left[-\left(\frac{y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta}}{\sigma}\right)\right]\right\} - 1\right)^{-2} \Bigg\} \\
& + \sum_{i \in C} \log\left\{ 1 - \mathrm{erf}\left[\left(\exp\left\{\exp\left[-\left(\frac{y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta}}{\sigma}\right)\right]\right\}\right)^{-1} \right] \right\},
\end{aligned} \tag{17}$$

where q is the observed number of failures. The ML estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be obtained by maximizing the Equation (17). Using the adjusted model (16), the sf of Y_i can be estimated by

$$\widehat{S}(y_i; \widehat{\sigma}, \widehat{\boldsymbol{\beta}}^\top) = 1 - \mathrm{erf}\left[\left(\exp\left\{\exp\left[-\left(\frac{y_i - \boldsymbol{x_i}^\top \widehat{\boldsymbol{\beta}}}{\widehat{\sigma}}\right)\right]\right\}\right)^{-1} \right].$$

Under general regularity conditions, the asymptotic distribution of $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ can be

approximated by the multivariate normal $N_{p+1}(0, J(\boldsymbol{\theta})^{-1})$, where $J(\boldsymbol{\theta}) = \partial^2 \ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^\top \partial\boldsymbol{\theta}$ is the $(p+1)\times(p+1)$ observed information matrix. Statistical inference procedures for the parameter vector $\boldsymbol{\theta}$ can be made based on the asymptotic normality. In particular, an $100(1-\alpha)\%$ asymptotic confidence interval for each parameter θ_s is given by

$$\mathrm{ACI}_s = (\theta_s - z_{\alpha/2}\sqrt{\widehat{J}^{s,s}}, \theta_s + z_{\alpha/2}\sqrt{\widehat{J}^{s,s}}),$$

where $\widehat{J}^{s,s}$ denotes the sth diagonal element of the inverse of the estimated observed information matrix $J(\widehat{\boldsymbol{\theta}})^{-1}$ and $z_{\alpha/2}$ is the quantile $1-\alpha/2$ of the standard normal distribution.

4. Some mathematical properties

From now on, we present the process of obtaining the mathematical properties of the new model.

4.1 Quantile function

The quantile function (qf) of the erf-G distribution is obtained in an explicit form by inverting (2)

$$Q_F(u) = Q_G\left(\frac{\Phi^{-1}(\frac{u+1}{2})}{\sqrt{2}+\Phi^{-1}(\frac{u+1}{2})}\right), \tag{18}$$

where Q_G is the baseline quantile function and Φ^{-1} is the standard normal quantile function. Beyond to allow defining important quantiles (e.g., the median), (18) may also be used as a random variables generator, adopting uniform outcomes as inputs.

4.2 Ordinary and incomplete moments

Let X be a random variable following erf-G distribution. From Equation (11), the rth moment of X may be written as

$$\mathrm{E}(X^r) = \sum_{k,m=0}^{\infty} a_{m,k}\,\mathrm{E}(Y_{m+2k+1}^r),$$

where Y_{m+2k+1} follows the exponentiated distribution at the power parameter $m+2k+1$. Another way to represent the rth moment is through of the quantile function as follow:

$$\mathrm{E}(X^r) = \sum_{k,m=0}^{\infty} a_{m,k}\int_0^1 \left[Q_G\big(u^{\frac{1}{m+k+1}}\big)\right]^r \mathrm{d}u.$$

The rth incomplete moment of X can be given as follow

$$\mathrm{T}_r(z) = \int_{-\infty}^{z} x^r f(x)dx = \sum_{k,m=0}^{\infty} a_{m,k}\,\mathrm{T}_r^*(z),$$

where $\mathrm{T}_r^*(z)$ is the rth incomplete moment of the Y_{m+2k+1}. A second manner to obtain the rth incomplete moment of X is by using the quantile function, we have

$$\mathrm{T}_r(z) = \int_{-\infty}^{z} x^r f(x)dx = \sum_{m,k=0}^{\infty} a_{m,k} \int_0^{[G(z)]^{m+2k+1}} \left[Q_G\left(u^{\frac{1}{m+2k+1}}\right)\right]^r \mathrm{d}u.$$

4.3 Moment generating function

By using the Equation (11), the mgf of X can be expressed as

$$\mathrm{M}(t) = \sum_{m,k=0}^{\infty} b_{m,k}\, \mathrm{M}_{m+2k+1}(t),$$

where $\mathrm{M}_{m+2k+1}(t)$ is the mgf of Y_{m+2k+1} given by

$$\mathrm{M}_{m+2k+1}(t) = \int_{-\infty}^{\infty} \exp(tx)(m+2k+1)g(x)[G(x)]^{m+2k}\mathrm{d}x.$$

Another form to obtain an expansion of the mgf of X is by using the qf. We have

$$\mathrm{M}(t) = \sum_{m,k=0}^{\infty} (m+2k+1)\, b_{m,k} \int_0^1 \exp\left[t\, Q_G(u)\right] u^{m+2k}\mathrm{d}u.$$

4.4 Entropy

Two well-known variability measures are the Shannon and Rényi entropies. Determining their expressions consist an important task to quantify disorder in stochastic systems. In what follows, we derive these measures for the erf-G family. First consider the expansion: Assuming that $|z| < 1$ and $\rho > 0$,

$$(1-z)^{-\rho} = \sum_{j=1}^{\infty} w_j z_j, \quad w_j = \frac{\Gamma(\rho+j)}{j!\,\Gamma(\rho)}. \tag{19}$$

Considering the Taylor expansion and (19) an expression to the erf-G Rényi entropy is (for $\delta > 0$ and $\delta \neq 1$)

$$\begin{aligned}
\mathrm{I_R}(\delta) &= \frac{1}{1-\delta}\log\left(\int_0^{\infty}[f(x)]^{\delta}\mathrm{d}x\right) \\
&= \frac{1}{1-\delta}\log\left[\frac{2^{\delta}}{\pi^{\delta/2}}\sum_{k=0}^{\infty}\sum_{j=1}^{\infty}\frac{(-\delta)^k\, w_j}{k!}\int_0^{\infty}\left[g(x)\right]^{\delta}\left[G(x)\right]^{2k+j}\mathrm{d}x\right] \\
&= \frac{1}{1-\delta}\left\{\delta\log(2) - \frac{\delta}{2}\log(\pi) + \log\left(\sum_{k=0}^{\infty}\sum_{j=1}^{\infty}\frac{(-\delta)^k\, w_j}{k!}\int_0^{\infty}\left[g(x)\right]^{\delta}\left[G(x)\right]^{2k+j}\mathrm{d}x\right)\right\},
\end{aligned}$$

where $w_j = \dfrac{\Gamma[2(\delta+1)+j]}{j!\,\Gamma[2(\delta+1)]}$.

The Shannon entropy is defined as $\mathrm{E}\{-\log[f(X)]\}$ and it can be obtained from the Rnyi entropy doing $\delta \uparrow 1$. Note that

$$\mathrm{E}\big\{-\log[f(X)]\big\} = -2\log(2)+\frac{1}{2}\log(\pi)-\mathrm{E}\big[\log(X)\big]+\mathrm{E}\left\{\left[\frac{G(X)}{1-G(X)}\right]^2\right\}-2\mathrm{E}\big[\log(1-g(X))\big].$$

After some algebraic manipulations, we obtain

$$\mathrm{E}\big[\log(X)\big] = \sum_{m,k=0}^{\infty} b_{m,k}(m+2k+1)\int_0^1 u^{m+2k}\log\big[g(Q_G(u))\big]\mathrm{d}u,$$

$$\begin{aligned}\mathrm{E}\left\{\left[\frac{G(X)}{1-G(X)}\right]^2\right\} &= \int_{-\infty}^{\infty}\mathrm{E}\left\{\left[\frac{G(X)^2}{[1-G(X)]^2}\right]^2\right\}f(x)\mathrm{d}x \\ &= \sum_{m,k=0}^{\infty} b_{m,k}(m+2k+1)\int_0^1 \frac{u^{m+2k+2}}{(1-u)^2}\mathrm{d}u\end{aligned}$$

and

$$\mathrm{E}\big[\log(1-g(X))\big] = -\sum_{i=0}^{\infty}\sum_{m,k=0}^{\infty}\frac{b_{m,k}(m+2k+1)}{(i+1)(m+2k+i+2)}.$$

5. Numerical applications

In order to assess the performance of estimation procedures, we carry out a Monte Carlo study and two real data set applications.

5.1 A Monte Carlo study

This section aims to quantify the performance of ML estimators for erf-G parameters distribution. To that end, we consider the exponential (exp), Levy and Maxwell (Max) models, after we specify the following baseline models: erf{Exp, Levy, Max} using equation (3). The PDF's of the Exp, Levy and Max distributions are given, respectively, by

$$f(x,\lambda) = \lambda\exp(-\lambda x), \quad x>0, \quad \lambda>0,$$

$$f(x,\lambda) = \sqrt{\frac{\lambda}{2\pi}}\frac{\exp(-\frac{a}{2x})}{x^{\frac{3}{2}}}, \quad x>0, \quad \lambda>0$$

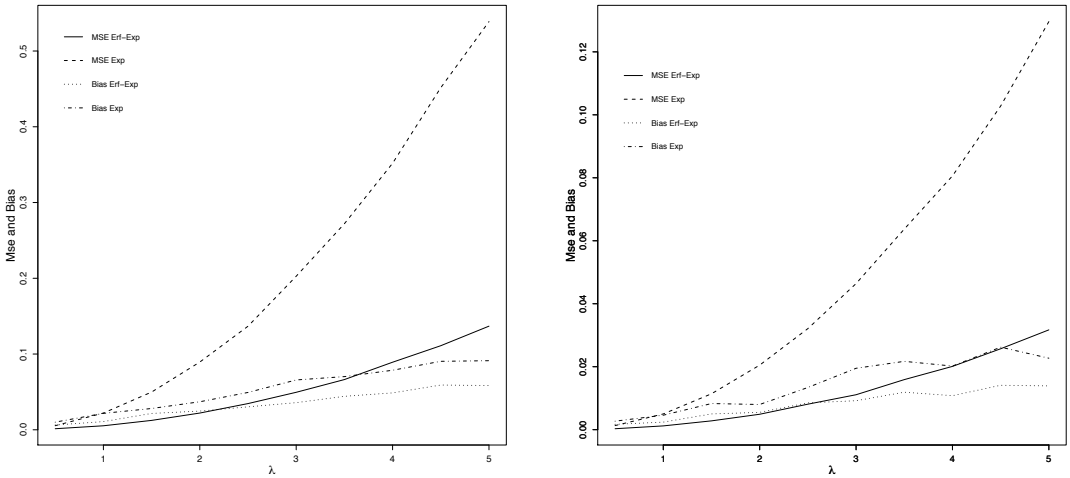
and

$$f(x;a) = \sqrt{\frac{2}{\pi}}a^{\frac{3}{2}}x^2\exp\left(\frac{1}{2}ay^2\right), \quad y>0, \quad a>0.$$

We make a Monte Carlo study with $10,000$ replications such that, for several baseline parameter values and sample sizes $n\in\{50,200\}$, two comparison criteria are quantified:

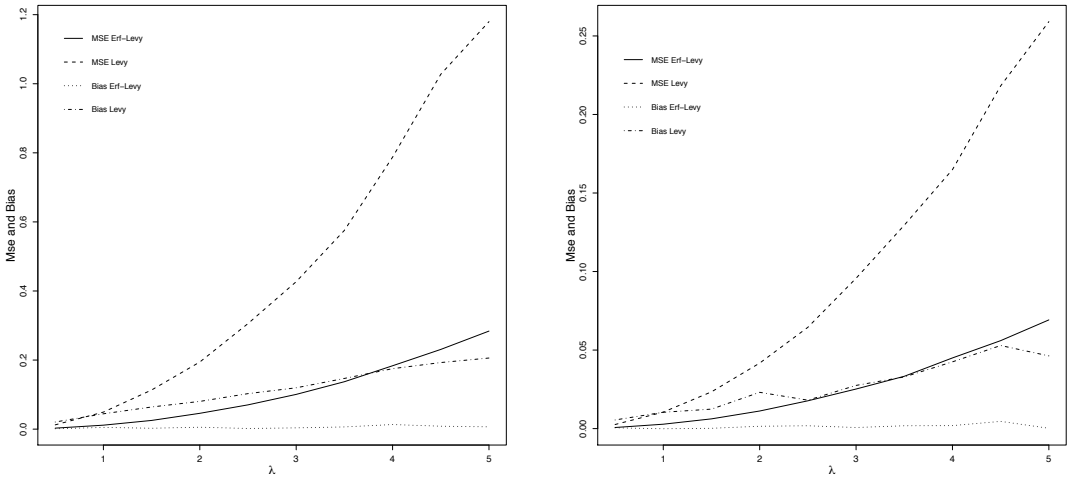
biases and root mean squared error (RMSE). All computations are implemented using the `R` programming language, which has numerous advantages, perhaps the main one being the fact that it is distributed free of charge through the so-called *GNU Public license*. For more information about `R`, visit the `https://www.r-project.org` website. To ensure the reproducibility of this experiment, the following comments are needed: It was utilized the `maxLik(.)` function of the `R` package `maxLik` . Specifically, the BFGS iterative method was used in the optimization process.

Simulation results are presented in Figures 8, 9 and 10. Based on these plots, we conclude that: (i) As expected, the biases and RMSE decreases as the sample size increases; (ii) The erf-G models has superior performance when compared to their respective baseline models.



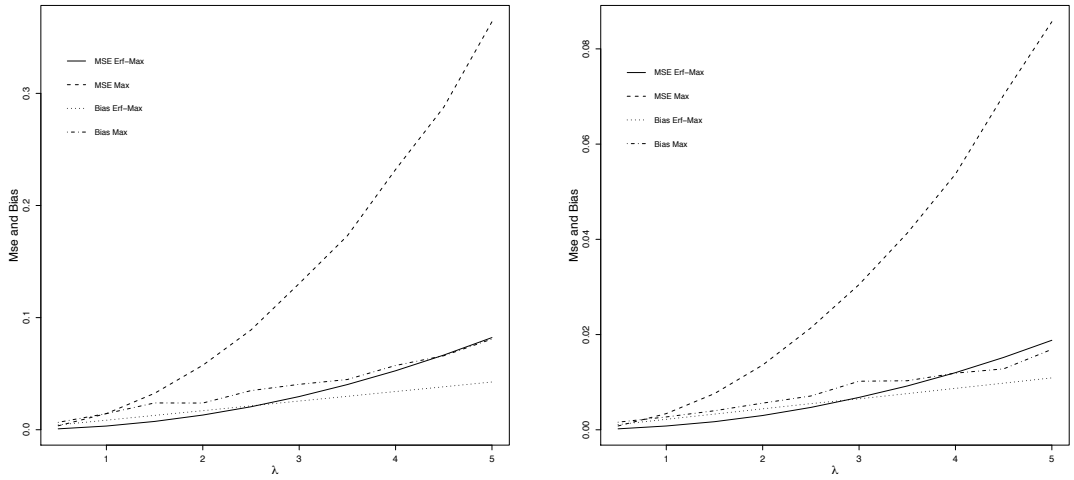
(a) $n = 50$ (b) $n = 200$

Figure 8. RMSEs and biases of $\widehat{\lambda}$ for the erfExp and Exp models.



(a) $n = 50$ (b) $n = 200$

Figure 9. RMSEs and biases of $\widehat{\lambda}$ for the erfLevy and Levy models.



(a) $n = 50$ (b) $n = 200$

Figure 10. RMSEs and biases of $\widehat{\lambda}$ for the erfMax and Max models.

5.2 Real data applications

Two applications to real data illustrate the performance of proposed models. First we describe a set of lifetime data by means of some erf-G models comparatively to the corresponding G distributions. Second the llr model performance is quantified and compared.

5.2.1 Unconditioned model

This section addresses an application to a real data set to illustrate the usefulness of the proposed family.

To that end, we consider three baseline distributions: exponential (Exp), Kumaraswamy (K) and Weibull (W). The main objective is to show that the distributions extended from the erf-G family perform better when compared with their baseline distributions.

We use a data set obtained in Proschan (1963) and corresponds to the time of successive failures of the air conditioning system of jet airplanes. These data were also studied by Dahiya and Gurland (1972), Gleser (1989) and Kuş (2007), among others. The data are

194, 413, 90, 74, 55, 23, 97, 50, 359, 50, 130, 487, 102, 15, 14, 10, 57, 320, 261, 51, 44, 9 , 254, 493, 18, 209, 41, 58, 60, 48, 56, 87, 11, 102, 12, 5, 100, 14, 29, 37, 186, 29, 104, 7, 4, 72, 270, 283, 7, 57, 33, 100, 61, 502, 220, 120, 141, 22, 603, 35, 98, 54, 181, 65, 49, 12, 239, 14, 18, 39, 3, 12, 5, 32, 9, 14, 70, 47, 62, 142, 3, 104, 85, 67, 169, 24, 21, 246, 47, 68, 15, 2, 91, 59, 447, 56, 29, 176, 225, 77, 197, 438, 43, 134, 184, 20, 386, 182, 71, 80, 188, 230, 152, 36, 79, 59, 33, 246, 1, 79, 3, 27, 201, 84, 27, 21, 16, 88, 130, 14, 118, 44, 15, 42, 106, 46, 230, 59, 153, 104, 20, 206, 5, 66, 34, 29, 26, 35, 5, 82, 5, 61, 31, 118, 326, 12, 54, 36, 34, 18, 25, 120, 31, 22, 18, 156, 11, 216, 139, 67, 310, 3, 46, 210, 57, 76, 14, 111, 97, 62, 26, 71, 39, 30, 7, 44, 11, 63, 23, 22, 23, 14, 18, 13, 34, 62, 11, 191, 14, 16, 18, 130, 90, 163, 208, 1, 24, 70, 16, 101, 52, 208, 95.

Some descriptive statistics for these data are given in Table 1. Note that the mean is greater than the median and the asymmetry coefficient is positive, i.e., the empirical distribution from data is positively asymmetric. There is a lot of variability in the data and they are overdispersed. Further, from the kurtosis coefficient, the distribution of the data is platykurtic.

Table 2 provides the ML estimates of considered model parameters (corresponding standard errors in parentheses) and the values of some goodness-of-fit measures: the Akaike information criterion (AIC), Bayesian information criterion (BIC) and consistent Akaike

information criterion (CAIC). In general, it is considered that the lower values (AIC, BIC and CAIC) indicate better fits. In all the situations, the proposed models outperform the corresponding baselines.

Table 1. Descriptive statistics for the air conditioning system of airplanes data.

Statistic	
Mean	93.141
Median	57
Variance	11398.471
Minimum	1
Maximum	603
Skewness	2.322
Kurtosis	3.692

Table 2. The ML estimates (standard errors in parentheses) and the AIC, BIC and CAIC for the phosphorus concentration data.

Distribution	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\gamma}$	$\widehat{\delta}$	$\widehat{\eta}$	Cramr	K-S	AD	AIC	BIC	CAIC
BGP	10.778	1.031	22.346	28.890	–	0.302	0.079	2.044	2386.988	2400.433	2387.180
	(0.791)	(0.356)	(1.549)	(0.872)							
KumaBXII	16.190	6.810	5.761	0.057	0.100	0.215	0.069	1.487	2381.423	2398.230	2381.713
	(3.375)	(1.831)	(2.008)	(0.019)	(0.059)						
Gama-Gama	10.997	0.001	22.975	–	–	0.851	0.122	5.085	2475.640	2485.724	2475.755
	(0.227)	(0.000)	(0.002)								
erf-We	0.043	0.524	–	–	–	0.475	0.109	2.857	2390.732	2397.455	2390.789
	(0.006)	(0.025)									

As qualitative comparison sources, plots ofnthe empirical and estimated PDF and CDF of the under discussion models are displayed in Figures 11. Results indicate the fitted erfW, erfExp and erfK models are better than the associated baselines for phosphorus concentration data. These are first practical evidences in favor of the use of the proposed family.

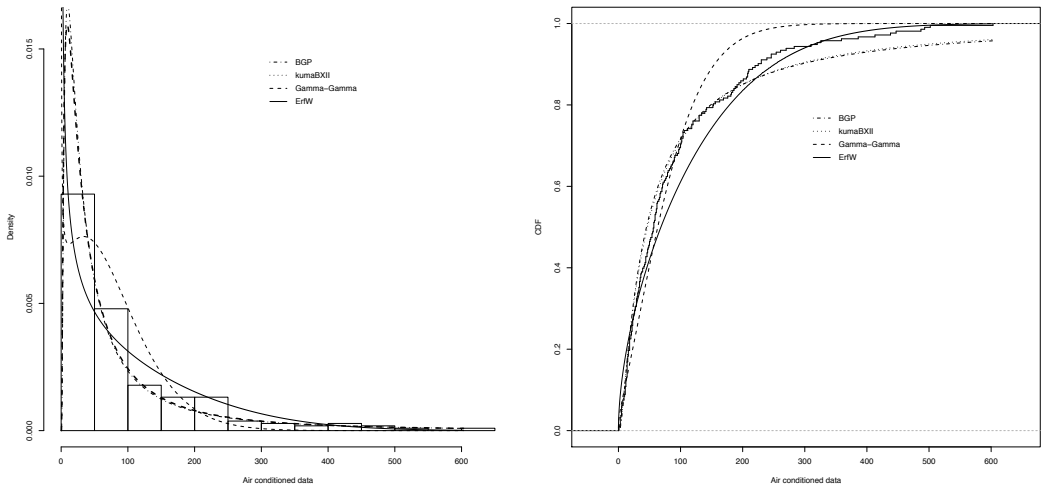


Figure 11. Plots of the fitted BGP, KumaBXII, Gamma-Gamma and erfW PDFs (left) and of the estimated CDFs of the BGP, KumaBXII, gamma-gamma and erfW models (right)s.

.

5.2.2 Regression model

Now, consider results obtained from a lifetime test experiment on 76 specimens of a type of electrical insulating fluid subjected to constant voltage stress, say x, at seven levels, $x = 26, 28, 30, 32, 34, 36$ and 38 kV. The time period until each sample has failed (or "broke"), say breaking time Y, was observed. Such study was firstly performed by Nelson (1972) and Vanegas et al. (2012). Now, we aim to investigate how the voltage level influences the failure time. Does the erf$G$ structure present advantage in the regression context likewise that for uncorrelated distributions?

To that end, we compare the lerfF and log-Frechet (L-F) regression models. Table 3 presents results for ML estimates of the adjusted models as well as their respective significance and standard error measures. We also provide values of the AIC, BIC and CAIC statistics as comparison means. From results of individual confidence intervals for β_i, one has that both considered slopes (and, as a consequence, used predictive variables) are meaningful at the level 5%, employing the asymptotic distribution of the t statistic for $\mathcal{H} : \beta_1 = 0$. From comparison point-of-view, lerfF regression model outperforms L-F, illustrating the importance of the erf-G family in the regression context.

Table 3. ML estimates of the parameters from some fitted regression models to the Minutes to breakdown data set, the corresponding standard errors (in parentheses), p-value (in brackets) and the AIC, BIC and CAIC measures.

Model	β_0	β_1	σ	AIC	BIC	CAIC
lerfF	13.5272	-0.3329	3.1597	297.5957	304.5879	297.9290
	(1.9256)	(0.0586)	(0.3053)			
	[<0.0001]	[<0.0001]	[<2e-16]			
L-F	7.1364	-0.1790	1.7176	320,9327	327,9249	321,2660
	(2.3629)	(0.0697)	(0.1601)			
	[0.0025]	[0.0102]	[<2e-16]			

6. Conclusions and future works

In this paper, we propose and study a new class of distributions called efr-G family. This family is based on the known error function and does not add parameters to its resulting models regard to the baseline distribution. As other advantage, the erf-G family seems to solve or at least to improve estimates based on flat likelihoods. We derive some of its mathematical properties, such as quantile function, ordinary and incomplete moments, moment generating function and Shannon and Rényi entropy measures. A log-linear regression model in the new family is also proposed. Simulation studies and real data applications illustrate the usefulness of the our proposals. For future works, new regression models and a complete study of residual analysis for the proposed models will be developed.

Acknowledgements

We thank two anonymous referees and the associate editor for their valuable suggestions, which certainly contributed to the improvement of this paper. Additionally, Thiago A. N. de Andrade is grateful the financial support from CAPES (Brazil), through its program to encourage post-doctoral researches. He also thanks to the statistics department of the Federal University of Pernambuco.

References

Afify, A., Altun, E., Alizadeh, M., Ozel Kadilar, G., and Hamedani, G., 2017. The odd exponentiated half-logistic-g family: Properties, characterizations and applications. Chilean Journal of Statistics, 8:65–91.

Chevillard, S., 2012. The functions erf and erfc computed with arbitrary precision and explicit error bounds. Information and Computation, 216:72 –95.

Cordeiro, G.M., Aristizabal, W.D., Suárez, D.M., and Lozano, S., 2015. The gamma modified Weibull distribution. Chilean Journal of Statistics, 6:37-48.

Cordeiro, G.M. and de Castro, M., 2011. A new family of generalized distributions. Journal of Statistical Computation and Simulation, 81:883–893.

Cordeiro, G.M., Ortega, E.M.M., and Cunha, D.C.C., 2013. The exponentiated generalized class of distributions. Journal of Data Science, 11:1–27.

da Silva, L.C.M., de Andrade, T.A.N., Maciel, D.B.M., Campos, R.P.S., and Cordeiro, G.M., 2013. A new lifetime model: the gamma extended Frechet distribution. Journal of Statistical Theory and Applications, 12:39–54.

Dahiya, R.C. and Gurland, J., 1972. Goodness of fit tests for the gamma and exponential distributions. Technometrics, 14:791–801.

de Andrade, T.A.N., Bourguignon, M., and Cordeiro, G.M., 2016. The exponentiated generalized extended exponential distribution. Journal of Data Science, 14:393–414.

de Andrade, T.A.N., Rodrigues, H., Bourguignon, M., and Cordeiro, G.M., 2015. The exponentiated generalized Gumbel distribution. Revista Colombiana de Estadística, 38:123–143.

Eugene, N., Lee, C., and Famoye, F., 2002. Beta-normal distribution and its applications. Communications in Statistics: Theory and Methods, 31:497–512.

Gleser, L.J., 1989. The gamma distribution as a mixture of exponential distributions. The American Statistician, 43:115–117.

Kuş, C., 2007. A new lifetime distribution. Computational Statistics and Data Analysis, 51:4497–4509.

Leao, J., Saulo, H., Bourguignon, M., Cintra, R., Rego, L., and Cordeiro, G.M., 2013. On some properties of the beta inverse Rayleigh distribution. Chilean Journal of Statistics, 4:111–131.

Marshall, A.N. and Olkin, I., 1997. A new method for adding a parameter to a family of distributions with applications to the exponential and Weibull families. Biometrika, 84:641–652.

Nadarajah, S., Cordeiro, G.M., and Ortega, E.M.M., 2015. The Zografos-Balakrishnan-G family of distributions: Mathematical properties and applications. Communications in Statistics: Theory and Methods, 44:186–215.

Nelson, W., 1972. Graphical analysis of accelerated life test data with the inverse power law model. IEEE Transactions on Reliability, 21:1–10.

Proschan, F., 1963. Theoretical explanation of observed decreasing failure rate. Technometrics, 5:375–383.

Ristic, M.M. and Balakrishnan, N. (2012). The gamma exponentiated exponential distribution. Journal of Statistical Computation and Simulation, 82:1191–1206.

Tahir, M.H. and Nadarajah, S., 2015. Parameter induction in continuous univariate distributions: Well-established G families. Anais da Academia Brasileira de Ciencias, 87:539–568.

Vanegas, L., Rondon, L., and Cordeiro, G.M., 2012. Diagnostic tools in generalized Weibull linear regression models. Journal of Statistical Computation and Simulation, 83:1–24.

Zografos, K. and Balakrishnan, N., 2009. On families of beta-and generalized gamma-generated distributions and associated inference. Statistical Methodology, 6:344–362.

Chilean Journal of Statistics
Vol. 11, No. 1, April 2020, 25–39

Categorical Data
Research Paper

Detecting outliers in $I \times J$ tables through the level of susceptibility

Thodur Parthasarathy Sripriya[1,*], Mamandur Rangaswamy Srinivasan[1], and Meenakshisundaram Subbiah[2]

[1]Department of Statistics, University of Madras, Chennai, India,
[2]CEO, Acroama Gnan Vikas Pvt. Ltd., Chennai, India

(Received: 15 November 2019 · Accepted in final form: 30 April 2020)

Abstract

Detecting outliers in two-way contingency tables is an important and interesting statistical problem. There is no clear objective procedure available in literature to handle outliers in categorical data unlike other data types. Therefore, this study envisages a two-step procedure, to first indicate and then to identify outliers in two-dimensional contingency tables. The approach deals with enhancing the summary measure to indicate the presence of possible outlying cells followed by residual approaches supplemented by boxplot in identifying the outliers. The fundamental definition of outlying cell as "markedly deviant" cell is clearly exploited in this two-step procedure. A simulation study has been carried out to examine the consistency of the proposed methods and later applied to a large collection of real datasets from various applications of social sciences.

Keywords: Boxplot · Contingency tables · Outlying cells · Residuals · Summary measures.

Mathematics Subject Classification: Primary 62H17 · Secondary 97K80.

1. Introduction

The phenomenal growth of availability of data, in recent years, has drawn the attention of researchers in the identification of unusual observations (outliers) in data for its own significance and its impact on the data analysis. Outliers may be errors, or else accurate but unexpected observations, which could shed new light on the phenomenon under study (Barnett and Lewis (1978)). On the other hand, it is possible that an outlier is simply a manifestation of the inherent variability of the data.

Unlike in metric case, there exists no clarity in the definition of outliers for categorical data, as the cells are purely frequencies or counts of a contingency table. Hence, the problem is to first identify a pivotal element and then markedly deviated cells are detected as outliers. In continuous data, mean or quartiles are considered as pivot and the metric

*Corresponding author. Email: sri.chocho@gmail.com

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

such as $Q_3 \pm 1.5Q_1$, $\mu \pm K\sigma$, etc., are used to identify the outliers (Park et al., 2019; Kim, 2015). However, it is challenging to establish exact criteria for deciding on an observation to be unusual, denoted as an outlier, in contingency tables. Hence, an attempt has been made to provide a set of statistical rules enabling the experimenter to look closely for causes of an outlier to really exist, and then to decide on its plausible acceptability.

The existence of one or two outliers in a sample can badly distort the summary indications and analyses of data. In the detection of outliers in contingency tables, residual based approach has been widely used (Haberman, 1973; Brown, 1974; Simonoff, 1988; Fuchs and Kenett, 1980; Bradu and Hawkins, 1982; Yick and Lee, 1998).

The use of residual approach may cause masking and swamping and a method resistant to it has been studied by Kotze and Hawkins (1984) and Lee and Yick (1999). But, residuals play an important role in detecting outliers in two-way contingency tables, and an extensive review is presented in Kateri (2014). Graphical display of contingency table can be made with plots such as association plot, sieve plot, and mosaic plot (Friendly (2000)) which are based on independence of the row and column variables. Velez and Marmolejo-Ramos (2017) proposed an extension of a graphical diagnostic test for contingency tables using polygraph. Kuhnt (2004) described a procedure to identify outliers based on the tails of the Poisson distribution and declared a cell as outlier if the actual count falls in the tails of the distribution.

Rapallo (2012) studied the pattern of outliers by fitting log-linear model and tests the goodness of fit to specify the notion of outlier with the use of algebraic statistics. Sripriya and Srinivasan (2018a) and Sripriya and Srinivasan (2018b) have suggested a new approach in the detection of outliers in categorical tables of order $I \times J$, based on Poisson log-linear model. Kuhnt et al. (2014) detected outliers through subsets of cell counts called minimal patterns for the independence model.

The principal interest in the analysis of $I \times J$ contingency tables is to test the independence between the two categories. The Pearson chi-square and the log-likelihood ratio statistics (Agresti (2002)) are the long standing techniques in testing independence under multinomial set up. Literature is abundant to show that the residual test statistic converges approximately to the chi-square distribution (Song, 2007; McCullagh and Nelder, 1989). Following Agresti (2002) and Sangeetha et al. (2014) proposed the reversal pattern of association (RAP) to understand deeply the association between attributes in high dimensional tables. Sripriya and Srinivasan (2018a), Sripriya and Srinivasan (2018b) adapted the RAP to detect the outliers based on chi-square statistic through an iterative algorithm. Indeed, there are many procedures like residuals based approach, pattern based approach, and test based approach which are more heuristic in nature as pointed by Simonoff (2003) leading us to the present study based on the characteristics of the contingency table.

In this paper, an attempt has been made to explain the fundamental meaning of "markedly deviant" by answering; which cell, from where and, by how much. To realize the definition, there is a need for a measure which captures the deviation from the pivotal element. Thus, a measure based on the generic characteristics of the table has been considered as a pivotal element for detection of outliers.

The purpose of the present study is to detect possible outliers for a two-way contingency table in a more generic way by a two-step procedure, firstly through an indicator followed by an exact identifier. The first step involves the enhancement of summary measures for categorical data, and a methodical way to indicate susceptibility to outliers by explaining the characterization of contingency tables through three different methods. In step two, potential outliers are detected by using theoretical approach of residuals supported by the boxplot in explaining the deviation of residuals. Lastly, a simulation study has been carried out by contaminating the cell values to determine the stability of the results for detecting outliers through the proposed method.

The paper is organized as follows. In Section 2, we define our two-step procedure and discuss the classification of level of susceptibility. The results of simulation study in Section 3 reveals that the two-step approach performs well in detecting outliers. Section 4 presents few applications to real data in detecting the outliers in two-way contingency tables. Finally, some concluding remarks are given in Section 5.

2. Two-Step Procedure

Let X and Y denote two categorical response variables, X with I categories $R_1, \ldots, R_I$ and Y with J categories $C_1, \ldots, C_J$ leading to IJ possible combinations. When the cells contain frequency (n_{ij}) of outcomes from a sample, the table is called a contingency table, or cross-classification table.

Sparseness in contingency tables often occurs in practice and detecting outliers in the sparse data is a challenging one. The remedial actions for sparseness in categorical data such as collapsibility of cells with small frequencies, or dropping the tables altogether lead to loss of information (Baglivo et al. (1988)). However, this study considers the detection of outlier in $I \times J$ contingency tables without considering the sparseness index but in terms of polarization and its underlying issues.

Further, polarization of cell counts is one of the major problem when it comes to outlier detection. Polarization is basically a highly uneven distribution of counts in $I \times J$ tables. Polarization in contingency tables involves presence of counts/frequencies of disparate in nature, such as zero counts, low counts, high counts, and extreme values, etc. Suppose a table consists of more number of zero counts and very few high counts forming unusual clusters which could affect the inference of $I \times J$ tables, in addition to detection of outliers. Thus, the structure and nature of cell counts in a contingency table play an important role in the data analysis with the cell counts ranging from zero to very high frequencies (Sangeetha et al. (2014)). The relevance of sparseness on summary measure and the sensitivity of analysis in 2×2 tables have been discussed by Subbiah and Srinivasan (2008).

The prevailing researches on the characteristics of $I \times J$ tables are: Order of k, numerical issues (aberration/zero width intervals ZWI), polarization of cell counts, low cell count, sparseness and computational complexity. However, the present study is concerned with the detection of unusual observations or outliers in contingency table. The two step process considered in this study as follows:

Step 1: Indicator – Identify whether the table contains outlier cells through the level of susceptibility
Step 2: Identifier – Detect the exact outlying cells using boxplot of residuals

---

The detailed two step procedure is as follows:
Step I: Contingency tables are often summarized by its size $I \times J$ $(= k)$ and total frequency $N = \sum_i \sum_j n_{ij}$ (Agresti and Yang (1987)). However, there can be other characteristics of contingency table which can be captured and included in the summary measures, such as

Z_C: Number of zero counts in a $I \times J$ table
P_Z: Proportion of zero counts in a table $= Z_C/k$
L_C: Number of low counts in a table
P_L: Proportion of low counts in a table $= L_C/k$
H_C: Number of high counts in a table
P_H: Proportion of high counts in a table $= H_C/k$
R: Range of the cell counts
T: $T = N/k$
Q: $Q = \text{Range}/k$

The three defined measures T, Q and P (P_Z, P_L, P_H) can be considered as an enhancement of the summary measures apart from k and N and could constitute an important component of contingency tables and in particular to indicate the presence of outliers in a table. In an ideal table, all the observations are expected to be closer to the pivot element and thereby expected values are closer with smaller residuals. Suppose all the k cells are quite closer to T, then one may not suspect outlier(s) to be present, except in the heuristic residual approach. Hence, T can be perceived as an Pivot element, for example, a table with $k = 36$ cells, $N = 366$, and $T = 10.16667$ yields all the cells counts to be pretty closer to T and the expected values are closer to each other. Following Agresti and Yang (1987), the present study considers the classification of P, T and Q for the detection of outliers with Low (L), Moderate (M), and High (H) categories as follows

$$P_Z = \begin{cases} \text{Low}, & 0 \le P_Z \le 0.10; \\ \text{Moderate}, & 0.10 < P_Z \le 0.20; \\ \text{High}, & P_Z > 0.20; \end{cases}$$

$$P_L(n_{ij} < 6) = \begin{cases} \text{Low}, & 0 \le P_L \le 0.20; \\ \text{Moderate}, & 0.20 < P_L \le 0.40; \\ \text{High}, & P_L > 0.40; \end{cases}$$

$$P_H(n_{ij} > T) = P_L(n_{ij} < T) = \begin{cases} \text{Low}, & 0 \le P_H, P_L \le 0.45; \\ \text{Moderate}, & 0.45 < P_H, P_L \le 0.55; \\ \text{High}, & P_H, P_L > 0.55. \end{cases}$$

Similarly, T and Q have been classified as

$$T = \begin{cases} \text{Low}, & 0 \le T \le 20; \\ \text{Moderate}, & 20 < T \le 250; \\ \text{High}, & T > 250; \end{cases}$$

$$Q = \begin{cases} \text{Low}, & 0 \le Q \le 10; \\ \text{Moderate}, & 10 < Q \le 100; \\ \text{High}, & Q > 100. \end{cases}$$

Table 1. Categorization of susceptibility

Susceptibility	(T, P_Z, P_L)	(Q, P_Z, P_L)	(P_Z, P_L, P_H)
High	8	12	12
Moderate	10	9	12
Low	9	6	3

Our study proposed three methods (i) (T, P_Z, P_L) (ii) (Q, P_Z, P_L) and (iii) (P_Z, P_L, P_H) based on the above classification to identify the susceptibility to outliers in $I \times J$ tables. Thus there will be a total of 27 combinations for each method under consideration. Suppose a table with (T, P_Z, P_L) is (L, L, L), then, there will be a less chance of outliers being present and hence denote the $I \times J$ table as of low susceptibility to outliers. Correspondingly, a table with (T, P_Z, P_L) is (H, L, L), then there may be few markedly deviant cells to exist in the table and denoted as highly susceptible to outliers. Similarly, the combination of M and L is taken to be moderately susceptible to outliers. Thus the 27 combinations of L, M, and H are categorized for susceptibility under the three proposed methods and presented in Table 1. The categorization of susceptibility is based on the direction provided by Agresti and Yang (1987), but could be suitably modified based on T, Q, and P

and accordingly susceptibility to outliers will also vary. In the same way, method 2 has been categorized under the three levels; H, M, and L, whereas in the third method, LLL is taken to be highly susceptible to outliers based on the above mentioned classification. Thus, 27 combinations are categorized into three methods as presented in the following table. Consider a 5×5 table constructed by Simonoff (1988) for the detection of outliers. Based on the approach outlined earlier, with $k = 25$, $N = 558$, $T = 22.32$, $Q = 1$, $P_Z = 0$, $P_L(n_{ij} < 6) = 0$, $P_L(n_{ij} < T) = 0.8$, and $P_H = 0.2$ reveals the table is highly susceptible to outliers. Thus, the study basically affirms the approach to be capable of indicating the presence of outliers. After due classification of $I \times J$ table, the next step is to identify the outlying frequencies in the table.

Step II: Residual techniques have been carried out by many researchers in order to identify the outlying cells in a table by considering "large" residual. But many of them failed to justify "how large" the residual should be considered for an observation as an outlier. The usual residual based methods of outlier detection methods are devoid of contingency table characteristics. In the heuristic approach, outliers are identified irrespective of the polarization of cell frequencies and order of the tables. To overcome this, the box plot of the following three types of residuals has been considered to identify the outlying cell:
(i) [Pearson residual]

$$r_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}, \quad e_{ij} = (n_{i+} \times n_{+j})/N.$$

(ii) [Adjusted residual; Haberman (1973)]

$$\tilde{r}_{ij} = \frac{r_{ij}}{AF}, \quad \text{AF} = (1 - n_{i+}/N)(1 - n_{+j}/N).$$

(iii) [Deleted residual; Simonoff (1988)]

$$r_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}, \quad e_{ij} = (n_{i+} - n_{ij})(n_{+j} - n_{ij})/(N - n_{i+} - n_{+j} + n_{ij}).$$

Thus, the two step process provides a systematic approach of identifying outliers under conditions of polarity for varying order of k. The following section deals with examining the robustness of susceptibility criteria as envisaged through a simulation study.

3. Simulation Study

Simulating a two way contingency table situation can be achieved using varying combinations of its total frequency, levels in each of the categories, cell probabilities, and the test statistic used to analyze the independence. Thus, the present study considers two scenarios of generating $I \times J$ tables where the cell entries are from (i) bi-variate normal distribution with the assumption of independence, and (ii) multinomial distribution as in Agresti (2002) since it models the probability of counts in each categories for n independent trails.

Bivariate normal distribution The simulation study starts with generating the entries of $I \times J$ table from bi-variate normal distribution with different correlation structures. In this scenario, the study considered correlation ρ and the size of the table $k(= I \times J)$ as the potential parameters. Here, we consider four different values of k, (9, 16, 25, and 100) with five different correlation structures (0.5, 0.6, 0.7, 0.8, and 0.9) to evaluate the performance

of proposed susceptibility methods by contaminating each cell at a time with a constant α $(=0.5, 1, 1.5, 2)$ and repeated 500 times. The results of this simulation are summarized in Table 2. The following are the observations based on the simulation presented in Table 2:

(i) The pattern of susceptibility level remains unchanged for $k = 9, 16$ and with changes in $k = 25$ and 100 irrespective of ρ and α in methods 1-2.
(ii) When k increases, the susceptibility level increases only when the correlation is 0.5 when $\alpha = 0.5$. However, it shows few fluctuations due to outliers in other correlation structures with different α considered.
(iii) Susceptibility level fluctuates largely for all k in all methods irrespective of ρ and α.
(iv) As α increases, the susceptibility level shows similar pattern for all order of k with $\rho = 0.5, 0.6, 0.8$, and 0.9. However, fluctuations are visible between the contamination α for all k with $\rho = 0.7$.
(v) The variability in the susceptibility is largely observed from method 3 as it gives poor results for all k irrespective of correlation structure and α.

Table 2. Susceptibility to outliers (in %) for scenario 1

Order		Method 1 (T, P_Z, P_L)				Method 2 (Q, P_Z, P_L)				Method 3 (P_Z, P_L, P_H)			
$I \times J$	ρ	α											
		0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2
3×3	0.5	76	75	73	72	78	77	75	74	45	44	42	39
	0.6	74	74	73	73	76	75	74	73	73	72	71	70
	0.7	75	52	60	54	79	53	59	52	68	53	52	51
	0.8	66	58	47	41	63	49	45	44	64	57	52	41
	0.9	78	65	51	49	74	62	53	47	72	63	51	50
4×4	0.5	77	74	72	70	75	74	72	71	47	45	40	38
	0.6	76	73	70	69	74	72	71	68	71	69	68	65
	0.7	77	55	62	53	75	51	59	52	69	51	49	49
	0.8	60	52	40	40	62	54	44	43	60	58	51	39
	0.9	76	60	55	52	72	61	52	50	70	64	55	54
5×5	0.5	78	74	71	70	76	71	69	68	49	46	41	37
	0.6	59	56	51	47	62	57	53	48	60	55	51	49
	0.7	72	54	56	52	74	51	58	49	65	50	45	43
	0.8	56	42	38	34	53	45	41	37	50	49	42	36
	0.9	66	50	45	42	62	51	42	40	60	54	49	44
10×10	0.5	81	79	76	71	79	74	71	69	51	49	47	46
	0.6	67	64	60	53	66	61	57	50	63	52	51	50
	0.7	75	57	63	55	72	53	65	51	69	54	49	42
	0.8	61	58	47	44	59	47	45	40	60	57	51	44
	0.9	76	64	54	51	69	57	54	48	67	55	47	46

Following susceptibility, Table 3 presents the results of the simulation involving the identification of outliers based on three residual methods under different levels of contamination. The following are the observations based on simulation presented in Table 3:

(i) The identification of exact outlying cell for all k shows similar trend irrespective of α and ρ in all the three residuals considered in this simulation scenario.
(ii) As α increases, the identification level also increases for all k irrespective of the correlation structure ρ.

(iii) Stability of level of identifying the outlier cell increases as ρ increases for $k = 9, 16, 25$. However, for $k = 100$, yields poorer results for all the three residual approaches.

Table 3. Identification of outliers (in %)

		Pearson				Adjusted				Deleted			
$I \times J$	N	α											
		0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2
3×3	0.5	33.3	34	34.3	34.5	35	35.5	36	36	34	35.5	36.3	36.5
	0.6	36	36.5	36.7	37	36.5	37	37.7	38	37	37.5	38	38.3
	0.7	36.2	37	37.5	37.7	36	37	37.7	38.2	38.2	38.7	39.2	39.7
	0.8	37	37.7	38.2	38.5	36.7	37.5	38	39	37.7	38.5	39.7	40
	0.9	38.2	38.5	39	39	38	38.7	39	39.7	38	38.7	39.5	40
4×4	0.5	43	44.5	45	45.5	45	45	46	46	46	46.7	47	47.5
	0.6	45	46	46.5	47.5	47	47.7	48	48.5	47	48.5	48.7	49
	0.7	46.5	47.2	48.5	48.7	46	47.7	48.2	48.7	48.5	49	49.5	49.5
	0.8	46.7	46.7	47.2	48	47.7	48	48.5	49	48.7	49	49.2	49.7
	0.9	48	48.7	49.2	49.5	48	48	49.5	49.7	48	48.5	49	50
5×5	0.5	44.4	45	45.4	46.5	45	45.5	46	46	45	45.5	46.4	46.5
	0.6	46	46.5	46.7	47	46.5	47	47.7	48	47	47.5	48	48.4
	0.7	46.2	47	47.5	47.7	46	47	47.7	48.2	48.2	48.7	49.2	49.7
	0.8	47	47.7	48.2	48.5	46.7	47.5	48	49	47.7	48.5	49	49.5
	0.9	48.2	48.5	49	49	48	48.7	49	49.7	48	48.7	49.5	49.7
10×10	0.5	25	25.5	26.2	26.5	25	25.5	26	26	24	25.5	26.2	26.5
	0.6	26	26.5	26.2	22	26.5	22	22.2	28	22	22.5	28	28.2
	0.7	26.2	22	22.5	22.2	26	22	22.2	28.2	28.2	28.2	29.2	29.2
	0.8	27	27.2	28.2	28.5	26.2	26.5	28	29	22.2	28.5	29	29.2
	0.9	28.2	28.5	29	29	28	28.2	29	29.2	28	28.2	29.5	30.5

The associations between the two categorical variables are identified generally using the chi-square distribution. Here, the p-value of the chi-square distribution is used to identify the independence of the two categorical outcomes and found that there is no change in the independence assumption even after contaminating the cell entries. Moreover, the data generation process in simulation in no way alters the independence assumption. The percentage of identification of outliers in this scenario yield poor results since the data generated from bi-variate normal distribution with the parameter lambda where lambda is the parameter used to change the continuous bi-variate normal random variables to count variables. Thus, a more appropriate data generation rule using multinomial distribution is considered and is explained below.

Multinomial distribution The simulation study considers two potential parameters k; the size and N; the total frequency of the table and $X_1, X_2, \ldots, X_k \sim$ Multinomial$(N, (p_1, \ldots, p_k))$ where the probability $p_i \sim U(0,1); i = 1, \ldots, k$. The probability range between 0 and 1 is automatically maintained in multinom function in R. The study of over 100 real time datasets from various fields of social sciences has shown that polarization is largely observed in tables of order more than 4 and larger tables $(I, J > 10)$ occurs occasionally and are not discussed in the simulation study. Hence our simulation study is restricted to $k = 9, 16, 20$ and 56 with $N = 50, 350, 950, 2150$, and 4550 providing a varied cross section of the contingency table to examine the susceptibility to outliers. The process starts by contaminating the cell frequencies with alpha (α) for each cell at a time and then covering the entire table k times. Four different level of contamination α

$(=0.5, 1, 1.5, 2)$ are considered and repeated 500 times. The results of simulation based on the above procedure are summarized in Table 4.

The following are the observations based on the simulation presented in Table 4:

(i) Susceptibility level remains unchanged for $k = 9, 16$ and minor fluctuations in $k = 20$ and 56 irrespective of N and α in method 1.
(ii) When k increases, irrespective of α, there exists small changes due to outliers in method 2 for moderate N of size 350 and 950.
(iii) Susceptibility level fluctuate largely for all k except for a lower order of k $(=9)$, in method 3 irrespective of N and α.
(iv) As α increases, the level of susceptibility remains constant for all order of k and for small and large values of N under method 1. However, fluctuations are visible for moderate values of N and higher order of k.
(v) Susceptibility level remains constant as α increases for all k and for large values of N under method 2. However, fluctuations are visible for low and moderate values of N irrespective of k.
(vi) In method 3, as α increases, the susceptibility level remains constant for a small order of k and moderate to large N and the instability in susceptibility are observed from rest of k and N.

Table 4. Susceptibility to outliers (in %)

Order		Method 1 (T, P_Z, P_L)				Method 2 (Q, P_Z, P_L)				Method 3 (P_Z, P_L, P_H)			
$I \times J$	N	α											
		0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2
3×3	50	98	92.6	92.8	87.2	63.2	62.2	62.2	61.6	71.2	71.2	71.2	71
	350	100	100	100	100	100	100	100	100	100	100	100	100
	950	100	98.6	98.6	98	100	100	100	100	100	100	100	100
	2150	100	100	100	100	100	100	100	100	100	100	100	100
	4550	100	100	100	100	100	100	100	100	100	100	100	100
4×4	50	100	100	100	100	69.8	69.8	69.8	69.8	50.2	50.2	48.6	49
	350	99.4	95.4	95.4	95.4	99.4	95.4	95.4	95.4	90.2	89.6	89.6	80
	950	100	100	100	100	100	100	100	100	100	100	100	100
	2150	100	100	100	100	100	100	100	100	99.4	99.4	99.4	99
	4550	100	100	100	100	100	100	100	100	100	100	100	100
5×4	50	100	100	100	100	86.4	86.4	86.4	86.4	65.2	55.4	55.4	55
	350	85.6	77	77	70.6	86.2	79.4	77.6	71.2	87	64.6	59.4	54
	950	96.8	94.8	94.8	94.6	96.8	94.8	94.8	94.8	69.8	69.8	64.8	64
	2150	100	100	100	100	100	100	100	100	95	90.6	88.2	85
	4550	100	100	100	100	100	100	100	100	100	100	100	100
7×8	50	100	100	100	100	100	100	100	100	100	100	99.2	99
	350	91	80	80	80	99.4	99.4	99.4	99.4	91.4	99	93.2	93
	950	97.6	97.6	89.2	97.6	97.6	97.6	97.6	97.6	55.8	55.8	55.8	52
	2150	100	100	100	100	100	100	100	100	94	94	90.8	91
	4550	100	100	100	100	100	100	100	100	87.4	87.4	87.4	87

As outlined in Section 2, following susceptibility, next step involves identification of outliers based on three residual methods under different levels of contamination. The results of the simulation are presented in Table 5.

The following are the observations based on simulation presented in Table 5:

(i) Identification of exact outlying cell remains same for all k irrespective of α and N and a few fluctuations are observed in moderate to high N in Pearson

and Adjusted residual approach whereas in the case of Deleted residual, the simulation yields inconclusive results.

(ii) As α increases, the identification level decreases for all k and it remains constant when N varies from moderate to high in Pearson and Adjusted residual approach whereas in Deleted residual approach, the identification level decreases as α increases for all k except for $k = 16$ irrespective of N.

(iii) Stability of level of identifying the outlier cell oscillates as N increases irrespective of k and α for all the three residual approaches.

Table 5. Identification of outliers (in %)

		Pearson				Adjusted				Deleted			
$I \times J$	N	α											
		0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2
3×3	50	95.8	93.8	91.4	86	96.4	93.8	92	89.2	83	72	92	54.4
	350	92.6	89.6	87.5	85	94.4	89.4	86.3	84.2	99.4	96.4	94.2	91.6
	950	99.2	99	99	99	99.3	99	99	99	91.2	92.3	90	90
	2150	100	100	100	100	100	100	100	100	97	96	95	95
	4550	95	94.9	100	95	97	98	100	98	93	90	100	91
4×4	50	94.8	92.8	91	88	95.8	92.8	92	89	90	87	70	69
	350	93	93	89	89	94	92	90	88	89	88	89	80
	950	98.8	99	99	99	99	99	99	99	92	92.3	90	91
	2150	99	99	99	99	99	99	99	99	98	97	95	96
	4550	96	94	93	91	97	95	92	90	100	100	96	99
5×4	50	96	94	90	87	95	93	89	89	92	87	83	86
	350	93	93	90	87	94	92	91	88	89	86	85	80
	950	100	100	100	100	100	100	100	100	89	86	82	78
	2150	92	90	91	92	93	92	93	92	83	79	77	75
	4550	93	92	94	90	94	93	95	91	90	91	89	89
7×8	50	94	93	91	89	95	94	92	92	89	82	77	76
	350	93	93	89	90	94	92	90	91	89	88	85	83
	950	95	94	90	89	96	95	91	90	88	85	83	78
	2150	100	100	100	100	100	100	100	100	89	87	83	72
	4550	92	100	97	96	93	95	98	97	84	86	88	78

In summary, even though the level of susceptibility fluctuate in few cases in all the methods, the identification level of exact outlying cells in all the residual approaches show that our two-step procedure could be a best alternative in the detection of outliers in $I \times J$ tables. The results based on the simulation study have paved the way to examine the application of two-step process of detection of outliers in contingency table to real time datasets.

4. Data Analysis

In this section, we illustrate our two-step procedure to six datasets from literature by assuming the nature of the data as nominal. Kotze and Hawkins (1984) considered a dataset with $k = 196$, $N = 775$ and identified 15 most outlying cells by adding 0.5 to zero cells using elimination method. The mosaic display of the data is presented in Figure 1.

The present approach, with $T = 3.95$, $Q = 0.27$, $P_Z = 0.26$, $P_L(n_{ij} < 6) = 0.52$, $P_L(n_{ij} < T) = 0.39$, and$P_H = 0.35$, shows low susceptibility in method 1 and 2 and high susceptibility in method 3. Also, boxplot for residuals as presented in Figure 2 identified

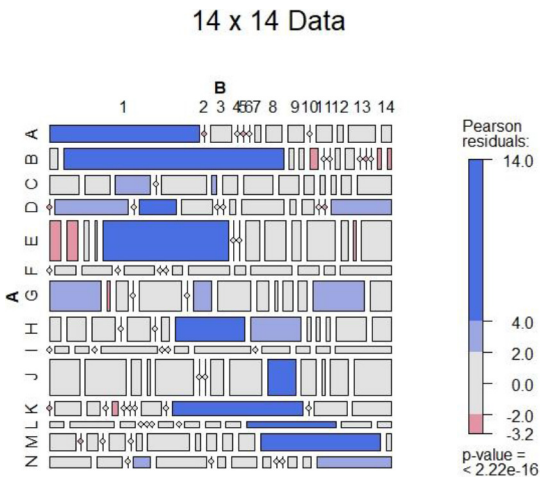


Figure 1. Mosaic Plot for 14×14 data

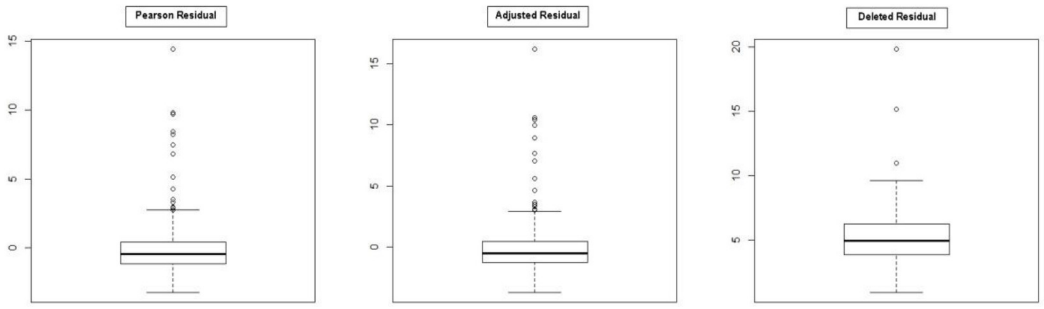


Figure 2. Boxplots for Kotze and Hawkins Data

the same 14 cells as possible outliers in the case of Pearson and Adjusted residuals and only 3 cells in the case of Deleted residuals.

Yick and Lee (1998) considered the archaeological data and artificial data by Simonoff (1988) in identifying outliers. For the artificial 5×5 data, three cells (2, 1), (1, 2) and (1, 3) are identified as outliers and the cell (1, 1) being swamped in the perturbation approach. In our method, with $k = 25$, $N = 558$, $T = 22.32$, $Q = 1$, $P_Z = 0$, $P_L(n_{ij} < 6) = 0$, $P_L(n_{ij} < T) = 0.8$, and $P_H = 0.2$, this dataset is found to be moderately susceptible to outliers and the residual boxplot identifies exactly the same cells as outliers as in perturbation approach.

For the archeological data, the perturbation approach identified three cells (2, 3), (11, 5) and (18, 1) as outliers out of which two cells have extreme residuals and these two extreme cells are identified correctly in our two step procedure with $k = 114$, $N = 3297$, $T = 28.92$, $Q = 3.42$, $P_Z = 0.07$, $P_L(n_{ij} < 6) = 0.21$, $P_L(n_{ij} < T) = 0.65$, $P_H = 0.72$ and the method show that the data is moderately susceptible to outliers. The mosaic display and boxplot of residuals for these two data is presented in Figures 3, 4 and 5.

Yick and Lee (1998) considered the 7×8 student enrolment data from seven community schools from Northern Territory, Australia and identified the cells (1, 5), (1, 6), (2, 4) and (2, 5) as potential outliers using perturbation diagnostics. The mosaic display of the data is presented in Figure 6.

In our proposed method, the datasets is highly susceptible to outliers with $k = 56$, $N = 5248$, $T = 93.71$, $Q = 2.9$, $P_Z = 0$, $P_L(n_{ij} < 6) = 0$, $P_L(n_{ij} < T) = 0.625$, $P_H = 1$ and identified the cells (2, 4) and (1, 6) as potential outliers using boxplot of all the residuals and boxplot are presented in Figure 7.

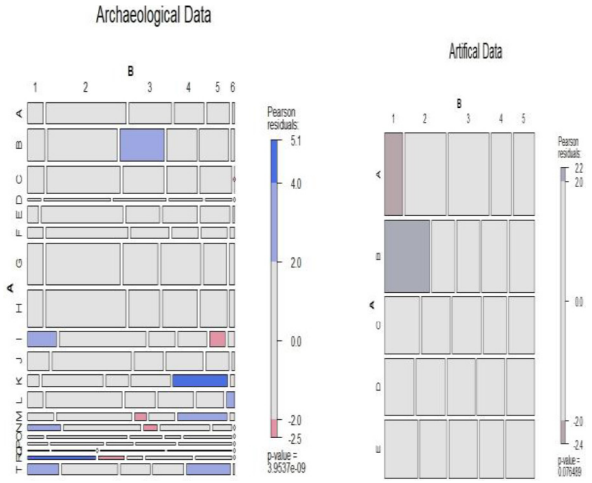


Figure 3. Mosaic Plot for Archaeological and Artificial data

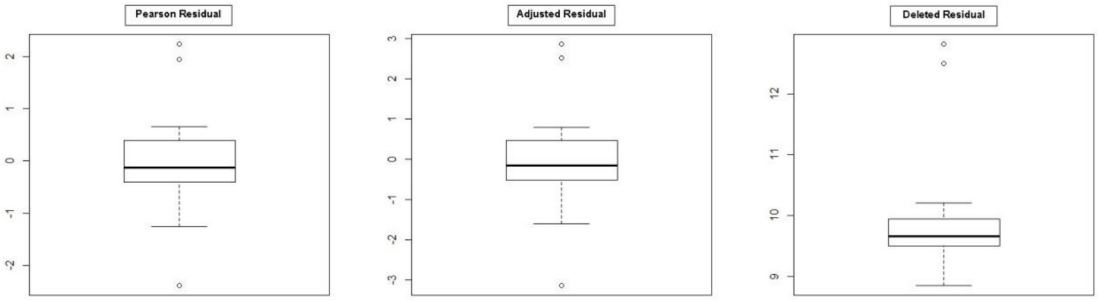


Figure 4. Boxplots for Artificial Data

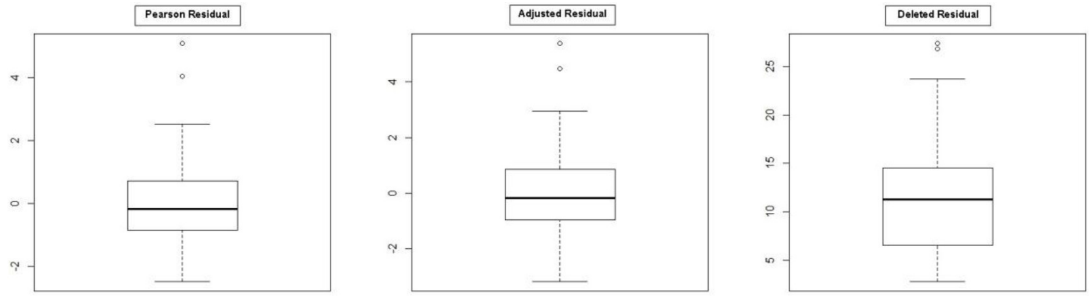


Figure 5. Boxplots for archeological data

Kuhnt et al. (2014) considered 3×3 table of social mobility in Britain and 4×4 table of artifacts in Nevada and detected outliers using three different algorithms. For the social mobility data, all the three algorithms doesn't give satisfactory results and detected, (i) all the cell counts, (ii) only diagonal cells and (iii) cells (1, 1), (3, 1), (1, 3) and (3, 3) as outliers, whereas in our method the table shows highly susceptible to outliers with $k = 9$, $N = 3494$, $T = 366.33$, $Q = 67$, $P_Z = 0$, $P_L(n_{ij} < 6) = 0$, $P_L(n_{ij} < T) = 0.44$, $P_H = 0.56$, and detected the cells (1, 1), (3, 1) and (2, 2) as outliers with the help of boxplot of residuals and the mosaic display is presented in Figure 8.

For the Artifacts in Nevada data, the author identified two cells as outliers but our methods gave inconclusive decision in susceptibility with $k = 16$, $N = 164$, $T = 10.25$, $Q = 3.77$, $P_Z = 0$, $P_L(n_{ij} < 6) = 0.43$, $P_L(n_{ij} < T) = 0.68$, $P_H = 0.32$, and no outliers are detected using boxplot of residuals. The boxplot for these datasets are presented in Figure 9 and 10.

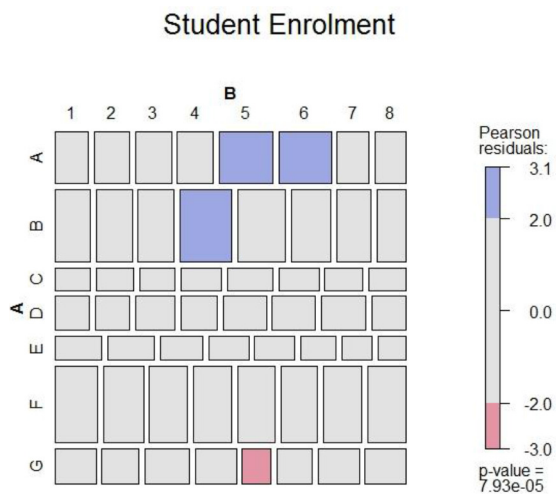


Figure 6. Mosaic Plot for Student Enrolment data

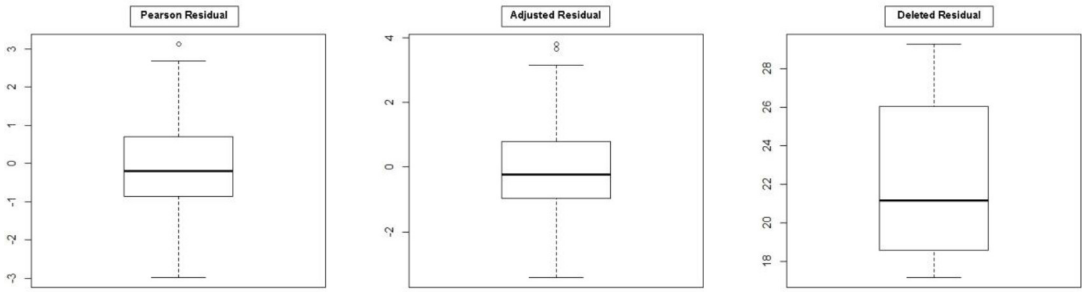


Figure 7. Boxplots for Student Enrolment Data

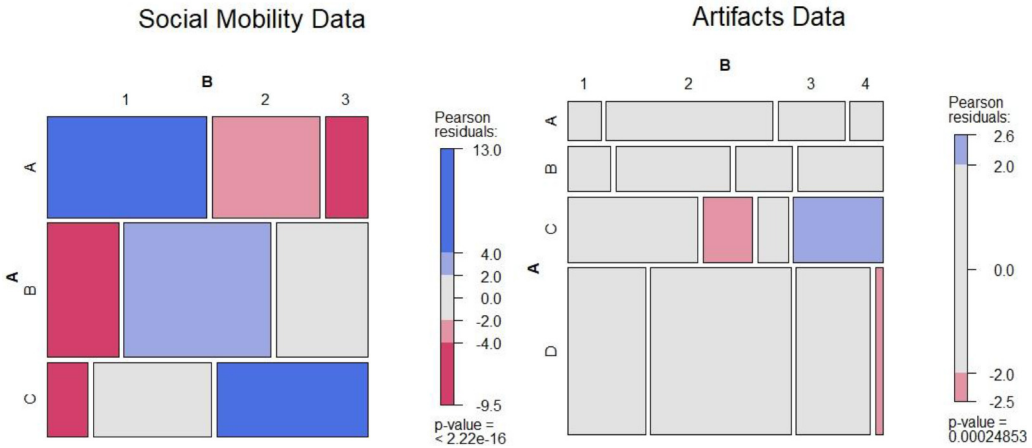


Figure 8. Mosaic Plot for Social Mobility and Artifacts data

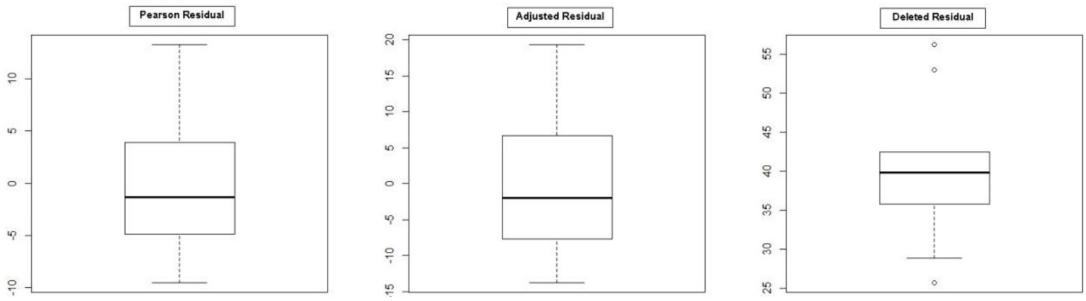


Figure 9. Boxplots for social mobility data

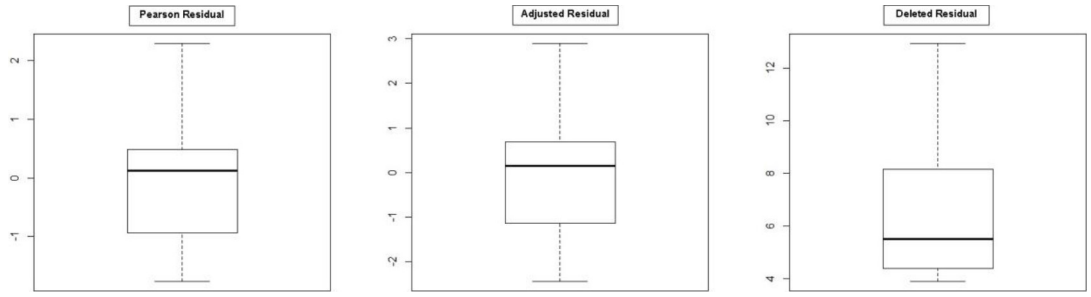


Figure 10. Boxplots for Artifacts in Nevada data

In addition, our study considered 50 other datasets of varied characteristics ranging from $k = 6$ to 196 cells based on the literature to identify the feasibility of our methods and the results are presented in Table 6. Most of the researchers nullified the zero cells in a table by adding constants, but our two-step method helps to identify the outlying cells even in the presence of zero cells in a table.

Table 6. Identification of outlying cells through Boxplot

	(T, P_Z, P_L)			(Q, P_Z, P_L)			(P_Z, P_L, P_H)		
	T	I	NI	T	I	NI	T	I	NI
Highly susceptible	25	25(100%)	–	27	25(92.5%)	2	41	38(92.6%)	3
Moderately susceptible	7	6(85.7%)	1	9	8(88.8%)	1	9	2(22.2%)	7
Low susceptible	18	9(50)	9	14	7(50)	7	–	–	–

T–Total; I–Identified; NI–Not-Identified

The above table clearly shows that method 1 performs better in highly susceptible category and method 2 performs better in moderately susceptible category, method 1 & 2 equally performs better in low susceptible category. The classification of datasets under method 2 also contains the datasets under method 1. On the whole, method 3 appears to be more stringent in identifying outliers since it classifies almost all datasets as highly susceptible to outliers.

5. Conclusions

The problem of identification of outliers in $I \times J$ contingency tables has been examined through the ambiguous notion of "markedly deviant" nature of cells from which the other cell values deviate greatly. However, in this paper a simple measure T has been introduced as a pivotal element to explain the deviation of other cells in the table. In this direction, a two-step procedure is devised to first examine the nature of the table through susceptibility followed by identification of outliers through box plot techniques. The stability of our proposed methods towards the identification of outliers is examined through a simulation study. The results have revealed that methods (T, P_Z, P_L) and (Q, P_Z, P_L) are found to be more consistent based on two simulation scenarios. Moreover, it is evident from the results that a triplet with the pivot element along with proportion of zero and low counts provide an idea of polarization in the table, and is found to be useful in detecting outliers.

Based on the numerical results, we conclude that the two-step approach as a combination of summary measures and boxplot for residuals could be a feasible approach to identify outlier cells in contingency table. However, as pointed out in the earlier section, a judicious choice is necessary in some cases of ambiguity. Further, even if the boxplot or the residual approach fails in some cases, summary measure will indicate clearly whether the table contains high, moderate, or low outlying cells. The practicality of two pronged approach has been well corroborated by an extensive amount of data sets for its efficacy and its usefulness in identifying outlying cells.

Acknowledgement

The authors wish to thank the Editors and Reviewers for their constructive comments on an earlier version of this manuscript.

References

Agresti, A. and Yang, M.C., 1987. An empirical investigation of some effects of sparseness in contingency tables. Computational Statistics and Data Analysis, 5, 9–21.

Agresti, A., 2002. Categorical Data Analysis. Wiley, New York.

Baglivo, J. Olivier, D. and Pagano, M., 1988. Methods for the analysis of contingency tables with data, large and small cell counts. Journal of the American Statistical Association, 83, 1006–1013.

Barnett, V.D. and Lewis, T., 1978. Outliers in Statistical Data. Wiley, New York.

Bradu, D. and Hawkins, D.M., 1982. Location of multiple outliers in two-way tables using tetrads. Technometrics, 24, 103–108.

Brown, B.M., 1974. Identification of the sources of significance in two-way contingency tables. Journal of the Royal Statistical Society C, 23, 405–413.

Friendly, M., 2000. Visualizing Categorical Data. Cary, NC: SAS Institute.

Fuchs, C. and Kenett, R., 1980. A test for detecting outlying cells in the multinomial distribution and two-way contingency tables. Journal of the American Statistical Association, 75, 395–398.

Haberman, S.J., 1973. The analysis of residuals in cross-classified tables. Biometrics, 29, 205–220.

Kateri, M., 2014. Contingency Table Analysis. Springer, New York.

Kim, S.S., 2015. Variable selection and outlier detection for automated K-means clustering. Communications for Statistical Applications and Methods, 22, 55–67.

Kotze, T.J.W. and Hawkins, D.M., 1984. The identification of outliers in two-way contingency tables using 2 x 2 subtables. Applied Statistics, 33, 215–223.

Kuhnt, S., 2004. Outlier identification procedures for contingency tables using maximum likelihood and L1 estimates. Scandinavian Journal of Statistics, 31, 431–442.

Kuhnt, S. Rapallo, F. and Rehage, A., 2014. Outlier detection in contingency tables based on minimal patterns. Statistics and Computing, 24, 481–491.

Lee, A.H. and Yick J.S., 1999. A perturbation approach to outlier detection in two-way contingency tables. Australian and New Zealand Journal of Statistics, 41, 305–314.

McCullagh P. and Nelder, J., 1989. Generalized Linear Models. Chapman and Hall/CRC, Boca Raton, FL, US.

Park, J.S. Park, C.G. and Lee, K.E., 2019. Simultaneous outlier detection and variable selection via difference-based regression model and stochastic search variable selection. Communications for Statistical Applications and Methods, 26, 149-161.

Rapallo, F., 2012. Outliers and patterns of outliers in contingency tables with algebraic statistics. Scandinavian Journal of Statistics, 39, 784-797.

Sangeetha, U. Subbiah, M. Srinivasan, M.R. and Nandram, B., 2014. Sensitivity analysis of Bayes factor for categorical data with emphasis on sparse multinomial data. Journal of Data Science, 12, 339–357.

Simonoff, J.S., 1988. Detecting outlying cells in two-way contingency tables via backwards stepping. Technometrics, 30, 339–345.

Simonoff, J.S., 2003. Analyzing Categorical data. Springer, New York.

Song, P.X.K., 2007. Correlated Data Analysis: Modeling, Analytics, and Applications. Springer, New York

Sripriya, T.P. and Srinivasan, M.R., 2018a. Detection of outliers in categorical data using model based diagnostics. Special Proceedings of 20th Annual Conference of SSCA held at Pondicherry University, Puducherry, 2018, 35-43.

Sripriya, T.P. and Srinivasan, M.R., 2018b. Detection of outlying cells in two-way contingency tables. Statistics and Applications, 16,103–113.

Subbiah, M. and Srinivasan, M.R., 2008. Classification of 2 x 2 sparse data with zero cells. Statistics and Probability Letters, 78, 3212–3215.

Velez, J.I. and Marmolejo-Ramos, F., 2017. Extension of a graphical diagnostic test for contingency tables. Chilean Journal of Statistics, 8, 53-65.

Yick, J.S. and Lee, A.H., 1998. Unmasking outliers in two-way contingency tables. Computational Statistics and Data Analysis, 29, 69–79.

Chilean Journal of Statistics
Vol. 11, No. 1, April 2020, 41–51

Multivariate statistical inference
Research Paper

A likelihood ratio test for correlated paired multivariate samples

Adolphus Wagala*

[1]Departamento de Probabilidad y Estadística, Centro de Investigación en Matemáticas, AC, Guanajuato, Mexico

(Received: 02 August 2019 · Accepted in final form: 20 April 2020)

Abstract

Many laboratory experiments in the fields of biological sciences usually involve two main groups say the healthy and infected subjects. In one of these kind of experiments, each specimen from each group can be divided in two portions; one portion is stimulated while the other remains unstimulated. Consequently resulting into two main groups with paired measurements that are correlated. For all the groups, p genes are measured for expression. The stimulation in this case can be done by introducing a known infection causing micro-organism like the group A streptococcus which is usually associated with the acute rheumatic fever. An important question in such experiment would be to statistically test for the differences in the differences in means for the healthy and the infected groups. That is, the difference in the means of the healthy group (stimulated and unstimulated) is tested against the difference in the means of the infected (stimulated and unstimulated) group. In this paper, a likelihood ratio test statistic is developed for such kind of problems. The developed statistics and the Hotelling T^2 statistic are both applied to the data are simulated from real biological situations and their performances are compared. The simulated data exhibit the correlation structure similar to that of real biological data obtained from experiments involving the milliplex analyst biomarker data sets. The results indicate that the proposed test statistic give the same conclusions for the hypotheses tested as those of the Hotelling T^2 test. However, the proposed test is intuitively more appealing since it takes care of the correlations between the pairs in the data. The simulation study confirms that the test statistics follow a chi-square distribution. This research contributes a theoretical analysis of paired correlated samples motivated by a practical problem for which the existing statistical methods in use have seldomly taken into account the correlation structure of the data.

Keywords: Correlated pairs· Likelihood ratio test · Multivariate samples

Mathematics Subject Classification: Primary62H15 · Secondary 62J15.

1. Introduction

Consider an experiment involving two groups of subjects namely the healthy (H) and the infected (I) donors. Each group is further divided into two sub-groups whereby one subgroup is stimulated using some infection causing organism for example group A streptococcus (GAS) which causes the acute rheumatic fever (ARF). The other subgroup remains unstimulated. As a result, we end up with paired samples for the H and also another paired

*Corresponding author. Email:adolphus.wagala@cimat.mx

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

samples for the I, resulting into two groups with paired measurements that are correlated. The samples from all these groups are then sequenced to measure the expression levels for the p genes under consideration. The genes whose expression levels are measured are the same for all the paired groups. It is expected that the GAS stimulation of H and I subjects can help in understanding how the GAS affects the H and I subjects thereby possibly able to identify the biomarkers associated with the ARF. The effect of GAS stimulation/unstimulation can lead to changes in the genes with regards to up or down regulations or no change. Assuming that the sample sizes for H and I subjects are m and k respectively and that p genes are considered in the experiment. It is easy to see that the m paired measurements for the H are correlated and at the same time the k paired measurements for the I are correlated while H and I groups are independent. Furthermore, since the genes usually act in a group, the p genes are expected to be correlated.

The main goal therefore is to develop a statistical framework for testing the changes in expression levels in the different sets of genes between the two main groups which have the properties of independence between them but paired correlation within the subjects. The observations are independent and identically distribute (IID). We use the well known likelihood ratio theory to formally derive a new test for formally testing for the difference in the differences of the mean expression levels for the healthy and infected subjects.

The remainder of this paper is organized as follows, Section 2 gives a brief review of the likelihood ratio testing. The proposed likelihood ratio test statistic for multivariate paired, correlated samples is presented in Section 3 while the simulation study is given in 4. Finally, the summary and conclusions are given in Section 5.

2. The likelihood ratio test

The theory of the likelihood ratio test (LRT) is well understood and has been utilized extensively in the field of statistical inference. Most standard multivariate statistics books like for example Anderson (2003), Seber (2004), Mardia et al. (1980), Johnson and Wichern (2007) to mention but a few, contain comprehensive treatment of this subject matter.

To review, the LRT, we start by letting $\boldsymbol{\theta}$ be the parameter vector for the likelihood function $L(\boldsymbol{\theta})$ with observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ with a density function given by $f(\boldsymbol{x}; \boldsymbol{\theta})$. If the parameter space is given by $\boldsymbol{\Theta}$ and suppose that we want to test the null hypothesis $H_o : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ where $\boldsymbol{\Theta}_0$ is a subset of $\boldsymbol{\Theta}$. The parameter space $\boldsymbol{\theta}$ is unconstrained while $\boldsymbol{\theta}_0$ is constrained. The LRT statistic is given by

$$\Lambda = \frac{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_o} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta})}.$$

The null hypothesis H_o is rejected when $\Lambda < C$, where C is a critical value depending on the type-I error. The LRT has good power properties asymptotically and usually is as good or better than many other test statistics Seber (2004). The LRT statistic under general conditions and with large samples are approximately $\chi^2_{(d)}$ distributed where d is the degree of freedom which in general is given by the total number of variables under consideration. The LRT is given by

$$-2\text{Log}\Lambda = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \{-2\text{Log L}(\boldsymbol{\theta})\} - \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \{-2\text{Log L}(\boldsymbol{\theta})\}.$$

Some common problems that have been tackled in the said standard multivariate statistics analysis setting with regards to the LRT include the following.

- Suppose we have N observations on $\boldsymbol{X}$ that is multivariate normally distributed accord-

ing to $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, a test statistic is derived to test for the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ when Σ is unknown. The obvious MLE for Σ in this case is the sample covariance. The resultant test statistics is the T^2 statistics which follows the T^2 Hotelling distribution. This test can be used for testing the hypothesis about the mean vector $\boldsymbol{\mu}$ of the population and obtaining the confidence region for the unknown vector $\boldsymbol{\mu}$ see (Anderson, 2003; Seber, 2004; Mardia et al., 1980; Johnson and Wichern, 2007).

- The two sample problem with unequal covariance matrices has also been addressed. In this case, let $\left\{\boldsymbol{y_j^{(i)}}\right\}, j = 1, \ldots, N$ be samples from $\mathcal{N}(\boldsymbol{\mu}^{(i)}, \Sigma_i), i = 1, 2$ a test statistic for testing $H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$ is developed. The distribution for the respective sample mean vectors is given by $\mathrm{E}(\bar{\boldsymbol{y}}^{(1)} - \bar{\boldsymbol{y}}^{(2)}) = \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}$ while the covariance for the difference $\mathrm{Cov}(\bar{\boldsymbol{y}}^{(1)} - \bar{\boldsymbol{y}}^{(2)}) = \Sigma_1/N_1 + \Sigma_2/N_2$. It is shown that when $N_1 \neq N_2$ and assuming that $N_1 < N_2$ then a suitable test would be a T^2 test with $[N_1 - 1)$ degrees of freedom; see (Anderson, 2003).
- When Σ_1 and Σ_2 are assumed to be equal and unknown, then a pooled sample covariance is used as an estimate. The test statistic is found to be the usual T^2 which follows the T^2 distribution; see (Anderson, 2003; Seber, 2004).
- The topic of paired comparisons is also treated especially in Johnson and Wichern (2007) in which for the paired samples, the difference between them is calculated. The T^2 test is then applied to the differences.
- Most of the likelihood problems tackled only compare two mean vectors and the resultant statistic is the T^2 with a certain degree of freedom depending on the problem set-up.

In other related type of studies, Varuzza and Pereira (2010) developed an exact significance test for comparing digital expression profiles which took in to the asymptotic properties unlike the χ^2 test. Furthermore Lim et al. (2010) developed LRT to compare multiple multivariate normally correlated samples.

3. Proposed LRT statistic

Following the illustration in Section 1, for the healthy subjects, suppose that each gene has m paired measurements $[(h_{u1}, h_{s1}), (h_{u2}, h_{s2}), \ldots, (h_{um}, h_{sm})]$ where h symbolizes one of the groups, say healthy while the subscripts u and s stand for unstimulated and stimulated respectively. Therefore first measurement is for the expression level for the unstimulated specimen, while the second one is for a stimulated one for the same subject. In a similar manner let a represent the second group, say the infected subjects. Assume that each of the p genes has k paired measurements $[(a_{u1}, a_{s1}), (a_{u2}, a_{s2}), \ldots, (a_{uk}, a_{sk})]$ for the unstimulated and stimulated specimens in each pair respectively.

The m measurements from healthy subjects are assumed to be IID from a multivariate normal distribution $\begin{pmatrix} \boldsymbol{h}_u \\ \boldsymbol{h}_s \end{pmatrix} \sim \mathcal{N}_{2p}\left[\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_s \end{pmatrix}, \boldsymbol{\Sigma}\right]$ and the k measurements from the infected subjects are also assumed to be IID from a multivariate normal distribution $\begin{pmatrix} \boldsymbol{a}_u \\ \boldsymbol{a}_s \end{pmatrix} \sim \mathcal{N}_{2p}\left[\begin{pmatrix} \boldsymbol{\nu}_u \\ \boldsymbol{\nu}_s \end{pmatrix}, \boldsymbol{\Sigma}\right]$. Here, $\boldsymbol{\mu}_u$ and $\boldsymbol{\mu}_s$ represent the mean vectors for unstimulated and stimulated healthy subjects respectively. On the other hand, $\boldsymbol{\nu}_u$ and $\boldsymbol{\nu}_s$ denote mean vectors for unstimulated and stimulated infected subjects respectively while $\boldsymbol{\Sigma}$ is the covariance matrix which is assumed to be the same for the two groups of healthy and infected.

The hypotheses to be tested are:

$$H_0 : (\boldsymbol{\mu}_u - \boldsymbol{\mu}_s) = (\boldsymbol{\nu}_u - \boldsymbol{\nu}_s) \text{ versus } H_a : (\boldsymbol{\mu}_u - \boldsymbol{\mu}_s) \neq (\boldsymbol{\nu}_u - \boldsymbol{\nu}_s).$$

Case 1: Assuming the covariance matrix Σ is known For m healthy subjects denote a $2p \times 1$ vector of parameters $\boldsymbol{\mu} = \binom{\boldsymbol{\mu}_u}{\boldsymbol{\mu}_s}$ for the random vector $\boldsymbol{h} = \binom{\boldsymbol{h}_u}{\boldsymbol{h}_s}$ where the first p elements represent the elements of $\boldsymbol{h}_u$ while the remaining p represents the $\boldsymbol{h}_s$. Similarly for the k infected subjects we have the vector of parameters $\boldsymbol{\nu} = \binom{\boldsymbol{\nu}_u}{\boldsymbol{\nu}_s}$ and is associated with random variables $\boldsymbol{a} = \binom{\boldsymbol{a}_u}{\boldsymbol{a}_s}$ and $\boldsymbol{\nu}$ is of $2p \times 1$ dimension.

The joint probability density function is given as

$$f(\boldsymbol{h}, \boldsymbol{a}) = (2\pi)^{-p}|\boldsymbol{\Sigma}|^{-1}\exp\Big(-\frac{1}{2}\big[(\boldsymbol{h}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{h}-\boldsymbol{\mu}) + (\boldsymbol{a}-\boldsymbol{\nu})'\Sigma^{-1}(\boldsymbol{a}-\boldsymbol{\nu})\big]\Big).$$

A reduced -2log of the likelihood function in terms of sufficient statistics is given by

$$-2\text{Log L}(\boldsymbol{\mu}, \boldsymbol{\nu}) = B + m(\bar{\boldsymbol{h}}-\boldsymbol{\mu})'\Sigma^{-1}(\bar{\boldsymbol{h}}-\boldsymbol{\mu}) + k(\bar{\boldsymbol{a}}-\boldsymbol{\nu})'\Sigma^{-1}(\bar{\boldsymbol{a}}-\boldsymbol{\nu}), \tag{1}$$

where B is a constant that does not contain the parameters under consideration and vanishes during the optimization.

The MLEs under H_o are obtained by considering the parameter space given by $\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{\nu} : -\infty < -\infty < \boldsymbol{\mu}, \boldsymbol{\nu} < \infty\}$ and then optimizing the constrained log-likelihood function using the Lagrangian $S(\boldsymbol{\Theta}, \boldsymbol{\lambda}) = -2\text{LogL}(\boldsymbol{\mu}, \boldsymbol{\nu}) + \boldsymbol{\lambda}'(\boldsymbol{\mu}_u - \boldsymbol{\mu}_s - \boldsymbol{\nu}_u + \boldsymbol{\nu}_s)$. The constraint $\boldsymbol{\lambda}'(\boldsymbol{\mu}_u - \boldsymbol{\mu}_s - \boldsymbol{\nu}_u + \boldsymbol{\nu}_s)$ is conveniently expressed in a matrix form as $A(\boldsymbol{\mu}-\boldsymbol{\nu}) = 0$ where $A = (\boldsymbol{I}, -\boldsymbol{I})$ and $\boldsymbol{I}$ is a $p \times p$ identity matrix. The constraint added to Equation (1) is of the form $2(\boldsymbol{\mu}-\boldsymbol{\nu})'A'\boldsymbol{\lambda} = 2[\boldsymbol{\lambda}'A(\boldsymbol{\mu}-\boldsymbol{\nu})]'$. The partial derivatives of the constrained function with respect to each unknown parameter are given as

$$\frac{\partial S(\boldsymbol{\Theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\mu}} = -2m\Sigma^{-1}(\bar{\boldsymbol{h}}-\boldsymbol{\mu}) + 2A'\boldsymbol{\lambda}, \tag{2}$$

$$\frac{\partial S(\boldsymbol{\Theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\nu}} = -2k\Sigma^{-1}(\bar{\boldsymbol{a}}-\boldsymbol{\nu}) - 2A'\boldsymbol{\lambda}, \tag{3}$$

$$\frac{\partial S(\boldsymbol{\Theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = 2A(\boldsymbol{\mu}-\boldsymbol{\nu}). \tag{4}$$

Now, equating (2), (3) and (4) to zero and simplifying, we get

$$\Sigma^{-1}(\bar{\boldsymbol{h}}-\boldsymbol{\mu}) - \frac{1}{m}A'\boldsymbol{\lambda} = 0, \tag{5}$$

$$\Sigma^{-1}(\bar{\boldsymbol{a}}-\boldsymbol{\nu}) + \frac{1}{k}A'\boldsymbol{\lambda} = 0, \tag{6}$$

$$A(\boldsymbol{\mu}-\boldsymbol{\nu}) = 0.$$

Subtracting Equation (5) from (6) and with some algebraic manipulations results in

$$\begin{aligned}
\Sigma^{-1}(\bar{\boldsymbol{a}} - \boldsymbol{\nu} - \bar{\boldsymbol{h}} + \boldsymbol{\mu}) + \Big[\frac{1}{m} + \frac{1}{k}\Big] A'\boldsymbol{\lambda} &= 0 \\
(\bar{\boldsymbol{a}} - \boldsymbol{\nu} - \bar{\boldsymbol{h}} + \boldsymbol{\mu}) &= -\Big[\frac{1}{m} + \frac{1}{k}\Big]\Sigma A'\boldsymbol{\lambda} \\
A(\boldsymbol{\mu} - \boldsymbol{\nu}) + A(\bar{\boldsymbol{a}} - \bar{\boldsymbol{h}}) &= -\Big[\frac{m+k}{mk}\Big] A\Sigma A'\boldsymbol{\lambda} \\
A(\bar{\boldsymbol{h}} - \bar{\boldsymbol{a}}) &= \Big[\frac{m+k}{mk}\Big] A\Sigma A'\boldsymbol{\lambda} \\
\boldsymbol{\lambda} &= \Big[\frac{mk}{m+k}\Big](A\Sigma A')^{-1} A(\bar{\boldsymbol{h}} - \bar{\boldsymbol{a}}) \\
\boldsymbol{\lambda} &= \Big[\frac{mk}{m+k}\Big](A\Sigma A')^{-1}\Delta \qquad (7)
\end{aligned}$$

where $\Delta = A\bar{\boldsymbol{h}} - A\bar{\boldsymbol{a}}$. From Equations (5) and (6), we get

$$\hat{\boldsymbol{\mu}}_0 = \bar{\boldsymbol{h}} - \frac{1}{m}\Sigma A'\boldsymbol{\lambda} \qquad (8)$$

$$\hat{\boldsymbol{\nu}}_0 = \bar{\boldsymbol{a}} + \frac{1}{k}\Sigma A'\boldsymbol{\lambda} \qquad (9)$$

The MLEs under the alternative hypothesis H_a are obtained by maximizing the unconstrained likelihood function are given by; $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{h}}$ and $\hat{\boldsymbol{\nu}} = \bar{\boldsymbol{a}}$.

Now, let $\boldsymbol{\theta}$ be the parameter vector for the likelihood function $L(\boldsymbol{\theta})$ with observations from the paired samples of healthy and infected subjects. Consider the parameter space given by $\boldsymbol{\Theta}$; we wish to test the null hypothesis $H_o : \boldsymbol{\theta} \in \boldsymbol{\Theta}$ versus the alternative $H_a : \boldsymbol{\theta} \notin \boldsymbol{\Theta}$.

Substituting the MLEs under H_0 (Equations (8) and (9)) into the log likelihood function given by Equation (1) we get

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_o}\{-2\text{Log L}(\boldsymbol{\theta})\} \\
&= B + m\left(\frac{1}{m}\Sigma A'\boldsymbol{\lambda}\right)'\Sigma^{-1}\left(\frac{1}{m}\Sigma A'\boldsymbol{\lambda}\right) + k\left(\frac{1}{k}\Sigma A'\boldsymbol{\lambda}\right)'\Sigma^{-1}\left(\frac{1}{k}\Sigma A'\boldsymbol{\lambda}\right) \\
&= B + \frac{1}{m}\left(\boldsymbol{\lambda}'A\Sigma\right)\Sigma^{-1}\left(\Sigma A'\boldsymbol{\lambda}\right) + \frac{1}{k}\left(\boldsymbol{\lambda}'A\Sigma\right)\Sigma^{-1}\left(\Sigma A'\boldsymbol{\lambda}\right) \\
&= B + \frac{1}{m}\left(\boldsymbol{\lambda}'A\Sigma A'\boldsymbol{\lambda}\right) + \frac{1}{k}\left(\boldsymbol{\lambda}'A\Sigma A'\boldsymbol{\lambda}\right) \\
&= B + \frac{[k+m]}{mk}\left(\boldsymbol{\lambda}'A\Sigma A'\boldsymbol{\lambda}\right). \qquad (10)
\end{aligned}$$

We now substitute for the expression of $\boldsymbol{\lambda}$ from (7) into Equation (10) to get

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_o}\{-2\text{Log L}(\boldsymbol{\theta})\} = \\
&B + \frac{[k+m]}{mk}\left(\frac{mk}{[m+k]}(A\Sigma A')^{-1}\Delta\right)'(A\Sigma A')^{-1}\left(\frac{mk}{[m+k]}(A\Sigma A')^{-1}\Delta\right) \\
&= B + \frac{mk}{[m+k]}\Delta'(A\Sigma A')^{-1}\Delta.
\end{aligned}$$

Under the unconstrained hypothesis $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{-2\text{Log L}(\boldsymbol{\theta})\} = B$. The log LRT is therefore given as

$$\begin{aligned} 2\text{Log}\Lambda &= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_o}\{-2\text{Log L}(\boldsymbol{\theta})\} - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{-2\text{Log L}(\boldsymbol{\theta})\} \\ &= \frac{mk}{[m+k]}\Delta'(A\Sigma A')^{-1}\Delta \end{aligned} \tag{11}$$

The distribution of $\Delta = A\bar{\boldsymbol{h}} - A\bar{\boldsymbol{a}}$ is $\Delta \sim N\left((A(\boldsymbol{\mu}-\boldsymbol{\nu}), \frac{(k+m)}{mk}(A\Sigma A')^{-1}\right)$. If H_0 is true then $A(\boldsymbol{\mu}-\boldsymbol{\nu}) = 0$ so that $\Delta \sim N\left(0, \frac{(k+m)}{mk}(A\Sigma A')^{-1}\right)$. It is well known that given that $X \sim N_p(0,V)$ then $V^{-\frac{1}{2}} \sim N(0,I)$ implying that $(V^{-\frac{1}{2}}X)^T(V^{-\frac{1}{2}}X) \sim \chi^2_{(p)}$ and so $X^TV^{-1}X \sim \chi^2_{(p)}$, thus

$$-2\text{Log}\Lambda = \frac{mk}{(m+k)}\Delta'(A\Sigma A')^{-1}\Delta \sim \chi^2_{(p)}. \qquad \blacksquare$$

Case 2: Assuming the covariance matrix Σ is unknown We estimate the covariance matrix by first rewriting the -2log likelihood as

$$\begin{aligned} l = mp\log(2\pi) + m\log|\Sigma| + \text{tr}\Sigma^{-1}\boldsymbol{S}_h + \text{tr}\Sigma^{-1}(\bar{\boldsymbol{h}}-\boldsymbol{\mu})(\bar{\boldsymbol{h}}-\boldsymbol{\mu})' \\ +kp\log(2\pi) + k\log|\Sigma| + \text{tr}\Sigma^{-1}\boldsymbol{S}_a + \text{tr}\Sigma^{-1}(\bar{\boldsymbol{a}}-\boldsymbol{\nu})(\bar{\boldsymbol{a}}-\boldsymbol{\nu})', \end{aligned} \tag{12}$$

where $\boldsymbol{S}_h = \sum_{i=1}^m(\boldsymbol{h}_i-\bar{\boldsymbol{h}})(\boldsymbol{h}_i-\bar{\boldsymbol{h}})'$ and $\boldsymbol{S}_a = \sum_{j=1}^k(\boldsymbol{a}_j-\bar{\boldsymbol{a}})(\boldsymbol{a}_j-\bar{\boldsymbol{a}})'$. We obtain the partial derivative of l (12) with respect to Σ^{-1}, then equate the result to zero. The estimator for the variance-covariance matrix is then obtained as

$$\hat{\Sigma} = \frac{1}{[m+k]}\left[\boldsymbol{S}_h + \boldsymbol{S}_a + m(\bar{\boldsymbol{h}}-\hat{\boldsymbol{\mu}})(\bar{\boldsymbol{h}}-\hat{\boldsymbol{\mu}})' + k(\bar{\boldsymbol{a}}-\hat{\boldsymbol{\nu}})(\bar{\boldsymbol{a}}-\hat{\boldsymbol{\nu}})'\right].$$

By substituting the plug-in estimators for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\nu}}$ which are $\bar{\boldsymbol{h}}$ and $\bar{\boldsymbol{a}}$ respectively, we get the plug-in estimator for the covariance matrix as

$$\hat{\Sigma} = \frac{1}{[m+k]}\left[\boldsymbol{S}_h + \boldsymbol{S}_a\right].$$

The estimator $\hat{\Sigma}$ is then plugged-in into the LRT statistic given in Equation (11) which has $\chi^2_{(p)}$ distribution to get

$$-2\text{Log}\Lambda = \frac{mk}{[m+k]}\Delta'(A\hat{\Sigma}A')^{-1}\Delta. \tag{13}$$

Proposition Denote Equation (11) by $\Lambda_1 = \frac{mk}{[m+k]}\Delta'(A\Sigma A')^{-1}\Delta$ and (13) by $\Lambda_2 = \frac{mk}{[m+k]}\Delta'(A\hat{\Sigma}A')^{-1}\Delta$ and noting that $\hat{\Sigma}$ is a consistent estimator of Σ. Since $\Lambda_1 \overset{d}{\sim} \chi^2_{(p)}$ then $\Lambda_2 \overset{a}{\sim} \chi^2_{(p)}$, where $\overset{d}{\sim}$ means exactly distributed while $\overset{a}{\sim}$ stands for asymptotically distributed.

PROOF Since $\hat{\Sigma} \xrightarrow{p} \Sigma$ as $n \to \infty$ where $n = m + k$ and the fact that $(A\Sigma A')$ is positive definite, we had shown in Case 1 that $A(\bar{\boldsymbol{h}} - \bar{\boldsymbol{a}}) \overset{d}{\sim} \mathrm{N}\left(0, \frac{m+k}{mk} A\Sigma A'\right)$ under H_0 then it follows that in a similar manner $A(\bar{\boldsymbol{h}} - \bar{\boldsymbol{a}}) \overset{a}{\sim} \mathrm{N}\left(0, \frac{m+k}{mk} A\hat{\Sigma} A'\right)$ under H_0. Consequently the LRT statistic $\frac{mk}{[m+k]} \Delta' (A\hat{\Sigma} A')^{-1} \Delta \overset{a}{\sim} \chi^2_{(p)}$. ■

REMARK We note that the world applications, p is usually less than n, that is, $p < n$. In such a case, the derived statistic in Equation (13) becomes untenable because the matrix $(A\hat{\Sigma} A')$ is singular. In order to overcome this problem, the usage of the general inverse as in Ben-Israel and Greville (2003) of the covariance matrix is instead used.

4. Simulation study

In this section, a synthetic data are generated and then analyzed using the proposed LRT method and the well known Hotelling T^2 statistic. All the simulations and data analysis were done using the `R` software (R Core Team, 2020). The data are simulated with the following different set-ups.

- The mean vector for the "healthy unstimulated" is obtained by first simulating p uniform random variables in the range of $(0, 0.5)$ to the vector $\boldsymbol{\mu}_u$.
- Similarly we generate p uniform random variables in the interval $(0.6, 0.75)$ to create $\boldsymbol{\mu}_s$ which is the "healthy unstimulated".
- For the "infected unstimulated", the values for simulation of $\boldsymbol{\nu}_u$ used to generate uniform random variables of dimension p is $(0, 0.55)$.
- The $\boldsymbol{\nu}_s$ are obtained by generating a p uniform random variables of the interval $(0.001, 0.2)$ to obtain the mean vector for the "infected stimulated".

For each of the category, we assume that all the two paired measurements we generate at randomly a $2p \times 2p$ positive definite covariance matrix V. The number of subjects for the healthy group is arbitrarily set at 20 while the infected group is set at 19.

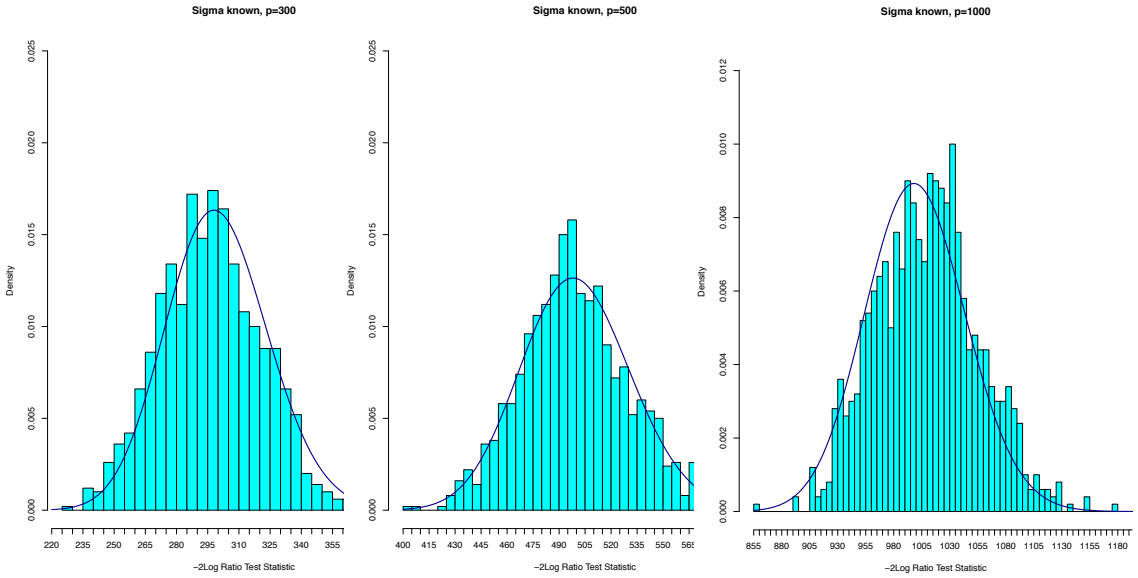
The data are simulated for four different values of p namely $p = \{300, 500, 1000\}$ while the sample sizes were fixed at $m = 20$ and $k = 19$. A LRT statistic and the corresponding p-value are calculated when Σ is assumed to be unknown and when it is known. A resampling distribution is then obtained from which an approximate p-value is then computed. The results are shown in Table 1 in addition to the plots in Figure 1.

The proposed statistic is applied to the simulated data. The results presented in Table 1 reveal that both the calculated p-value and the one obtained from resampling lead to the same conclusions regarding the hypothesis testing. In this case, for all the cases, the difference in the means was statistically significant at 5% level.

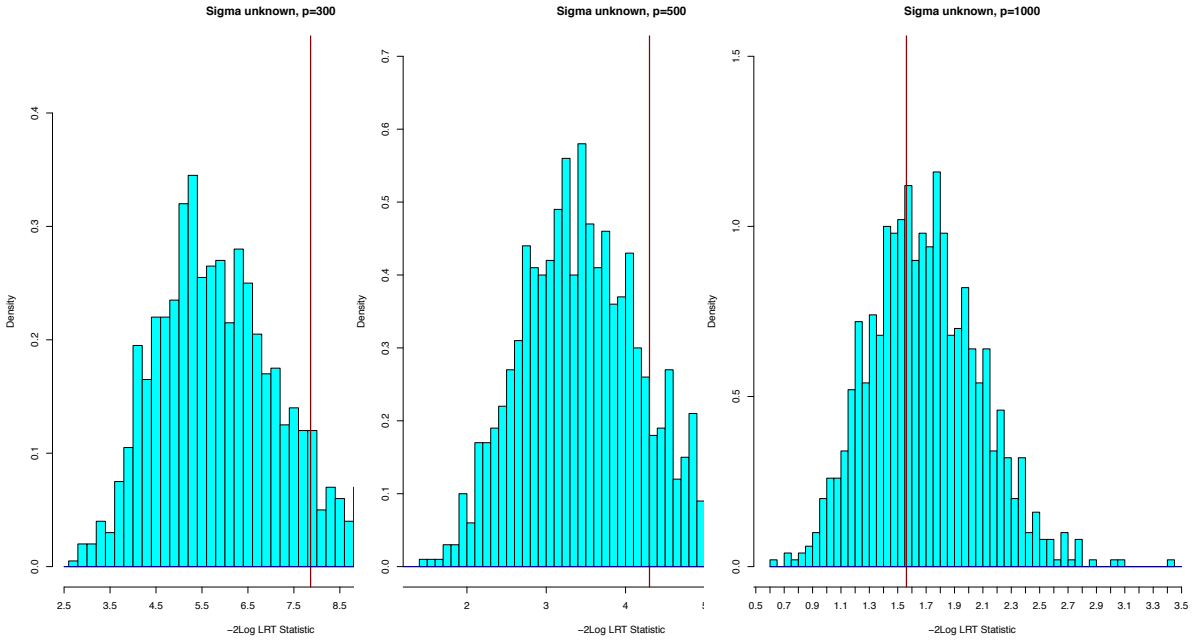
Table 1.: Calculated LRT statistic and the p-values for the simulation experiment 1.

	$p = 300$		$p = 500$		p =1000	
	Σ known	Σ Unknown	Σ known	Σ unknown	Σ known	Σ unknown
Log LRT	373.61	7.86	617.43	4.3	1202.05	1.56
calculated p-value*	0.00025	1.00	0.002	1.00	0.00	1.00
p-value from resampling	0.001	0.396	0.001	0.166	0.00	0.601

* p-values calculated from the exact $\chi^2_{(p)}$ distribution.



(a) p =300, Σ-known (b) p =500, Σ-known (c) p =1000, Σ-known



(d) p =300, Σ-unknown (e) p =500, Σ-unknown (f) p =1000, Σ-unknown

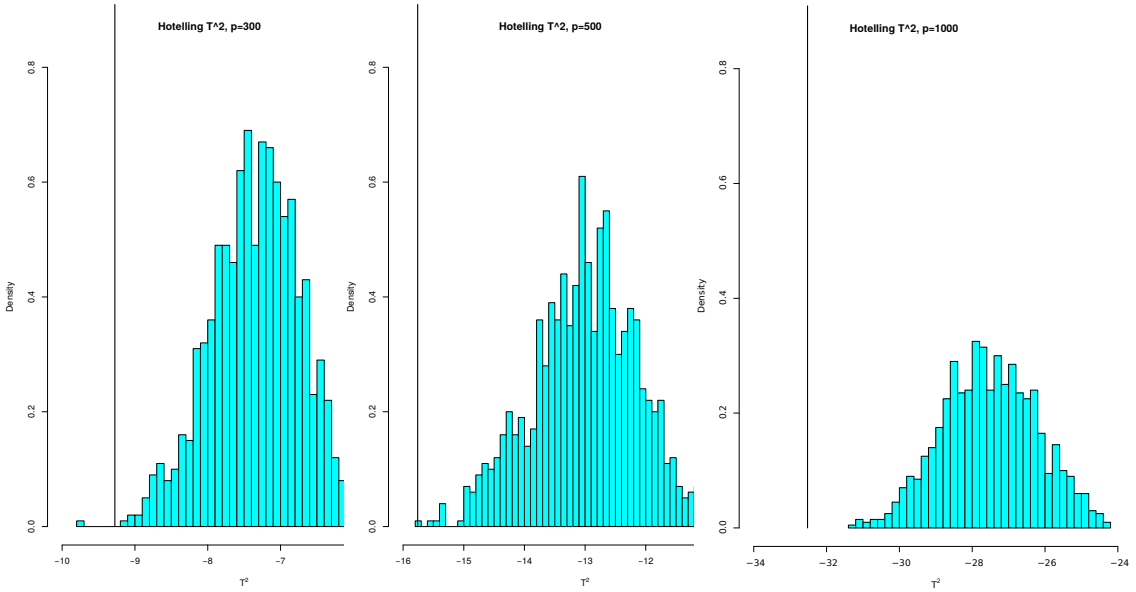
Figure 1.: Histograms for the proposed LRT from a the permutation of the statistic for p =300, 500 and 1000, Σ known and unknown.

In Figure 1, the curves are the chi-squared densities for the corresponding degrees of freedom p. The plots show that the distributions for the -2 log likelihood test statistic follow a chi-square distribution and are also positive skewed. However, as the number of p increases, the distributions look like normal distribution and the skewness is less when the degree of freedom is higher. The normal looking distribution are still a chi-squared, for they approach $\mathrm{N}(p, 2p)$ distribution as the degree of freedom gets large. The red vertical lines (when shown) indicates the position of the computed statistic for the un-resampled data. The plots without the vertical lines are the ones whose computed statistic is far too small beyond the scale used in plotting.

The simulated data are analyzed using the Hotelling T^2 statistic (Hotelling, 1931) in order to compare the performance of our proposed method with it. During the computation, when number of variables p is much greater than the number of samples n, then the covariance matrix is estimated using the shrinkage approach of Schäfer and Strimmer (2005). The results are presented in Table 2 . The results indicate that there is a significant difference in the means at 5% significance level. The permutations for Hotelling T^2 statistic is done and the different values plotted on an histogram shown in Figure 2 which reveals that the statistic is chi-square distributed for all the different values of p. The results are consistent with the one obtained by the proposed algorithm.

Table 2.: Hotelling T^2 values for the simulated data.

	p=300	p=500	p=1000
Hotelling T^2 value	-9.28	-15.76	-32.52
p-value from resampling	0.001	0.00	0.00



(a) p =300, Σ-known (b) p =500, Σ-unknown (c) p =1000, Σ-unknown

Figure 2.: Histograms of the distribution of the permuted test statistics for Hotelling T^2 when p =300, 500 and 1000, Σ unknown.

5. Conclusions

In this research, we have considered two main groups (say, healthy and infected specimens) with paired measurements that are correlated. We aim to provide a proper statistical framework for testing the difference in the difference in the means for the healthy and infected subjects. We have shown that this is not a trivial problem and so derived a likelihood ratio test for these differences. The derived test do follow a chi-square distribution with p degrees of freedom when the variance-covariance matrix is known. We have assumed that the observed measurements follow a multivariate normal distribution with a known variance-covariance matrix which can be deduced from the prior network that has been chosen. Finally, a likelihood ratio test statistic has been derived when the variance-covariance matrix is unknown. A simulation study has been done and demonstrated that the developed tests can be useful when applied to other cases which have similar problem set-ups. The study demonstrated that the proposed test statistic give the same conclusions for the hypotheses tested as those of the Hotelling T^2 test. However, the proposed test is intuitively more appealing since it takes care of the correlations between the pairs in the experiments. This research contributes a theoretical analysis of paired correlated samples motivated by a practical problem for which no formal statistical method is in use.

Acknowledgements

This research was done during my academic visit at the Speed Lab as a PhD candidate (partly supported by the Mexico's Consejo Nacional de Ciencias y Tecnología (CONACyT) scholarship number 384101), Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research (WEHI), Melbourne, Australia. I am very grateful to Prof. Terence Speed for his guidance through out this period. Much appreciation to Dr. Jo Keeble and Prof. Ian Wicks for their support and discussions on the Acute Rheumatic Fever (ARF) project that lead to the conception of ideas relating to this work. Much appreciation to Dr. Graciela González Farías of CIMAT, Gto, México for the useful comments that helped in improving this article. This research was partially supported by the Mexico's Consejo Nacional de Ciencias y Tecnología (CONACyT) project number 252996 through Dr. Graciela González Farías.

References

Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis. Wiley, New York.

Ben-Israel, A. and Greville, T., 2003. Generalized Inverses:Theory and Applications. Springer, New York.

Hotelling, H., 1931. The generalization of student's ratio. Annals of Mathematical Statistics, 2:360–378.

Johnson, R.A. and Wichern, D.W., 2007. Applied Multivariate Analysis. Prentice Hall, New Jersey.

Lim, J., Li, E., and Lee., S.J., 2010. Likelihood ratio tests of correlated multivariate samples. Journal of Multivariate Analysis, 101:541–554.

Mardia, K.V., Kent, J.T., and Bibby, J.M., 1980. Multivariate Analysis. Academic Press, New York.

Schäfer, J. and Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology, 4:32.

Seber, A. 2004. Multivariate Observations. Wiley, New York.

R Core Team 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Varuzza, L. and Pereira, C., 2010. Significance test for comparing digital gene expression profiles: Partial likelihood application. Chilean Journal of Statistics, 1:91–102.

Chilean Journal of Statistics
Vol. 11, No. 1, April 2020, 53–67

Distribution Theory
Research Paper

A new one-parameter unit-Lindley distribution

Josmar Mazucheli[1], Sudeep R. Bapat[2,*], and André Felipe B. Menezes[1]

[1]Department of Statistics, Universidade Estadual de Maringá, DEs, PR, Brazil
[2]Department of Statistics and Applied Probability, University of California, Santa Barbara, USA

(Received: 28 May 2020 · Accepted in final form: 11 December 2019)

Abstract

A large number of useful distributions for data analysis are obtained by transforming different random variables. An example is the one-parameter unit-Lindley distribution, obtained by transforming a random variable which has a Lindley distribution. In this paper, we introduce a new one-parameter unit-Lindley distribution, useful for data analysis in the interval (0,1]. It follows some interesting properties such as having closed form expressions for the moments, belonging to the exponential family. We also analyze a practical application having covariates, by setting up a suitable regression and show that our model fits much better than both unit-Lindley and beta regressions.

Keywords: Maximum likelihood estimation · Proportion data · Regression model · Unit-Lindley distribution · Unit interval.

Mathematics Subject Classification: Primary 60E05 · Secondary 62F10.

1. Introduction

In many practical applications, one encounters data which is spread out in a bounded interval. Moreover this interval happens to be $(0, 1)$, where the data would be certain proportions, ratios or standardized scores. Some of the well known distributions having supports in $(0, 1)$ are uniform, beta and Kumaraswamy. However all of these contain at least 2 parameters and hence it becomes tedious when it comes to estimation. Further, the beta distribution doesn't have closed form expressions for the cumulative distribution function (CDF), whereas the Kumaraswamy distribution fails to have a closed form expression for the moments. Some of the only one-parameter distributions in $(0, 1)$ are the Topp-Leone distribution (Topp and Leone, 1955) and the newly proposed unit-Lindley distribution by Mazucheli et al. (2019), where the authors have transformed a suitable Lindley distribution. One of the outlook in recent times has been to transform some existing distributions to get more useful distributions having specific properties. A lot of work has been done related to the Lindley distribution in the last few years. Some of the prominent works include the quasi-Lindley distribution by Shanker and Mishra (2013), the log-Lindley distribution by Gómez-Déniz et al. (2014), the power-Lindley distribution or the generalized-Lindley distribution by Nadarajah et al. (2011).

*Corresponding author. Email: bapat@pstat.ucsb.edu

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

In this paper we propose a new unit-Lindley (NUL) distribution which is a modification to the existing unit-Lindley distribution, by picking a different transformation. This NUL distribution enjoys several interesting properties such as existence of closed form expressions for the moments, the CDF and belonging to the exponential family. Due to its simple formula, one can incorporate a regression setup by involving several covariates in the mean to study their dependence on the response. The advantage of this NUL distribution over the existing unit-Lindley model can be clearly seen through the real-data application which we present in Section 4.

In Section 2 we propose the NUL distribution by providing the density and the distribution functions. We also focus on several interesting properties such as defining the moments, the HR function, the mean residual life function, the quantile function and others. Section 3 involves estimation properties including both method of moments and maximum likelihood (ML) estimators, where we also provide a bias-corrected ML estimators, in addition to a regression modeling. In Section 4, we provide the numerical applications of our work. Extensive simulation analyses are covered by taking a wide range of parameter values. We fit our proposed NUL model to a real-data from finance which involves a ratio of premiums plus uninsured losses and the total assets as the response whereas Section 5 provides brief conclusions.

2. Some Mathematical Results

In the following subsections, we provide a number of key properties of the NUL distribution.

2.1 The NUL distribution

Some probability distributions useful in analyzing data in the unit interval, such as Johnson S_B (Johnson, 1949), Johnson S'_B (Johnson, 1955), unit-Gamma Grassia (1977); Tadikamalla (1981), unit-Logistic (Tadikamalla and Johnson, 1982), log-Lindley (Gómez-Déniz et al., 2014), unit-Inverse-Gaussian (Ghitany et al., 2018), unit-Birnbaum-Saunders (Mazucheli et al., 2018a), unit-Weibull (Mazucheli et al., 2018b) are formulated by transforming specific random variables (RVs). It is important to note that beta and Kumaraswamy (Kumaraswamy, 1980) distributions also can be obtained by transformations.

A unit-Lindley distribution was proposed by Mazucheli et al. (2019) by considering the transformation $X = Y/[1+Y]$, where $Y \sim \text{Lindley}(\theta)$ (Lindley, 1958). Here we apply the transformation $X = 1/[1+Y]$, where $Y \sim \text{Lindley}(\theta)$ and propose the distribution of X to be the NUL distribution. One can easily derive its probability density function (PDF) and the CDF say, using the inverse transform method. These expressions are given respectively by

$$f(x|\theta) = \frac{\theta^2}{x^3\,[1+\theta]} \exp\left(-\theta\left[\frac{1-x}{x}\right]\right), \tag{1}$$

$$F(x|\theta) = \frac{[\theta+x]}{x\,[1+\theta]} \exp\left(-\theta\left[\frac{1-x}{x}\right]\right), \tag{2}$$

where $0 < x \leq 1$ and $\theta > 0$. Figure (1) shows the PDF of the unit-Lindley distribution for selected values of θ.

Unlike other distributions such as the unit-Lindley, here we have the possibility of having observations equal to 1 and from (1) the first derivative of $f(x|\theta)$ is

$$\frac{\mathrm{d}}{\mathrm{d}x} f(x|\theta) = \frac{\theta^2[\theta - 3x]}{[1+\theta]x^5} \exp\left(-\theta\left[\frac{1-x}{x}\right]\right),$$

which implies that the PDF is unimodal with maximum at $X_{\max} = \theta/3$ for all values of $\theta < 3$ and $X_{max} = 1$ for $\theta \geq 3$.

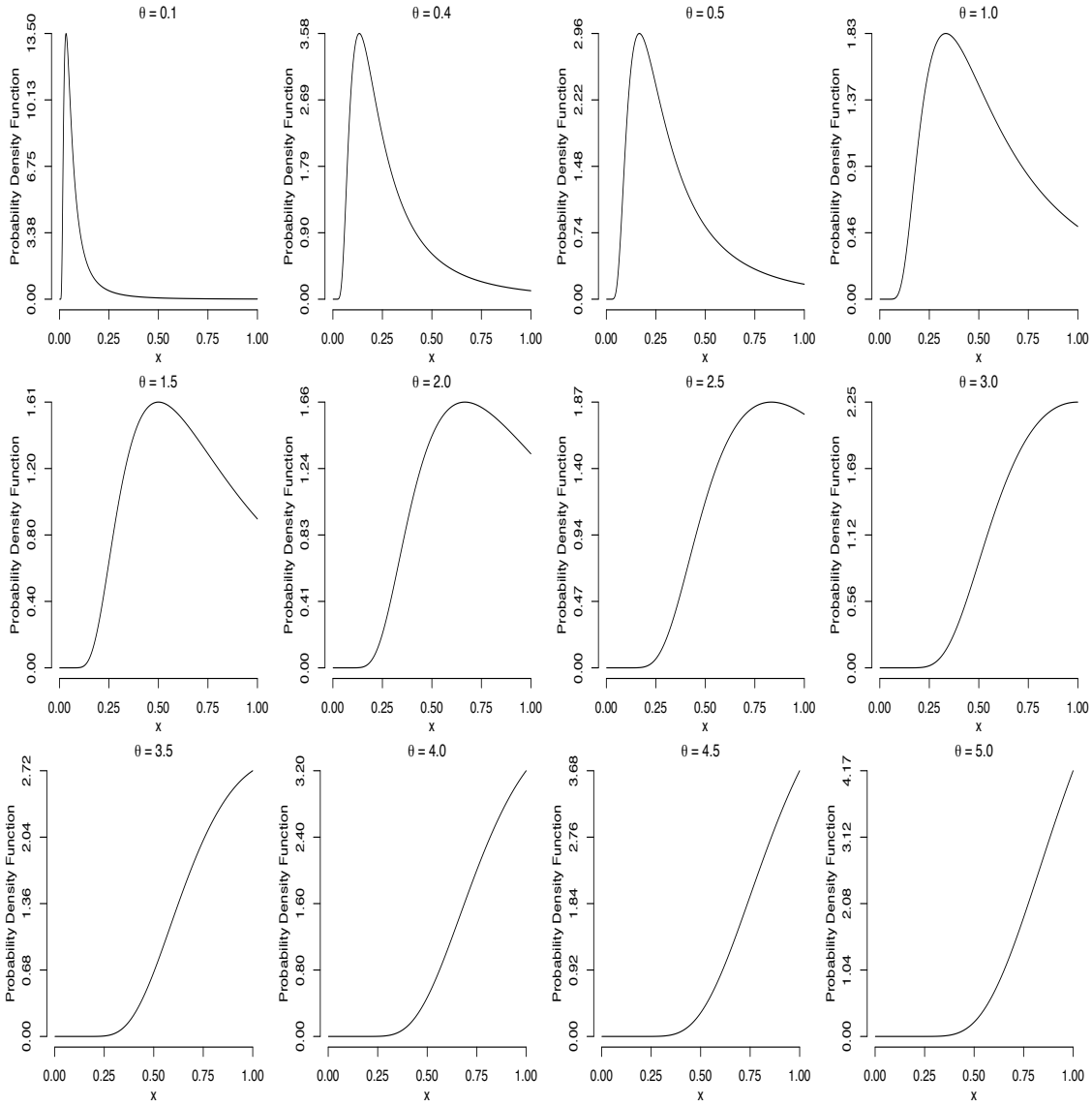


Figure 1. Probability density function of the NUL distribution for selected values of θ.

2.2 Convexity

Proposition 2.1 The CDF of the NUL is convex for $\theta > 3$.

Proof The second derivative of $F(x|\theta)$ is

$$F''(x|\theta) = \frac{\theta^2[\theta - 3x]}{x^5[1+\theta]} \exp\left(-\frac{\theta[1-x]}{x}\right).$$

This implies that for all x in $(0,1)$, $F''(x|\theta) < 0$ only if $\theta < 0$ therefore it can never be concave and $F''(x|\theta) > 0$ if $\theta > 3$. Hence $F(x \mid \theta)$ is a convex function of x for $\theta > 3$.

Proposition 2.2 The PDF of the unit-Lindley distribution is log-concave for all $0 < x \leq 1$ if $\theta > 3/2$.

Proof We know that $f(x \mid \theta)$ is log-concave (log-convex) function of x if for all x in $(0,1]$ $\frac{\mathrm{d}}{\mathrm{d}x} \log f(x|\theta)$ is a non-increasing (non-decreasing) function of x. Note that

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2} \log f(x|\theta) = \frac{\mathrm{d}}{\mathrm{d}x} \frac{f'(x \mid \theta)}{f(x \mid \theta)} = \frac{\mathrm{d}}{\mathrm{d}x} \frac{[\theta - 3x]}{x^2} = -\frac{2[\theta - 3x]}{x^3} - \frac{3}{x^2}.$$

This is always < 0 for all x in $(0,1]$ whenever $\theta > 3/2$. Hence $f(x \mid \theta)$ is log-concave for all $0 < x \leq 1$, if $\theta > 3/2$.

2.3 Hazard rate function

The hazard rate (HR) function of the unit-Lindley distribution is given by

$$h(x|\theta) = \frac{f(x|\theta)}{1 - F(x|\theta)} = \frac{\theta^2}{[\theta + x]x^2}, \quad 0 < x \leq 1.$$

Since $\mathrm{d}h(x|\theta)\mathrm{d}x = -[\theta^2(2\theta + 3x)]/[x^3(\theta + x)^2] < 0$ for all $\theta > 0$ the HR function is decreasing in x. Note that $\lim_{x\to 0} h(x|\theta) = \infty$ while $\lim_{x\to 1} h(x \mid \theta) = \theta^2/[1+\theta]$.

2.4 Moments

The k-th moment about origin of the unit-Lindley distribution can be obtained from

$$\mu'_k = \mathrm{E}\left(X^k\right) = \int_0^1 k x^{k-1} \left\{1 - \frac{[\theta + x]}{x\,[1+\theta]} \exp\left(-\theta\left[\frac{1-x}{x}\right]\right)\right\} \mathrm{d}x, \quad k = 1, 2, \ldots.$$

In particular, for $k = 1, 2, 3, 4$ we get

$$\mu'_1 = \tfrac{\theta}{1+\theta}, \qquad \mu'_2 = \tfrac{\theta^2 \exp(\theta) Ei(1,\theta)}{1+\theta},$$

$$\mu'_3 = \tfrac{\theta^2[1-\theta \exp(\theta) Ei(1,\theta)]}{1+\theta}, \qquad \mu'_4 = \tfrac{\theta^2[1-\theta+\theta^2 \exp(\theta) Ei(1,\theta)]}{2[1+\theta]},$$

where $\mathrm{Ei}(a,z) = \int_1^\infty x^{-a} \exp(-xz)\mathrm{d}x$ is the exponential integral function; see Abramowitz and Stegun (1974).

The k-th incomplete moment about origin is obtained from

$$T_k(t) = \mathrm{E}\left(X^k | X < t\right) = \frac{\theta^2}{[1+\theta]\, F\left(t|\theta\right)} \int_0^t x^{k-3} \exp\left(-\theta\left[\frac{1-x}{x}\right]\right) \mathrm{d}x, \quad k = 1, 2, \ldots.$$

and for for $k = 1, 2, 3, 4$ we have

$$T_1\left(t\right) = \tfrac{\theta t}{[\theta+t]}, \qquad T_2\left(t\right) = \tfrac{\theta^2 \exp\left(\theta\right) t Ei\left(1, \frac{\theta}{t}\right)}{[\theta+t] \exp\left(\theta[t-1]/t\right)},$$

$$T_3\left(t\right) = \tfrac{\theta^2 t\left[t - \theta Ei\left(1, \frac{\theta}{t}\right) \exp\left(\frac{\theta}{t}\right)\right]}{[\theta+t]}, \qquad T_4\left(t\right) = \tfrac{\theta^2 t\left[t[t-1] + \theta^2 Ei\left(1, \frac{\theta}{t}\right) \exp\left(\frac{\theta}{t}\right)\right]}{2[\theta+t]}.$$

2.5 Mean residual life function

For a nonnegative continuous RV X the mean residual life function is defined as $\mu(t|\theta) = \mathrm{E}(X - t | X > t)$ and is given by

$$\mu(t|\theta) = \frac{1}{S(t|\theta)} \int_t^\infty S(x \mid \theta) \mathrm{d}x.$$

For the NUL distribution, we get

$$\mu(t|\theta) = \frac{t\left\{\left[\left(1+\theta\right)t - \theta\right]\delta\left(t, \theta\right) - \exp\left(\theta\right)t\right\}\delta\left(-t, \theta\right)}{t\left[\theta + t\right]\delta\left(\frac{t}{t-1}, \theta\right) - \left[1+\theta\right]},$$

where $\delta\left(t, \theta\right) = \exp\left(\theta/t\right)$.

2.6 Stress strength reliability

Let X and Y be two independent NUL RVs with parameters θ_1, θ_2 respectively and having PDF's f_X and f_Y. Then the stress-strength reliability measure (Kotz and Pensky, 2003) is given by

$$R = P\left(Y < X\right) = \int_0^1 f_X(x|\theta_1)\, F_Y(x \mid \theta_2)\, \mathrm{d}x$$

$$= \frac{{\theta_1}^2\left[{\theta_1}^2\theta_2 + {\theta_1}^2 + \theta_1 + 2\,\theta_1{\theta_2}^2 + 4\,\theta_1\theta_2 + 3\,\theta_2 + {\theta_2}^3 + 3\,{\theta_2}^2\right]}{\left[1+\theta_2\right]\left[1+\theta_1\right]\left[\theta_1+\theta_2\right]^3}.$$

2.7 Quantile function

Let X be a NUL RV with CDF as given in (2). The quantile function, $Q(p) = F^{-1}(p)$, can then be written as

$$Q(p|\theta) = -\frac{\theta}{1 + W[-\exp\left(-(1+\theta)\right)p(1+\theta)]}, \tag{3}$$

such that $0 < p < 1$ and W is the Lambert W function which is a multivalued complex function defined as the solution of the equation $W(z)\, \exp[W(z)] = z$. For more details on

the Lambert W function, readers may refer to Corless et al. (1996), Jodrá (2010), Veberič (2012) and references cited therein.

2.8 Mean deviation

As pointed out, for example in Ghitany et al. (2008), the amount of scatter in a population is measured to some extent by the totality of deviations from the mean and the median. These are known as the mean deviation about the mean and the mean deviation about the median and are defined as

$$\delta\left(X\right)=\int_{x}^{\infty}\left|X-m\right|f(x\mid\theta)\mathrm{d}x=2\left[mF\left(m\right)-\int_{0}^{m}xf(x\mid\theta)\mathrm{d}x\right], \tag{4}$$

with $m=\mathrm{E}\left(X\right)$ or $m=\mathrm{Median}(X)$ respectively. Considering (2) and (1) in (4) we get

$$\delta(X)=\frac{2m\exp(\theta(m-1)/m)}{1+\theta}.$$

For $m=\mathrm{E}\left(X\right)$ we get $\delta\left(X\right)=2\theta\exp\left(-1\right)/(1+\theta)^{2}$. Considering $m=Q(0.5|\theta)$ we have the expression for the mean deviation about the median, where the expression for $Q(\cdot|\theta)$ is given in (3).

2.9 Exponential family

A distribution belongs to the exponential family (Dobson, 2001) if it is of the form

$$f(x|\theta)=\exp\left(Q(\theta)\,T(x\mid\theta)+D(\theta)+S(x\mid\theta)\right).$$

It can be easily seen that the proposed distribution belongs to the exponential family by rewriting the PDF given in (1) as

$$f(x|\theta)=\exp\left(-\frac{\theta(1-x)}{x}\right)\exp\left(\log\left(\frac{\theta^{2}}{1+\theta}\right)\right)\exp\left(\log(x^{-3})\right),$$

where $Q(\theta)=\theta,\quad T(x\mid\theta)=[1-x]/x,\quad D(\theta)=\log\left(\theta^{2}/[1+\theta]\right),\quad S(x\mid\theta)=\log(x^{-3})$. Therefore, $T(\mathbf{x})=\sum_{i=1}^{n}\left[1-x_{i}\right]/x_{i}$ is a complete sufficient estimator for θ based on a sample of size n from the proposed distribution. Besides that, since the distribution belongs to an exponential family, a minimum-variance unbiased estimator can be obtained by bias corrected ML estimator.

3. Estimation

In this section, we will derive the method of moments (MME) and ML estimators of parameter θ of a NUL distribution. For the ML estimator of θ we derive the closed-form expressions for the second order bias-correction. In addition, in this section, we consider regression modeling.

3.1 Maximum likelihood estimation

Let $X_1, \ldots, X_n$ be a random sample from the NUL distribution with PDF. (1). Then, for observed $\boldsymbol{x} = (x_1, \ldots, x_n)$, the log-likelihood function of θ can be written as

$$\ell(\theta|\boldsymbol{x}) \propto 2n \log(\theta) - n \log(1+\theta) - \theta t(\boldsymbol{x}).$$

The ML estimate $\widehat{\theta}$ of θ is obtained by solving the following linear equation

$$\frac{d}{d\theta}\, \ell(\theta|\boldsymbol{x}) = \frac{2\,n}{\theta} - \frac{n}{1+\theta} - t(\boldsymbol{x}) = 0$$

which gives

$$\widehat{\theta} = \frac{1}{2\,t(\boldsymbol{x})} \left[n - t(\boldsymbol{x}) + \sqrt{t(\boldsymbol{x})^2 + 6\,n\,t(\boldsymbol{x}) + n^2} \right].$$

Next

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\, \ell(\theta|\boldsymbol{x}) = \frac{n}{(1+\theta)^2} - \frac{2\,n}{\theta^2} < 0$$

for all θ, in particular for $\theta = \widehat{\theta}$.

Since $\mathrm{d}^2\ell(\theta|\boldsymbol{x})/\mathrm{d}\theta^2$ is data-independent, we have that $n\,\mathrm{E}[\mathrm{d}^2 \log f(X|\theta)/\mathrm{d}\theta^2] = \mathrm{d}^2\ell(\theta|\boldsymbol{x})/\mathrm{d}\theta^2$. Thus, the expected Fisher information is $\mathrm{I}(\widehat{\theta}) = 2\,n/\theta^2 - n/[1+\theta]^2$. From the large sample theory (Lehmann and Casella (1998, pp. 461-463)), the asymptotic distribution of ML estimator $\widehat{\theta}$ of θ is such that

$$\sqrt{n}\,(\widehat{\theta} - \theta) \xrightarrow{D} \mathrm{N}\left(0, \mathrm{V}(\widehat{\theta})\right),$$

where $\xrightarrow{D}$ denotes convergence in distribution and $\mathrm{V}(\widehat{\theta})$ is just the inverse of the expected Fisher information written as $\mathrm{V}(\widehat{\theta}) = \theta^2\,[1+\theta]^2/n\,[\theta^2 + 4\,\theta + 2]$. It is easy to see that for $\psi = g(\theta) = \mathrm{E}\,(X)$ $\widehat{\psi} = \widehat{\mathrm{E}}\,(X) = 1/[1+\widehat{\theta}]$ and $\mathrm{V}(\widehat{\psi}) = \theta^2/n\,[\theta^2 + 4\,\theta + 2]$. Hence, the asymptotic $100\,(1-\alpha)\%$ confidence intervals (CIs) for θ and ψ are given, respectively, by

$$\widehat{\theta} \pm z_{\alpha/2} \sqrt{\frac{\widehat{\theta}^2\,[1+\widehat{\theta}]^2}{n\,[\widehat{\theta}^2 + 4\,\widehat{\theta} + 2]}} \quad \text{and} \quad \frac{1}{1+\widehat{\theta}} \pm z_{\alpha/2} \sqrt{\frac{\widehat{\theta}^2}{n\,[\widehat{\theta}^2 + 4\,\widehat{\theta} + 2]}},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

It is important to note that for a Bayesian setup, we can use the Jeffreys invariant prior for θ, given by $\pi(\theta) \propto \sqrt{\mathrm{I}(\theta)}$. However we will not consider it further in this paper.

Cox and Snell (1968) provided a framework to estimate the bias, to $\mathcal{O}(n^{-1})$ for the ML estimates of parameters of regular densities. Hence, subtracting the estimated bias from the original ML estimator produces a bias-corrected estimator (BCE) that is unbiased to $\mathcal{O}(n^{-2})$. Following Cox and Snell (1968) the analytical expression for bias-correction of an scalar $\widehat{\theta}$, given by

$$\mathcal{B}\left(\widehat{\theta}\right) = \left(\kappa^{11}\right)^2\,[0.5\,\kappa_{111} + \kappa_{11,1}] + \mathcal{O}(n^{-2}),$$

where

$$\kappa^{11} = \mathrm{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\,\ell(\theta|\boldsymbol{x})\right]^{-1} = \frac{\theta^2\,(1+\theta)^2}{n\,(\theta^2+4\,\theta+2)},$$

$$\kappa_{11,1} = \mathrm{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\,\ell(\theta|\boldsymbol{x}) \times \frac{d}{d\theta}\,\ell(\theta|\boldsymbol{x})\right] = 0,$$

and

$$\kappa_{111} = \mathrm{E}\left[-\frac{d^3}{d\theta^3}\,\ell(\theta|\boldsymbol{x})\right] = \frac{2\,n\left(\theta^3+6\,\theta^2+6\,\theta+2\right)}{\theta^3\,(1+\theta)^3}.$$

Thus, the bias-corrected ML estimator $\widetilde{\theta}$ is

$$\widetilde{\theta} = \widehat{\theta} - \frac{\widehat{\theta}\left[1+\widehat{\theta}\right]\left[\widehat{\theta}^3+6\widehat{\theta}^2+6\widehat{\theta}+2\right]}{n\left[\widehat{\theta}^2+4\widehat{\theta}+2\right]^2},$$

where the right hand side is $\widehat{\mathcal{B}}\left(\widehat{\theta}\right)$.

Re-parameterizing (1) in terms of the mean $\mu = \theta/[1+\theta]$, the ML of μ is obtained as

$$\widehat{\mu} = \frac{1}{2n}\left[3\,n + t(\boldsymbol{x}) - \sqrt{t(\boldsymbol{x})^2 + 6\,n\,t(\boldsymbol{x}) + n^2}\right],$$

and the corresponding bias-corrected ML estimator $\widetilde{\mu}$ of μ as

$$\widetilde{\mu} = \widehat{\mu} - \frac{2\widehat{\mu}\,[\widehat{\mu}-1]^2}{n\,[\widehat{\mu}^2-2]^2}.$$

3.2 Method of moment estimation

Let $X_1, \ldots, X_n$ be a random sample from the unit-Lindley distribution with PDF (1). Then, the MME $\widehat{\theta}_{\mathrm{MME}}$ of θ is given by

$$\widehat{\theta}_{\mathrm{MME}} = \frac{\bar{X}}{1-\bar{X}} = \left[\frac{1}{\bar{X}} - 1\right]^{-1},$$

which is positively biased, that is, $\mathrm{E}(\widehat{\theta}) - \theta > 0$.

Proof Let $\widehat{\theta}_{\mathrm{MME}} = g(\overline{X})$ and $g(t) = t/[1-t]$ for $t > 0$. Since $g''(t) = -2/[t-1]^3 > 0$ for all $t < 1$, $g(t)$ is strictly convex. Thus, by Jensen's inequality, we have $\mathrm{E}(g(\overline{X})) > g(\mathrm{E}(\overline{X}))$. Since $g(\mathrm{E}(\overline{X})) = g\left(\theta/[1+\theta]\right) = \theta$ we get $\mathrm{E}(\widehat{\theta}) > \theta$.

3.3 Regression analysis

We will now present a real data analysis in order to showcase the applicability of the proposed distribution. Since the NUL distribution has a closed form expression for the

mean we are able to introduce a new regression model for bounded response variable. The re-parametrized PDF of the NUL distribution is given by

$$f(y|\mu) = \frac{\mu^2}{[1-\mu]\, y^3} \exp\left(-\frac{\mu\,[1-y]}{y\,[1-\mu]}\right), \tag{5}$$

where $0 < y \leq 1$ and $0 < \mu \leq 1$. Under this parametrization the mean and variance of NUL distribution are given by

$$\mathrm{E}(Y) = \mu \quad \text{and} \quad \mathrm{Var}(Y) = \frac{\mu^2}{1-\mu}\left[\mathrm{Ei}\left(1, \frac{\mu}{1-\mu}\right) \exp\left(\frac{\mu}{1-\mu}\right) + \mu - 1\right].$$

Let $Y_1, \ldots, Y_n$ be n independent RVs, where $Y_i \sim \mathrm{NUL}(\mu_i)$, $i = 1, \ldots, n$ with PDF. given by (5). The NUL regression model is defined assuming that the mean of Y_i can be written as

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{(p-1)})^\top$ is a p-dimensional vector of unknown regression coefficients $(p < n)$ and $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{i(p-1)})^\top$ denotes the observations on p known covariates. Note that the variance of Y_i is a function of μ_i and, as a consequence of the covariate values, which implies that non-constant response variances are naturally accommodated into the model.

We shall assume that the mean link function g is a strictly monotonic and twice differentiable function that maps $(0, 1)$ into R. Some of the most common link functions are:

(i) logit: $g(\mu_i) = \log\left(\mu_i/(1-\mu_i)\right)$;
(ii) probit: $g(\mu_i) = \Phi^{-1}(\mu_i)$, where Φ^{-1} is the standard normal quantile function;
(iii) complementary log-log: $g(\mu_i) = \log\left[-\log(1-\mu_i)\right]$.

Inferences about the regression coefficients $\boldsymbol{\beta}$ can be performed under the likelihood paradigm (Lehmann and Casella, 1998) . The log-likelihood function based on a sample of n independent observations is

$$\ell(\boldsymbol{\beta}) \propto 2\sum_{i=1}^{n} \log(\mu_i) - \sum_{i=1}^{n} \log(1-\mu_i) - \sum_{i=1}^{n} \frac{\mu_i\,[1-y_i]}{y_i\,[1-\mu_i]}, \tag{6}$$

where $\mu_i = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})$.

The ML estimates $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ are obtained by maximizing the log-likelihood function defined in (6) using standard optimization methods, such as Newton-Raphson or quasi-Newton. In this paper, the ML estimate were obtained by the quasi-Newton method available in the SAS/NLMIXED procedure (https://www.sas.com/).

For comparison purpose, we also considered the beta and unit-Lindley regression models. The PDF of the alternative regression models are:

- Beta regression (Cepeda-Cuervo, 2001; Ferrari and Cribari-Neto, 2004):

$$f(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\,\phi)\Gamma([1-\mu]\,\phi)} y^{\mu\phi-1}[1-y]^{[1-\mu]\phi-1}, \quad 0 < y < 1$$

where $0 < \mu < 1$ denotes the mean and $\phi > 0$ is a precision parameter.

- Unit-Lindley regression (Mazucheli et al., 2019):

$$f(y|\mu) = \frac{[1-\mu]^2}{\mu\,[1-y]^3}\,\exp\left(-\frac{y\,[1-\mu]}{\mu\,[1-y]}\right), \quad 0 < y < 1$$

where $0 < \mu < 1$ denotes the mean.

To discriminate and choose the best among the proposed models, the Akaike (AIC) (Akaike, 1974), Schwarz (BIC) (Schwarz, 1978) and corrected Akaike (AICC) (Cavanaugh, 1997) information criteria were used. These measures are defined as follows

$$\text{AIC} = 2\,p - 2\,\log\widehat{L}, \quad \text{BIC} = \log(n)\,p - 2\,\log\widehat{L}, \quad \text{AICC} = \frac{2n\,[p+1]}{n-p-2} - 2\,\log\widehat{L}$$

where $\widehat{L}$ is the likelihood evaluated at the ML estimates, p is the number of parameters in the model and n the number of observations. The decision rule, in all these criteria, is favorable to the model with the lowest value (Held and Sabanés Bové, 2014). To quantify the uncertainty associated with these criteria, the non-parametric Bootstrap approach was used to decide on the final model. We considered 10,000 independent runs and calculated the percentage of times each model was selected.

To assess the adequacy of the regression models we used the Cox-Snell residuals and examined the half-normal plot with simulated envelope (Atkinson, 1981). The Cox-Snell residuals are defined as

$$r_i = -\log\left(1-\widehat{F}(y_i)\right), \quad i = 1,\ldots,n,$$

where $\widehat{F}$ is the estimated CDF. A notable property of the Cox-Snell residuals is that if the regression model fits the data well, r_i's follow a standard exponential distribution.

4. Numerical results

In this section, we conduct a Monte Carlo simulation in order to evaluate and compare the finite-sample behavior of the ML estimators, its bias-corrected counterpart obtained by the Cox-Snell methodology (BCE) and the MME of the parameter θ of the NUL distribution. In addition, in this section, an empirical illustration is conducted.

4.1 Simulation study

We have generated samples ranging from 10 to 90 with a gap of 10 and $\theta = 0.1, 0.5, 1.0, 1.5, 2.0, 3.0$ and 4.0. To simulate observations from the proposed distribution we generated Y from Lindley distribution (see, `rlindley` function in LindleyR library) and then used the transformation $X = 1/[1+Y]$. The simulation experiment was repeated $M = 20,000$ times. The performance evaluation was done based on the estimated bias and estimated root mean squared error (RMSE).

Figure 2 shows that ML estimates and MME of θ are positively biased, while the BCE estimator achieve substantial bias reduction, especially for small and moderate sample sizes. It is also observed that the RMSE decreases as n increases, as expected. Additionally, the RMSE of the corrected estimates are smaller than those of the uncorrected estimates.

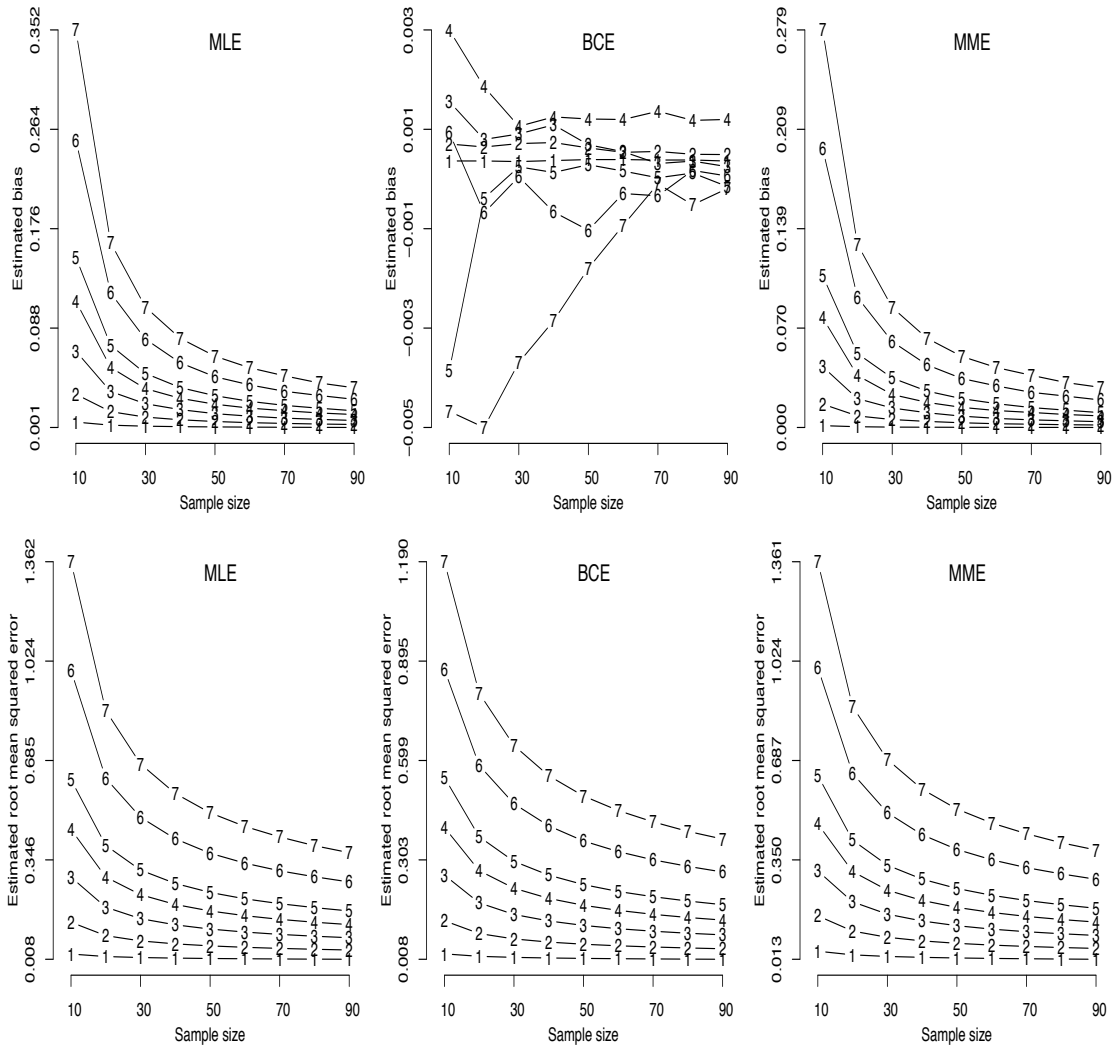


Figure 2. Upper Panel: Estimated bias. Lower Panel: Estimated root mean squared error. (1: $\theta = 0.1$, 2: $\theta = 0.5$, 3: $\theta = 1.0$, 4: $\theta = 1.5$, 5: $\theta = 2.0$, 6: $\theta = 3.0$ and 7: $\theta = 4.0$).

4.2 Empirical illustration

The real data set considered is presented by Schmit and Roth (1990), and corresponds to the 73 responses to a questionnaire sent to 374 risk managers of large North American organizations. The objective of Schmit and Roth (1990) was to evaluate the cost effectiveness with the management philosophy of controlling the company's exposure to various property losses and accidents, taking into account company characteristics such as size and type of industry.

The response variable y (Firm cost) is the firm-specific ratio of premiums plus uninsured losses divided by total assets. The covariates associated with this response variable are:

- X_1 (Assume): firm-specific ratio of the summation of per occurrence retention levels, as measured by the corporate risk manager.
- X_2 (Cap): 1 if the firm uses a captive and 0 otherwise.
- X_3 (Sizelog): log of the firm's total asset value.
- X_4 (Indcost): industry average of premiums plus uninsured losses divided by total assets, as measured by the 1985 Cost of Risk Survey (a measure of risk).
- X_5 (Central): importance of local manager in choosing local retention levels, as measured by the corporate risk manager.

- X_6 (Soph): importance of analytical tools in making risk management decisions, as measured by the corporate risk manager.

We assume that the regression structure for the mean is given by

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}, \quad i = 1, \ldots, 73,$$

where x_{ij} are the values of the covariate X_j.

The point estimates and the 95% confidence intervals for the parameters of the three regression models are given in Table 1. It is observed that the NUL and beta regression models have the same significant covariates to explain the response variable, which are the Sizelog and Indcost variables.

Table 1. The ML estimates and the 95% confidence intervals.

	NUL		UL		Beta	
Parameter	MLE	95% CI	MLE	95% CI	MLE	95% CI
β_0	4.3789	(2.6395, 6.1183)	3.0506	(0.8132, 5.2879)	1.8880	(-0.4096, 4.1855)
β_1	-0.0050	(-0.0273, 0.0173)	-0.0592	(-0.0830, -0.0354)	-0.0121	(-0.0394, 0.0151)
β_2	-0.0112	(-0.3780, 0.3556)	1.8972	(1.2649, 2.5295)	0.1780	(-0.2763, 0.6322)
β_3	-0.8943	(-1.0709, -0.7176)	-0.6606	(-0.8889, -0.4322)	-0.5115	(-0.7524, -0.2705)
β_4	1.7145	(1.0244, 2.4046)	4.5081	(2.8651, 6.1511)	1.2362	(0.3359, 2.1366)
β_5	-0.0538	(-0.1878, 0.0801)	0.0885	(-0.1143, 0.2912)	-0.0122	(-0.1836, 0.1593)
β_6	0.0012	(-0.0317, 0.0340)	-0.0846	(-0.1415, -0.0277)	-0.0037	(-0.0455, 0.0380)
ϕ	-	-	-	-	6.3305	(4.1300, 8.5311)

Table 2 gives the values of the likelihood-based statistics and one can see that the NUL regression model provides the best fit, since it has the lowest values of AIC, AICC and BIC. It is also observed that the NUL was selected approximately 68% of the times as opposed to the UL and beta models.

Table 2. The likelihood-based statistics of fit.

Criteria	NUL	UL	Beta
AIC (%)†	-224.9780 (68.26%)	-77.3946 (16.17%)	-159.4460 (15.57%)
AICC (%)	-223.2549 (68.44%)	-75.6715 (16.23%)	-157.1960 (15.33%)
BIC (%)	-208.9447 (69.16%)	-61.3614 (16.42%)	-141.1223 (14.42%)

†: % of times out of 10,000 non-parametric Bootstrap runs that the model is selected.

In Figure 3 we present the half-normal plots for the Cox-Snell residuals with simulated envelopes. It is observed for the NUL regression model that all points lie inside the envelopes, suggesting that there is no serious violation of the model assumptions. We can conclude that NUL regression model provides a good fit to these data and therefore can be used for inference purposes.

From the inference results of NUL model (see Table 1) it is observed that the mean of Firm cost is negatively related to the log of the firm's total asset value (Sizelog). In contrast, the measure of risk (Indcost) has a positive impact on the mean response.

5. Concluding remarks

The ideas in this paper stem from a recent work which proposed a unit-Lindley distribution by transforming a Lindley random variable appropriately. We applied a slightly

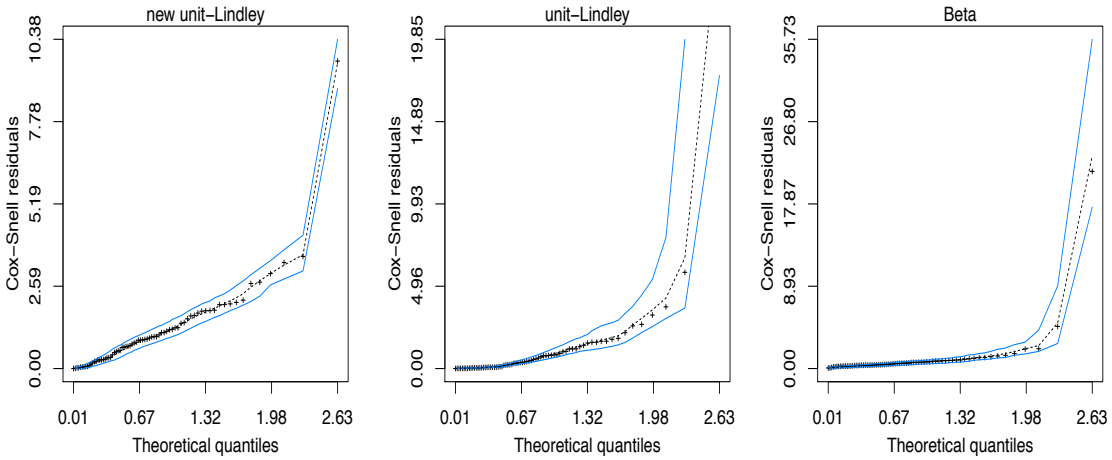


Figure 3. The half-normal plot with simulated envelope for the Cox-Snell residuals.

different transformation, yet again to a Lindley random variable and introduced a new one-parameter unit-Lindley distribution which is capable of describing data which is limited to the interval (0,1]. Several mathematical properties of the new distribution are presented in detail and parameter estimation is discussed considering the methods of maximum likelihood and moments. We also derived an analytical expression for the bias corrected maximum likelihood estimator. Using a simple re-parametrization of the new distribution we introduced a newer regression model to describe data in a bounded interval. An application of the proposed model to a real dataset from finance shows a better and more parsimonious fit than the classical beta regression model. As such we envisage that the new model attracts the attention of practitioners across all relevant fields of science.

A few related ideas for future work could be to provide a Fisher scoring algorithm for parameter estimation, and to check if this algorithm is equivalent to an iteratively re weighted least squares, as the model belongs to the exponential family.

Acknowledgments

The authors would like to sincerely thank the editors-in-chief of ChJS and the anonymous referees for their valuable and constructive comments which greatly improved an earlier manuscript.

References

Abramowitz, M. and Stegun, I.A., 1974. Handbook of Mathematical Functions with Formulas, graphs, and Mathematical Tables. National Bureau of Standards Applied Mathematics Series. Dover Pub- lications, Incorporated, New York.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716–723.

Atkinson, A.C., 1981. Two graphical displays for outlying and influential observations in regression. Biometrika, 68, 13–20.

Cavanaugh, J.E., 1997. Unifying the derivations for the Akaike and corrected Akaike information criteria. Statistics and Probability Letters, 33, 201–208.

Cepeda-Cuervo, E., 2001. Variability modeling in generalized linear models. Ph.D. thesis, Mathematics Institute, Universidade Federal do Rio de Janeiro.

Corless, R.M., Gonnet, G.H., Hare, D.E.G., Jeffrey, D.J., and Knuth, D.E., 1996. On the Lambert W function. Advances in Computational Mathematics, 5, 329–359.

Cox, D.R. and Snell, E.J., 1968. A general definition of residuals. Journal of the Royal Statistical Society, Series B, 30, 248–275.

Dobson, A.J., 2001. An Introduction to Generalized Linear Models, Second Edition. Chapman and Hall/CRC.

Ferrari, S. and Cribari-Neto, F., 2004. Beta regression for modeling rates and proportions. Journal of Applied Statistics, 31, 799–815.

Ghitany, M.E., Atieh, B., and Nadarajah, S., 2008. Lindley distribution and its application. Mathematics and Computers in Simulation, 78, 493–506.

Ghitany, M.E., Mazucheli, J., Menezes, A.F.B., and Alqallaf, F., 2018. The unit-inverse Gaussian distribution: A new alternative to two-parameter distributions on the unit interval. Communications in Statistics: Theory and Methods, 48, 1–19.

Gómez-Déniz, E., Sordo, M.A., and Calderin-Ojeda, E., 2014. The Log-Lindley distribution as an alternative to the beta regression model with applications in insurance. Insurance: Mathematics and Economics, 54:49–57.

Grassia, A., 1977. On a family of distributions with argument between 0 and 1 obtained by trans- formation of the Gamma distribution and derived compound distributions. Australian Journal of Statistics, 19, 108–114.

Held, L. and Sabanés Bové, D., 2014. Applied Statistical Inference-Likelihood and Bayes. Springer, New York.

Jodrá, P., 2010. Computer generation of random variables with Lindley or Poisson-Lindley distribution via the Lambert W function. Mathematics and Computers in Simulation, 81, 851–859.

Johnson, N.L., 1949. Systems of frequency curves generated by methods of translation. Biometrika 36, 149–176.

Johnson, N.L., 1955. Systems of frequency curves derived from the first law of Laplace. TEST, 5, 283–291.

Kotz, S. and Pensky, M., 2003. The Stress-Strength Model and its Generalizations: Theory and Applications. World Scientific, Singapore.

Kumaraswamy, P., 1980. A generalized probability density function for double-bounded random processes. Journal of Hydrology, 46, 79–88.

Lehmann, E.J. and Casella, G., 1998. Theory of Point Estimation. Springer, Berlin.

Lindley, D.V., 1958. Fiducial distributions and Bayes′theorem. Journal of the Royal Statistical Society B, 20, 102–107.

Mazucheli, J., Menezes, A.F.B., and Chakraborty, S., 2019. On the one parameter unit-Lindley distribution and its associated regression model for proportion data. Journal of Applied Statistics, 46, 700–714.

Mazucheli, J., Menezes, A.F.B., and Dey, S., 2018a. The unit-Birnbaum-Saunders distribution with applications. Chilean Journal of Statistics, 1, 47–57.

Mazucheli, J., Menezes, A.F.B., and Ghitany, M.E., 2018b. The unit-Weibull distribution and associated inference. Journal of Applied Probability and Statistics, 13, 1–22.

Nadarajah, S., Bakouch, H.S., Tahmasbi, R., 2011. A generalized Lindley distribution. Sankhya B, 73, 331–359.

Schmit, J.T. and Roth, K., 1990. Cost effectiveness of risk management practices. Journal of Risk and Insurance, 57, 455–470.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics, 6, 461–464.

Shanker, R. and Mishra, A., 2013. A quasi Lindley distribution. African Journal of Mathematics and Computer Science Research, 6, 64–71.

Tadikamalla, P.R., 1981. On a family of distributions obtained by the transformation of the Gamma distribution. Journal of Statistical Computation and Simulation, 13, 209–214.

Tadikamalla, P.R. and Johnson, N.L., 1982. Systems of frequency curves generated by transformations of Logistic variables. Biometrika, 69, 461–465.

Topp, C.W. and Leone, F.C., 1955. A family of J-Shaped frequency functions. Journal of the American Statistical Association, 50, 209–219.

Veberič, D., 2012. Lambert W function for applications in physics. Computer Physics Communications, 183, 2622–2628.

Information for authors

The editorial board of the Chilean Journal of Statistics (ChJS) is seeking papers, which will be refereed. We encourage the authors to submit a PDF file of the manuscript in a free format to Editors of the ChJS (E-mail: `chilean.journal.of.statistics@gmail.com`). Submitted manuscripts must be written in English and contain the name and affiliation of each author followed by a leading abstract and keywords. The authors must include a"cover letter" presenting their manuscript and mentioning: "We confirm that this manuscript has been read and approved by all named authors. In addition, we declare that the manuscript is original and it is not being published or submitted for publication elsewhere".

Preparation of accepted manuscripts

Manuscripts accepted in the ChJS must be prepared in Latex using the ChJS format. The Latex template and ChJS class files for preparation of accepted manuscripts are available at `http://chjs.mat.utfsm.cl/files/ChJS.zip`. Such as its submitted version, manuscripts accepted in the ChJS must be written in English and contain the name and affiliation of each author, followed by a leading abstract and keywords, but now mathematics subject classification (primary and secondary) are required. AMS classification is available at `http://www.ams.org/mathscinet/msc/`. Sections must be numbered 1, 2, etc., where Section 1 is the introduction part. References must be collected at the end of the manuscript in alphabetical order as in the following examples:

Arellano-Valle, R., 1994. Elliptical Distributions: Properties, Inference and Applications in Regression Models. Unpublished Ph.D. Thesis. Department of Statistics, University of São Paulo, Brazil.

Cook, R.D., 1997. Local influence. In Kotz, S., Read, C.B., and Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences, Vol. 1., Wiley, New York, pp. 380-385.

Rukhin, A.L., 2009. Identities for negative moments of quadratic forms in normal variables. Statistics and Probability Letters, 79, 1004-1007.

Stein, M.L., 1999. Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Tsay, R.S., Peña, D., and Pankratz, A.E., 2000. Outliers in multivariate time series. Biometrika, 87, 789-804.

References in the text must be given by the author's name and year of publication, e.g., Gelfand and Smith (1990). In the case of more than two authors, the citation must be written as Tsay et al. (2000).

Copyright