



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

REVSTAT

Statistical Journal



Catálogo Recomendada

REVSTAT. Lisboa, 2003-
Revstat : statistical journal / ed. Instituto Nacional
de Estatística. - Vol. 1, 2003- . - Lisboa I.N.E.,
2003- . - 30 cm
Trimestral. - Continuação de : Revista de Estatística =
ISSN 0873-4275. - edição exclusivamente em inglês
ISSN 1645-6726 ; e-ISSN 2183-0371

CREDITS

- | | |
|--|---|
| <ul style="list-style-type: none">- EDITOR-IN-CHIEF<ul style="list-style-type: none">- <i>Isabel Fraga Alves</i>- CO-EDITOR<ul style="list-style-type: none">- <i>Giovani L. Silva</i>- ASSOCIATE EDITORS<ul style="list-style-type: none">- <i>Marília Antunes</i>- <i>Barry Arnold</i>- <i>Narayanaswamy Balakrishnan</i>- <i>Jan Beirlant</i>- <i>Graciela Boente</i>- <i>Paula Brito</i>- <i>Valérie Chavez-Demoulin</i>- <i>David Conesa</i>- <i>Charmaine Dean</i>- <i>Fernanda Figueiredo</i>- <i>Jorge Milhazes Freitas</i>- <i>Alan Gelfand</i>- <i>Stéphane Girard</i>- <i>Marie Kratz</i>- <i>Victor Leiva</i>- <i>Artur Lemonte</i>- <i>Shuangzhe Liu</i>- <i>Maria Nazaré Mendes-Lopes</i>- <i>Fernando Moura</i>- <i>John Nolan</i>- <i>Paulo Eduardo Oliveira</i>- <i>Pedro Oliveira</i>- <i>Carlos Daniel Paulino (2019-2021)</i>- <i>Arthur Pewsey</i>- <i>Gilbert Saporta</i>- <i>Alexandra M. Schmidt</i>- <i>Julio Singer</i> | <ul style="list-style-type: none">- <i>Manuel Scotto</i>- <i>Lisete Sousa</i>- <i>Milan Stehlik</i>- <i>María Dolores Ugarte</i>- FORMER EDITOR-IN-CHIEF<ul style="list-style-type: none">- <i>M. Ivette Gomes</i>- FORMER CO-EDITOR<ul style="list-style-type: none">- <i>M. Antónia Amaral Turkman</i>- EXECUTIVE EDITOR<ul style="list-style-type: none">- <i>José A. Pinto Martins</i>- FORMER EXECUTIVE EDITOR<ul style="list-style-type: none">- <i>Maria José Carrilho</i>- <i>Ferreira da Cunha</i>- SECRETARIAT<ul style="list-style-type: none">- <i>José Cordeiro</i>- <i>Olga Bessa Mendes</i>- PUBLISHER<ul style="list-style-type: none">- <i>Instituto Nacional de Estatística, I.P. (INE, I.P.)</i>- <i>Web site: http://www.ine.pt</i>- COVER DESIGN<ul style="list-style-type: none">- <i>Mário Bouçadas, designed on the stain glass window at INE by the painter Abel Manta</i>- LAYOUT AND GRAPHIC DESIGN<ul style="list-style-type: none">- <i>Carlos Perpétuo</i>- PRINTING<ul style="list-style-type: none">- <i>Instituto Nacional de Estatística, I.P.</i>- EDITION<ul style="list-style-type: none">- <i>140 copies</i>- LEGAL DEPOSIT REGISTRATION<ul style="list-style-type: none">- <i>N.º 191915/03</i>- PRICE [VAT included]<ul style="list-style-type: none">- <i>€ 9,00</i> |
|--|---|




INDEX

Asymmetric Kernels for Boundary Modification in Distribution Function Estimation	
<i>Habib Allah Mombeni, Behzad Mansouri and Mohammad Reza Akhoond</i>	463
Production Processes with Different Levels of Risk: Addressing the Replacement Option	
<i>Francisco Stefano de Almeida, Cláudia Nunes and Carlos Oliveira</i>	485
A Truncated General-G Class of Distributions with Application to Truncated Burr-G Family	
<i>Farrukh Jamal, Hassan S. Bakouch and M. Arslan Nasir</i>	513
An Efficient Mixed Randomized Response Model for Sensitive Characteristic in Sample Survey	
<i>Amod Kumar, Gajendra K. Vishwakarma and G.N. Singh</i>	531
A Regression Model for Positive Data Based on the Slashed Half-Normal Distribution	
<i>Yolanda M. Gómez, Diego I. Gallardo and Mário de Castro</i>	553
Comparison of the Likelihood Ratios of Two Diagnostic Tests Subject to a Paired Design: Confidence Intervals and Sample Size	
<i>José Antonio Roldán-Nofuentes and Saad Bouh Sidaty-Regad</i>	575

ASYMMETRIC KERNELS FOR BOUNDARY MODIFICATION IN DISTRIBUTION FUNCTION ESTIMATION

Authors: HABIB ALLAH MOMBENI
– Statistics Department, Shahid Chamran University of Ahvaz,
Ahvaz, Iran
habiballamombeni@gmail.com

BEHZAD MANSOURI 
– Statistics Department, Shahid Chamran University of Ahvaz,
Ahvaz, Iran
b.mansouri@scu.ac.ir

MOHAMMAD REZA AKHOOND 
– Statistics Department, Shahid Chamran University of Ahvaz,
Ahvaz, Iran
mra.biostat@gmail.com

Received: December 2017

Revised: April 2019

Accepted: July 2019

Abstract:

- Kernel-type estimators are popular in density and distribution function estimation. However, they suffer from boundary effects. In order to modify this drawback, this study has proposed two new kernel estimators for the cumulative distribution function based on two asymmetric kernels including the Birnbaum–Saunders kernel and the Weibull kernel. We show the asymptotic convergence of our proposed estimators in boundary as well as interior design points. We illustrate the performance of our proposed estimators using a numerical study and show that our proposed estimators outperform the other commonly used methods. The illustration of our proposed estimators to a real data set indicates that they provide better estimates than those of the formerly-known methodologies.

Keywords:

- *cumulative distribution function; boundary effects; kernel-type estimators; asymmetric kernels.*

AMS Subject Classification:

- 62G05, 62G20.

1. INTRODUCTION

Suppose that X_1, X_2, \dots, X_n be a set of continuous random variables with unknown cumulative distribution function $F(x)$ which we wish to estimate. The Empirical distribution function provides a uniformly consistent estimate of the cumulative distribution function. However, estimations which are provided by the Empirical distribution are not smooth. Another approach for estimating the cumulative distribution function is to use Kernel-type estimators. Kernel-type estimators for distribution estimation, based on symmetric kernels, have been introduced by authors such as Nadaraya [14] and Watson and Leadbetter [21], and their asymptotic properties have been investigated by Singh *et al.* [18]. Asymptotical superiority of Kernel-type estimators to the empirical distribution function at a single point in density estimation was shown by Reiss [15] and Falk [5].

Although the symmetric kernels are popular and commonly used in Kernel-type estimators, they are not efficient for those distribution (density) functions which have a compact support due to the boundary bias. This problem is known as *boundary effects* and several approaches have so far been proposed to deal with it in regression and density estimation tasks (Gasser and Muller [6], Rice [16], Gasser *et al.* [7] and Muller [12]). In a similar manner, Tenreiro [19] proposed some boundary kernels for estimating a cumulative distribution function with a finite interval support. These approaches, hereafter called the Boundary kernel methods or briefly the B-K methods, are based on symmetric kernels.

Asymmetric kernel functions were introduced by Chen [2] as an alternative approach to the boundary correction in kernel density estimation. He proposed the beta kernel density estimator to estimate a density with support on $[0, 1]$. Chen [3] considered the gamma kernel density estimator to estimate a density with support on $[0, \infty)$. In order to provide a boundary-free estimation for the density function $f(x)$ with support on $[0, \infty)$ by the gamma kernel density estimator, Zhang [22] has shown that having a shoulder at $x = 0$, whose derivative of $f(x)$ is zero at $x = 0$, is a necessary condition. For densities not satisfying this condition, the gamma kernel density estimator suffers from severe boundary problems. This approach was extended for estimating a density with support on $[0, \infty)$ using other asymmetric kernels (Jin and Kawczak [10], Scaillet [17], Hirukawa and Sakudo [8] and Hirukawa and Sakudo [9]).

So far, the boundary effects in density estimation have attracted the attention of many researchers. Accordingly, several methods, using symmetric and asymmetric kernels, have been proposed to solve the problem. However, in the cumulative distribution estimation, the boundary effects have received little if any attention.

In this paper, we have focused on estimating those distribution functions with support on $[0, \infty)$ and proposed a new Kernel-type estimator for the cumulative distribution function based on asymmetric kernels. Our estimator at the design point x has the following form:

$$(1.1) \quad \hat{F}_n(x) = n^{-1} \sum_{i=1}^n \bar{K}_{x,b}(X_i),$$

where $\bar{K}_{x,b}(t) = \int_t^\infty k_{x,b}(u) du$ and $k(\cdot)$ is an asymmetric kernel function on $[0, \infty)$ with the smoothing parameter b . Thus, the kernel has the same support as the true distribution function.

We introduce two estimators by considering two asymmetric kernels including the Birnbaum–Saunders (B-S) kernel and the Weibull kernel. In the next section, we demonstrate the asymptotic properties of our proposed estimator based on the B-S kernel, hereafter called the B-S kernel estimator. We investigate the rate of convergence of the B-S kernel estimator both in the interior and the boundary points. In Section 3, we have run the same study for our second estimator which is based on the Weibull kernel, hereafter called the Weibull kernel estimator. The rest of the paper is organised as follows. Section 4 is dedicated to illustrating the performance of our proposed estimators. We conducted a comprehensive numerical study and considered various cumulative distribution functions to estimate and compare the performance of our estimators with other existing methods. In Section 5 we have illustrated the performance of our proposed estimators on a real data set. Finally, Section 6 is devoted to discussions and conclusions.

In this paper, we assume that the cumulative distribution function $F(x)$ satisfies the following assumptions:

Assumption 1. The cumulative distribution function $F(x)$ is absolutely continuous with respect to Lebesgue measure on $(0, \infty)$ and has two continuous and bounded derivatives.

Assumption 2. The smoothing parameter $b = b_n > 0$ satisfies $b \rightarrow 0$, as $n \rightarrow \infty$.

Assumption 3. The following integrals

$$(1.2) \quad \int_0^\infty (x f(x))^2 dx \quad \text{and} \quad \int_0^\infty (x^2 f'(x))^2 dx$$

are finite.

Following Hirukawa and Sakudo [9] and ‘In order to describe different asymptotic properties of an asymmetric kernel estimator across positions of the design point $x > 0$ ’, we denote by ‘interior x ’ and a sequence of points converging to the boundary or ‘boundary x ’ a design point x that satisfies $x/b \rightarrow \infty$ and $x/b \rightarrow k$ for some $0 < k < \infty$ as $n \rightarrow \infty$, respectively.

2. ASYMMETRIC CUMULATIVE DISTRIBUTION FUNCTION ESTIMATION USING B-S KERNEL

In this section, we aim at demonstrating the asymptotic convergence of our first proposed estimator: Equation (1.1) based on the B-S kernel, i.e. the B-S kernel estimator. To forward this end we will show that the B-S kernel estimator is asymptotically unbiased and consistent. We will obtain an appropriate smoothing parameter for our estimator through minimizing the mean integrated square error. In addition, we will discuss the convergence rate of the B-S kernel estimator in the boundary points.

2.1. Asymptotic properties of the B-S kernel estimator

Consider the Birnbaum–Saunders kernel given by

$$(2.1) \quad \bar{K}_{B-S}(t; \beta, \alpha) = 1 - \Phi \left(\frac{\left(\sqrt{\frac{t}{\beta}} - \sqrt{\frac{\beta}{t}} \right)}{\alpha} \right), \quad t > 0, \alpha > 0, \beta > 0,$$

where $\Phi(\cdot)$ is the Standard Normal distribution function. Let $\alpha = \sqrt{b}$ and $\beta = x$, where x and b denote the design point and the smoothing parameters, respectively. The B-S kernel estimator for the cumulative distribution function is defined as:

$$(2.2) \quad \hat{F}_1(x) = n^{-1} \sum_{i=1}^n \bar{K}_{B-S}(X_i; x, \sqrt{b}).$$

In what follows we will obtain two approximate expressions for the bias and variance for $\hat{F}_1(x)$ in Lemma 2.1 and Lemma 2.2, respectively. First consider that for the two continuous distribution functions F and G and their corresponding density function f and g , it is easy to show that:

$$(2.3) \quad E_g(F(X)) = 1 - E_f(G(X)),$$

where $E_g(F(X))$ is the expectation of $F(X)$, when X is a random variable following the distribution G .

Lemma 2.1. *Suppose that Assumptions 1–3 hold. Then we have:*

$$(2.4) \quad E(\hat{F}_1(x)) = F(x) + \frac{b}{2} (x f(x) + x^2 f'(x)) + O(b^2).$$

Proof: Since X_i 's are identical, we have

$$(2.5) \quad E_f(\hat{F}_1(x)) = E_f(\bar{K}_{B-S}(T; x, \sqrt{b})),$$

where T is a random variable following the distribution F . Using equation (2.3) and Taylor expansion, we have:

$$(2.6) \quad \begin{aligned} E_f(\bar{K}_{B-S}(T; x, \sqrt{b})) &= E_f(1 - K_{B-S}(T; x, \sqrt{b})) = E_k(F(T)) \\ &= F(x) + f(x) E(T - x) + \sum_{j=1}^{\infty} \frac{f^{(j)}(x)}{j!} E(T - x)^{j+1}, \end{aligned}$$

where $f^{(j)}(\cdot)$ is the j -th derivative of $f(x)$ and now $T \sim k_{x, \sqrt{b}}(t)$, where

$$(2.7) \quad k_{x, \sqrt{b}}(t) = \frac{t^{-\frac{3}{2}}(t+x)}{\sqrt{2\pi b x}} \exp \left\{ -\frac{1}{2b} \left(\frac{t}{x} + \frac{x}{t} - 2 \right) \right\}, \quad t > 0, x > 0, b > 0.$$

Using the results of Johnson *et al.* [11], we have:

$$\begin{aligned}
 E(T - x) &= b \frac{x}{2}, \\
 E(T - x)^2 &= \frac{bx^2}{2} (2 + 3b), \\
 E(T - x)^3 &= \frac{9b^2 x^3}{2} (3 + 5b), \\
 \implies E(F_1(x)) &= F(x) + f(x) \left(b \frac{x}{2} \right) + \frac{f'(x)}{2} \left(\frac{bx^2}{2} (2 + 3b) \right) + \frac{f^{(2)}(x)}{6} \left(\frac{9b^2 x^3}{2} (3 + 5b) \right) + \dots \\
 &= F(x) + \frac{b}{2} \left(x f(x) + x^2 f'(x) \right) + O(b^2),
 \end{aligned}
 \tag{2.8}$$

where $f'(\cdot)$ is the first derivative of $f(x)$. □

So, for the interior points, the bias of the B-S kernel estimator is of order $O(b)$. Although this rate of convergence to zero seems disappointing, one should be aware that the smoothing parameter is a function of n . In the remainder of this section, we will show that by taking this relation into account and considering the rate of convergence based on n , the bias of the B-S kernel estimator is normal (not too bad). We defer a detailed discussion of this matter until later in Section 3 where we provide a comparison between the bias of the B-S kernel estimator and the Weibull kernel estimator. In addition, in the numerical study, we will see that the overall performance of the B-S kernel estimator is not only satisfactory but also better than the other competitors. This achievement is the result of a reduction in the variance of the B-S kernel estimator, as we will see in Lemma 2.2, and what is the so-called trade-off between the variance and the bias.

Now we turn to the variance of the B-S kernel estimator. The following lemma shows that the variance of $\hat{F}_1(x)$ resembles the variance of the Empirical distribution function to some extent but it involves a negative term which can lead to its superiority over the Empirical distribution function since it has a smaller variance.

Lemma 2.2. *Suppose that Assumptions 1–3 hold. Then variance of the B-S kernel estimator can be obtained as:*

$$\text{Var}(\hat{F}_1(x)) = n^{-1} F(x) (1 - F(x)) - n^{-1} b^{\frac{1}{2}} \pi^{-\frac{1}{2}} x f(x) + O(n^{-1}b).
 \tag{2.9}$$

Proof: First consider that

$$\begin{aligned}
 E\left(\bar{K}_{\text{B-S}}^2(T; x, \sqrt{b})\right) &= \int_0^\infty \bar{K}_{\text{B-S}}^2(t; x, \sqrt{b}) f(t) dt \\
 &= \int_0^\infty F(t) \left(2 k_{\text{B-S}}(t; x, \sqrt{b}) \bar{K}_{\text{B-S}}(t; x, \sqrt{b}) \right) dt \quad (\text{using integral by part}) \\
 &= F(x) + f(x) E(Z - x) + \frac{1}{2} f'(x) E(Z - x)^2 + \dots,
 \end{aligned}
 \tag{2.10}$$

where $Z \sim 2 k_{\text{B-S}}(z; x, \sqrt{b}) \bar{K}_{\text{B-S}}(z; x, \sqrt{b})$ (a skew probability density function) and

$$k_{\text{B-S}}(z; x, \sqrt{b}) = \frac{z^{-\frac{3}{2}}(z+x)}{\sqrt{2\pi b x}} \exp\left\{-\frac{1}{2b} \left(\frac{z}{x} + \frac{x}{z} - 2\right)\right\}, \quad z > 0, \quad x > 0, \quad b > 0.
 \tag{2.11}$$

By extending the results of Vilca and Leiva [20], we have:

$$\begin{aligned}
 E(Z - x) &= \frac{b^{\frac{1}{2}} x}{2} \left(\omega_1 + b^{\frac{1}{2}} \gamma_2 \right), \\
 E(Z - x)^2 &= \frac{b x^2}{2} \left(2 \gamma_2 + \gamma_4 + b^{\frac{1}{2}} x \omega_3 \right),
 \end{aligned}
 \tag{2.12}$$

where $\gamma_r = E(W^r)$ and $\omega_r = E(W^r \sqrt{b W^2 + 4})$. In addition, W is a random variable with a Skewed Normal distribution, i.e. $W \sim \text{SN}(0, 1, -1)$.

Using the Taylor expansion for $W \sqrt{b W^2 + 4}$ and $W^3 \sqrt{b W^2 + 4}$, we obtain

$$W \sqrt{b W^2 + 4} = 2 W + \frac{1}{4} b W^3 - \frac{1}{64} b^2 W^5 + O(b^3),$$

and

$$W^3 \sqrt{b W^2 + 4} = 2 W^3 + \frac{1}{4} b W^5 - \frac{1}{64} b^2 W^7 + O(b^3).$$

Nadarajah and Kotz [13] show that $E(W) = -\frac{1}{\sqrt{\pi}}$, $E(W^3) = -\sqrt{\frac{5}{4\pi}}$, thus we can deduce that

$$\gamma_2 = 1, \quad \gamma_4 = 3, \quad \omega_1 \approx -\frac{2}{\sqrt{\pi}}, \quad \omega_3 \approx -\sqrt{\frac{5}{\pi}}.$$

By substituting $\gamma_2, \gamma_4, \omega_1$ and ω_3 in (2.12) and then substituting (2.12) in (2.10), we obtain

$$E\left(\bar{K}_{\text{B-S}}^2(T; x, \sqrt{b})\right) = F(x) - \sqrt{\frac{b}{\pi}} x f(x) + O(b).$$

Using this result and the result of Lemma 2.1, we have:

$$\begin{aligned}
 \text{Var}(\hat{F}_1(x)) &= \text{Var}\left(n^{-1} \sum_{i=1}^n \bar{K}_{\text{B-S}}(X_i; x, \sqrt{b})\right) = n^{-1} \text{Var}\left(\bar{K}_{\text{B-S}}(T; x, \sqrt{b})\right) \\
 &= n^{-1} \left\{ E\left(\bar{K}_{\text{B-S}}^2(T; x, \sqrt{b})\right) - E^2\left(\bar{K}_{\text{B-S}}(T; x, \sqrt{b})\right) \right\} \\
 &= n^{-1} \left\{ F(x) - b^{\frac{1}{2}} \pi^{-\frac{1}{2}} x f(x) + O(b) \right\} \\
 &\quad - n^{-1} \left\{ \left(F(x) + \frac{b}{2} (x f(x) + x^2 f'(x)) + O(b^2) \right)^2 \right\} \\
 &= n^{-1} F(x) (1 - F(x)) - n^{-1} \left(b^{\frac{1}{2}} \pi^{-\frac{1}{2}} x f(x) \right) + O(n^{-1} b). \quad \square
 \end{aligned}
 \tag{2.14}$$

Using Lemma 2.1 and Lemma 2.2, we can derive an estimate of the mean integrated square error (MISE) for the B-S kernel estimator as follows:

$$\begin{aligned}
 \text{MISE}_{\text{B-S}}(\hat{F}_1(x)) &= \int_0^\infty \text{MSE}(\hat{F}_1(x)) dx \\
 &\approx n^{-1} \int_0^\infty F(x) (1 - F(x)) dx - n^{-1} b^{\frac{1}{2}} \pi^{-\frac{1}{2}} \int_0^\infty x f(x) dx \\
 &\quad + \frac{b^2}{4} \int_0^\infty (x f(x) + x^2 f'(x))^2 dx.
 \end{aligned}
 \tag{2.15}$$

This result gives rise to the following proposition.

Proposition 2.1. *The optimal smoothing parameter for the B-S kernel estimator based on minimizing the MISE is*

$$\begin{aligned}
 (2.16) \quad b_{\text{B-S}}^{\text{MISE}} &= \underbrace{\arg \min}_{b>0} \left(\text{MISE}_{\text{B-S}}(\hat{F}_1(x)) \right) \\
 &\approx \left\{ \int_0^\infty x f(x) dx \right\}^{\frac{2}{3}} \left\{ \pi^{\frac{1}{2}} \int_0^\infty (x f(x) + x^2 f'(x))^2 dx \right\}^{-\frac{2}{3}} n^{-\frac{2}{3}}.
 \end{aligned}$$

This indicates that the optimal smoothing parameter is of order $O(n^{-2/3})$. By substituting $b_{\text{B-S}}^{\text{MISE}}$ in (2.15), we have:

$$\begin{aligned}
 \text{MISE}_{\text{B-S}}(\hat{F}_1(x)) &= n^{-1} \int_0^\infty F(x) (1 - F(x)) dx \\
 &\quad - \frac{3}{4} n^{-\frac{4}{3}} \pi^{-\frac{2}{3}} \left\{ \int_0^\infty x f(x) dx \right\}^{\frac{4}{3}} \left\{ \int_0^\infty (x f(x) + x^2 f'(x))^2 dx \right\}^{-\frac{1}{3}} \\
 &\quad + O(n^{-\frac{5}{3}}) \\
 \implies \text{MISE}_{\text{B-S}}(\hat{F}_1(x)) &= n^{-1} \int_0^\infty F(x) (1 - F(x)) dx - O(n^{-\frac{4}{3}}).
 \end{aligned}$$

2.2. The performance of the B-S kernel estimator at near boundary points

In order to delve in asymptotic properties of the B-S kernel estimator at the boundary points and compare the rate of its convergence at the boundary points and the interior points, we consider two specific cases for the design point x :

- a) In the case where $x = 0$, the B-S kernel is zero, i.e. $\bar{K}_{\text{B-S}}(t; x, \sqrt{b}) = 0$ and, therefore, in this case $\hat{F}_1(0) = 0$ which is remarkable because the ordinary kernel estimator does not satisfy this property.
- b) For the case where $x = cb$, where $0 < c < 1$, we have:

$$(2.17) \quad E(\hat{F}_1(x)) = F(x) + \frac{cb^2}{2} f(x) + O(b^3),$$

and

$$(2.18) \quad \text{Var}(\hat{F}_1(x)) = n^{-1} F(x) (1 - F(x)) - n^{-1} b^{\frac{3}{2}} \pi^{-\frac{1}{2}} f(x) + O(n^{-1} b^2).$$

Therefore, we can compute the mean square error (MSE) for the B-S kernel estimator at the boundary points as follows:

$$(2.19) \quad \text{MSE}_{\text{B-S}}(\hat{F}_1(x)) \approx n^{-1} F(x) (1 - F(x)) - n^{-1} b^{\frac{3}{2}} \pi^{-\frac{1}{2}} f(x) + \frac{c^2 b^4}{4} f^2(x).$$

Comparing the bias and variance terms of the B-S kernel estimator at the near boundary and interior points (in equations (2.19) and (2.15), respectively) shows that the bias term is smaller at the near boundary points at the expense of increasing the variance term.

Because at the near boundary points, the rate of convergence to zero of the negative portion of variance, which is the gain of smoothing technique over the empirical distribution function, is smaller than that of interior points.

Now it is easy to show that the optimal smoothing parameter which minimizes the MSE is

$$(2.20) \quad b_{B-S}^{MSE} = O(n^{-\frac{2}{5}}).$$

By substituting (2.20) in (2.19), we have:

$$(2.21) \quad \text{MSE}_{B-S}(\hat{F}_1(x)) = n^{-1} F(x)(1 - F(x)) + O(n^{-\frac{8}{5}}).$$

3. ASYMMETRIC CUMULATIVE DISTRIBUTION FUNCTION ESTIMATION USING WEIBULL KERNEL

In the previous section, we introduced the B-S kernel estimator and demonstrated its asymptotic consistency. In this section, we will run a similar study and introduce another cumulative distribution function estimator based on the Weibull kernel, i.e. the Weibull kernel estimator.

3.1. Asymptotic properties of the Weibull kernel estimator

Consider the Weibull kernel given by

$$(3.1) \quad \bar{K}_{wbl}(t; \alpha, \beta) = \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\}, \quad t \geq 0, \alpha > 0, \beta > 0.$$

Since $T \sim \text{Weibull}(\alpha, \beta)$ then we have:

$$(3.2) \quad E(T^k) = \beta^k \Gamma\left(1 + \frac{k}{\alpha}\right), \quad k = 1, 2, \dots,$$

where $\Gamma\left(1 + \frac{k}{\alpha}\right) = 1 - \frac{k\gamma}{\alpha} + \frac{k^2}{12\alpha^2} (\pi^2 + 6\gamma^2) + O(\alpha^3)$ and $\gamma = 0.57721$ is the Euler's constant. Hirukawa and Sakudo [9] proposed an expansion for $\Gamma\left(1 + \frac{2}{\alpha}\right) \Gamma^{-2}\left(1 + \frac{1}{\alpha}\right)$ as follows:

$$(3.3) \quad \Gamma\left(1 + \frac{2}{\alpha}\right) \Gamma^{-2}\left(1 + \frac{1}{\alpha}\right) = 1 + \frac{\pi^2}{6\alpha^2} + \frac{\gamma\pi^2 - 3\gamma^3}{2\alpha^3} + O(\alpha^{-4}).$$

Similarly, it is easy to show that

$$(3.4) \quad \Gamma\left(1 + \frac{3}{\alpha}\right) \Gamma^{-3}\left(1 + \frac{1}{\alpha}\right) = 1 + \frac{\pi^2}{2\alpha^2} + 2\frac{\gamma\pi^2 - 3\gamma^3}{\alpha^3} + O(\alpha^{-4}).$$

Let $(\alpha, \beta) = (1/b, x/\Gamma(1 + \alpha^{-1}))$ where x and b denote the design point and the smoothing parameters, respectively. Our second asymmetric Kernel-type estimator, i.e. the Weibull kernel estimator, is defined as follows:

$$(3.5) \quad \hat{F}_2(x) = n^{-1} \sum_{i=1}^n \bar{K}_{wbl}\left(X_i; 1/b, x/\Gamma(1 + b)\right).$$

The Weibull kernel estimator $\hat{F}_2(x)$ is nonnegative and appropriate to estimate cumulative distribution functions with support on $[0, \infty)$. In what follows, we present the theoretical properties of $\hat{F}_2(x)$ and we will obtain an appropriate smoothing parameter for this estimator through minimizing the mean integrated square error. We will obtain approximate expressions for the bias and variance for $\hat{F}_2(x)$ in Lemma 3.1 and Lemma 3.2, respectively. In addition, we will discuss the convergence rate of the Weibull kernel estimator in the boundary points.

Lemma 3.1. *Suppose that Assumptions 1–3 hold. Then the expectation value of $\hat{F}_2(x)$ can be obtained as:*

$$(3.6) \quad E(\hat{F}_2(x)) = F(x) + b^2 \frac{\pi^2 x^2 f'(x)}{12} + O(b^3).$$

Proof: The proof is analogous with the proof of Lemma 2.1. Using equation (2.3) and Taylor expansion, we have:

$$\begin{aligned} E(\hat{F}_2(x)) &= E\left(\bar{K}_{\text{wbl}}\left(T; 1/b, x/\Gamma(1+b)\right)\right) = E_k(F(T)) \\ &= F(x) + f(x) E(T-x) + \sum_{j=1}^{\infty} \frac{f^{(j)}(x)}{j!} E(T-x)^{j+1}, \end{aligned}$$

where T is a random variable with Weibull($1/b, x/\Gamma(1+b)$) probability density function. Using equations (3.3) and (3.4), we have:

$$\begin{aligned} E(T-x) &= 0, \\ E(T-x)^2 &= \frac{(xb\pi)^2}{6} + x^2 b^3 \left(\frac{\gamma\pi^2}{2} - 3\gamma^3\right) + O(b^4), \\ E(T-x)^3 &= (xb)^3 \left(\frac{\gamma\pi^2}{2} + 3\gamma^3\right) + O(b^4). \end{aligned}$$

Now we can conclude that

$$\begin{aligned} E(\hat{F}_2(x)) &= F(x) + \frac{1}{2} f'(x) \left(\frac{(xb\pi^2)}{6} + x^2 b^3 \left(\frac{\gamma\pi^2}{2} - 3\gamma^3\right)\right) \\ &\quad + \frac{f^{(2)}(x)}{6} \left((xb)^3 \left(\frac{\gamma\pi^2}{2} + 3\gamma^3\right)\right) + \dots \\ &= F(x) + b^2 \frac{\pi^2 x^2 f'(x)}{12} + O(b^3). \quad \square \end{aligned}$$

Note that, for the interior points, the bias of the Weibull kernel estimator is of order $O(b^2)$. However, by considering the smoothing parameter as a function of n in Remark 3.1, we will see that in the sense of convergence rate of bias, the Weibull kernel estimator is the same as the B-S kernel estimator. The following lemma provides an approximation for the variance of the Weibull kernel estimator.

Lemma 3.2. *Suppose that Assumptions 1–3 hold. Then the variance of $\hat{F}_2(x)$ can be obtained as:*

$$(3.7) \quad \text{Var}(\hat{F}_2(x)) = n^{-1} F(x) (1 - F(x)) - n^{-1} b \ln(2) x f(x) + O(n^{-1} b^2).$$

where $\ln(\cdot)$ is the natural logarithm.

Proof: First note that

$$\begin{aligned}
 E\left(\bar{K}_{\text{wbl}}^2\left(T; 1/b, x/\Gamma(1+b)\right)\right) &= \int_0^\infty \bar{K}_{\text{wbl}}^2\left(t; 1/b, x/\Gamma(1+b)\right) f(t) dt \\
 (3.8) \qquad \qquad \qquad &= \int_0^\infty F(t) \left\{ 2k_{\text{wbl}}\left(t; 1/b, x/\Gamma(1+b)\right) \bar{K}_{\text{wbl}}\left(t; 1/b, x/\Gamma(1+b)\right) \right\} dt \\
 &= F(x) + f(x) E(Z - x) + \frac{1}{2} f'(x) E(Z - x)^2 + \dots,
 \end{aligned}$$

where $Z \sim 2k_{\text{wbl}}(z; 1/b, x/\Gamma(1+b)) \bar{K}_{\text{wbl}}(z; 1/b, x/\Gamma(1+b))$, $z > 0, b > 0, x > 0$. It is easy to show that Z is random variable with Weibull $(\alpha, \frac{\beta}{2^\alpha})$ density function.

Since $2^{-\frac{1}{\alpha}} = 1 - \frac{\ln(2)}{\alpha} + \frac{\ln^2(2)}{2\alpha^2} + O(\alpha^{-3})$, we have:

$$\begin{aligned}
 (3.9) \qquad E(Z - x) &= -bx \ln(2) + \frac{(bx \ln(2))^2}{6} + O(b^3), \\
 E(Z - x)^2 &= (xb)^2 \left(\ln(2)^2 + \frac{\pi^2}{6} \right) + O(b^3).
 \end{aligned}$$

By substituting (3.9) in (3.8), we obtain

$$(3.10) \qquad E\left(\bar{K}_{\text{wbl}}^2\left(T; 1/b, x/\Gamma(1+b)\right)\right) = F(x) - bx \ln(2) f(x) + O(b^2).$$

Using (3.10) and Lemma 3.1, we can deduce that:

$$\begin{aligned}
 (3.11) \qquad \text{Var}(\hat{F}_2(x)) &= \text{Var}\left(n^{-1} \sum_{i=1}^n \bar{K}_{\text{wbl}}\left(X_i; 1/b, x/\Gamma(1+b)\right)\right) \\
 &= n^{-1} \text{Var}\left(\bar{K}_{\text{wbl}}\left(T; 1/b, x/\Gamma(1+b)\right)\right) \\
 &= n^{-1} \left\{ E\left(\bar{K}_{\text{wbl}}^2\left(T; 1/b, x/\Gamma(1+b)\right)\right) - E^2\left(\bar{K}_{\text{wbl}}\left(T; 1/b, x/\Gamma(1+b)\right)\right) \right\} \\
 &= n^{-1} \left\{ F(x) - bx \ln(2) f(x) + O(b^2) - \left(F(x) + b^2 \frac{\pi^2 x^2 f'(x)}{12} + O(b^3) \right)^2 \right\} \\
 &= n^{-1} F(x) (1 - F(x)) - n^{-1} b \ln(2) x f(x) + O(n^{-1} b^2). \qquad \square
 \end{aligned}$$

Using Lemma 3.1 and Lemma 3.2, we can derive an estimate of the MISE for the Weibull kernel estimator as follows:

$$\begin{aligned}
 (3.12) \qquad \text{MISE}_{\text{wbl}}(\hat{F}_2(x)) &\approx n^{-1} \int_0^\infty F(x) (1 - F(x)) dx - n^{-1} b \ln(2) \int_0^\infty x f(x) dx \\
 &\quad + b^4 \frac{\pi^4}{144} \int_0^\infty (x^2 f'(x))^2 dx.
 \end{aligned}$$

Now we can select the optimal smoothing parameter based on minimizing the MISE.

Proposition 3.1. *The optimal smoothing parameter for the Weibull kernel estimator based on minimizing the MISE of $\hat{F}_2(x)$ in (3.12) is*

$$\begin{aligned}
 (3.13) \qquad b_{\text{wbl}}^{\text{MISE}} &= \underbrace{\arg \min}_{b>0} \left(\text{MISE}_{\text{wbl}}(\hat{F}_2(x)) \right) \\
 &= \left\{ 36 \ln(2) \int_0^\infty x f(x) dx \right\}^{\frac{1}{3}} \left\{ \pi^4 \int_0^\infty (x^2 f'(x))^2 dx \right\}^{-\frac{1}{3}} n^{-\frac{1}{3}}.
 \end{aligned}$$

Note that the optimal smoothing parameter is of order $O(n^{-1/3})$. By substituting $b_{\text{wbl}}^{\text{MISE}}$ in (3.12), we have:

$$\begin{aligned}
 \text{MISE}_{\text{wbl}}(\hat{F}_2(x)) &= n^{-1} \int_0^\infty F(x)(1 - F(x)) dx \\
 (3.14) \quad &- 2.4764 (n\pi)^{-\frac{4}{3}} (\ln(2))^{\frac{4}{3}} \left\{ \int_0^\infty xf(x) dx \right\}^{\frac{4}{3}} \left\{ \int_0^\infty (x^2 f'(x))^2 dx \right\}^{-\frac{1}{3}} \\
 &+ O(n^{-\frac{5}{3}}) \\
 \implies \text{MISE}_{\text{wbl}}(\hat{F}_2(x)) &= n^{-1} \int_0^\infty F(x)(1 - F(x)) dx - O(n^{-\frac{4}{3}}).
 \end{aligned}$$

Remark 3.1. From the two equations (2.16) and (3.13), the optimal smoothing parameter of the B-S kernel estimator and the Weibull kernel estimator are of order $O(n^{-2/3})$ and $O(n^{-1/3})$, respectively. Therefore, in terms of the rate of convergence to zero, we have $b_{\text{B-S}}^{\text{MISE}} \approx (b_{\text{wbl}}^{\text{MISE}})^2$. Thus from (2.4) and (3.6), we can conclude that for the interior points, the bias of the B-S kernel estimator has the same rate of convergence to zero as the bias of the Weibull kernel estimator.

3.2. The performance of the Weibull kernel estimator near boundary points

In this subsection, we run a similar study like what we have done in Section 2.2 in order to investigate the asymptotic properties of the Weibull kernel estimator at the boundary points. This helps us to compare the rate of convergence at the boundary points and the interior points. We consider two specific cases for the design point x :

- a) In the case where $x = 0$, we have $\bar{K}_{\text{wbl}}(T; 1/b, x/\Gamma(1 + b)) = 0$, so in this case, unlike the ordinary kernel estimator, $\hat{F}_2(0) = 0$.
- b) For the case where $x = cb$, where $0 < c < 1$, we have:

$$(3.15) \quad E(\hat{F}_2(x)) = F(x) + \frac{c^2 b^4}{2} f(x) + O(b^5)$$

and

$$(3.16) \quad \text{Var}(\hat{F}_2(x)) = n^{-1} F(x)(1 - F(x)) - n^{-1} cb^2 \ln(2) f(x) + O(n^{-1} b^3).$$

So we can compute the MSE for the Weibull kernel estimator at the boundary points as follows:

$$(3.17) \quad \text{MSE}_{\text{wbl}}(\hat{F}_2(x)) \approx n^{-1} F(x)(1 - F(x)) - n^{-1} cb^2 \ln(2) f(x) + \frac{c^4 b^8}{4} f^2(x).$$

Comparing the two equations (3.17) and (3.14) shows a trade-off between the bias and the variance terms for the Weibull kernel estimator. This is something like what we have seen for the B-S kernel estimator in Section 2. The bias term is again smaller at the near boundary points at the expense of increasing the variance term.

Now it is easy to show that the optimal smoothing parameter which minimizes the above-mentioned MSE is

$$(3.18) \quad b_{\text{wbl}}^{\text{MSE}} = O(n^{-\frac{1}{6}}).$$

By substituting (3.18) in (3.17), we have:

$$(3.19) \quad \text{MSE}_{\text{wbl}}(\hat{F}_2(x)) = n^{-1}F(x)(1 - F(x)) + O(n^{-\frac{4}{3}}).$$

Remark 3.2. From the two equations (2.20) and (3.18), the optimal smoothing parameter of the B-S kernel estimator and the Weibull kernel estimator are of order $O(n^{-2/5})$ and $O(n^{-1/6})$, respectively. By substituting back these values into the corresponding bias terms of the two estimators, we can deduce that for the near boundary points, the bias of the B-S kernel estimator is of order $O(n^{-4/5})$ while the bias of the Weibull kernel estimator is of order $O(n^{-2/3})$.

4. NUMERICAL STUDY

In this section, we illustrate the performance of the proposed estimators (the B-S kernel estimator and the Weibull kernel estimator) through a simulation study. We compare our proposed estimators with the ordinary kernel method (O-K method), the B-K method and the Empirical distribution method. In both the O-K method and the B-K method, we use the Epanechnikov kernel. In order to select an appropriate bandwidth for the O-K and the B-K methods, we use the optimal bandwidth proposed by Altman and Leger [1] and Tenreiro [19], respectively.

We generated 1000 samples of size $n = 256$ and 1024 from eight various distributions including, 1: Burr(1, 3, 1), 2: Gamma(0.6, 2), 3: Gamma(4, 2), 4: Generalized Pareto(0.4, 1, 0), 5: Halfnormal(0, 1), 6: Lognormal(0, 0.75), 7: Weibull(1.5, 1.5) and 8: Weibull(3, 2). In order to estimate the smoothing parameter for the B-S kernel estimator and the Weibull kernel estimator, we used Gamma density $f(x) = \frac{x^{\alpha-1} \exp(-\frac{x}{\beta})}{\beta^\alpha \Gamma(\alpha)}$ as a referenced density in equations (2.16) and (3.13), respectively. The parameters (α, β) have been estimated by the method of maximum likelihood estimation.

In order to evaluate the performance of our proposed estimators and compare their functionality with other existing methods, we considered the integrated squared error $\text{ISE}_i = \int_0^\infty (\hat{F}_i(x) - F(x))^2 dx$ as an error metrics, where $\hat{F}_i(x)$, $i = 1, 2, \dots, 5$, stands for the B-S kernel estimator, the Weibull kernel estimator, the O-K method, the B-K method and the Empirical distribution method, respectively. In our setting, we approximated the integral with summation.

Table 1 shows the mean and standard deviation of the ISE for the eight distributions and the two sample sizes over one thousand repetitions. In all cases, the mean and standard deviation of the ISE decreased as the sample size increased. The simulation results show that based on the ISE, regardless of the sample size, our proposed estimators perform better than the other three methods. The only exception is distribution 5: Halfnormal(0, 1) with a sample size of 256 for which the B-K method has a smaller mean of ISE than that of the Weibull

kernel estimator. However, even for this case, when the sample size is increased to 1024, both the B-S kernel estimator and the Weibull kernel estimator have a better performance. The comparison between the B-S kernel estimator and the Weibull kernel estimator indicates the superiority of the B-S kernel estimator. This is true, surprisingly, even in estimating two distributions Weibull(1.5, 1.5) and Weibull(3, 2).

Table 1: The mean and standard deviation of the ISE in estimating eight distributions via five methods (see the text for explanation) for $n = 256$ and 1024.

Value ($\times 10^{-4}$)		B-S		Weibull		Ordinary		Boundary		Empirical	
N	Example	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
256	1	1.33	1.17	1.37	1.20	1.63	1.27	1.59	1.26	1.53	1.12
	2	4.24	4.17	4.50	4.28	6.37	5.14	5.46	5.14	4.96	4.24
	3	1.37	0.94	1.43	0.98	1.64	1.08	1.64	1.08	1.55	0.91
	4	4.44	4.59	4.70	4.74	6.52	5.46	5.33	5.47	5.14	4.64
	5	2.90	2.77	2.99	2.85	3.60	3.02	2.96	3.03	3.33	2.82
	6	5.89	5.90	6.10	6.05	8.05	6.82	7.59	6.82	6.17	5.51
	7	3.29	3.02	3.37	3.12	4.20	3.34	3.65	3.29	3.74	3.03
	8	2.62	2.40	2.68	2.46	3.09	2.58	3.03	2.57	2.98	2.39
1024	1	0.34	0.28	0.35	0.29	0.46	0.33	0.45	0.32	0.38	0.28
	2	1.18	1.14	1.22	1.17	2.38	1.55	1.78	1.56	1.30	1.14
	3	0.28	0.31	0.28	0.29	0.74	0.55	0.73	0.54	0.28	0.30
	4	1.18	1.21	1.22	1.24	2.18	1.51	1.53	1.54	1.28	1.19
	5	0.74	0.72	0.76	0.73	1.14	0.82	0.81	0.82	0.81	0.71
	6	0.63	0.56	0.64	0.58	1.04	0.70	0.91	0.69	0.67	0.52
	7	0.91	0.81	0.93	0.81	1.30	0.90	1.12	0.89	1.00	0.80
	8	0.69	0.62	0.70	0.62	0.87	0.68	0.86	0.68	0.75	0.62

In order to provide a better comparison between the aforementioned methods, we have presented the boxplots of the ISE for the case $n = 1024$ in Figure 1. In this figure, we consider eight boxplots for eight divers’ distributions. In the boxplots, the vertical axis shows the ISE and the horizontal axis contains the methods. The dotted line in each of the boxplots shows the lowest median of the ISEs. The overall superiority of the B-S kernel estimator in all cases is obvious. The overall performance of the Weibull kernel estimator is better than the B-K method and the Empirical distribution method. The O-K method shows the worst performance as is expected.

In Figure 2, we provide the results on the mean squared error (MSE) at various points of the support of the considered distributions in 1000 repetitions for the sample size (1024). This helps one to see the performance of the compared methods depending on the point where the distribution function is estimated. To increase the visibility and better compare other kernel-type estimators, we have ignored the Empirical distribution in this figure. The poor performance of the O-K method at near boundary region is obvious. At the points far from the boundary, the O-K method and the B-K method almost match. Although the amount of MSE is dependent on the design point and the distribution which we want to estimate, the overall performance of the two proposed estimators are better than both the O-K and the B-K methods. Note that, the shape of asymmetric kernels changes with the design point and for the points, those are far enough from the boundary, they become symmetric, and finally all the methods almost match in Figure 2.

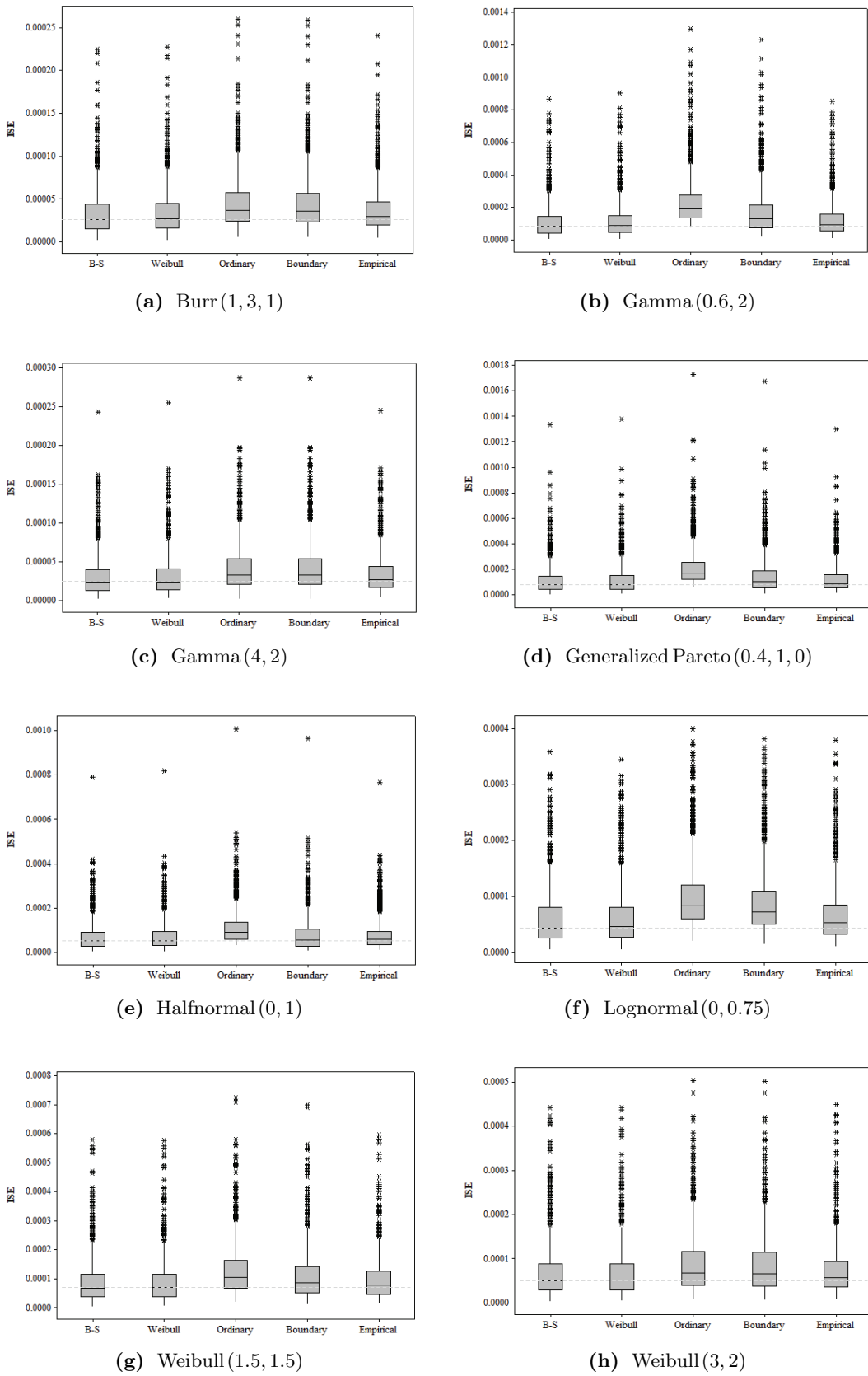


Figure 1: The boxplots of the ISE in estimating eight distribution functions via five methods in 1000 repetitions ($n = 1024$) (see text for further explanation).

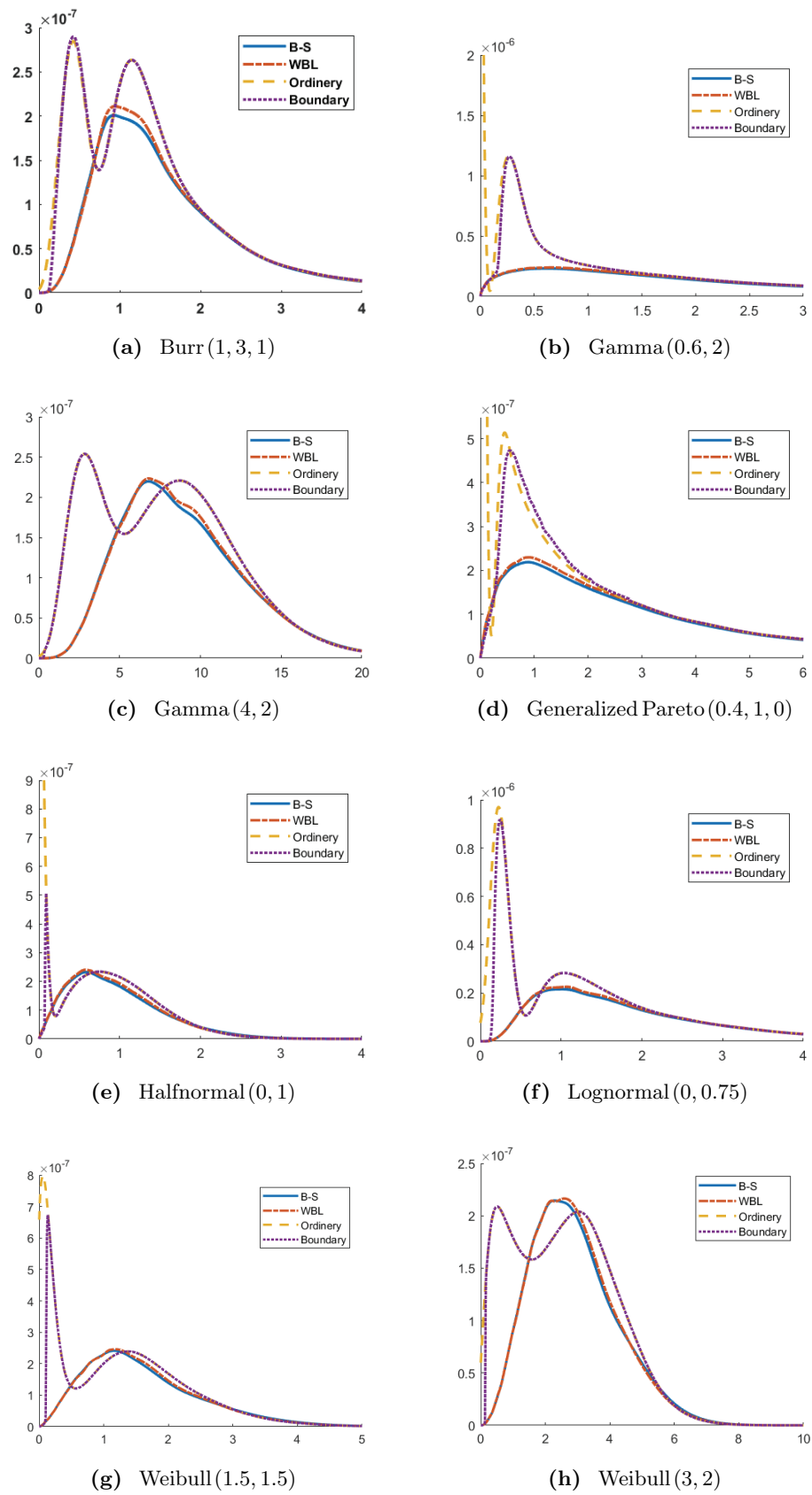
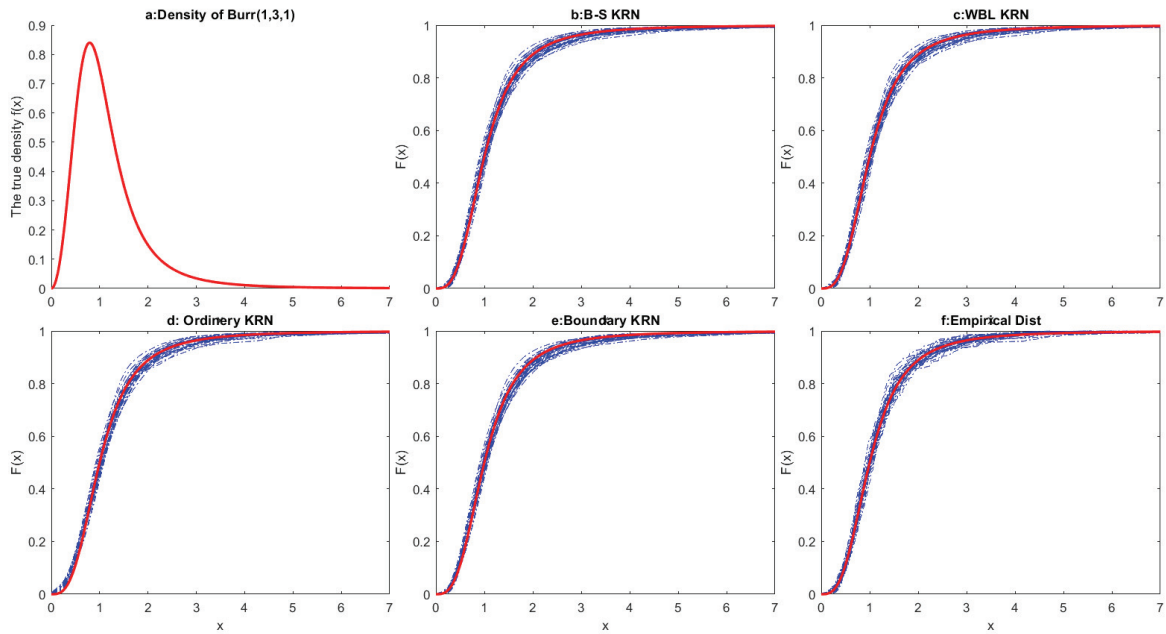
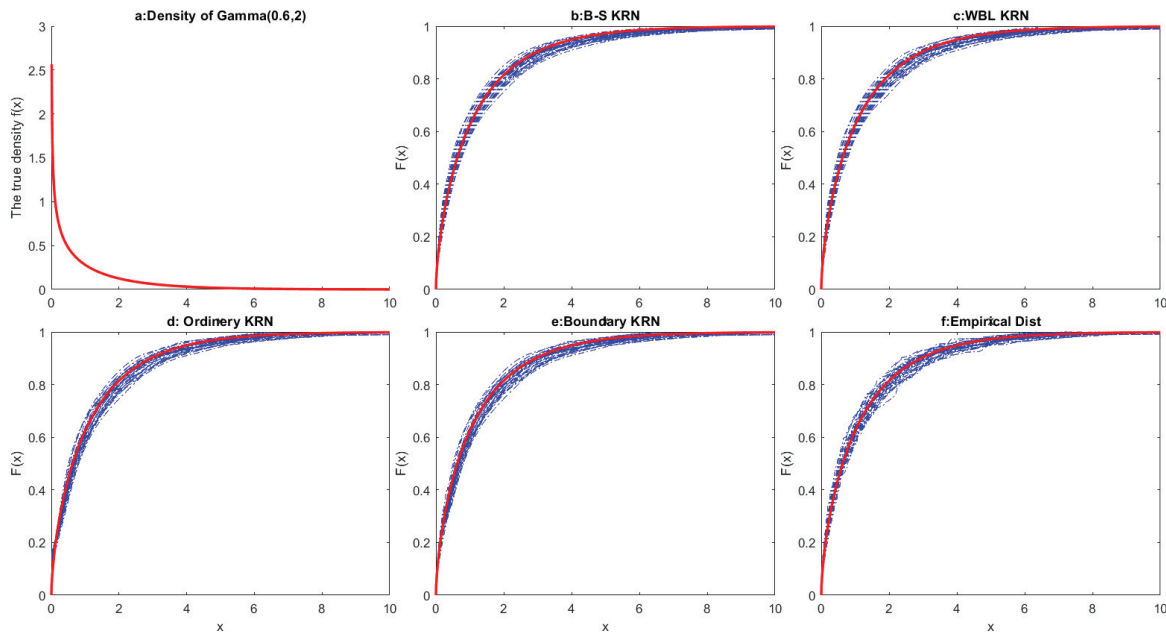


Figure 2: The Plot of the MSE in estimating eight distribution functions via five methods in 1000 repetitions ($n = 1024$) (see the text for further explanation).

Figures 3 to 6 illustrate 30 estimates in blue along with the true distribution in red for the eight different distributions ($n = 256$) via five methods. The density function of these distributions is plotted as well in the top left corner of each image. The boundary bias of the O-K method is obvious. The B-K method remedies this drawback but not completely.



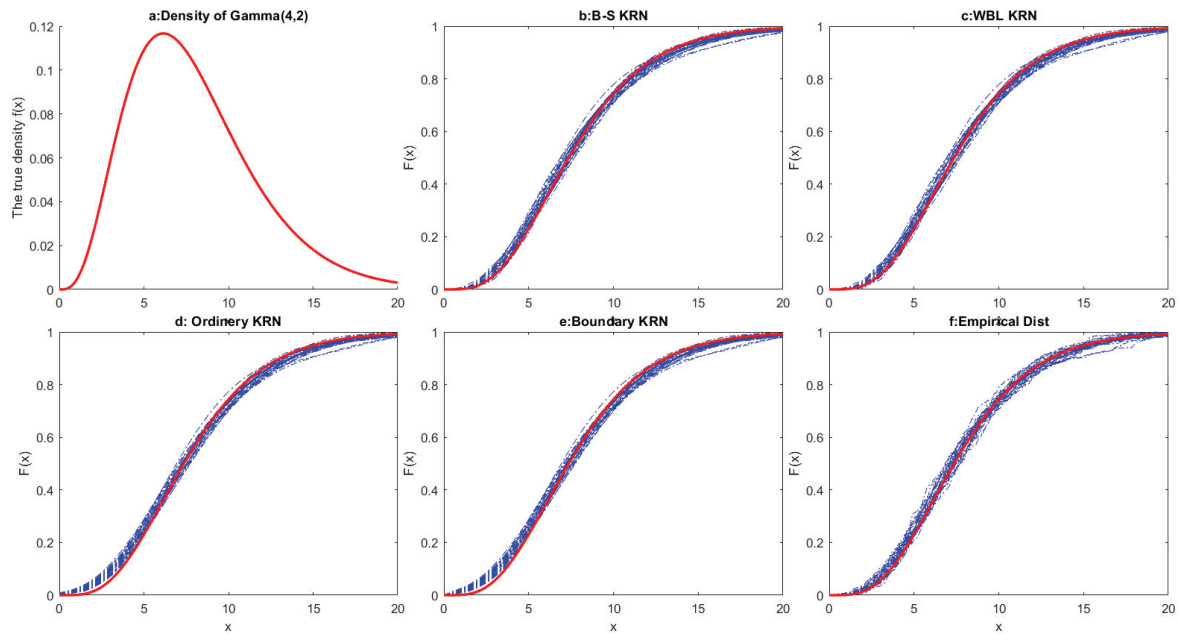
(a) Burr (1, 3, 1)



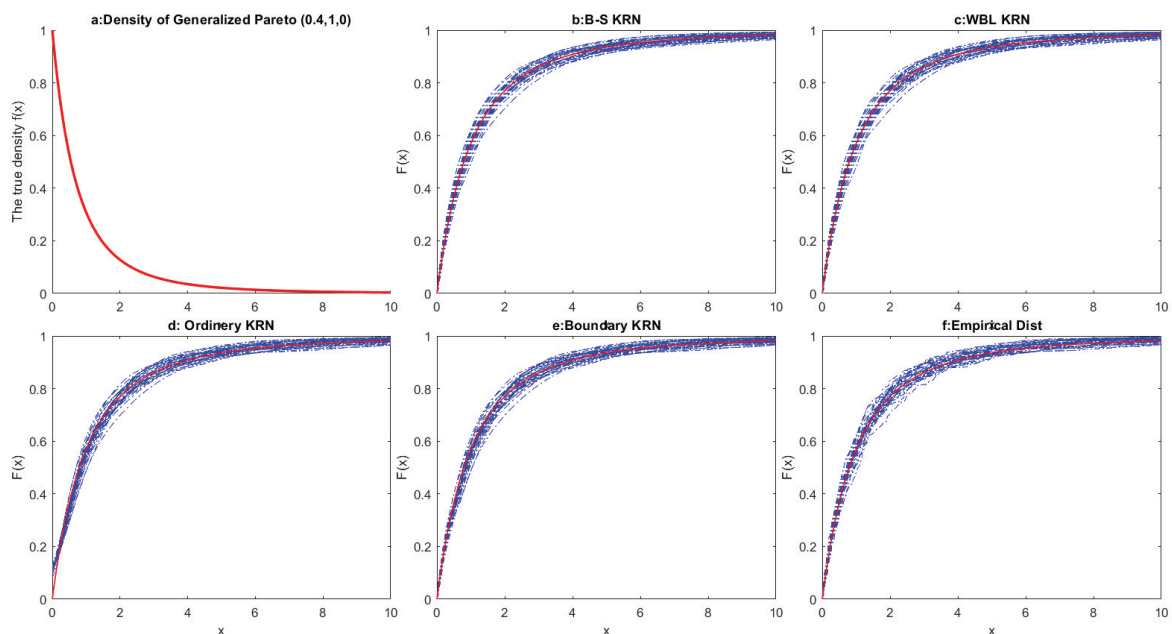
(b) Gamma (0.6, 2)

Figure 3: Plots of 30 estimates (in blue) of Burr (1, 3, 1) and Gamma (0.6, 2) via five methods: (b) B-S kernel estimator (top mid), (c) Weibull kernel estimator (top right), (d) O-K method (Bottom left), (e) B-K method (Bottom mid) and (f) Empirical distribution (bottom right). The true distribution is shown in red and sample size $n = 256$. The top left (a) shows the density function of each distribution.

In particular, a careful inspection of the figures, especially Gamma(4, 2) and Weibull(3, 2), for near boundary points, shows that the B-K method suffers from over-estimation. It seems that this problem depends on the shape of the distribution which we wish to estimate.



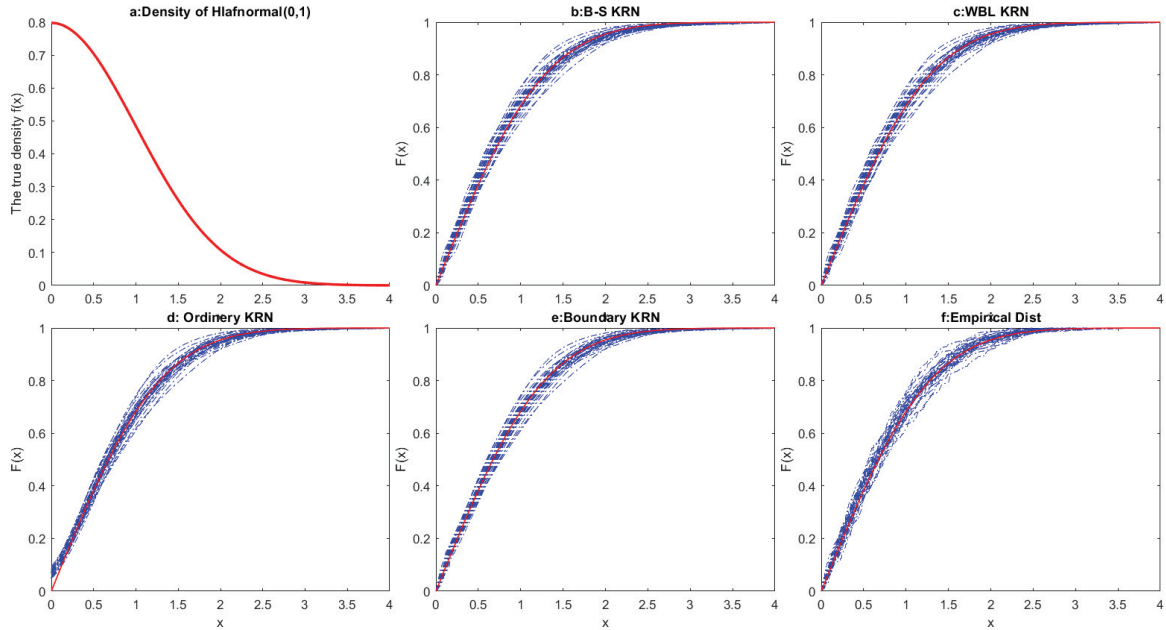
(a) Gamma (4, 2)



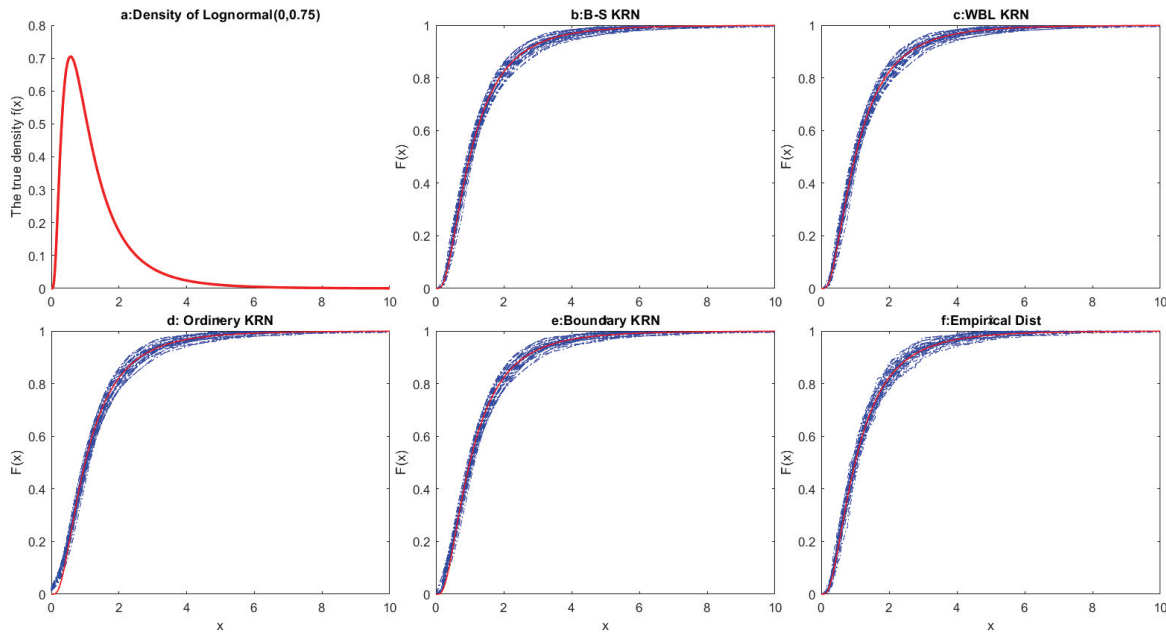
(b) Generalized Pareto(0.4, 1, 0)

Figure 4: Plots of 30 estimates (in blue) of Gamma(4, 2) and Generalized Pareto(0.4, 1, 0) via five methods: (b) B-S kernel estimator (top mid), (c) Weibull kernel estimator (top right), (d) O-K method (Bottom left), (e) B-K method (Bottom mid) and (f) Empirical distribution (bottom right). The true distribution is shown in red and sample size $n = 256$. The top left (a) shows the density function of each distribution.

Another striking point is that the Empirical distribution could not provide smooth estimates. In general, the performance of our proposed estimators is satisfying.

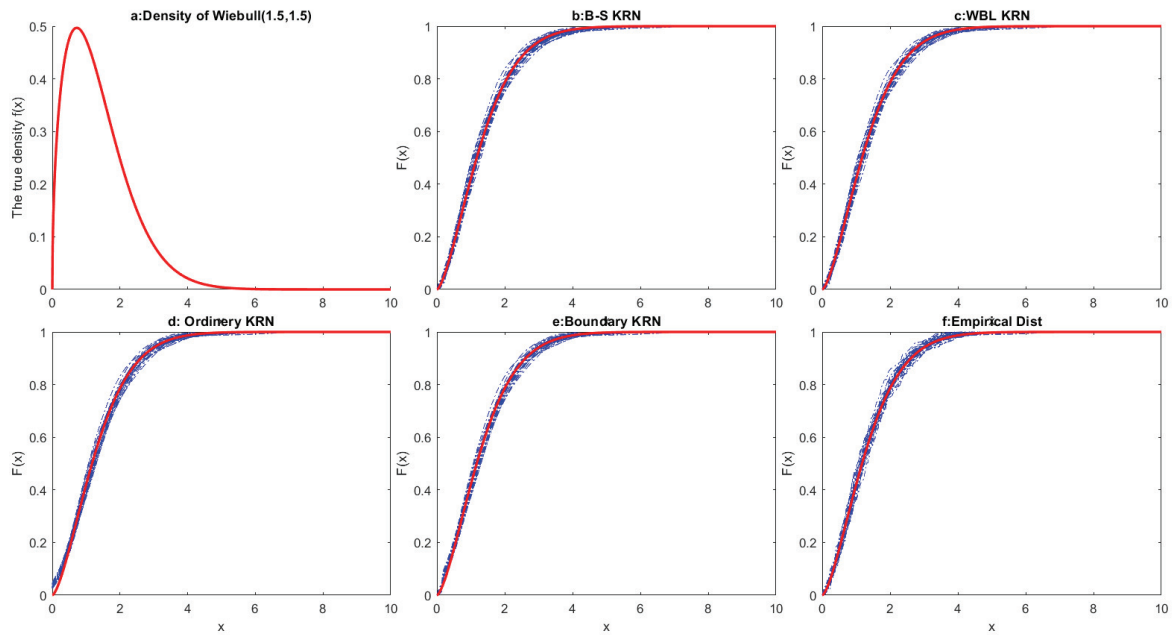


(a) Halfnormal(0, 1)

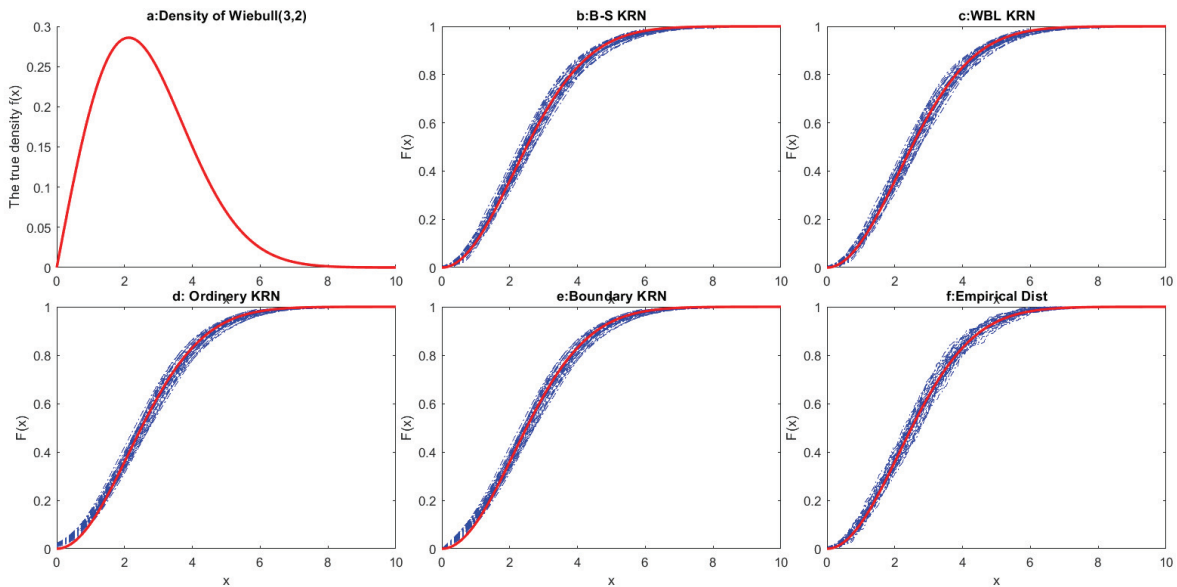


(b) Lognormal(0, 0.75)

Figure 5: Plots of 30 estimates (in blue) of Halfnormal(0, 1) and Lognormal(0, 0.75) via five methods: (b) B-S kernel estimator (top mid), (c) Weibull kernel estimator (top right), (d) O-K method (Bottom left), (e) B-K method (Bottom mid) and (f) Empirical distribution (bottom right). The true distribution is shown in red and sample size $n = 256$. The top left (a) shows the density function of each distribution.



(a) Weibull(1.5, 1.5)



(b) Weibull(3, 2)

Figure 6: Plots of 30 estimates (in blue) of Weibull(1.5, 1.5) and Weibull(3, 2) via five methods: (b) B-S kernel estimator (top mid), (c) Weibull kernel estimator (top right), (d) O-K method (Bottom left), (e) B-K method (Bottom mid) and (f) Empirical distribution (bottom right). The true distribution is shown in red and sample size $n = 256$. The top left (a) shows the density function of each distribution.

5. ILLUSTRATION WITH A REAL DATA SET

In this section, we apply our two proposed estimators to a real dataset. The data are the time distance between marriage to the first childbirth. This dataset is a result of a field research performed by Choromzadeh *et al.* [4] to study the factors that influence childbirth behavioral patterns of women aged 15–49 in a sample of size $n = 1106$ in Ahwaz, Iran. Due to the traditions, many families tend to have children immediately after marriage. Therefore, the data has a natural peak in 9–18 months after marriage. There are rare cases of childbirth in 1–8 months, which are probably the result of pregnancy before marriage. Figure 7 shows the histogram of this dataset. On the other hand, due to the changes in socioeconomic and cultural statuses, there are few families that give birth to their first child in a considerable time after their marriage. Also, there are some families whose delayed first birth is due to sterility problems. Thus, a long tail with sparse data is another considerable feature in the distribution of this dataset.

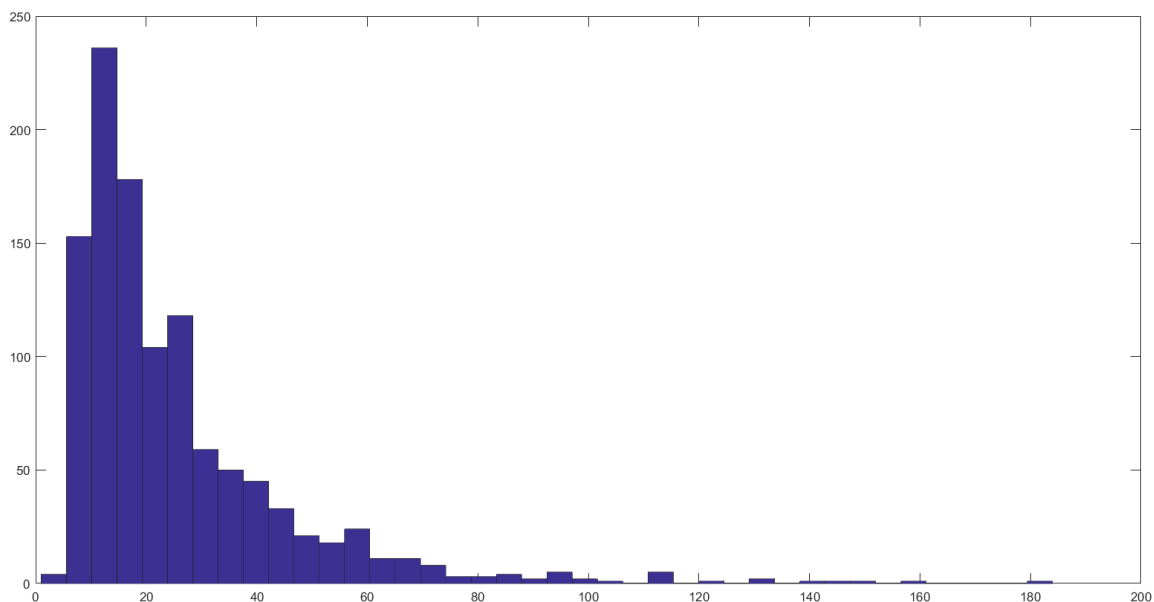


Figure 7: Histogram of the months after marriage before the first childbirth.

Figure 8 illustrates 5 estimates of the distribution of this data via five methods. The methods of choosing the smoothing parameter for various estimators are described in Section 4. Figure 8(a) shows that estimates mainly differ at the near origin. In order to provide a better insight, we separately illustrate the estimates in the first 9 months in Figure 8(b). In comparison with the Empirical distribution, the estimates created by the O-K method and the B-K method are similar. It seems they rise too early. In the simulation study, we have seen that these two estimators suffer from over-estimating for near boundary points, especially for those distributions that have the same shape as in Figure 7. The B-S kernel estimate and the Weibull kernel estimate are very close, and the more consistent they are with the Empirical distribution and for this dataset, the more realistic they seem to be.

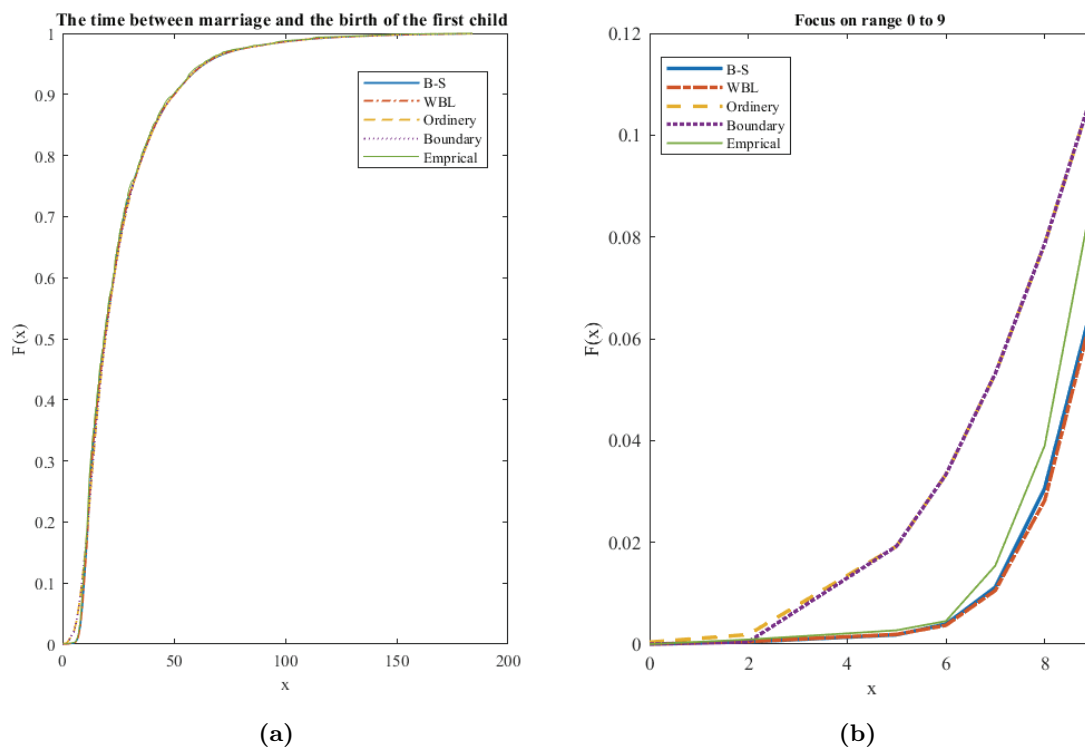


Figure 8: Five estimates of the distribution of first childbirth via five methods: B-S kernel estimator (solid-blue), Weibull kernel estimator (dashed-red), O-K method (dashed-yellow), B-K method (dotted-purple) and Empirical distribution (solid-green).

6. CONCLUSION AND DISCUSSION


This paper is devoted to proposing some appropriate estimators for the cumulative distribution functions with non-negative support. To achieve this goal, we proposed a general asymmetric Kernel-type estimator and introduced two asymmetric estimators for the cumulative distribution function. We demonstrated the asymptotic consistency of our proposed estimators and we showed that they are free from boundary effects as well. Comparing our estimators based on the rate of convergence at the boundary points, we found that the B-S kernel estimator was better than the Weibull kernel estimator. In our setting, we estimated the bandwidths of the two estimators based on minimizing the MISE. In order to evaluate the performance of our estimators and compare them with other existing methods, we conducted a numerical study. The results of the numerical study show that both the B-S kernel and the Weibull kernel estimators are superior to the B-K method proposed by Tenreiro [19]. In the numerical study, the B-S kernel estimator achieved the best results and outperformed the Weibull kernel estimator. This is consistent with the good asymptotic properties of the B-S kernel estimator. In this research, we used the B-S kernel and the Weibull kernel as the asymmetric kernels in our general estimator. As a path for future research, one can try other existing asymmetric kernels. Another area for future research can be the estimation of those cumulative distributions with a finite interval support, for instance $[a, b]$. In addition, application of this type of cumulative distribution estimator in several other fields such as the survival analysis and the copula methods is an interesting topic for future research.


REFERENCES

- [1] ALTMAN, N. and LEGER, C. (1995). Bandwidth selection for kernel distribution function estimation, *Journal of Statistical Planning and Inference*, **46**(2), 195–214.
- [2] CHEN, S. (1999). Beta kernel estimators for density functions, *Computational Statistics and Data Analysis*, **31**(2), 131–145.
- [3] CHEN, S. (2000). Probability density function estimation using Gamma kernels, *Annals of the Institute of Statistical Mathematics*, **52**(3), 471–480.
- [4] CHOROMZADEH, R.; AKHOUND, M.R. and RASEKH, A. (2015). Factors affecting women's birth intervals: the case of women referred to health centers in Ahwaz, *Journal of Hayat*, **20**(4), 35–50.
- [5] FALK, M. (1983). Relative efficiency and deficiency of kernel type estimators of smooth distribution functions, *Statistica Neerlandica*, **37**(2), 73–83.
- [6] GASSER, T. and MULLER, H. (1979). Kernels estimation of regression functions, *Lecture Notes in Mathematics*, **757**, 23–68.
- [7] GASSER, T.; MULLER, H. and MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation, *Journal of the Royal Statistical Society, Series B*, **47**(2), 238–252.
- [8] HIRUKAWA, M. and SAKUDO, M. (2014). Nonnegative bias reduction methods for density estimation using asymmetric kernels, *Computational Statistics and Data Analysis*, **75**, 112–123.
- [9] HIRUKAWA, M. and SAKUDO, M. (2015). Family of the generalised gamma kernels: a generator of asymmetric kernels for nonnegative data, *Journal of Nonparametric Statistics*, **27**(1), 41–63.
- [10] JIN, X. and KAWCZAK, J. (2003). Birnbaum–Saunders and lognormal kernel estimators for modelling durations in high frequency financial data, *Annals of Economics and Finance*, **4**(1), 103–124.
- [11] JOHNSON, N.; KOTZ, S. and BALAKRISHNAN, N. (1995). *Continuous Univariate Distribution*, Volume 2, 2nd Ed., Wiley, New York.
- [12] MULLER, H. (1991). Smooth optimum kernel estimators near endpoints, *Biometrika*, **78**(3), 521–530.
- [13] NADARAJAH, S. and KOTZ, S. (2003). Skewed distribution generated by normal kernel, *Statistics and Probability Letters*, **65**(3), 269–277.
- [14] NADARAYA, E.A. (1964). Some new estimate for distribution function, *Theory of Probability and Its Applications*, **9**(3), 497–500.
- [15] REISS, R.D. (1981). Nonparametric estimation of smooth distribution functions, *Scandinavian Journal of Statistics*, **8**(2), 116–119.
- [16] RICE, J. (1984). Boundary modification for kernel regression, *Communications in Statistics – Theory and Methods*, **13**(7), 893–900.
- [17] SCAILLET, O. (2004). Density estimation using inverse and reciprocal inverse Gaussian kernels, *Journal of Nonparametric Statistics*, **16**(1), 217–226.
- [18] SINGH, R.S.; GASSER, T. and PRASAD, B. (1983). Nonparametric estimates of distribution functions, *Communications in Statistics – Theory and Methods*, **12**(18), 2095–2108.
- [19] TENREIRO, C. (2013). Boundary kernels for distribution function estimation, *REVSTAT – Statistical Journal*, **11**(2), 169–190.
- [20] VILCA, F. and LEIVA, V. (2006). A new fatigue life model based on the family of skew-elliptical distributions, *Communications in Statistics – Theory and Methods*, **35**(2), 229–244.
- [21] WATSON, G. and LEADBETTER, M. (1964). Hazard analysis II, *Sankhya: The Indian Journal of Statistics, Series A*, **21**(1), 229–244.
- [22] ZHANG, S. (2010). A note on the performance of the gamma kernel estimators at the boundary, *Statistics and Probability Letters*, **80**(7), 548–557.

PRODUCTION PROCESSES WITH DIFFERENT LEVELS OF RISK: ADDRESSING THE REPLACEMENT OPTION

Authors: FRANCISCO STEFANO DE ALMEIDA
– Mercer Portugal, Lisboa, Portugal

CLÁUDIA NUNES 
– Department of Mathematics and CEMAT, Instituto Superior Técnico,
Universidade de Lisboa,
Av. Rovisco Pais, 1049-001 Lisboa, Portugal

CARLOS OLIVEIRA  *
– CEMAPRE, ISEG – School of Economics and Management, Universidade de Lisboa,
Rua do Quelhas 6, 1200-781 Lisboa, Portugal
carlosoliveira@iseg.ulisboa.pt

Received: April 2019

Revised: August 2019

Accepted: August 2018

Abstract:

- It is often found that a company has the opportunity to change its original production process to a different one. Here we consider two different situations: in the first one, the new production process may lead to larger losses (in case the demand decreases), but may also lead to larger profits (in case the demand increases); so we increase the risk. In the second one, the opposite holds (and we decrease the risk).

We derive the optimal replacement strategy, and we study the impact of the drift and the volatility in the decision. Afterward, we include the option to exit the market and we compare both situations (with and without the exit option), concluding that in case the investment is in a less risky process, the impact of the exit option is different, depending on a relation involving the costs and the drift parameters.

Keywords:

- *risk management; replacement policies; exit decision; real options; optimal stopping.*

*Corresponding author.

1. INTRODUCTION

Last decades have been characterized by relevant changes in the global macroeconomic scenario. Naturally, this has affected the way managers make decisions about the future of firms. Real options theory provides efficient tools to analyze what is the best decision to make, taking into account all the options faced by the firm. In particular, we can highlight the following options/problems: the strategy options (see Huisman and Kort (2003) [11], Pawlina and Kort (2006) [18], and Brealey *et al.* (2012) [3]), valuation of options for real assets (see Dixit and Pindyck (1994) [6]), investment options (see Bjerksund and Ekern (1990) [2], Dixit and Pindyck (1994) [6], Majd and Pindyck (1987) [13], and McDonald and Siegel (1985) [14]), technology adoption problem (see Farzin *et al.* (1998) [8], and Hagspiel *et al.* (2016) [10]), or abandonment problem (see Brennan and Schwartz (1985) [4], and Myers and Majd (2001) [15]).

During the last economic crisis, many firms felt the need to adjust their production process, to face declining markets and to avoid large losses. An example of a strategy used by decision-makers to decrease the costs associated with the production process is the layoff. Companies like Merck, Yahoo, General Electric, Xerox, Pratt & Whitney, Goldman Sachs, Whirlpool, Bank of America, Alcoa and Coca-Cola implemented layoff periods, to reduce costs and face adverse market conditions¹. The main goal of firms adopting this type of strategies is to reduce the risk of having large losses by adopting a production process which results in a “flat payoff function”: the profits would not be very large if the demand is large but in case the demand decreases, the firm faces also small losses; the resulting is a sort of a compromise situation. The idea of flat payoff function is already used in Decision Theory, with a similar meaning; see, for instance, Pannell (2006) [17]. On the other side of the scale, you also find companies with a more *aggressive* behavior, meaning, in this particular framework, that the firm adapts its production process in order to obtain large profits for high levels of demand, even if the losses may be large, for small levels of demand. There are, of course, many strategies which lead to intermediate payoff functions: between the flat payoff functions (less risky, in terms of potential losses) and the more aggressive ones (more risky, in terms of potential losses).

The changes in the profit function may be due to several causes, such as technology innovation or improvement in the production process. Indeed, nowadays, companies face many challenges, as the markets are very competitive, and technology innovations can radically change the costs and profits. Technology innovation may change the production costs, as it gets more advanced, prices drop and products get better. However, it can exist some drawbacks, such as the costs associated with the technological process or even a chance that the switch to the new technology does not lead to positive profits but leads to losses, due to declining markets, for instance. These challenges amplify with the large uncertainty that is inherent to the market, as Ward *et al.* (1995) [20] refer.

There are several examples of such a situation. One of such examples occurs in the area of IT (information technology), the decision of where and when to allocate resources to IT programs is risky, as although there are many positive outcomes, the executives struggle with the massive costs and high uncertainty. According to Clemons and Weber (1990) [5],

¹Uchitelle, L. (2008, October 26). U.S. layoffs increase as businesses confront the crisis, *The New York Times*. Retrieved from <http://www.nytimes.com/2008/10/26/business/worldbusiness/26iht-layoffs.1.17246245.html>

IT can confer advantage under appropriate conditions, and equally important, even when it fails to confer advantage, it may still prove crucial. The same authors mention the case of Manufacturers Hanover, that in the early 80's invested 300 million dollars in a telecommunication network. The actual volumes reached only 50% of the estimates, well below the capacity, and leading to massive losses, as they could not recover the system cost. See Benaroch (2002) [1].

Another such example is the present situation of ASML: the largest supplier in the world of photolithography systems for the semiconductor industry. ASML is one of the 8th foreign companies that have sales of at least 1 billion dollars in South Korea. Recent investments in the next-generation technologies have allowed ASML to reduce their potential costs by 30% or 40%. But a serious flare-up between North and South Korea would cause a huge disruption to commerce. And if operations in the country were suspended or set back for a long time due to the destruction of facilities, that would disrupt the supply chain of companies around the world. And ASML, which is vulnerable to this situation, would then face major losses.² Therefore investments in this area of the planet, although lead to potentially large profits, also may lead the massive losses.

The last example that we provide is related to the use of statistical process control (SPC) charts to monitor quality. Control charts are used to keep a process in statistical control, where the output quality is at a target level; the design of the control chart is usually known as economic design (see Lorenzen and Vance (1986) [12]). But the implementation of statistical control can be quite expensive, as Nembhard *et al.* (2002) [16] refer. But, on the other hand, if a control chart is not used, the manufacturer may not be aware that the system is producing low-quality parts. And this may have a cost, as these products may be returned, with extra replacement costs. Therefore the choice between implementing a production scheme with or without a rigorous statistical control is a relevant decision in terms of profits and losses, and the decision must take into account the dynamics in the market conditions.

These examples show a common feature: firms have the opportunity to change their production systems, due to several reasons, but when deciding about it they need to balance between potential losses and gains, as these investments do not lead only to larger profits. Our main objective is to study the time at which the firm should optimally change its production system. Reporting to the literature of real options, this problem falls into the category of single-switch or replacement problems, a problem that is crucial from the management viewpoint. We will be mainly concerned with the implications of adjusting the current production process in a risky or less risky way, where we use the following interpretation:

- The risk increases if when compared with the current profit, the gains of the firm increase when the demand is sufficiently high, but the losses also increase in case the demand is not sufficiently high;
- The risk decreases if, when compared with the current profit, the losses of the firm decrease when the demand reaches sufficiently small levels, but the gains also decrease in case the demand becomes sufficiently high.

Throughout the paper, we use the terms *replacement* and *investing* indistinctly, in the sense that they both mean that the firm will change its original production process (leading to a profit function Π_1) by a different production process (leading to a profit function Π_2).

²Wong, S. and Miller, L.J. (2017, August 20). These are the most vulnerable foreign companies in Korea, *Bloomberg Politics*. Retrieved from <https://www.bloomberg.com/news/articles/2017-08-20/in-shadow-of-red-line-companies-with-a-lot-to-lose-in-korea>

Besides the option to change its production process, a firm may still decide to abandon the market, in case the conditions are no longer favorable in terms of its profits. Therefore we also analyze the situation where after the investment in the second production process, the firm may decide to exit. Moreover, we compare the impact of the abandonment option in the invest moment in the second production market and, as we will see, this impact depends on whether the firm intends to increase the risk or not, and the relation between the involved costs and the parameters of the demand process. Here we assume that abandonment only happens after investing in the second production process, which is equivalent to say that abandonment out of the first production process is equally costly as first investing in the second production process and then abandon. This assumption is also considered in chapter 7 of Dixit and Pindyck (1994) [6].

The rest of the paper is organized as follows: in Section 2 we describe the model, along with some considerations about the economical meaning; in Section 3 we present the Hamilton–Jacobi–Bellman equation for the optimization problem. In Section 4 we derive the solution of the problem and in Section 5 we present comparative statics results. Finally, in Section 6 we consider the option to abandon the market, after investing in the second production process. The proofs of the propositions and corollaries can be found in Appendix A.

2. MODEL

In this paper, we consider a firm that produces an established product in a stochastic environment, which is characterized by the stochastic demand process $X = \{X_t : t \geq 0\}$, defined on a complete filtered space $(\Omega, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$. Moreover, we assume that X follows a geometric Brownian motion, solution of the stochastic differential equation:

$$dX_t = \mu X_t dt + \sigma X_t dW_t,$$

where $X_0 = x$, $\mu \in \mathbb{R}$ is the drift, the volatility is equal to $\sigma > 0$, and $\{W_t : t \geq 0\}$ is a Brownian motion.

Currently, the profit of the firm is Π_1 , that depends on X , and the firm has the option to change its profit function to Π_2 , but staying in the same market (and thus the uncertainty process, X , does not change its dynamics as a consequence of this change). If the firm decides to materialize this option at time τ , then its value is given by

$$\begin{aligned} J(x, \tau) &= E_x \left[\int_0^\tau e^{-\gamma s} \Pi_1(X_s) ds - e^{-\gamma \tau} R + \int_\tau^\infty e^{-\gamma s} \Pi_2(X_s) ds \right] \\ &= E_x \left[\int_0^\tau e^{-\gamma s} \Pi_1(X_s) ds + \int_\tau^\infty e^{-\gamma s} (\Pi_2(X_s) - \gamma R) ds \right], \end{aligned}$$

where $R \geq 0$ is the cost of adjusting its production process, $\gamma > 0$ is the interest rate and E_x represents the conditional expectation when $X_0 = x$. Defining by \mathcal{S} the set of all admissible $\{\mathcal{F}_t\}$ -stopping times, we are looking for the right moment of changing the production process. Thus, we define the value function \mathcal{V} , given by:

$$(2.1) \quad \mathcal{V}(x) = \sup_{\tau \in \mathcal{S}} J(x, \tau) = J(x, \tau^*).$$

If in the problem (2.1), one has $\Pi_1(x) \leq \Pi_2(x) - \gamma R$, for all $x > 0$, then the decision is trivial: $\tau^* = 0$, and therefore the firm must change immediately. On the other hand, when $\Pi_2(x) - \gamma R \leq \Pi_1(x)$, for all $x > 0$, then the decision is also trivial: $\tau^* = \infty$, and therefore the firm never takes the decision to invest in the second production process. However, the most interesting situation is illustrated in Figure 1. In fact, assuming that there is $c > 0$, such that $\Pi_1(c) = \Pi_2(c) - \gamma R \equiv d > 0$, then, we have the following situations:

- a) $\Pi_1(x) < \Pi_2(x) - \gamma R$ if and only if $x > c$. In this case, for lower values of the demand process, Π_1 leads to larger profits or lower losses than Π_2 , whereas for large values of demand, Π_2 is more profitable. For this reason, we say that in this situation the *risk increases*, when we switch from Π_1 to Π_2 ;
- b) $\Pi_1(x) > \Pi_2(x) - \gamma R$ if and only if $x < c$. Then Π_2 leads to smaller losses/smaller earnings in case the demand decreases/increases, when compared with Π_1 . For this reason, we say that in this situation the *risk decreases*.

Here, we use isoelastic profit functions, with some constant linear factor:

$$\Pi_i(x) = a_i x^{\theta_i} - b_i, \quad \text{with } \theta_i \geq 1, \quad a_i, b_i \geq 0,$$

where θ_i is the elasticity coefficient and b_i denotes a fixed cost.

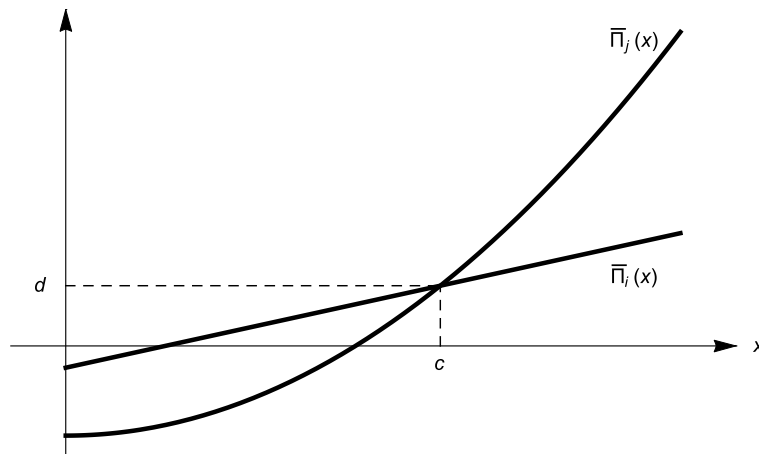


Figure 1: Representation of the functions $\bar{\Pi}_s$, with $s = i, j$ and $i \neq j \in \{1, 2\}$, where $\bar{\Pi}_s(x) = \Pi_1(x)$ if $s = 1$ or $\bar{\Pi}_s(x) = \Pi_2(x) - \gamma R$, if $s = 2$, for all $x > 0$.

Additionally, we will discuss how the option to abandon definitely the market after the replacement influences the value of the firm as well as the economic mechanisms behind the decisions. Then, the problem can be re-stated as follows:

$$(2.2) \quad \begin{aligned} \mathcal{W}(x) &= \sup_{\tau_1 \leq \tau_2 \in \mathcal{S}} E \left[\int_0^{\tau_1} e^{-\gamma s} \Pi_1(X_s) ds - e^{-\gamma \tau_1} R + \int_{\tau_1}^{\tau_2} e^{-\gamma s} \Pi_2(X_s) ds - e^{-\gamma \tau_2} S \right] \\ &\equiv \sup_{\tau_1 \leq \tau_2 \in \mathcal{S}} I(x, \tau_1, \tau_2), \end{aligned}$$

where τ_1 is the time to replace Π_1 by Π_2 , and τ_2 is the time to abandon the market. In (2.2), S represents the abandonment cost, when S is positive (meaning that the firm needs to pay

to abandon the market) or a salvage value/disinvestment subsidy, when S is negative (meaning that the firm receives money upon the exit of the market).

In order to have a well-posed problem, in the sense that the next integrability condition holds:

$$(2.3) \quad E_x \left[\int_0^\infty e^{-\gamma s} |\Pi_i(X_s)| ds \right] < \infty, \quad \text{for } i = 1, 2,$$

we assume the following relation on the parameters:

$$\gamma > \frac{\sigma^2}{2} (\theta_i - 1) \theta_i + \theta_i \mu \equiv \mu_{\theta_i}, \quad \text{for } i = 1, 2.$$

See Guerra *et al.* (2016) [9] for further mathematical explanations about the integrability condition (2.3). Additionally, for (γ, μ, σ) fixed, let β_1 and β_2 denote the two roots of the quadratic equation

$$\gamma = \frac{\sigma^2}{2} (y - 1) y + \mu y,$$

with $\beta_1 < 0 < \beta_2$. We notice that the condition (2.3) implies that $\beta_2 > \theta > 1$.

Although the natural economic modeling of this problem relies on the set of parameters (r, μ, σ) , it can be, equivalently, modeled by using the set of parameters $(\beta_1, \beta_2, \sigma)$, since

$$\gamma = -\frac{\sigma^2}{2} \beta_1 \beta_2 \quad \text{and} \quad \mu = \frac{\sigma^2}{2} (1 - \beta_1 - \beta_2).$$

For future reference, we note that the functions $(\mu, \sigma) \rightarrow \beta_i(\mu, \sigma)$, with $i = 1, 2$, are such that the function $\beta_1(\cdot, \sigma)$, $\beta_2(\cdot, \sigma)$ and $\beta_2(\mu, \cdot)$ ³ are decreasing, while $\beta_1(\mu, \cdot)$ is increasing. This follows in view of the following derivatives:

$$\frac{\partial \beta_i}{\partial \sigma} = (-1)^{i+1} \frac{\sigma \beta_i (\beta_i - 1)}{\sqrt{(\mu - \frac{1}{2} \sigma^2)^2 + 2 \sigma^2 \gamma}} \quad \text{and} \quad \frac{\partial \beta_i}{\partial \mu} = (-1)^{i+1} \frac{\beta_i}{\sqrt{(\mu - \frac{1}{2} \sigma^2)^2 + 2 \sigma^2 \gamma}}.$$

3. HAMILTON–JACOBI–BELLMAN EQUATIONS

In this section, we introduce the HJB equations that lead to the solution of the optimization problems. We start by noticing that for the replacement problem, we may write the functional J as follows:

$$\begin{aligned} J(x, \tau) &= E_x \left[\int_0^\tau e^{-\gamma s} (\Pi_1(X_s) - \Pi_2(X_s) + \gamma R) ds \right] + E_x \left[\int_0^\infty e^{-\gamma s} (\Pi_2(X_s) - \gamma R) ds \right] \\ &= E_x \left[\int_0^\tau e^{-\gamma s} (a_1 X_s^{\theta_1} - a_2 X_s^{\theta_2} - b + \gamma R) ds \right] + a_2 \frac{x^{\theta_2}}{\gamma - \mu_2} - \frac{b_2 + \gamma R}{\gamma}, \end{aligned}$$

³Along this paper, we use $f(\cdot, y)$ to denote the function f as a function of the first variable, keeping the second fixed and equal to y .

for every $(x, \tau) \in]0, \infty[\times \mathcal{S}$, with $b = b_1 - b_2$.⁴ Then, for all $x > 0$,

$$(3.1) \quad \mathcal{V}(x) = V(x) + a_2 \frac{x^{\theta_2}}{\gamma - \mu_2} - \frac{b_2 + \gamma R}{\gamma},$$

with

$$(3.2) \quad V(x) = \sup_{\tau \in \mathcal{S}} E_x \left[\int_0^\tau e^{-\gamma s} \left(a_1 X_s^{\theta_1} - a_2 X_s^{\theta_2} - b + \gamma R \right) ds \right],$$

and the remaining part of the right-hand side of Equation (3.1) representing the net present value associated to the second production process. Thus, henceforward, we will be concerned about the optimal stopping problem defined in (3.2).

In light of the classical Theory of Optimal Stopping (see, for instance, Peskir and Shiryaev (2006) [19]), V satisfies the HJB equation:

$$\min \left\{ \gamma v(x) - \mu x v'(x) - \frac{\sigma^2}{2} x^2 v''(x) - \left(\Pi_1(x) - \Pi_2(x) + \gamma R \right), v(x) \right\} = 0.$$

From this equation, it follows that $V(x) \geq 0$, for $x > 0$. Additionally, if there is $x_0 > 0$ such that $V(x_0) > 0$, then V should satisfy the ODE

$$(3.3) \quad \gamma v(x) - \mu x v'(x) - \frac{\sigma^2}{2} x^2 v''(x) - \left(\Pi_1(x) - \Pi_2(x) + \gamma R \right) = 0,$$

in the set $\{x > 0: |x - x_0| < \epsilon\}$, for some $\epsilon > 0$. Equation (3.3) is an Euler–Cauchy differential equation and admits as solution the function

$$(3.4) \quad v(x) = Ax^{\beta_1} + Bx^{\beta_2} + \alpha x^{\theta_1} - \beta x^{\theta_2} - \frac{b}{\gamma} + R,$$

with

$$\alpha = \frac{a_1}{\gamma - \mu_{\theta_1}} \quad \text{and} \quad \beta = \frac{a_2}{\gamma - \mu_{\theta_2}},$$

for every $A, B \in \mathbb{R}$.

When we consider the exit option after investing in the second process production process, one may see that standard arguments (see, for instance, Duckworth and Zervos (2000) [7]) allow us to get an equivalent expression to (2.2), that is:

$$(3.5) \quad \begin{aligned} \mathcal{W}(x) &= \sup_{\tau_1 \in \mathcal{S}} E \left[\int_0^{\tau_1} e^{-\gamma s} \left(\Pi_1(X_s) + \gamma R + \gamma S \right) ds + e^{-\gamma \tau_1} \tilde{W}(X_{\tau_1}) \right] - R - S \\ &\equiv \sup_{\tau_1 \leq \tau_2 \in \mathcal{S}} \tilde{I}(\tau_1, \tau_2, x) - R - S, \end{aligned}$$

where

$$\tilde{W}(x) = \sup_{\tau \in \mathcal{S}} E \left[\int_0^\tau e^{-\gamma s} \left(\Pi_2(X_s) + \gamma S \right) ds \right].$$

Thus the corresponding HJB equation is the following:

$$\min \left\{ \gamma w(x) - \mu x w'(x) - \frac{\sigma^2}{2} x^2 w''(x) - \Pi_1(x) - \gamma(R + S), w(x) - \tilde{W}(x) \right\} = 0$$

where, in its turn, \tilde{W} is a solution of the HJB equation corresponding to the exit problem:

$$\min \left\{ \gamma \tilde{W}(x) - \mu x \tilde{W}'(x) - \frac{\sigma^2}{2} x^2 \tilde{W}''(x) - \Pi_2(x) - \gamma S, \tilde{W}(x) \right\} = 0.$$

⁴From now on we will use the notation $a = a_1 - a_2$.

4. THE REPLACEMENT OPTION

In this section we present the solution to the problem (3.2), assuming that $\theta_i = 1$ and $\theta_j = \theta \geq 1$, for $i \neq j \in \{1, 2\}$. With this assumption, we may derive analytical expressions for the relevant quantities. Recall that $c > 0$ is such that $\Pi_1(c) - \Pi_2(c) + \gamma R = 0$; moreover, we let $d = \Pi_1(c) = \Pi_2(c) - \gamma R$.

To solve the problem (3.2), we need to use the smooth pasting conditions in order to find the unknown terms of (3.4), and its domain. Therefore we need to propose a continuation region. In fact, depending on the sign of $\Pi_1 - \Pi_2 + \gamma R$, the geometry of the problem is different and, consequently, the continuation region is also distinct.

In case of the increasing risk, we expect that the continuation region is of the form $\mathcal{C} = \{x > 0: x < \delta\}$, with $\delta \geq c$, as in that case one should only invest in the more risky production process when the demand is high (and higher than c , because for $x < c$, $\Pi_2(x) - \gamma R - \Pi_1(x) < 0$). But if the risk decreases, then we expect the continuation region to be $\mathcal{C} = \{x > 0: x > \zeta\}$, with $\zeta \leq c$, since in that case the replacement should be undertaken when the levels of demand are low. Therefore we need to study the two cases separately, as we present in the next sections.

4.1. INCREASING RISK

Here, we assume that the profit functions Π_1 and Π_2 are given by:

$$(4.1) \quad \Pi_1(x) = a_1x - b_1 \quad \text{and} \quad \Pi_2(x) = a_2x^\theta - b_2,$$

with $\theta > 1$, and

$$(4.2) \quad b_1 \leq b_2 + \gamma R.$$

Note that this inequality may be interpreted as follows: the fixed cost when using Π_1 must be lower than or equal to the sum of the investment rate cost plus the fixed cost of using Π_2 . If this condition does not hold, then replacement would be optimal right away (i.e., the optimal time would be zero).

Proposition 4.1. *Let Π_i , with $i = 1, 2$ be given by (4.1). Then, the solution of (3.2) is as follows:*

$$(4.3) \quad V(x) = \begin{cases} Bx^{\beta_2} + \frac{a_1}{\gamma - \mu}x - \frac{a_2}{\gamma - \mu\theta}x^\theta - \frac{b - \gamma R}{\gamma}, & x < \delta, \\ 0, & x \geq \delta, \end{cases}$$

where B is given by

$$(4.4) \quad B = \left(\frac{a_2 \delta^\theta}{\gamma - \mu\theta} - \frac{a_1 \delta}{\gamma - \mu} + \frac{b - \gamma R}{\gamma} \right) \delta^{-\beta_2} \geq 0.$$

Additionally, δ is the unique positive solution to

$$(4.5) \quad f(x) := \frac{a_2(\beta_2 - \theta)}{\gamma - \mu_\theta} x^\theta - \frac{a_1(\beta_2 - 1)}{\gamma - \mu} x + \beta_2 \frac{b - \gamma R}{\gamma} = 0,$$

and verifies $\delta \geq c$. The result remains true when $\theta = 1$, $a_1 < a_2$ and $b_1 < b_2 + \gamma R$.

Taking into account the explanations provided in Section 3, it follows that

$$\mathcal{V}(x) = \begin{cases} Bx^{\beta_2} + \frac{a_1}{\gamma - \mu} x - \frac{b_1}{\gamma}, & x < \delta, \\ \frac{a_2}{\gamma - \mu_\theta} x^\theta - \frac{b_2 + \gamma R}{\gamma}, & x \geq \delta, \end{cases}$$

which means that for large levels of demand ($x > \delta$) it is always optimal to switch from the actual production process to the new one. This reinforces the idea that this type of strategy may be useful in markets that are in expansion. While the terms $\frac{a_1}{\gamma - \mu} - \frac{b_1}{\gamma}$ and $\frac{a_2}{\gamma - \mu_\theta} - \frac{b_2}{\gamma}$ represent the net present value associated to the first and second production process, respectively, the term Bx^{β_2} gives the value associated with the replacement option when the current value of the demand is x .

Corollary 4.1. *If $\tilde{b} \equiv b - \gamma R = 0$, then the replacement threshold δ can be explicitly given by:*

$$\delta_0 \equiv \delta \Big|_{\tilde{b}=0} = \theta^{-1} \sqrt[\theta]{\frac{a_1}{a_2} \frac{\beta_2 - 1}{\gamma - \mu} \frac{\gamma - \mu_\theta}{\beta_2 - \theta}} = \theta^{-1} \sqrt[\theta]{\frac{a_1}{a_2} \frac{\beta_1 - \theta}{\beta_1 - 1}}.$$

If $\theta = 1$, $a_1 < a_2$ and $b_1 < b_2 + \gamma R$, then δ can be explicitly given by:

$$\delta \Big|_{\theta=1} = \frac{b - \gamma R}{a} \left(\frac{\gamma - \mu}{\gamma} \frac{\beta_2}{\beta_2 - 1} \right) = \frac{b - \gamma R}{a} \left(1 - \frac{1}{\beta_1} \right) \geq \frac{b - \gamma R}{a} = c.$$

For future reference, one can note that δ_0 is a lower bound to δ since the function $\tilde{b} \rightarrow \delta(\tilde{b})$ is decreasing and consequently $\delta_0 \leq \delta$. Indeed, in light of the calculations presented in the proof of Lemma A.1, we get

$$\frac{\partial \delta}{\partial \tilde{b}}(\tilde{b}) = -\frac{\beta_2}{\gamma} f'(\delta) < 0.$$

4.2. DECREASING RISK

Consider now the case:

$$(4.6) \quad \Pi_1(x) = a_1 x^\theta - b_1 \quad \text{and} \quad \Pi_2(x) = a_2 x - b_2,$$

with $\theta > 1$, and

$$b_1 \geq b_2 + \gamma R.$$

Similarly to the previous situation, the interpretation of this condition is also clear. In order to have a non-trivial problem, we need to impose that the fixed cost associated with Π_1 is larger than the investment cost rate plus the fixed cost of Π_2 . Otherwise, replacement would never be optimal and we would have the optimal time equal to ∞ .

Proposition 4.2. *The value function defined by (3.2) is given by:*

$$(4.7) \quad V(x) = \begin{cases} 0, & x < \zeta, \\ Ax^{\beta_1} + \frac{a_1}{\gamma - \mu\theta} x^\theta - \frac{a_2}{\gamma - \mu} x - \frac{b - \gamma R}{\gamma}, & x \geq \zeta, \end{cases}$$

where A is given by

$$(4.8) \quad A = \left(\frac{a_2 \zeta}{\gamma - \mu} - \frac{a_1 \zeta^\theta}{\gamma - \mu\theta} + \frac{b - \gamma R}{\gamma} \right) \zeta^{-\beta_1} \geq 0$$

and ζ is the unique positive solution to

$$(4.9) \quad g(x) := \frac{a_1(\theta - \beta_1)}{\gamma - \mu\theta} x^\theta - \frac{a_2(1 - \beta_1)}{\gamma - \mu} x + \beta_1 \frac{b - \gamma R}{\gamma} = 0,$$

and verifies $\zeta \leq c$. The result remains true when $\theta = 1$, $a_1 > a_2$ and $b_1 > b_2 + \gamma R$.

In this case, the value function \mathcal{V} is given by

$$\mathcal{V}(x) = \begin{cases} \frac{a_2}{\gamma - \mu} x - \frac{b_2 + \gamma R}{\gamma}, & x < \zeta, \\ Ax^{\beta_1} + \frac{a_1}{\gamma - \mu\theta} x^\theta - \frac{b_1}{\gamma}, & x \geq \zeta, \end{cases}$$

and, consequently, it is always optimal to reduce the risk associated with the production process when the demand is sufficiently small ($x < \zeta$). This strategy may be very useful in declining markets, since it allows the firm to protect itself against the possibility of having large losses. The term Ax^{β_1} represents the value of the replacement option when the current value is $x > \zeta$; otherwise is zero.

Corollary 4.2. *If $\tilde{b} \equiv b - \gamma R = 0$, then, the replacement threshold ζ can be given by:*

$$(4.10) \quad \zeta_0 \equiv \zeta \Big|_{\tilde{b}=0} = \theta^{-1} \sqrt[\theta]{\frac{a_2}{a_1} \frac{1 - \beta_1}{\gamma - \mu} \frac{\gamma - \mu\theta}{\theta - \beta_1}} = \theta^{-1} \sqrt[\theta]{\frac{a_2}{a_1} \frac{\beta_2 - \theta}{\beta_2 - 1}}.$$

If $\theta = 1$, $a_1 > a_2$ and $b_1 > b_2 + \gamma R$, then, the replacement threshold ζ can be given by:

$$(4.11) \quad \zeta \Big|_{\theta=1} = \frac{b - \gamma R}{a} \left(\frac{\gamma - \mu}{\gamma} \frac{\beta_1}{\beta_1 - 1} \right) = \frac{b - \gamma R}{a} \left(1 - \frac{1}{\beta_2} \right) \leq \frac{b - \gamma R}{a} = c.$$

For future reference, we note that the function $\tilde{b} \rightarrow \zeta(\tilde{b})$ admits the derivative

$$\frac{\partial \zeta}{\partial \tilde{b}}(\tilde{b}) = -\frac{\beta_1}{\gamma} g'(\zeta) > 0,$$

which means that $\zeta \geq \zeta_0$.

5. COMPARATIVE STATICS

In this section, we assess the impact of changing the demand parameters μ and σ on the decision strategy. We expect that this behavior depends on whether the replacement leads to higher or lower risks. We also analyze the effect of increasing /decreasing, even more, the risk. This will be analyzed by studying the movement of the respective threshold when a_i is replaced by $a_i + \Delta$ and $b_i \equiv b_i(a_i) = a_i c - d$ is replaced by $b_i(a_i + \Delta)$, where i is such that $\Pi_i(x; a_i) = a_i x - b_i(a_i)$. This is illustrated in Figure 2.

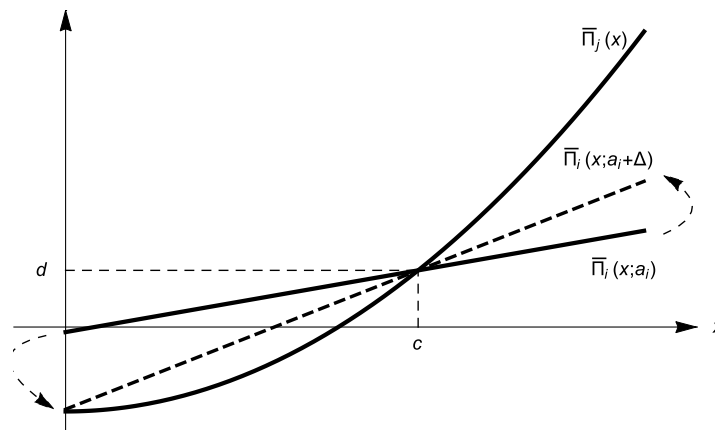


Figure 2: Representation of the functions $\bar{\Pi}_j(x) = a_j x^\theta - \bar{b}_j$ and the function $\bar{\Pi}_i(x; a) = a x^1 - \bar{b}_i(a)$, when $a = a_i$ and $a = a_i + \Delta$, $b_s(a)$, with $s = i, j$ and $i \neq j \in \{1, 2\}$, verifies $\bar{b}_s(a) = b_1(a)$ if $s = 1$ or $\bar{b}_s(a) = b_2(a) + \gamma R$ if $s = 2$.

In Proposition 5.1 we show that when the market becomes more uncertain, the firm waits longer until makes the decision of adjusting the production process. This is coherent with the classical Theory of Real Options, which postulates that more uncertainty postpones decisions. Furthermore, when the market becomes more attractive, i.e., the trend associated with the demand process increases, the decision of replacing the production process reacts in two ways: if the firm intends to increase the risk then it anticipates the decision, otherwise, it postpones the decision.

Proposition 5.1. *Let δ and ζ be implicitly defined by Equations (4.5) and (4.9). Then, the functions $(\mu, \sigma^2) \rightarrow \delta(\mu, \sigma^2)$ and $(\mu, \sigma^2) \rightarrow \zeta(\mu, \sigma^2)$ are such that*

$$\frac{\partial \delta}{\partial \mu}(\mu, \sigma) \leq 0 \quad \text{and} \quad \frac{\partial \zeta}{\partial \mu}(\mu, \sigma) \leq 0,$$

$$\frac{\partial \delta}{\partial \sigma}(\mu, \sigma) \geq 0 \quad \text{and} \quad \frac{\partial \zeta}{\partial \sigma}(\mu, \sigma) \leq 0.$$

First of all, we materialize the situation described in Figure 2 by setting that one of the following situations happen: (a) $i = 1$ and $j = 2$ or (b) $j = 1$ and $i = 2$. In the situation (a), changing a_i to $a_i + \Delta$ makes the scenario of adjusting the production process less risky than the original one. Consequently, when we decrease the slope of Π_1 , the replacement is

even riskier. In the case (b) by changing a_i to $a_i + \Delta$, the second production process becomes a bit riskier, and, consequently, such adjustment would be more contained in terms of gains and losses. Therefore, we can say that all the process of adjustment comes riskier.

We prove that for $\theta > 1$, the riskier the replacement process the later is made the decision of replacement. Note that in the case $\theta = 1$ (i.e., both Π_1 and Π_2 are linear functions), changing the risk does not have any impact on the thresholds, as in this case both δ and ϵ depend on a_1, a_2, b_1 and b_2 through c , which we assume to be constant.

Proposition 5.2. *Let δ and ζ be implicitly defined by Equations (4.5) and (4.9). Then the functions $(a_1, b_1) \rightarrow \delta(a_1, b_1)$ and $(a_2, b_2) \rightarrow \zeta(a_2, b_2)$ are such that*

$$\begin{aligned} \frac{\partial \delta}{\partial a_1}(a_1, a_1c - d; \theta) < 0, \quad \text{and} \quad \frac{\partial \zeta}{\partial a_2}(a_2, a_2c - d; \theta) < 0 \quad \text{for all } \theta > 1, \\ \frac{\partial \delta}{\partial a_1}(a_1, a_1c - d; \theta=1) = 0, \quad \text{and} \quad \frac{\partial \zeta}{\partial a_2}(a_2, a_2c - d; \theta=1) = 0. \end{aligned}$$

6. THE EFFECT OF THE EXIT OPTION

In this section we discuss how the abandonment option may influence the replacement decision. We denote by α the exit threshold, and thus, once the firm invests in the second production process, the firm stays active as long as the demand is above α ; then it abandons the market. To avoid trivial problems we assume that

$$(6.1) \quad b_2 > \gamma S,$$

which means that the abandonment problem is not trivial, in the sense that the time to abandon is finite, as the fixed cost (in the second production process) is larger than the exit rate cost.

For future reference, assuming that Π_2 is such that $\Pi_2(x) = a_2x^{\theta_2} - b_2$, with $\theta_2 \geq 1$, then

$$\tilde{W}(x) = \begin{cases} 0, & x \leq \alpha, \\ \tilde{A}x^{\beta_1} + \frac{a_2}{\gamma - \mu\theta_2}x^{\theta_2} - \frac{b_2 - \gamma S}{\gamma}, & x > \alpha, \end{cases}$$

where

$$(6.2) \quad \tilde{A} = -\frac{1}{\beta_1} \frac{a_2}{\gamma - \mu\theta_2} \alpha^{\theta_2 - \beta_1} > 0 \quad \text{and} \quad \alpha = \sqrt[\theta_2]{\frac{b_2 - \gamma S}{a_2} \left(1 - \frac{\theta_2}{\beta_2}\right)}.$$

These results follow in light of the Propositions 4.1 and 4.2 presented in the previous section. Additionally, the firm postpones the exit decision when either the uncertainty or the drift of the demand process increase. One can obtain such conclusions noticing that the function $(\mu, \sigma) \rightarrow \alpha(\mu, \sigma)$ verifies

$$\frac{\partial \alpha}{\partial \eta}(\mu, \sigma) = \frac{\alpha^{1-\theta_2}}{\beta_2^2} \frac{b_2 - \gamma S}{a_2} \frac{\partial \beta_2}{\partial \eta} < 0, \quad \text{with } \eta = \mu, \sigma.$$

Next we analyze separately the two cases: increasing and decreasing risk.

6.1. Increasing risk

In this section we consider the framework described in Section 4.1. In addition to the conditions (4.2) and (6.1), we also assume that

$$b_1 \leq \gamma S + \gamma R,$$

which means that the replacement followed by the abandonment is more costly than the fixed cost in the less risky production process, and therefore the time to invest is strictly positive.

The optimal strategy is depicted in Figure 3, and should be interpreted as follows: the firm stays in the first production process as long as the demand is below $\tilde{\delta}$. Then, as soon as it reaches this value, the firm replaces the production process, investing in the risky one. If the demand decreases below α , the firm exits the market.

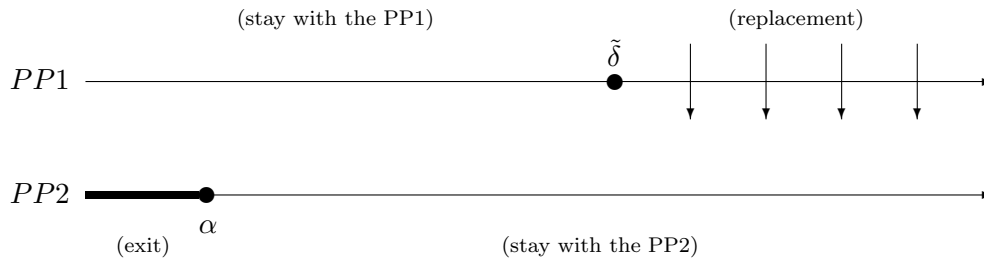


Figure 3: Replacement and abandonment strategy, when investing in the risky market.

Note that in this case the firm will stay in production after replacement for a strictly positive time, as $\tilde{\delta} > \alpha$. Thus, the value function is such that

$$(6.3) \quad \mathcal{W}(x) = \begin{cases} \tilde{B}x^{\beta_2} + \frac{a_1}{\gamma - \mu}x - \frac{b_1}{\gamma}, & x < \tilde{\delta}, \\ \tilde{A}x^{\beta_1} + \frac{a_2}{\gamma - \mu\theta}x^\theta - \frac{b_2}{\gamma} - R, & x \geq \tilde{\delta}, \end{cases}$$

where \tilde{A} is as in Equation (6.2), when we assume that $\theta_2 = \theta$, and \tilde{B} is given by

$$(6.4) \quad \tilde{B} = \left(\tilde{A}\tilde{\delta}^{\beta_1} + \frac{a_2\tilde{\delta}^\theta}{\gamma - \mu\theta} - \frac{a_1\tilde{\delta}}{\gamma - \mu} + \frac{b}{\gamma} - R \right) \tilde{\delta}^{-\beta_2}.$$

Additionally, $\tilde{\delta}$ satisfies the following equation

$$(6.5) \quad h(x) := \tilde{A}(\beta_2 - \beta_1)x^{\beta_1} + \frac{a_2(\beta_2 - \theta)}{\gamma - \mu\theta}x^\theta - \frac{a_1(\beta_2 - 1)}{\gamma - \mu}x + \beta_2 \frac{b - \gamma R}{\gamma} = 0.$$

As in Section 4, the terms $\frac{a_1}{\gamma - \mu}x - \frac{b_1}{\gamma}$ and $\frac{a_2}{\gamma - \mu\theta}x^\theta - \frac{b_2}{\gamma}$ represent the net present value associated with the first and second production processes, respectively. Additionally, the terms $\tilde{B}x^{\beta_2}$ and $\tilde{A}x^{\beta_1}$ represent, respectively, the value added by the replacement and exit options when the demand is x .

Proposition 6.1. *Let Π_i , with $i = 1, 2$ be given by (4.1) and \tilde{A} and α be defined as in Equation (6.2) by setting that $\theta_2 = \theta$. Then, the solution of (3.5), \mathcal{W} , is given by (6.3), with $\tilde{B} > 0$ given by (6.4). Additionally, $\tilde{\delta}$ is the unique positive solution to the Equation (6.5) satisfying $\tilde{\delta} > \alpha$.*

In the next proposition we discuss the influence of the exit option in the investment threshold. As expected, in the case we invest in a more risky production process, the decision is anticipated in case we still have the option to abandon the market. The proof of next proposition is trivial since $\tilde{A}(\beta_2 - \beta_1) x^{\beta_1} > 0$.

Proposition 6.2. *Let δ be the unique positive solution to Equation (4.5) and $\tilde{\delta}$ is the unique solution of Equation (6.5) such that $\tilde{\delta} > \alpha$. Then, $\tilde{\delta} < \delta$.*

Additionally, we can say that, as it holds when there is no option to abandon the market, a risky scenario, in the sense that a_1 is replaced by $a_1 - \Delta$ and $b_i \equiv b_i(a_i) = a_i c - d$ is replaced by $b_i(a_i - \Delta)$, postpones the replacement decision, when compared with the initial situation. The proof of this result follows in light of the proof of Proposition 5.2.

Proposition 6.3. *Let $\tilde{\delta}$ be implicitly defined by Equation (6.5). Then, the function $(a_1, b_1) \rightarrow \tilde{\delta}(a_1, b_1)$ is such that*

$$\frac{\partial \tilde{\delta}}{\partial a_1}(a_1, a_1 c - d) < 0.$$

The following table presents a numerical example which illustrates that although both replacement thresholds (δ , without the abandonment option, and $\tilde{\delta}$, with the abandonment option) increase with risk, the pace is not the same: $\tilde{\delta}$ increases faster with increasing risk (here measured by Δ) than δ .

Table 1: Thresholds δ and $\tilde{\delta}$ considering the parameters: $\mu = 0.001$, $\sigma^2 = 0.005$, $\gamma = 0.01$, $a_1 = 1$, $b_1 = 1$, $a_2 = 1$, $b_2 = 10$, $\theta = 2$, $R = 10$, $S = 110$.

Δ	$\tilde{\delta}(a_1 - \Delta)$	$\delta(a_1 - \Delta)$	$\delta(a_1 - \Delta) - \tilde{\delta}(a_1 - \Delta)$
0	5.046	5.171	0.125
0.1	5.194	5.311	0.117
0.2	5.342	5.451	0.109
0.3	5.488	5.591	0.103

6.2. Decreasing risk

In this section we consider the set up introduced in Section 4.2. From conditions $b_1 \geq b_2 + \gamma R$ and $b_2 > \gamma S$, trivially follows that:

$$b_1 \geq \gamma R + \gamma S.$$

This condition means that replacement occurs in finite time, as the fixed cost rate, before replacement, is larger than the total cost of replacement and abandonment.

We find two different strategies according to the value of the replacement cost. On the one hand, when R is sufficiently large, meaning that $R > R^*$, where

$$(6.6) \quad R^* \equiv \frac{1}{\beta_1} \left(a_1 \frac{\theta - \beta_1}{\gamma - \mu\theta} \alpha^\theta + \beta_1 \frac{b - \gamma S}{\gamma} \right),$$

the optimal strategy is depicted in Figure 4. In this case the value function takes the form

$$(6.7) \quad \mathcal{W}(x) = \begin{cases} -R - S, & x \leq \tilde{\zeta}, \\ \tilde{A}_1 x^{\beta_1} + \frac{a_1}{\gamma - \mu\theta} x^\theta - \frac{b_1}{\gamma}, & x > \tilde{\zeta}, \end{cases}$$

where

$$(6.8) \quad \tilde{A}_1 = -\frac{1}{\beta_1} \frac{a_1}{\gamma - \mu} \tilde{\zeta}^{\theta - \beta_1} \quad \text{and} \quad \tilde{\zeta} = \sqrt[\theta]{\frac{b_1 - \gamma(R + S)}{a_1} \left(1 - \frac{1}{\beta_2} \right)}.$$

Here, $\tilde{A}_1 x^{-\beta_1}$ represents the value of the abandonment option. Therefore, the firm produces using the first production process for large values of the demand, as long as they are above $\tilde{\zeta}$. Once the demand hits $\tilde{\zeta}$, it decides to abandon the market, paying a sunk cost equal to $R + S$. In this case, the firm does not actually produce with the second production process, as the time elapsed between replacement and abandonment is zero.



Figure 4: Abandonment strategy, when investing in the less risky market.

On the other hand, when $R < R^*$, the firm decides either to replace its production process by a second one when the demand reaches any level in $]\alpha, \zeta]$, or to abandon the market when the demand is smaller than or equal to the level α . The optimal strategy, in this case, is depicted in Figure 5. Furthermore, the value function is given by

$$(6.9) \quad \mathcal{W}(x) = \begin{cases} -R - S, & x \leq \alpha, \\ \tilde{A} x^{\beta_1} + \frac{a_2}{\gamma - \mu} x - \frac{b_2}{\gamma} - R, & \alpha < x \leq \zeta, \\ \tilde{A}_2 x^{\beta_1} + \frac{a_1}{\gamma - \mu\theta} x^\theta - \frac{b_1}{\gamma}, & x > \zeta, \end{cases}$$

where \tilde{A} and α are defined in Equation (6.2) by setting that $\theta_2 = 1$, and

$$(6.10) \quad \tilde{A}_2 - \tilde{A} = \frac{1}{\beta_1} \left(\frac{a_2}{\gamma - \mu} \zeta - \frac{a_1 \theta}{\gamma - \mu\theta} \zeta^\theta \right) \zeta^{-\beta_1}$$

and ζ is the unique solution to the equation (4.9).

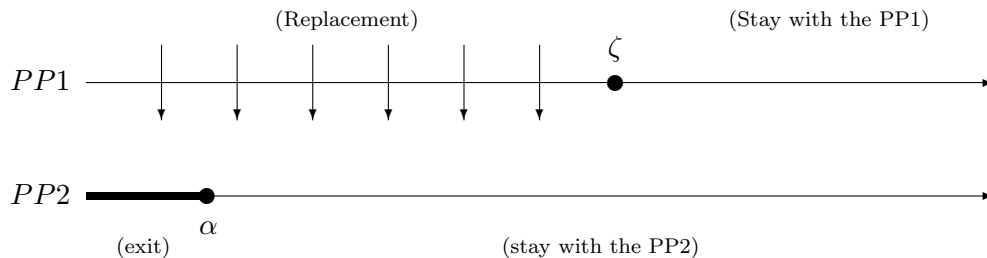


Figure 5: Replacement and abandonment strategy, when investing in the less risky market.

We start by noting that $\tilde{A}x^{\beta_1}$ represents the value of the abandonment option, while $\tilde{A}_2x^{\beta_1}$ represents the value of the replacement option when the demand is x . This representation is coherent with the classical theory since we are assuming that the exit option is only available after the replacement. Therefore in each moment until the scrapping, it is only possible to make one decision.

Proposition 6.4. *Let Π_i , with $i = 1, 2$ be given by (4.6) and \tilde{A} and α be defined as in Equation (6.2) by setting that $\theta_2 = 1$. Then, the solution of (3.5) is as follows:*

- *When $R \geq R^*$, the value function, \mathcal{W} , is given by (6.7), and $\tilde{A}_1 > 0$ given by (6.8);*
- *When $R < R^*$, the value function, \mathcal{W} , is given by (6.9), and $\tilde{A}_2 > 0$ given by (6.10). Additionally, ζ is the unique positive solution to Equation (4.9) satisfying $\zeta > \alpha$.*

Finally, we study the impact of changing the drift and/or the volatility in the parameter R^* . As we show in the next proposition, the situation depicted in Figure 5 is more likely to occur than the situation depicted in Figure 4 with increasing the drift or the volatility.

Proposition 6.5. *Consider $R^*(\mu, \sigma) \equiv R^*$, with R^* defined as in (6.6). Then the functions $R^*(\cdot, \sigma)$ and $R^*(\mu, \cdot)$ are both decreasing.*

7. CONCLUSION

This paper considers the problem of a producing firm that has the option to replace its current production process by a riskier /less risky one. The concept of risk here considered relies on the structure of the running payoff function, as described before.

Our main result is that the time until the decision of replacement increases when the risk associated with the replacement option increases. Additionally, if the firm evaluates the replacement option taking into account the abandonment option, then its decision regarding replacement is anticipated. But not only the timing changes, but also there is a clear change in the structure of the values of the economic indicator that lead to the decision. In fact, if, on the one hand, when we increase the level of risk of the alternative production process the replacement is optimal for large levels of the economic indicator, on the other hand, if we decrease, the replacement is optimal for small levels of the economic indicator.

A. APPENDIX – Proofs

Unless otherwise stated, we assume, without loss of generality, that $R = 0$.

A.1. Section 4

Before we prove Propositions 4.1 and 4.2, we state an auxiliary lemma, which will simplify the proof of this proposition.

Lemma A.1. Equation (4.5) (resp., (4.9)) has a unique solution, δ (resp., ζ).

Proof: To prove that δ is the unique root of Equation (4.5), we calculate f' :

$$(A.1) \quad f'(x) = \frac{a_2 \theta (\beta_2 - \theta)}{\gamma - \mu \theta} x^{\theta-1} - \frac{a_1 (\beta_2 - 1)}{\gamma - \mu}.$$

Then, $f'(x) \geq 0$, for $x \in [x_1, \infty[$, where x_1 is the unique zero of $f'(x)$, given by

$$x_1 = \left(\frac{a_1}{\theta a_2} \frac{\beta_2 - 1}{\gamma - \mu} \frac{\gamma - \mu \theta}{\beta_2 - \theta} \right)^{\frac{1}{\theta-1}}.$$

Furthermore, as

$$f(0) = \frac{\beta_2 b}{\gamma} \leq 0 \quad \text{and} \quad \lim_{x \rightarrow \infty} f(x) = \infty,$$

we conclude that there is a unique positive solution to the equation $f(x) = 0$, denoted by δ .

To prove that there is a unique positive solution ζ to Equation (4.9), we can follow the same strategy. For future reference, we note that $g'(x) \geq 0$ for $x \in [x_2, \infty[$, where x_2 is the unique zero of $g'(x)$. Furthermore, as

$$g(0) = \frac{\beta_1 b}{\gamma} \leq 0 \quad \text{and} \quad \lim_{x \rightarrow \infty} g(x) = \infty,$$

we conclude that there is a unique positive solution to the equation $g(x) = 0$, denoted by ζ . \square

Proof of Proposition 4.1: To find the parameter B and the threshold δ we use the smooth pasting conditions

$$\begin{aligned} \frac{a_1}{\gamma - \mu} \delta - \frac{a_2}{\gamma - \mu \theta} \delta^\theta + B \delta^{\beta_2} - \frac{b}{\gamma} &= 0, \\ \frac{a_1}{\gamma - \mu} - \frac{a_2}{\gamma - \mu \theta} \theta \delta^{\theta-1} + \beta_2 B \delta^{\beta_2-1} &= 0. \end{aligned}$$

Consequently, we obtain B given by (4.4) and δ as a solution to Equation (4.5). Additionally, Lemma A.1 states that δ is the unique solution to Equation (4.5).

To prove that the function V defined by (4.3) satisfies the HJB equation, we need to prove the following relationships:

$$(A.2) \quad \gamma V(x) - \mu x V'(x) - \frac{\sigma^2}{2} x^2 V''(x) - \Pi_1(x) + \Pi_2(x) \geq 0, \quad \text{for all } x \geq \delta,$$

$$(A.3) \quad V(x) \geq 0, \quad \text{for all } x \leq \delta.$$

First, we note that the inequality in (A.2) may be written as

$$(A.4) \quad f_1(x) := \Pi_1(x) - \Pi_2(x) \leq 0, \quad \text{for all } x \geq \delta,$$

as for $x \geq \delta$, $V = 0$. Since $f_1'(x) = a_1 - a_2 \theta x^{\theta-1}$, then f_1 is increasing for $x < (\frac{a_2}{a_1 \theta})^{\frac{1}{\theta-1}}$ and decreasing for $x > (\frac{a_2}{a_1 \theta})^{\frac{1}{\theta-1}}$. Taking into account that

$$f_1(0) = -b \geq 0 \quad \text{and} \quad \lim_{x \rightarrow +\infty} f_1(x) = -\infty,$$

then (A.4) holds true if and only if

$$(A.5) \quad \Pi_1(\delta) - \Pi_2(\delta) \leq 0.$$

To prove this, we note that

$$\Pi_1(\delta) - \Pi_2(\delta) = -\frac{1}{2} \sigma^2 (\delta)^2 V''(\delta),$$

where the equality follows because $\gamma V(\delta) - \mu \delta V'(\delta) - \frac{\sigma^2}{2} \delta^2 V''(\delta) - \Pi_1(\delta) + \Pi_2(\delta) = 0$ and the smooth pasting conditions. Additionally, we can calculate

$$V''(\delta) = \beta_2 (\beta_2 - 1) B \delta^{\beta_2 - 2} - \frac{a_2 \theta (\theta - 1)}{\gamma - \mu \theta},$$

which, combined with the smooth pasting conditions, allow us to obtain

$$\begin{aligned} -\frac{1}{2} \sigma^2 \delta^2 V''(\delta) &= -\frac{1}{2} \sigma^2 \left[\frac{a_2 (\beta_2 - \theta) \theta}{\gamma - \mu \theta} \delta^\theta - \frac{a_1 (\beta_2 - 1)}{\gamma - \mu} \delta \right] \\ &< -\frac{1}{2} \sigma^2 \left[\frac{a_2 (\beta_2 - \theta)}{\gamma - \mu \theta} \delta^\theta - \frac{a_1 (\beta_2 - 1)}{\gamma - \mu} \delta \right] = \frac{1}{2} \sigma^2 \frac{\beta_2 b}{\gamma} \leq 0. \end{aligned}$$

This proves (A.5) and allow us to conclude that

$$\delta \geq c.$$

Finally, to prove the inequality in (A.3), we notice that, in light of the relationship $f(\delta) = 0$, the parameter B can be written as

$$B = -\frac{1}{\beta_2} \left[\frac{a_1}{\gamma - \mu} \delta - \frac{a_2 \theta}{\gamma - \mu \theta} \delta^\theta \right] \delta^{-\beta_2}.$$

Now, calculating the derivative of the function $f_2(x) := -\frac{1}{\beta_2} \left(\frac{a_1}{\gamma - \mu} x - \frac{a_2 \theta}{\gamma - \mu \theta} x^\theta \right)$, we obtain $f_2'(x) = -\frac{1}{\beta_2} \left(\frac{a_1}{\gamma - \mu} - \frac{a_2 \theta^2}{\gamma - \mu \theta} x^{\theta-1} \right)$. Consequently, the function f_2 is increasing for $x \in]\delta_1, \infty[$, where δ_1 is the unique positive root of f_2' . Combining this with the fact that $f_2(\delta_0) = \beta_2 \frac{a_1 \delta_0}{\gamma - \mu} \frac{\theta - 1}{\beta_2 - \theta} > 0$, then $\delta_1 < \delta_0 \leq \delta$, and, consequently, $B \geq 0$.

Taking into account Equation (4.3) and the smooth pasting conditions, we have that

$$(A.6) \quad V(0) = -\frac{b}{\gamma} > 0 \quad \text{and} \quad V(\delta) = V'(\delta) = 0.$$

Additionally,

$$V'(x) = \beta_2 Bx^{\beta_2-1} + \frac{a_1}{\gamma - \mu} - \frac{a_2 \theta}{\gamma - \mu} x^{\theta-1},$$

and, consequently, $V'(0) = \frac{a_1}{\gamma - \mu} > 0$. Since

$$V''(x) = \left(\beta_2(\beta_2 - 1) Bx^{\beta_2-\theta} - \frac{a_2 \theta(\theta - 1)}{\gamma - \mu \theta} \right) x^{\theta-2},$$

then V' is decreasing for all $x \in]0, \delta_2[$ and increasing for all $x \in]\delta_2, \infty[$, where δ_2 is the unique positive root of the equation $V''(x) = 0$. This means that one of two situations may happen: (1) $V'(x) > 0$ for all $x \in]0, \delta[$ or (2) $V'(x) > 0$ for all $x \in]0, \delta_2[$ and $V'(x) < 0$ for all $x \in]\delta_2, \delta[$. The situation (1) cannot happen in light of (A.6). Naturally, this implies that $V \geq 0$. \square

Proof of Proposition 4.2: In order to determine values for A and ζ , we use the smooth pasting conditions

$$\begin{aligned} \frac{a_1 \zeta^{-\theta}}{\gamma - \mu \theta} - \frac{a_2 \zeta}{\gamma - \mu} + A \zeta^{\beta_1} - \frac{b}{\gamma} &= 0, \\ \frac{a_1 \theta}{\gamma - \mu \theta} \zeta^{\theta-1} - \frac{a_2}{\gamma - \mu} + \beta_1 A \zeta^{\beta_1-1} &= 0, \end{aligned}$$

which allow us to obtain the parameter A , as defined in (4.8), and ζ as a solution to Equation (4.9). Additionally, Lemma A.1 states that ζ is the unique solution to Equation (4.9).

To prove that the function V defined by (4.7) satisfies the HJB equation, we need to prove the following relationships:

$$(A.7) \quad \gamma V(x) - \mu x V'(x) - \frac{\sigma^2}{2} x^2 V''(x) - \Pi_1(x) + \Pi_2(x) \geq 0, \quad \text{for all } x \leq \zeta,$$

$$(A.8) \quad V(x) \geq 0, \quad \text{for all } x \geq \zeta.$$

First, we note that the inequality in (A.7) can be written as

$$(A.9) \quad \Pi_1(x) - \Pi_2(x) \leq 0, \quad \text{for all } x \leq \zeta.$$

In fact, a similar argument to the one used to prove the inequality in (A.2) proves that the inequality in (A.9) is satisfied. Additionally, we get that

$$\zeta \leq c.$$

To prove the inequality in (A.8), we note that, in light of the relationship $g(\zeta) = 0$, the parameter A can be written as

$$(A.10) \quad A = -\frac{1}{\beta_1} \left(\frac{a_1 \theta}{\gamma - \mu \theta} \zeta^\theta - \frac{a_2}{\gamma - \mu} \zeta \right) \zeta^{-\beta_1}.$$

Additionally, $V(\zeta) = 0$ in light of the smooth pasting conditions. Taking into account Equation (A.10), we can calculate

$$\begin{aligned} V'(x) &= a_1\theta \frac{x^{\theta-1}}{\gamma - \mu_\theta} - \frac{a_2}{\gamma - \mu} + A_1\beta_1 x^{\beta_1-1} \\ &= \frac{a_1\theta}{\gamma - \mu_\theta} x^{\theta-1} - \frac{a_2}{\gamma - \mu} - \left(\frac{a_1\theta}{\gamma - \mu_\theta} \zeta^{\theta-1} - \frac{a_2}{\gamma - \mu} \right) \frac{x^{\beta-1}}{\zeta^{\beta_1-1}} \geq 0, \end{aligned}$$

where the last inequality follows from:

$$x \rightarrow \frac{a_1\theta}{\gamma - \mu_\theta} x^{\theta-1} - \frac{a_2}{\gamma - \mu} \text{ is an increasing function, and } \frac{x^{\beta-1}}{\zeta^{\beta_1-1}} \leq 1 \text{ for all } x \geq \zeta.$$

As a consequence, the inequality (A.8) holds true, because V is increasing. To finish the proof, we just need to verify that $A > 0$. Consider the function

$$(A.11) \quad g_1(x) := \frac{a_1\theta}{\gamma - \mu_\theta} x^\theta - \frac{a_2}{\gamma - \mu} x.$$

Taking into account that $g'_1(x) := \frac{a_1\theta^2}{\gamma - \mu_\theta} x^{\theta-1} - \frac{a_2}{\gamma - \mu}$, then g_1 is increasing in $]\zeta_1, \infty[$, where ζ_1 is the unique positive root of g'_1 . The results follow in light of the facts:

$$g(0) = 0, \quad g(\zeta_0) = \zeta_0 \frac{a_2}{\gamma - \mu} \frac{\beta_1(1 - \theta)}{\theta - \beta_1} > 0 \quad \text{and} \quad \zeta \geq \zeta_0. \quad \square$$

A.2. Section 5

Proof of Proposition 5.1: By using the Implicit Function Theorem, we obtain that

$$\frac{\partial \delta}{\partial \mu}(\mu) = -\frac{\partial f}{\partial \mu}(\delta; \mu) \left(\frac{\partial f}{\partial \delta} \right)^{-1}(\delta; \mu) \quad \text{and} \quad \frac{\partial \zeta}{\partial \mu}(\mu) = -\frac{\partial g}{\partial \mu}(\zeta; \mu) \left(\frac{\partial g}{\partial \zeta} \right)^{-1}(\zeta; \mu).$$

Taking into account Lemma A.1, we note that $\frac{\partial f}{\partial \delta}(\delta; \mu) > 0$ and $\frac{\partial g}{\partial \delta}(\delta; \mu) > 0$, and consequently, we just need to study the sign of $\frac{\partial f}{\partial \mu}(\delta; \mu)$ and $\frac{\partial g}{\partial \mu}(\delta; \mu)$. Taking into account the smooth pasting conditions we get, after some simplifications,

$$\begin{aligned} \frac{\partial f}{\partial \mu}(\delta; \mu) &= \frac{a_2\theta}{\gamma - \mu_\theta} \left(\frac{1}{\beta_2} \frac{\partial \beta_2}{\partial \mu} + \frac{\beta_2 - \theta}{\gamma - \mu_\theta} \right) \delta^\theta - \frac{a_1}{\gamma - \mu} \left(\frac{1}{\beta_2} \frac{\partial \beta_2}{\partial \mu} + \frac{\beta_2 - 1}{\gamma - \mu} \right) \delta := p_1(\delta; \theta), \\ \frac{\partial g}{\partial \mu}(\zeta; \mu) &= \frac{a_1\theta}{\gamma - \mu_\theta} \left(-\frac{1}{\beta_1} \frac{\partial \beta_1}{\partial \mu} + \frac{\theta - \beta_1}{\gamma - \mu_\theta} \right) \zeta^\theta + \frac{a_2}{\gamma - \mu} \left(\frac{1}{\beta_1} \frac{\partial \beta_1}{\partial \mu} - \frac{1 - \beta_1}{\gamma - \mu} \right) \zeta := p_2(\zeta; \theta). \end{aligned}$$

Assume for now that (i) $\theta = 1$, $a_1 < a_2$ and $b_1 < b_2$ and (ii) $\theta = 1$, $a_1 > a_2$ and $b_1 > b_2$. Then, we can calculate explicitly the following derivatives (see Corollaries 4.1 and 4.2):

$$(i) \quad \frac{\partial \delta}{\partial \mu}(\mu) = \frac{b}{a} \frac{1}{\gamma} \frac{\frac{\partial \beta_2}{\partial \mu}}{(\beta_2 - 1)^2} \leq 0 \quad \text{and} \quad (ii) \quad \frac{\partial \zeta}{\partial \mu}(\mu) = \frac{b}{a} \frac{1}{\gamma} \frac{\frac{\partial \beta_1}{\partial \mu}}{(\beta_1 - 1)^2} \leq 0.$$

Combining these derivatives with the expressions of $\frac{\partial f}{\partial \mu}(\delta; \mu)$ and $\frac{\partial g}{\partial \mu}(\zeta; \mu)$, it is easy to note that

$$(A.12) \quad \frac{1}{\beta_2} \frac{\partial \beta_2}{\partial \mu} + \frac{\beta_2 - 1}{\gamma - \mu} \geq 0 \quad \text{and} \quad \frac{1}{\beta_1} \frac{\partial \beta_1}{\partial \mu} - \frac{1 - \beta_1}{\gamma - \mu} \leq 0.$$

Indeed, the previous inequalities do not depend on a_1 neither on a_2 , and thus this means that the result is true for every $a_1, a_2 > 0$. Additionally, returning to the case $\theta > 1$, it is a matter of calculations to see that $\frac{\partial \delta_0}{\partial \mu}(\mu) \leq 0$. In fact, this implies that $0 \leq \frac{\partial f}{\partial \mu}(\delta_0; \mu)$, and, consequently,

$$\frac{1}{\beta_2} \frac{\partial \beta_2}{\partial \mu} + \frac{\beta_2 - \theta}{\gamma - \mu\theta} \geq 0.$$

Furthermore, noticing that $\frac{\theta - \beta_1}{\gamma - \mu\theta} = \frac{\theta - \beta_1}{-\frac{1}{2}\sigma^2(\theta - \beta_1)(\theta - \beta_2)} = \frac{2}{\sigma^2} \frac{1}{(\beta_2 - \theta)}$ and that

$$\frac{\partial}{\partial \theta} \left(\frac{\theta - \beta_1}{\gamma - \mu\theta} \right) = \frac{2}{\sigma^2} \frac{1}{(\beta_2 - \theta)^2} > 0,$$

we obtain:

$$-\frac{1}{\beta_1} \frac{\partial \beta_1}{\partial \mu} + \frac{(\theta - \beta_1)}{\gamma - \mu\theta} \geq -\frac{1}{\beta_1} \frac{\partial \beta_1}{\partial \mu} + \frac{(1 - \beta_1)}{\gamma - \mu} \geq 0.$$

Note now that the function $p_1(x; \theta)$ has two roots: $x = 0$ and $x = a^*$, where a^* is its unique positive root. Additionally, it is a matter of calculations to see that there is a unique b^* such that $\frac{\partial p_1}{\partial x}(b^*; \theta) = 0$ and that, in light of Equation (A.12), $\frac{\partial p_1}{\partial x}(0; \theta) < 0$. Therefore, $p_1(x; \theta)$ is increasing for all $x > b^*$, and decreasing for all $x < b^*$, and, consequently, $0 \leq p_1(\delta_0; \theta) \leq p_1(\delta, \theta)$, since $\frac{\partial \delta}{\partial b} < 0$. Finally, we can observe that $p_2(x, \theta) = 0$ if and only if $x = 0$ and $x = c^* > 0$, where

$$c^* = \theta^{-1} \sqrt{\frac{a_2}{a_1 \theta} \frac{\gamma - \mu\theta}{\gamma - \mu} \left(-\frac{1}{\beta_1} \frac{\partial \beta_1}{\partial \mu} + \frac{1 - \beta_1}{\gamma - \mu} \right)} \leq \zeta_0 \theta^{-1} \sqrt{\frac{-\frac{1}{\beta_1} \frac{\partial \beta_1}{\partial \mu} + \frac{1 - \beta_1}{\gamma - \mu}}{-\frac{1}{\beta_1} \frac{\partial \beta_1}{\partial \mu} + \frac{\theta - \beta_1}{\gamma - \mu\theta}}} \leq \zeta_0.$$

Furthermore, $\frac{\partial p_2}{\partial x}(x; \theta) < 0$ for all $x < d^*$ and $\frac{\partial p_2}{\partial x}(x; \theta) > 0$ for all $x > d^*$, where d^* is the unique root of the function $x \rightarrow \frac{\partial p_2}{\partial x}(x; \theta)$. Therefore, $p_2(x; \theta)$ is an increasing function in x , for a fixed θ , if $x \geq d^*$, with $c^* > d^*$. Combining this with the roots to the equation $p_2(x; \theta)$ we get that $\zeta > \zeta_0 \geq c^* > d^*$ and, consequently,

$$0 = p_2(c^*; \theta) \leq p(\zeta; \theta),$$

which concludes this part of the proof.

To finish the proof, we use the Implicit Function Theorem

$$\frac{\partial \delta}{\partial \sigma}(\sigma) = -\frac{\partial f}{\partial \sigma}(\delta; \sigma) \left(\frac{\partial f}{\partial \delta} \right)^{-1}(\delta; \sigma) \quad \text{and} \quad \frac{\partial \zeta}{\partial \sigma}(\sigma) = -\frac{\partial g}{\partial \sigma}(\zeta; \sigma) \left(\frac{\partial g}{\partial \zeta} \right)^{-1}(\zeta; \sigma).$$

From the previous considerations, one just need to discuss the signs of $\frac{\partial f}{\partial \sigma}(\delta; \sigma)$ and $\frac{\partial g}{\partial \sigma}(\zeta; \sigma)$. By using the smooth pasting conditions, one can prove that

$$\begin{aligned} \frac{\partial f}{\partial \sigma}(\delta; \sigma) &= \frac{a_2 \theta}{\gamma - \mu\theta} \left(\frac{1}{\beta_2} \frac{\partial \beta_2}{\partial \sigma} + \frac{(\beta_2 - \theta) \sigma (\theta - 1)}{\gamma - \mu\theta} \right) \delta^\theta - \frac{a_1}{\gamma - \mu} \left(\frac{1}{\beta_2} \frac{\partial \beta_2}{\partial \sigma} \right) \delta := q_1(\delta), \\ \frac{\partial g}{\partial \sigma}(\zeta; \sigma) &= \frac{a_1 \theta}{\gamma - \mu\theta} \left(-\frac{1}{\beta_1} \frac{\partial \beta_1}{\partial \sigma} + \frac{(\theta - \beta_1) \sigma (\theta - 1)}{\gamma - \mu\theta} \right) \zeta^\theta + \frac{a_2}{\gamma - \mu} \left(\frac{1}{\beta_1} \frac{\partial \beta_1}{\partial \sigma} \right) \zeta := q_2(\zeta). \end{aligned}$$

To show that $\frac{\partial f}{\partial \sigma} \leq 0$, we note that, since $\beta_2 \geq 0$ and $\frac{\partial \beta_2}{\partial \sigma} \leq 0$, then $-\frac{a_1}{\gamma - \mu} \frac{1}{\beta_2} \frac{\partial \beta_2}{\partial \sigma} \geq 0$. Assuming now that $b = 0$, then $\delta = \delta_0$, it is a matter of calculations to see that $\frac{\partial \delta_0}{\partial \sigma} \geq 0$.

Consequently, $0 \geq q(\delta_0)$, thus $\frac{1}{\beta_2} \frac{\partial \beta_2}{\partial \sigma} + \frac{(\beta_2 - \theta) \sigma (\theta - 1)}{\gamma - \mu \theta} \leq 0$. Trivial calculations allow us to conclude that $q_1(x)$ is decreasing for all $x > e^*$ and increasing for all $x < e^*$, where e^* is the unique positive root of the function $x \rightarrow q'(x)$. Since there is $x^* > 0$ such that $x = 0$ and $x = x^* > 0$ are the unique non-negative roots of the function $x \rightarrow q(x)$, it follows that $0 \geq q(\delta_0) \geq q(\delta)$.

Now, one can note that $q_2(x) = 0$ if and only if $x = 0$ and $x = m^* > 0$, where

$$m^* = \sqrt[\theta-1]{\frac{a_2}{a_1 \theta} \frac{\gamma - \mu \theta}{\gamma - \mu} \frac{-\frac{1}{\beta_1} \frac{\partial \beta_1}{\partial \sigma}}{-\frac{1}{\beta_1} \frac{\partial \beta_1}{\partial \sigma} + \frac{(\theta - \beta_1) \sigma (\theta - 1)}{\gamma - \mu \theta}}} < \sqrt[\theta-1]{\frac{a_2}{a_1 \theta} \frac{\gamma - \mu \theta}{\gamma - \mu}} < \zeta_0 < \zeta.$$

The first inequality follows because $\frac{\partial \beta_1}{\partial \sigma} > 0$. Moreover, calculating the derivative of q_2 , in order to x , we can conclude that $q(x)$ is increasing for $x \geq n^*$, where n^* is such that $q_2'(n^*) = 0$. Combining all these facts we have

$$0 = q_2(m^*) \leq q_2(\zeta),$$

which ends the proof. □

Proof of Proposition 5.2: We will focus our attention in the case $\theta > 1$. To prove Proposition 5.2 we note that

$$f(x; a_1) = \frac{a_2(\beta_2 - \theta)}{\gamma - \mu \theta} x^\theta - \frac{a_1(\beta_2 - 1)}{\gamma - \mu} - \frac{\beta_2(b_1(a_1) - b_2)}{\gamma},$$

then,

$$f(x; a_1 + \Delta) - f(x; a_1) = \Delta \left(\frac{\beta_2}{\gamma} c - \frac{\beta_2 - 1}{\gamma - \mu} x \right).$$

Therefore, $f(x; a_1 + \Delta) > f(x; a_1)$ for every $x < \tilde{y}$, where

$$\tilde{y} = c \frac{\beta_2}{\gamma} \frac{\gamma - \mu}{1 - \beta_2}.$$

Note that

$$\begin{aligned} f(\tilde{y}) &= a_2 c^\theta \left(\frac{\beta_2 - \theta}{\gamma - \mu \theta} \left(\frac{\beta_2}{\gamma} \frac{\gamma - \mu}{\beta_2 - 1} \right)^\theta - \frac{\beta_2}{\gamma} \right) \\ &= a_2 c^\theta \left(\frac{1}{\sigma^2/2(\theta - \beta_1)} \left(1 - \frac{1}{\beta_1} \right)^\theta + \frac{1}{\beta_1 \sigma^2/2} \right) \\ &= a_2 c^\theta \frac{1}{\sigma^2/2(\theta - \beta_1)} \left(\left(1 - \frac{1}{\beta_1} \right)^\theta - \left(1 - \frac{\theta}{\beta_1} \right) \right), \end{aligned}$$

where we have used the following relationships:

$$\frac{\beta_2}{\gamma} \frac{\gamma - \mu}{\beta_2 - 1} = \frac{\beta_1(\beta_2 - 1) + 1 - \beta_2}{\beta_1(\beta_2 - 1)} > 1 \quad \text{and} \quad \gamma - \mu \theta = -\frac{\sigma^2}{2} (\theta - \beta_1) (\theta - \beta_2).$$

To determine the sign of $f(\tilde{y})$, we define the function

$$(\theta; \beta_1, \sigma_2) \rightarrow n(\theta; \beta_1, \sigma_2) = \left(1 - \frac{1}{\beta_1} \right)^\theta - \left(1 - \frac{\theta}{\beta_1} \right).$$

Then, taking into account that

$$(A.13) \quad \frac{\partial n}{\partial \beta_1}(\theta; \beta_1, \sigma^2) = \frac{\theta}{\beta_1^2} \left(\left(1 - \frac{1}{\beta_1}\right)^{\theta-1} - 1 \right) > 0$$

and

$$(A.14) \quad \lim_{\beta_1 \rightarrow -\infty} n(\theta; \beta_1, \sigma^2) = 0,$$

it follows that $f(\tilde{y}) > 0$. Consequently, $f(x; a_1 + \Delta) - f(x; a_1) > 0$ for all $x < \tilde{y}$, which implies that $\delta(a_1 + \Delta, b(a_1 + \Delta)) < \delta(a_1, b(a_1))$. The case $\theta = 1$ follows in light of similar arguments, using the relation $n(1; \beta_1, \sigma^2) = 0$.

To finish the proof, we can apply the same type of arguments to the function g . In fact, $g(x; a_2 + \Delta) > g(x; a_2)$ for every $x < \tilde{x}$, where

$$\tilde{x} = -c \frac{\beta_1}{\gamma} \frac{\gamma - \mu}{1 - \beta}.$$

Taking into account that

$$g(\tilde{x}) = a_1 c^\theta \frac{1}{\sigma^2/2(\beta_2 - \theta)} \left(\left(1 - \frac{1}{\beta_2}\right)^\theta - \left(1 - \frac{\theta}{\beta_2}\right) \right),$$

by using similar arguments to the previous ones, we get that $g(\tilde{x}) > 0$. □

A.3. Section 6

Before we start the proofs, we note that the value function may be re-written as follows:

$$\mathcal{W}(x) = \sup_{\tau_1 \leq \tau_2 \in \mathcal{S}} \tilde{I}(\tau_1, \tau_2, x) - R - S,$$

where \tilde{I} is defined as in (2.2). Therefore, throughout this section we will use the following notation:

$$H(x) \equiv \sup_{\tau_1 \leq \tau_2 \in \mathcal{S}} \tilde{I}(\tau_1, \tau_2, x).$$

Additionally, we consider $R \geq 0$.

Lemma A.2. Equation (6.5) has a unique solution $\tilde{\delta}$, which satisfies $\tilde{\delta} > \alpha$, where α is defined as in Equation (6.2) by setting that $\theta_2 = \theta$.

Proof: To prove that $\tilde{\delta}$ is the unique root of Equation (6.5) satisfying $\tilde{\delta} > \alpha$, we calculate h'' :

$$h''(x) = \tilde{A}(\beta_2 - \beta_1) \beta_1(\beta_1 - 1) x^{\beta_1-2} + \frac{a_2 \theta (\theta - 1) (\beta_2 - \theta)}{\gamma - \mu_\theta} x^{\theta-2} > 0.$$

Taking into account that

$$\lim_{x \rightarrow 0^+} h(x) = \lim_{x \rightarrow +\infty} h(x) = +\infty,$$

the result follows in light of the calculations:

$$h(\alpha) = -\frac{a_1(\beta_2 - 1)}{\gamma - \mu} + \beta_2 \frac{b_1 - \gamma R - \gamma S}{\gamma} < 0,$$

where we have used the smooth pasting conditions (used to obtain \tilde{A} and α):

$$\tilde{A}\alpha^{\beta_1} + \frac{a_2}{\gamma - \mu_\theta} \alpha^\theta \frac{b_2 - \gamma E}{\gamma} = 0,$$

$$\tilde{A}\beta_1\alpha^{\beta_1} + \frac{a_2\theta}{\gamma - \mu_\theta} \alpha^\theta = 0. \quad \square$$

Proof of Proposition 6.1: The parameters \tilde{B} and $\tilde{\delta}$ may be obtained by using the smooth pasting conditions:

$$\tilde{B}\tilde{\delta}^{\beta_2} + \frac{a_1}{\gamma - \mu} \tilde{\delta} - \frac{b_1 - \gamma R - \gamma S}{\gamma} = \tilde{A}\tilde{\delta}_1^\beta + \frac{a_2}{\gamma - \mu_\theta} \tilde{\delta}^\theta - \frac{b_2 - \gamma S}{\gamma},$$

$$\tilde{B}\beta_2\tilde{\delta}^{\beta_2-1} + \frac{a_1}{\gamma - \mu} = \tilde{A}\beta_1\tilde{\delta}^{\beta_1-1} + \frac{a_2\theta}{\gamma - \mu_\theta} \tilde{\delta}^{\theta-1}.$$

Moreover, in light of Lemma A.2, $\tilde{\delta}$ is the unique positive solution Equation (6.5) satisfying the condition $\tilde{\delta} > \alpha$.

To prove that \mathcal{W} , where \mathcal{W} is defined by (6.7), is the solution to the optimal stopping problem (2.2), we need to verify that the function $H(x) = \mathcal{W}(x) + S + R$ satisfies the following inequalities:

$$(A.15) \quad \gamma H(x) - \mu x H'(x) - \frac{\sigma^2}{2} x^2 H''(x) - \left(\Pi_1(x) + \gamma R + \gamma S\right) \geq 0, \quad \text{for all } x \geq \tilde{\delta},$$

$$(A.16) \quad H(x) \geq \tilde{W}(x), \quad \text{for all } x \leq \tilde{\delta}.$$

First of all, we note that (A.15) can be written as

$$\Pi_1(x) - \Pi_2(x) + \gamma R \leq 0$$

because $H(x) = \tilde{W}(x)$ and

$$(A.17) \quad \gamma \tilde{W}(x) - \mu x \tilde{W}'(x) - \frac{\sigma^2}{2} x^2 \tilde{W}''(x) - \Pi_2(x) - \gamma S = 0$$

for $x \geq \tilde{\delta}$. Since the function $x \rightarrow \Pi_1(x) - \Pi_2(x) + \gamma R$ is increasing for $x < \left(\frac{a_1}{a_2\theta}\right)^{\frac{1}{\theta-1}}$ and decreasing for $x > \left(\frac{a_1}{a_2\theta}\right)^{\frac{1}{\theta-1}}$, we just need to prove that $\Pi_1(\tilde{\delta}) - \Pi_2(\tilde{\delta}) + \gamma R \leq 0$. Now, combining Equation (A.17) with

$$\gamma H(x) - \mu x H'(x) - \frac{\sigma^2}{2} x^2 H''(x) - \left(\Pi_1(x) + \gamma R + \gamma S\right) = 0,$$

we obtain the following equality:

$$-\frac{\sigma^2}{2} \tilde{\delta}^2 \left(H''(\tilde{\delta}) - \tilde{W}''(\tilde{\delta})\right) = \Pi_1(\tilde{\delta}) - \Pi_2(\tilde{\delta}) + \gamma R.$$

It is a matter of calculations to see that

$$H''(\tilde{\delta}) - \tilde{W}''(\tilde{\delta}) = h'(\tilde{\delta}) \tilde{\delta}^{-1} > 0,$$

where h' is the derivative of h , defined in (6.5), and the last inequality follows in light of the calculations in the proof of Lemma A.2.

To prove the inequality (A.16), we note that the function $x \rightarrow H(x)$ is increasing if $B > 0$. In case $B > 0$, as $H(0) = -\frac{b_1 - \gamma R - \gamma S}{\gamma} \geq 0$, this proves that $H(x) \geq \tilde{W}(x)$ for all $x \leq \alpha$. To see that $\alpha \leq x \leq \tilde{\delta}$, we note that $\tilde{\delta}$ is the unique solution to the equation $H(x) = \tilde{W}(x)$. Therefore, since $H(\alpha) - \tilde{W}(\alpha) = H(\alpha) > H(0) \geq 0$, the result is straightforward.

To see that $B > 0$, one can see that

$$B\tilde{\delta}^{\beta_2} = \frac{\tilde{\delta}}{\beta_2} \left(\tilde{A}\beta_1\tilde{\delta}^{\beta_1-1} + \frac{a_2\theta}{\gamma - \mu\theta} \tilde{\delta}^{\theta-1} - \frac{a_1}{\gamma - \mu} \right).$$

Additionally, the function $x \rightarrow \tilde{A}\beta_1x^{\beta_1-1} + \frac{a_2\theta}{\gamma - \mu\theta}x^{\theta-1} - \frac{a_1}{\gamma - \mu}$ is increasing and crosses zero once. By using the smooth pasting conditions (used to obtain \tilde{A} and α), we get

$$\tilde{A}\beta_1\alpha^{\beta_1-1} + \frac{a_2\theta}{\gamma - \mu\theta} \alpha^{\theta-1} - \frac{a_1}{\gamma - \mu} = -\frac{a_1}{\gamma - \mu} < 0.$$

Let \tilde{x} be such that $\tilde{A}\beta_1\tilde{x}^{\beta_1-1} + \frac{a_2\theta}{\gamma - \mu\theta} \tilde{x}^{\theta-1} - \frac{a_1}{\gamma - \mu} = 0$. Then

$$h(\tilde{x}) = \beta_2 \left(A(1 - \beta_1)\tilde{x}^{\beta_1} + \frac{a_2(1 - \theta)}{\gamma - \mu\theta} \tilde{x}^\theta + \frac{b - \gamma R}{\gamma} \right) \equiv \tilde{h}(\tilde{x}).$$

Once again, due to the smooth pasting condition, $\tilde{h}(\alpha) = 0$, and

$$\tilde{h}'(x) = \beta_2 \left(A(1 - \beta_1)\beta_1\tilde{x}^{\beta_1-1} + \frac{a_2(1 - \theta)\theta}{\gamma - \mu\theta} x^{\theta-1} \right) < 0.$$

It follows that $h(\tilde{x}) < 0$, and therefore $\tilde{x} < \tilde{\delta}$. Consequently $B > 0$. □

Proof of Proposition 6.4: We start by noticing that, since the terminal cost is $\tilde{W}(x)$, as one can see through (3.5), the smooth pasting conditions are different according to $\zeta > \alpha$ or $\zeta \leq \alpha$. Let g be defined as in (4.9). Then it is a matter of calculations to see that

$$g(\alpha) = a_1 \frac{\theta - \beta_1}{\gamma - \mu\theta} \alpha^\theta + \beta_1 \frac{b_1 - \gamma S - \gamma R}{\gamma}.$$

Taking into account the analysis made in Lemma A.1, $\zeta > \alpha \Leftrightarrow g(\alpha) < 0$, which means that

$$R < R^* \equiv \frac{1}{\beta_1} \left(a_1 \frac{\theta - \beta_1}{\gamma - \mu\theta} \alpha^\theta + \beta_1 \frac{b - \gamma S}{\gamma} \right).$$

The proof of Proposition 6.4 when $R \geq R^*$ follows in light of the arguments used in the proof of Proposition 4.2. From now on, we will treat the case $R < R^*$.

By using the smooth pasting conditions, we obtain the following equations

$$\tilde{A}_2\zeta^{\beta_1} + \frac{a_1}{\gamma - \mu\theta} \zeta^\theta - \frac{b_1 - \gamma S - \gamma R}{\gamma} = \tilde{A}\zeta^{\beta_1} + \frac{a_2}{\gamma - \mu} \zeta - \frac{b_2 - \gamma S}{\gamma},$$

$$\tilde{A}_2\beta_1\zeta^{\beta_1-1} + \frac{a_1\theta}{\gamma - \mu\theta} \zeta^{\theta-1} = \tilde{A}\beta_1\zeta^{\beta_1-1} + \frac{a_2}{\gamma - \mu}.$$

These equations allow us to obtain the expression of \tilde{A}_2 as in (6.10) and Equation (4.9). In light of Lemma A.1, there is a unique solution ζ to Equation (A.1) and $\zeta < c$.

To prove that \mathcal{W} , defined by (6.9), is the solution of the optimal stopping problem (2.2), we need to verify that $H(x) = \mathcal{W}(x) + S + R$ satisfies the following inequalities:

$$(A.18) \quad \gamma H(x) - \mu x H'(x) - \frac{\sigma^2}{2} x^2 H''(x) - \left(\Pi_1(x) + \gamma R + \gamma S \right) \geq 0, \quad \text{for all } x \leq \zeta,$$

$$(A.19) \quad H(x) \geq \tilde{W}, \quad \text{for all } x \geq \zeta.$$

In order to prove the inequality (A.18), we start by noting that, for $x < \alpha$, we can write this equation as

$$(A.20) \quad \Pi_1(x) + \gamma(S + R) \leq 0 \quad \text{for all } x \leq \alpha.$$

Since $\Pi_1(0) + \gamma(S + R) = -b_1 + \gamma(S + R) < 0$ and Π_1 is an increasing function, it follows that (A.20) holds true if and only if $\Pi_1(\alpha) + \gamma(S + R) \leq 0$. This is true because

$$0 \geq g(\alpha) = -\frac{\beta_1}{\gamma} \left(a_1 \frac{\theta - \beta_1}{\gamma - \mu\theta} \frac{\gamma}{(-\beta_1)} \alpha^\theta - (b_1 - \gamma S - \gamma R) \right) \geq -\frac{\beta_1}{\gamma} \left(\Pi_1(\alpha) + \gamma(S + R) \right),$$

where the last inequality follows in light of the fact $\frac{\theta - \beta_1}{\gamma - \mu\theta} \frac{\gamma}{(-\beta_1)} > 1$. For $\alpha < x < \zeta$, we use a similar argument to the one used in the proof of Proposition 6.1. Therefore, the inequality (A.18) can be written as

$$\gamma \tilde{W}(x) - \mu x \tilde{W}'(x) - \frac{\sigma^2}{2} x^2 \tilde{W}''(x) - \left(\Pi_1(x) + \gamma R + \gamma S \right) = \Pi_2(x) - \Pi_1(x) - \gamma R,$$

which means that we just need to show that $\Pi_2(x) - \Pi_1(x) - \gamma R \geq 0$, for all $\alpha < x \leq \zeta$. We can easily prove that the function $x \rightarrow \Pi_2(x) - \Pi_1(x) - \gamma R$ increases for $x < \frac{a_1}{\theta a_2}$ and decreases for $x > \frac{a_1}{\theta a_2}$. Combining this with the fact that $\Pi_2(0) - \Pi_1(0) - \gamma R = b - \gamma R \geq 0$, we need to prove that $\Pi_2(\zeta) - \Pi_1(\zeta) - \gamma R \geq 0$, which is true in light of Proposition 4.2.

To prove the inequality (A.19), we note that

$$H(x) - \tilde{W}(x) = (\tilde{A}_2 - \tilde{A}) x^{\beta_1} + \frac{a_1}{\gamma - \mu\theta} x^\theta - \frac{a_2}{\gamma - \mu} x - \frac{b - \gamma R}{\gamma}, \quad H(\zeta) - \tilde{W}(\zeta) = 0,$$

and

$$H'(x) - \tilde{W}'(x) = \left(\frac{a_2}{\gamma - \mu} - \frac{a_1 \theta}{\gamma - \mu\theta} \zeta^{\theta-1} \right) \left(\frac{x}{\zeta} \right)^{\beta_1-1} - \left(\frac{a_2}{\gamma - \mu} - \frac{a_1 \theta}{\gamma - \mu\theta} x^{\theta-1} \right).$$

Taking into account the proof of Proposition 4.2, we have that

$$\frac{a_2}{\gamma - \mu} - \frac{a_1 \theta}{\gamma - \mu\theta} \zeta^{\theta-1} = \beta_1 A_1 \zeta^{\beta_1-1} < 0.$$

Since A_1 is defined in (4.8) and verifies $A_1 > 0$, the result follows because $\left(\frac{x}{\zeta} \right)^{\beta_1-1} < 1$, for all $x > \zeta$, and the function $x \rightarrow \frac{a_2}{\gamma - \mu} - \frac{a_1 \theta}{\gamma - \mu\theta} x^{\theta-1}$ is decreasing. Additionally, we can conclude that $\tilde{A}_2 > 0$. □

Proof of Proposition 6.5: Noticing that R^* can be written as

$$R^*(\mu, \sigma) = \frac{a_1}{\gamma^2} \left(\frac{b_2 - \gamma S}{\gamma} \right)^\theta \frac{1}{\frac{\theta}{\beta_2} - 1} \left(1 - \frac{1}{\beta_2} \right)^\theta + \frac{b - \gamma S}{\gamma},$$

the result follow in light of the following calculations:

$$\frac{\partial}{\partial \eta} \left(\frac{1}{\frac{\theta}{\beta_2} - 1} \left(1 - \frac{1}{\beta_2} \right)^\theta \right) = \beta_2 \left(1 - \frac{1}{\beta_2} \right)^{\theta-1} \frac{\theta - 1}{(\theta - \beta_2)^2} \frac{\theta}{\beta_2^2} \frac{\partial \beta_2}{\partial \eta} < 0. \quad \square$$

ACKNOWLEDGMENTS

We thank an anonymous referee for comments and suggestions that enhanced our original manuscript.

Funding: This work was supported by the Portuguese Foundation for Science and Technology [grants number SFRH/BD/102186/2014, CEMAPRE – UID/MULTI/00491/2019 and FARO_PTDC/EGE-ECO/30535/2017].

REFERENCES


- [1] BENAROCH, MICHEL (2002). Managing information technology investment risk: a real options perspective, *Journal of Management Information Systems*, **19**(2), 43–84.
- [2] BJERKSUND, PETTER and EKERN, STEINAR (1990). Managing investment opportunities under price uncertainty: from “last chance” to “wait and see” strategies, *Financial Management*, **19**(3), 65–83.
- [3] BREALEY, RICHARD A.; MYERS, STEWART C.; ALLEN, FRANKLIN and MOHANTY, PITABAS (2012). *Principles of Corporate Finance*, Tata McGraw-Hill Education.
- [4] BRENNAN, M.J. and SCHWARTZ, E.S. (1985). Evaluating natural resource investments, *Journal of Business*, **58**(2), 135–157.
- [5] CLEMONS, ERIC K. and WEBER, BRUCE W. (1990). Strategic information technology investments: guidelines for decision making, *Journal of Management Information Systems*, **7**(2), 9–28.
- [6] DIXIT, AVINASH K. and PINDYCK, ROBERT S. (1994). *Investment Under Uncertainty*, Princeton University Press.
- [7] DUCKWORTH, J. KATE and ZERVOS, MIHAIL (2000). An investment model with entry and exit decisions, *Journal of Applied Probability*, **37**(2), 547–559.
- [8] FARZIN, Y. HOSSEIN; HUISMAN, KUNO J.M. and KORT, PETER M. (1998). Optimal timing of technology adoption, *Journal of Economic Dynamics and Control*, **22**(5), 779–799.
- [9] GUERRA, MANUEL; NUNES, CLÁUDIA and OLIVEIRA, CARLOS (2016). Exit option for a class of profit functions, *International Journal of Computer Mathematics*, pages 1–16.
- [10] HAGSPIEL, VERENA; HUISMAN, KUNO J.M.; KORT, PETER M. and NUNES, CLÁUDIA (2016). How to escape a declining market: capacity investment or exit? *European Journal of Operational Research*, **254**(1), 40–50.
- [11] HUISMAN, KUNO J.M. and KORT, PETER M. (2003). Strategic investment in technological innovations, *European Journal of Operational Research*, **144**(1), 209–223.
- [12] LORENZEN, THOMAS J. and VANCE, LONNIE C. (1986). The economic design of control charts: a unified approach, *Technometrics*, **28**(1), 3–10.
- [13] MAJD, S. and PINDYCK, R.S. (1987). Time to build, option value, and investment decisions, *Journal of Financial Economics*, **18**, 7–27.
- [14] McDONALD, ROBERT and SIEGEL, DANIEL (1985). Investment and the valuation of firms when there is an option to shut down, *International Economic Review*, **26**, 331–349.

- [15] MYERS, STEWART C. and MAJD, SAMAN (2001). *Abandonment value and project life*. In “Real Options and Investment under Uncertainty: Classical Readings and Recent Contributions” (Eduardo S. Schwartz and Lenos Trigeorgis, Eds.), Chapter 14, pages 295–312, MIT Press, Cambridge.
- [16] NEMBHARD, HARRIET BLACK; SHI, LEYUAN and AKTAN, MEHMET (2002). A real options design for quality control charts, *The Engineering Economist*, **47**(1), 28–59.
- [17] PANNELL, DAVID J. (2006). Flat earth economics: the far-reaching consequences of flat payoff functions in economic decision making, *Review of Agricultural Economics*, **28**(4), 553–566.
- [18] PAWLINA, GRZEGORZ and KORT, PETER M. (2006). Real options in an asymmetric duopoly: who benefits from your competitive disadvantage? *Journal of Economics & Management Strategy*, **15**(1), 1–35.
- [19] PESKIR, GORAN and SHIRYAEV, ALBERT (2006). *Optimal Stopping and Free-Boundary Problems*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, ISBN 978-3-7643-2419-3; 3-7643-2419-8.
- [20] WARD, ALLEN; LIKER, JEFFREY K.; CRISTIANO, JOHN J. and SOBEK, DURWARD K. (1995). The second toyota paradox: how delaying decisions can make better cars faster, *Sloan Management Review*, **36**(3), 43.

A TRUNCATED GENERAL-G CLASS OF DISTRIBUTIONS WITH APPLICATION TO TRUNCATED BURR-G FAMILY

Authors: FARRUKH JAMAL

– Department of Statistics, The Islamia of University Bahawalpur,
Bahawalpur, Pakistan
drfarrukh1982@gmail.com

HASSAN S. BAKOUCH 

– Department of Mathematics, Faculty of Science, Tanta University,
Tanta, Egypt

M. ARSLAN NASIR

– Department of Statistics, Govt. S.E College,
Bahawalpur, Pakistan

Received: May 2019

Revised: August 2019

Accepted: August 2019

Abstract:

- In this paper, we introduce a truncated general-G class of distributions. This class can be viewed as a weighted family of distributions with a general weight function, and also it generalizes the beta generator family proposed by Eugene *et al.* (2002). Some features of the class are stated with a comprehensive study to the truncated Burr-G (TB-G) family as one of the important sub-class of the introduced class. The study includes the mixture representation in terms of baseline distribution, moments, moment generating function, stochastic ordering, stress-strength parameter, entropies, estimation by the maximum likelihood. The applicability of some new sub-models of the TB-G family is shown using two practical data sets.

Keywords:

- *family of distributions; Burr distribution; quantile function; simulation; estimation; goodness-of-fit statistics.*

AMS Subject Classification:

- 60E05, 62E15.

1. INTRODUCTION

Over the last two decades, several extensions of the well-known lifetime distributions have been developed for modeling many types of practical data sets. This development is followed by many approaches for generating new families of (probability) distributions which increase chances of modeling data of various random nature. Among those families, we can mention: The beta generator (beta-G) by Eugene *et al.* (2002) [9], the gamma-G (type 1) by Zografos and Balakrishnan (2009) [19], the Kumaraswamy-G (Kw-G) by Cordeiro and de Castro (2011) [7], the gamma-G (type 2) by Ristic and Balakrishnan (2012) [13], the log-gamma-G by Amini *et al.* (2014) [4], beta weighted modified Weibull distribution using the beta generator by Saboor *et al.* (2016) [14], the generalized transmuted family of distributions by Alizadeh *et al.* (2017) [3], the odd-Burr generalized family of distributions by Alizadeh *et al.* (2017) [2], the odd Burr-III family of distributions by Jamal *et al.* (2017) [11], the extended odd family of probability distributions by Bakouch *et al.* (2019) [5] and mid-truncated Burr XII distribution and its applications in order statistics by Saran *et al.* (2019) [15].

In practical life problems, truncation arises in many fields, such as industry, biology, hydrology, reliability theory and medicine. An example of truncation is the progression of a disease which is not an increasing function, but will stabilize after time point. This point is called the truncation for the support of the variable of the interest which may be time, length, height etc. Therefore, many researchers are attracted to analyze such truncated data using truncated versions of the standard statistical distributions. For instance, the truncated Weibull distribution has been applied to analyze the tree diameter and height distributions in forestry, fire size and high-cycle fatigue strength prediction (see Zhang and Xie, 2011 [18]). In Zaninetti and Ferraro (2008) [17], the truncated Pareto distribution is compared to the Pareto distribution using astrophysics data and they concluded, generally, that the truncated Pareto distribution performs better than the Pareto. Burroughs and Tebbens (2002) [6] showed the suitability of truncated power law distributions for data sets of earthquake magnitudes and forest fire areas. Additional applications of the former distributions in hydrology and atmospheric science are given by Aban *et al.* (2006) [1].

Motivated by the importance of general families of distributions and truncation, we introduce a more flexible class of distributions with the cumulative distribution function (cdf)

$$(1.1) \quad F(x) = \int_0^{G(x,\xi)} r_T(t) dt = \int_0^{G(x,\xi)} \frac{r(t)}{R(1)} dt = \frac{R[G(x,\xi)]}{R(1)},$$

where $r_T(t)$ is the probability density function (pdf) of a random variable (rv) with support $[0, 1]$, hence it can be any truncated rv T on this support with a cdf, $R(\cdot)$ and $G(x, \xi)$ is the cdf of a real-valued rv X with pdf $g(x, \xi)$, ξ denoting the related parameter vector. Table 1 gives a list of some truncated distribution in the interval $[0,1]$. The associated pdf of (1.1) is

$$(1.2) \quad f(x) = \frac{r[G(x,\xi)] g(x,\xi)}{R(1)}, \quad x \in \mathbb{R},$$

and the survival function based on (1.1) is given as

$$(1.3) \quad h(x) = \frac{r[G(x,\xi)] g(x,\xi)}{R(1) - R[G(x,\xi)]}.$$

Further, the associated quantile function based on (1.1) having the form

$$(1.4) \quad Q_x(u) = G^{-1}\left\{R^{-1}[R(1) \times u]\right\},$$

where $u \sim \text{uniform}[0, 1]$.

Table 1: List of some truncated distribution in the interval [0,1].

S.r	Distribution	$r(t)$	$r_T(t)$
1.	Uniform	$F(x) = \frac{x}{\theta}$	$F(x) = x$
2.	Exponential	$F(x) = 1 - e^{-\theta x}$	$F(x) = \frac{1 - e^{-\theta x}}{1 - e^{-\theta}}$
3.	Weibull	$F(x) = 1 - e^{-ax^b}$	$F(x) = \frac{1 - e^{-ax^b}}{1 - e^{-a}}$
4.	Gamma	$F(x) = \frac{\gamma(a, \frac{x}{b})}{\Gamma(a)}$	$F(x) = \frac{\gamma(a, \frac{x}{b})}{\gamma(a, \frac{1}{b})}$
5.	Lomax	$F(x) = 1 - \left(1 + \frac{x}{a}\right)^{-b}$	$F(x) = \frac{1 - \left(1 + \frac{x}{a}\right)^{-b}}{1 - \left(1 + \frac{1}{a}\right)^{-b}}$
6.	log-logistic	$F(x) = 1 - \left(1 + \frac{x^c}{a}\right)^{-1}$	$F(x) = \frac{1 - \left(1 + \frac{x^c}{a}\right)^{-1}}{1 - \left(1 + \frac{1}{a}\right)^{-1}}$
7.	Burr XII	$F(x) = 1 - (1 + x^c)^{-k}$	$F(x) = \frac{1 - (1 + x^c)^{-k}}{1 - 2^{-k}}$
8.	Burr III	$F(x) = (1 + x^{-c})^{-k}$	$F(x) = \frac{(1 + x^{-c})^{-k}}{2^{-k}}$
9.	Frechet	$F(x) = \exp\left[-\left(\frac{a}{x}\right)^b\right]$	$F(x) = \frac{\exp\left[-\left(\frac{a}{x}\right)^b\right]}{\exp[-a^b]}$
10.	Power function	$F(x) = \left(\frac{x}{\theta}\right)^k$	$F(x) = x^k$
11.	Log normal	$F(x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right)$	$F(x) = \frac{\Phi\left(\frac{\ln x - \mu}{\sigma}\right)}{\Phi\left(\frac{-\mu}{\sigma}\right)}$

Some additional motivations of the class defined by (1.2) are as follows. The class (1.2) can be interpreted as weighted family of distributions, for $g(x, \xi)$, with the general weight function $w(X) = r(G(x, \xi))$ and normalizing constant $R(1) = E\{w(X)\}$. Also, the introduced class generalizes the beta generator family (Eugene *et al.*, 2002 [9]) as beta distribution is a sub-model of $r_T(t)$.

As it can be seen from (1.2), we have a truncated general-G class of distributions and the only sub-model we aware of is the truncated Weibull G family proposed by Najarzagdegan *et al.* (2017) [12] as a powerful alternative to beta-G family of distributions. Because of having two composite general functions $R(\cdot)$ and $G(\cdot)$, we can not investigate more analytic properties and therefore we aim to study extensively the truncated Burr-G (TB-G) family of distributions by considering $R(\cdot)$ as the cdf of Burr distribution and $G(\cdot)$ is a general cdf. The reason of using Burr is due to its ability of analyzing hydrologic, environmental, survival and reliability data. Another aim is to provide an empirical evidence on the great flexibility of sub-models of the TB-G family to fit practical data from different domains and this is investigated in the application section.

Rest of the paper is outlined as follows. Section 2 concerns with some general mathematical properties of the TB-G family, including mixture representation in terms of baseline distribution, moments, incomplete moments, moment generating function, stochastic ordering of the random variables following such family, stress-strength parameter and entropies (Shannon and Rényi). Also, some new special models of the generated family are considered. In Section 3, estimation of the parameters of the family is implemented through maximum likelihood method with application to two practical data sets. Section 4 gives a simulation study for a sub-model of the family.

2. THE TRUNCATED BURR-G FAMILY: SOME PROPERTIES AND SUB-MODELS

This section gives some general mathematical properties of the TB-G family, including moments, incomplete moments, moment generating function, stochastic ordering, stress-strength parameter and entropies. Further, some new sub-models of the family are obtained.

2.1. The truncated Burr-G family

In this section, we introduce the TB-G family of distributions and give its mixture representation in terms of baseline distribution.

Recall that the Burr distribution has the cdf

$$(2.1) \quad R(x) = 1 - (1 + x^c)^{-k}, \quad x > 0,$$

using (1.1), the cdf of the TB-G family is expressed as

$$(2.2) \quad F(x) = \frac{1 - [1 + G^c(x, \xi)]^{-k}}{1 - 2^{-k}},$$

where c, k are the shape parameters of the family and $G(x, \xi)$ is a baseline cdf, which depends on a parameter vector ξ . Hereafter, for simplicity, we ignore mention of ξ in the functions of interest, e.g., we set $G(x) = G(x, \xi)$, $g(x) = g(x, \xi)$.

The pdf corresponding to (2.2) is given by

$$(2.3) \quad f(x) = \frac{c k g(x) G^{c-1}(x) [1 + G^c(x)]^{-k-1}}{1 - 2^{-k}}, \quad x \in \mathbb{R}.$$

The survival function and hazard rate are, respectively, given by

$$(2.4) \quad \bar{F}(x) = \frac{[1 + G^c(x)]^{-k} - 2^{-k}}{1 - 2^{-k}}$$

and

$$(2.5) \quad \tau(x) = \frac{c k g(x) G^{c-1}(x) [1 + G^c(x)]^{-k-1}}{[1 + G^c(x)]^{-k} - 2^{-k}}.$$

Also, the quantile function of the TB-G family has the form

$$(2.6) \quad Q_x(u) = G^{-1} \left[\left\{ \left[1 - (1 - 2^{-k}) u \right]^{-\frac{1}{k}} - 1 \right\}^{\frac{1}{c}} \right].$$

Further, the shapes of the density and hazard rate functions of the TB-G family can be described analytically using their critical points as follows. The critical points of the TB-G density are the roots of the equation:

$$\frac{g'(x)}{g(x)} + (c - 1) \frac{g(x)}{G(x)} - c(k + 1) \frac{g(x) G^{c-1}(x)}{1 - G^c(x)} = 0,$$

while the critical points of the hazard rate are the roots of the equation:

$$\frac{g'(x)}{g(x)} + (c - 1) \frac{g(x)}{G(x)} - c(k + 1) \frac{g(x) G^{c-1}(x)}{1 - G^c(x)} + k c \frac{g(x) G^{c-1}(x) [1 + G^c(x)]^{-k-1}}{[1 + G^c(x)]^{-k} - 2^{-k}} = 0.$$

Note that the equation above may have more than one root.

Now, we close this subsection by obtaining the mixture representation of the TB-G in terms of baseline distribution as follows.

Consider the series expansion, for $|z| < 1$,

$$(2.7) \quad (1 - z)^{-b} = \sum_{i=0}^{\infty} \binom{b + i - 1}{i} z^i,$$

the cdf in equation (2.2) can be written as

$$(2.8) \quad F(x) = \frac{1}{1 - 2^{-k}} \left[1 - \sum_{i=0}^{\infty} \binom{k + i - 1}{i} (-1)^i G^{ic}(x) \right].$$

Also, it can be rewritten in the form

$$(2.9) \quad F(x) = \sum_{l=0}^{\infty} b_l H_l(x),$$

where $b_l = \frac{1}{1 - 2^{-k}} \sum_{i=1}^{\infty} \sum_{j=l}^{\infty} \binom{k+i-1}{i} \binom{c i}{j} \binom{j}{l} (-1)^{i+j+l+1}$ and $H_l(x) = G^l(x)$ is the exp-G distribution function with power parameter l .

Similarly, simple derivation of the previous equation gives the pdf

$$(2.10) \quad f(x) = \sum_{l=0}^{\infty} b_l h_{l-1}(x),$$

where $h_{l-1}(x) = l \times g(x) G^{l-1}(x)$ is the exp-G density function with power parameter $l - 1$. Thus, some mathematical properties of the proposed family can be derived from (2.10) and those of exp-G properties. For example, the ordinary and incomplete moments and moment generating function (mgf) of X can be obtained from those exp-G quantities, see the next subsection.

2.2. Moments and moment generating function

In this subsection, we will discuss the r^{th} moments, m^{th} incomplete moments and moment generating function of the TB-G family.

The moments of the TB-G family of distributions can be obtained by using the infinite mixture representation

$$(2.11) \quad E(X^r) = \sum_{l=0}^{\infty} b_l \int_{-\infty}^{\infty} x^r h_{l-1}(x) dx,$$

where b_l and $h_{q-1}(x)$ are defined in (2.10).

The s^{th} incomplete moment of the TB-G family can be obtained as

$$(2.12) \quad T'_s(x) = \sum_{l=0}^{\infty} b_l \int_{-\infty}^x x^s h_{l-1}(x) dx.$$

The moment generating function of the TB-G family of distributions is

$$M_X(t) = \sum_{l=0}^{\infty} b_l \int_{-\infty}^{\infty} e^{tx} h_{l-1}(x) dx.$$

Bonferroni and Lorenz curves, defined for a given probability, π , by $B(\pi) = T'_1(q)/(\pi\mu'_1)$ and $L(\pi) = T'_1(q)/\mu'_1$, respectively, where $\mu'_1 = E(X)$, $T'_1(x) = \sum_{l=0}^{\infty} b_l \int_{-\infty}^x x h_{l-1}(x) dx$ and $q = Q(\pi)$ is the quantile function of X at π . These curves for the Truncated Burr log logistic (TBLL) distribution (see definition of TBLL in the next subsection) as functions of π , are plotted for some parameter values in Figure 1. These curves are very useful in economics, reliability, demography, insurance and medicine. The skewness and kurtosis measures can be calculated from the ordinary moments using well-known relationships from equation (2.11).

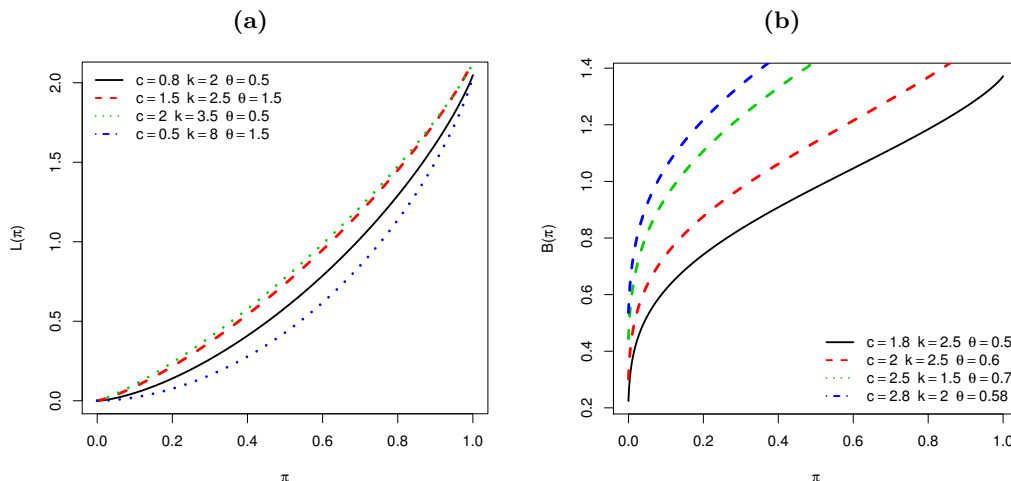


Figure 1: Plots of $B(\pi)$ and $L(\pi)$ versus π for the TB-LL distribution.

Plots of skewness and kurtosis of the TBLL distribution for $\theta = 1.5$ are displayed in Figure 2. Based on these plots, we conclude that, if c and k increase, the skewness and kurtosis decrease.

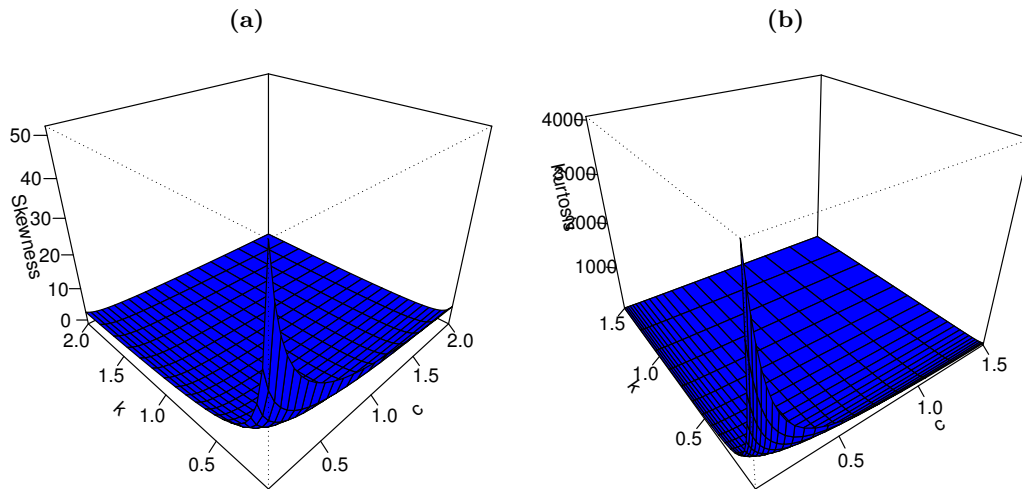


Figure 2: Plots for skewness and kurtosis of the TB-LL distribution.

2.3. Stochastic ordering and reliability parameter

Comparative behavior of random variables can be measured by stochastic ordering concept (Shaked and Shanthikumar, 1994 [16]) that is summarized in the next proposition.

Proposition 2.1. *Let $X_1 \sim \text{TB-G}(c, k_1, \xi)$ and $X_2 \sim \text{TB-G}(c, k_2, \xi)$, then the likelihood ratio $\frac{f(x)}{g(x)}$ is*

$$\frac{f(x)}{g(x)} = \frac{k_1}{k_2} [1 + G^c(x)]^{k_2 - k_1} \frac{1 - 2^{-k_2}}{1 - 2^{-k_1}}.$$

Taking derivative with respect to x , we have

$$\frac{d}{dx} \frac{f(x)}{g(x)} = \frac{k_1}{k_2} \frac{1 - 2^{-k_2}}{1 - 2^{-k_1}} [1 + G^c(x)]^{k_2 - k_1 - 1} (k_2 - k_1) c g(x) G^{c-1}(x),$$

then $\frac{d}{dx} \frac{f(x)}{g(x)} < 0$ for $k_2 < k_1$. So, the likelihood ratio exists and this implies that the random variable X_1 is a likelihood ratio order than X_2 , that is $X_1 \leq_{lr} X_2$. Other stochastic ordering behaviors follow using $X_1 \leq_{lr} X_2$, such as hazard rate order ($X_1 \leq_{hr} X_2$), mean residual life order ($X_1 \leq_{mrl} X_2$) and stochastically greater ($X_1 \leq_{st} X_2$).

The stress strength model is a common approach used in various applications of engineering and physics. Let X_1 and X_2 be two independent random variables with $\text{TB-G}(c, k_1, \xi)$ and $\text{TB-G}(c, k_2, \xi)$ distributions. Then the stress strength model is given by

$$R = \int_{-\infty}^{\infty} f_1(x) F_2(x) dx.$$

Now, by using the mixture representation given in (2.10) and (2.9), we have

$$R = \sum_{l=0}^{\infty} \sum_{m=0}^{\infty} b_l b_m \int_{-\infty}^{\infty} h_{l-1}(x) H_m(x) dx ,$$

where $h_{l-1}(x)$ and $H_m(x)$ are already defined by equations (2.9) and (2.10).

2.4. Entropies

The entropy of a random variable X with density function $f(x)$ is a measure of variation of the uncertainty of physical systems. Two popular entropy measures are due to Shannon entropy and Rényi entropy. A large value of the entropy may indicate the greater uncertainty in the data; conversely, a small entropy means less uncertainty. The Rényi entropy is defined by

$$(2.13) \quad I_{\delta} = \frac{1}{1-\delta} \log \left(\int_{-\infty}^{\infty} f^{\delta}(x) dx \right), \quad \delta > 0 \quad \text{and} \quad \delta \neq 1 .$$

Let $f(x)$ follow the TB-G family, then we have

$$f^{\delta}(x) = \frac{(ck)^{\delta} g^{\delta}(x) G^{\delta(c-1)}(x) [1 + G^c(x, \xi)]^{-\delta(k+1)}}{(1 - 2^{-k})^{\delta}} .$$

After some algebra, we get

$$f^{\delta}(x) = \left(\frac{ck}{1 - 2^{-k}} \right)^{\delta} \sum_{j=0}^{\infty} \binom{\delta(k+1) + j - 1}{j} (-1)^j g^{\delta}(x) G^{c(j+\delta)-\delta}(x) .$$

Rewriting the above expression as

$$f^{\delta}(x) = \sum_{j=0}^{\infty} w_j(\delta) g(x; \delta, c(j + \delta)) ,$$

where $w_j(\delta) = \left(\frac{ck}{1 - 2^{-k}} \right)^{\delta} \binom{\delta(k+1) + j - 1}{j} (-1)^j$ and $g(x; \delta, c(j + \delta)) = g^{\delta}(x) G^{c(j+\delta)-\delta}(x)$.

Now equation (2.13) becomes

$$I_{\delta} = \frac{1}{1-\delta} \log \left[\sum_{j=0}^{\infty} w_j(\delta) \int_{-\infty}^{\infty} g(x; \delta, c(j + \delta)) dx \right] .$$

The above expression depends only for any choice of baseline distribution.

On the other side, the Shannon entropy of the TB-G family can be obtained using its definition as

$$(2.14) \quad \eta = -E[\log f(X)] .$$

Using the pdf of the TB-G family, we have

$$(2.15) \quad -E[\log f(X)] = \log[1 - 2^{-k}] - \log(ck) - E[\log g(X)] - (c - 1) E[\log G(X)] + (k + 1) E[\log\{1 + G^c(X)\}].$$

Making use of the expansions, for $|x| < 1$,

$$\log(1 + x) = \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i} x^i, \\ \log x = \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i} (x - 1)^i,$$

we obtain

$$E[\log\{1 + G^c(X)\}] = \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i} E[G^{ci}(X)], \\ E[\log G(X)] = \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i} \sum_{j=0}^i \binom{i}{j} (-1)^j E(G^{i-j}(X)).$$

Hence, equation (2.15) becomes

$$-E[\log f(X)] = \log[1 - 2^{-k}] - \log(ck) - E[\log g(X)] - (c - 1) \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i} \sum_{j=0}^i \binom{i}{j} (-1)^j E(G^{i-j}(X)) + (k + 1) \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i} E[G^{ci}(X)].$$

The expression above depends only on an arbitrary choice of the baseline distribution.

2.5. Some sub-models

In this subsection, we present four sub-models of the TB-G family by selecting some baseline distributions and the plots of their density and hazard rate functions. The plots indicate various shapes for both functions which proves the flexibility of the family. This flexibility is also confirmed by comparing those sub-models with other competing distributions for some practical data in Section 3.

Truncated Burr Uniform (TBU) distribution

Consider the uniform distribution on $(0, \theta)$ as the baseline distribution with the pdf and cdf, $g(x, \theta) = \frac{1}{\theta}$ and $G(x, \theta) = \frac{x}{\theta}$, respectively. Then the pdf and cdf of the TBU distribution are given by

$$f(x; c, k, \theta) = \frac{ck}{\theta} \frac{\left(\frac{x}{\theta}\right)^{c-1}}{1 - 2^{-k}} \left[1 + \left(\frac{x}{\theta}\right)^c\right]^{-k-1}$$

and

$$F(x; c, k, \theta) = \frac{1 - \left[1 + \left(\frac{x}{\theta}\right)^c\right]^{-k}}{1 - 2^{-k}}, \quad 0 < x < \theta.$$

Figure 3 gives the plots of density and hrf of the TBU distribution.

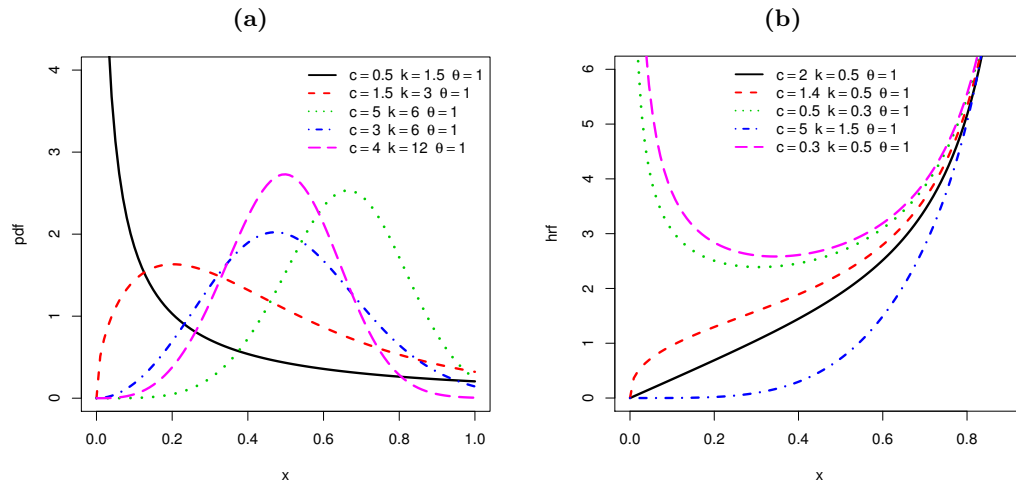


Figure 3: Plots for density and hrf of the TBU.

Truncated Burr Weibull (TBW) distribution

Let the Weibull distribution be the baseline one with the associated pdf and cdf, $g(x, a, b) = abx^{b-1}e^{-ax^b}$ and $G(x, a, b) = 1 - e^{-ax^b}$, respectively. Then the pdf and cdf of the TBW distribution are given by

$$f(x; c, k, a, b) = \frac{ckabx^{b-1}e^{-ax^b}}{1 - 2^{-k}} \frac{[1 - e^{-ax^b}]^{c-1}}{[1 + \{1 - e^{-ax^b}\}^c]^{k+1}}$$

and

$$F(x; c, k, a, b) = \frac{1 - [1 + \{1 - e^{-ax^b}\}^c]^{-k}}{1 - 2^{-k}}, \quad 0 < x < \infty.$$

Figure 4 displays the plots of density and hrf of the TBW distribution.

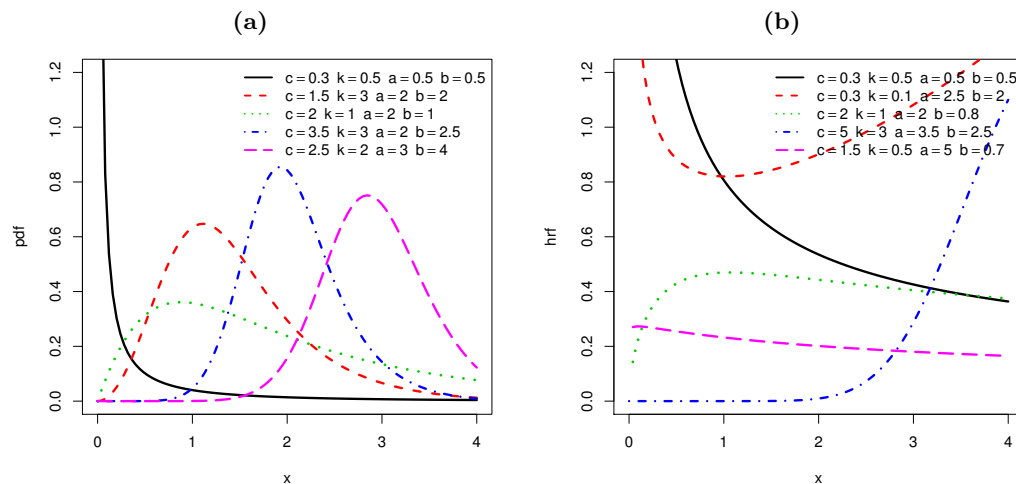


Figure 4: Plots for density and hrf of the TBW.

Truncated Burr Logistic (TBL) distribution

Consider the Logistic as the baseline distribution with associated pdf and cdf, $g(x, \theta) = \{1 - e^{-\theta x}\}^{-1}$ and $G(x, \theta) = \theta e^{-\theta x} \{1 - e^{-\theta x}\}^{-2}$, respectively. Then the pdf and cdf of the TBL distribution are given by

$$f(x; c, k, \theta) = \frac{c k \theta e^{-\theta x}}{[1 - 2^{-k}] \{1 - e^{-\theta x}\}^2} [1 - e^{-\theta x}]^{1-c} \left[1 + \{1 - e^{-\theta x}\}^{-c}\right]^{-k-1}$$

and

$$F(x; c, k, \theta) = \frac{1 - \left[1 + \left\{\left[1 - e^{-\theta x}\right]^{-c}\right\}^{-k}\right]}{1 - 2^{-k}}, \quad 0 < x < \infty.$$

In Figure 5 we give the plots of density and hrf of the TBL distribution.

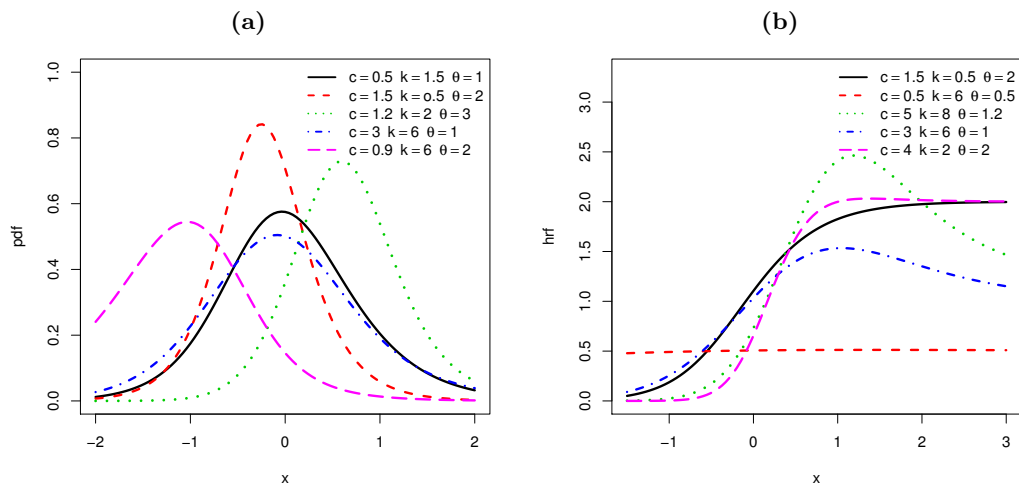


Figure 5: Plots for density and hrf of the TBL.

Truncated Burr log logistic (TBLL) distribution

Let log logistic be the baseline distribution with the associated pdf and cdf, $g(x, \theta) = \frac{\theta x^\theta}{(1+x^\theta)^2}$ and $G(x, \theta) = \frac{x^\theta}{1+x^\theta}$, respectively. Then the pdf and cdf of the TBLL distribution are given by

$$f(x; c, k, \theta) = \frac{c k \theta x^\theta}{[1 - 2^{-k}] (1 + x^\theta)^2} \left[\frac{x^\theta}{1 + x^\theta}\right]^{c-1} \left[1 + \left\{\frac{x^\theta}{1 + x^\theta}\right\}^c\right]^{-k-1}$$

and

$$F(x; c, k, \theta) = \frac{1 - \left[1 + \left\{\frac{x^\theta}{1 + x^\theta}\right\}^c\right]^{-k}}{1 - 2^{-k}}.$$

Figure 6 portrays the plots of density and hrf of the TBLL distribution.

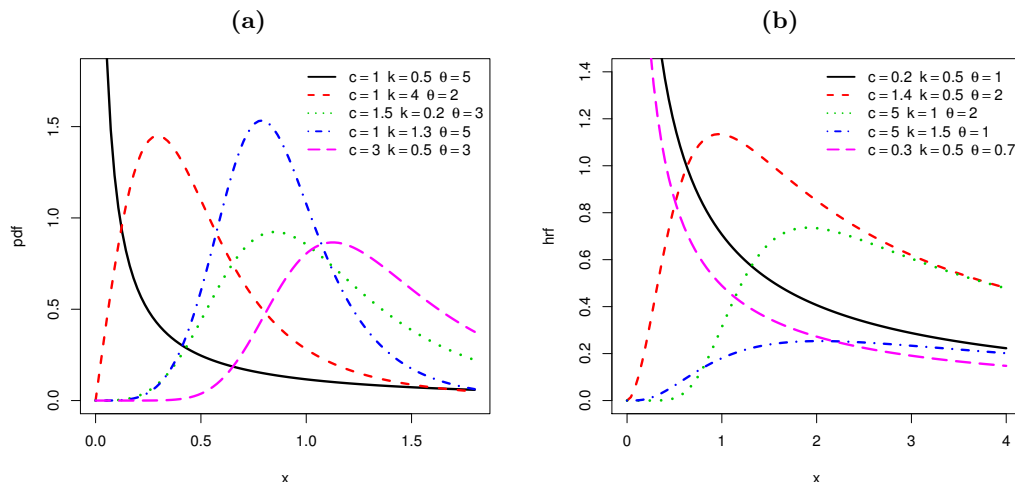


Figure 6: Plots for density and hrf of the TBLL.

3. ESTIMATION OF PARAMETERS WITH APPLICATIONS

In this section, we give the maximum likelihood estimators (MLEs) of the unknown parameters of the TB-G family for complete samples. Using those estimators we check the capability of some sub-models of this family for fitting some practical data sets. Let x_1, x_2, \dots, x_n be the observed values of a random sample of size n from the TB-G family given in equation (2.3). The log-likelihood function for the vector parameter $\Theta = [c, k, \xi]^T$ can be expressed as

$$\begin{aligned} \ell(\Theta) = & -n \log(1 - 2^{-k}) + n \log(c k) + \sum_{i=1}^n \log g(x_i) + (c - 1) \sum_{i=1}^n \log G(x_i) \\ & - (k + 1) \sum_{i=1}^n \log\{1 + G^c(x_i)\}. \end{aligned}$$

The components of score vector $U = (U_k, U_c, U_\xi)^T$ are given by

$$\begin{aligned} U_k &= -n \frac{2^{-k} \log 2}{1 - 2^{-k}} + \frac{n}{k} - \sum_{i=1}^n \log\{1 + G^c(x_i)\}, \\ U_c &= \frac{n}{c} + \sum_{i=1}^n \log G(x_i) - (k + 1) \sum_{i=1}^n \left[\frac{c g(x_i) G^{c-1}(x_i)}{1 + G^c(x_i)} \right], \\ U_\xi &= \sum_{i=1}^n \left[\frac{g^\xi(x_i)}{g(x_i)} \right] + (c - 1) \sum_{i=1}^n \left[\frac{G^\xi(x_i)}{G(x_i)} \right] - (k + 1) \sum_{i=1}^n \left[\frac{c G^\xi(x_i) G^{c-1}(x_i)}{1 + G^c(x_i)} \right]. \end{aligned}$$

The equations above are non-linear and hence can not be solved analytically, but can be solved numerically using software like R language. The rest of this section provides two applications of four sub-models of the TB-G family, namely, the TBW, TBLL, TBU and TBL distributions given in Subsection 2.5. Truncated Weibull-BXII (TW-BXII) and Truncated Weibull-Weibull (TW-W) introduced by Najarzagagan *et al.* (2017) [12] are used as competitive models for

those sub-models. For comparison purposes, we consider two practical data sets, one is taken from El-deeb (2015) [8] and another from Hinkley (1977) [10]. Description of both data sets is as follows.

Data set 1: This data set is given by El-deeb (2015) [8] and consists of failure times of (67) truncated Aircraft windshield. The windshield on an aircraft is a complex piece of equipment, comprised basically of several layers of material, all laminated under high temperature and pressure. Failures of these items are not structural failures. Instead, they typically involve damage or delimitation of the nonstructural outer ply or failure of the heating system. These failures do not result in damage to the aircraft, but do result in replacement of the windshield. The values of this data set are: 1.866, 2.385, 3.443, 1.876, 2.481, 3.467, 1.899, 2.610, 3.478, 1.911, 2.625, 3.578, 1.912, 2.632, 3.595, 1.070, 1.914, 2.646, 3.699, 1.124, 1.981, 2.661, 3.779, 1.248, 2.010, 2.688, 3.924, 1.281, 2.038, 2.82, 3, 3.000, 1.281, 2.085, 2.890, 1.303, 2.089, 2.902, 1.432, 2.097, 2.934, 1.480, 2.135, 2.962, 1.505, 2.154, 2.964, 1.506, 2.190, 3.000, 1.568, 2.194, 3.103, 1.615, 2.223, 3.114, 1.619, 2.224, 3.117, 1.652, 2.229, 3.166, 1.652, 2.300, 3.344, 1.757, 2.324, 3.376.

Data set 2: This data set is given by Hinkley (1977) [10] and consists of thirty successive values of March precipitation (in inches) in Minneapolis/St. Paul. In meteorology, precipitation is most commonly rainfall, but also includes hail, snow and other forms of liquid and frozen water falling to the ground and it is measured by inches in some time period. The data values are 0.77, 1.74, 0.81, 1.2, 1.95, 1.2, 0.47, 1.43, 3.37, 2.2, 3, 3.09, 1.51, 2.1, 0.52, 1.62, 1.31, 0.32, 0.59, 0.81, 2.81, 1.87, 1.18, 1.35, 4.75, 2.48, 0.96, 1.89, 0.9, 2.05.

For each distribution, the MLEs are computed using Quasi-Newton code for Bound Constrained Optimization (L-BFGS-B) and the log-likelihood function is evaluated. Consequently, the goodness-of-fit measures: Anderson–Darling (A^*), Cramer–von Mises (W^*), Akaike information criterion (AIC) and Bayesian information criterion (BIC) are computed. Lower values of those measures indicate better fit. The value for the Kolmogorov–Smirnov (KS) statistic and its p -value are also provided. The required computations are carried out using the R software.

The obtained results are presented in Tables 2–5. As we can see from Tables 2 and 4, the four sub-models of the TB-G family are strong competitor to the compared models.

Table 2: MLEs and their standard errors (in parentheses) for data set 1.

Distribution	c	k	θ	a	b
TBW	0.4564 (1.9144)	86.9870 (45.4333)	— —	9.1067 (2.1784)	7.9149 (3.2404)
TBLL	13.6258 (2.3252)	193.8078 (34.7291)	0.7890 (0.2350)	— —	— —
TBU	3.5954 (0.3412)	498.2935 (15.2232)	14.9104 (12.1123)	— —	— —
TBL	23.3433 (7.0993)	0.0024 (0.0018)	1.6699 (0.1944)	— —	— —
TW-BXII	1.2904 (0.3253)	11.4013 (13.4118)	32.4704 (35.6313)	37.8343 (40.8586)	3.4896 (2.4676)
TB-W	2.8676 (2.7877)	0.8444 (0.6816)	— —	31.2399 (2.1419)	6.7846 (8.0910)

Moreover, among all compared models, the TBLL distribution has the smallest values of the AIC, BIC, A^* , W^* , and KS, and the largest value of p -value. Thus, we can conclude that the TBLL distribution is the best fit among those models. Figures 7 and 8 display the plots of the fitted pdfs and cdfs of the compared distributions for visual comparison with the histogram and empirical cdf for both data sets. Those figures show the best fit of TBLL distribution.

Table 3: The Value, AIC, BIC, A^* , W^* , KS, P-Value values for data set 1.

Distribution	ℓ	AIC	BIC	A^*	W^*	KS	P-Value
TBW	75.1080	158.2162	167.0942	0.5552	0.0951	0.0992	0.5147
TBLL	74.8708	155.7418	162.4003	0.4637	0.0740	0.0808	0.7379
TBU	75.0909	156.1819	162.8404	0.5564	0.0954	0.0997	0.5080
TBL	76.2189	158.4378	165.0963	0.5855	0.0859	0.0927	0.6016
TW-BXII	75.0635	160.1271	171.2246	0.5051	0.0841	0.0893	0.6487
TW-W	75.0454	158.0909	166.9690	0.4889	0.0798	0.0835	0.7299

Table 4: MLEs and their standard errors (in parentheses) for data set 2.

Distribution	c	k	θ	a	b
TBW	0.3446 (2.8251)	30.8825 (17.3728)	- -	11.9180 (10.6096)	5.3663 (4.4130)
TBLL	8.6122 (6.0513)	123.2974 (12.2964)	0.4892 (0.4066)	- -	- -
TBU	1.8150 (0.2482)	259.5434 (12.1122)	40.3962 (33.2333)	- -	- -
TBL	7.7107 (2.1529)	0.5621 (3.0901)	1.3198 (0.3681)	- -	- -
TW-BXII	1.0579 (1.1048)	86.6647 (71.9193)	60.8969 (69.5585)	0.0024 (4.5165)	3.0599 (6.3469)
TB-W	9.7190 (12.7756)	6.2763 (9.6175)	- -	19.3190 (46.5365)	0.2883 (0.4437)

Table 5: The Value, AIC, BIC, A^* , W^* , KS, P-Value values for data set 2.

Distribution	ℓ	AIC	BIC	A^*	W^*	KS	P-Value
TBW	38.5661	85.1322	90.7370	0.1571	0.0203	0.0648	0.9996
TBLL	38.0934	82.1868	86.3904	0.1019	0.0137	0.0576	1
TBU	38.6334	83.2668	87.4701	0.1680	0.0217	0.0683	0.9990
TBL	38.9520	83.9040	88.1076	0.1466	0.0185	0.0692	0.9988
TW-BXII	38.0919	86.1839	93.1899	0.1037	0.0141	0.0605	0.9999
TW-W	38.6431	85.2862	90.8910	0.1690	0.0219	0.0688	0.9989

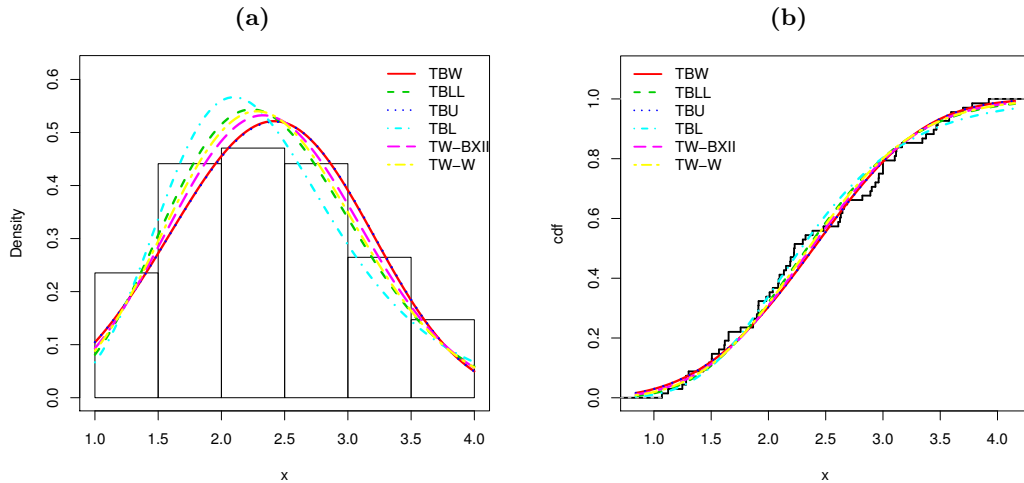


Figure 7: Estimated pdfs and cdfs for data set 1.

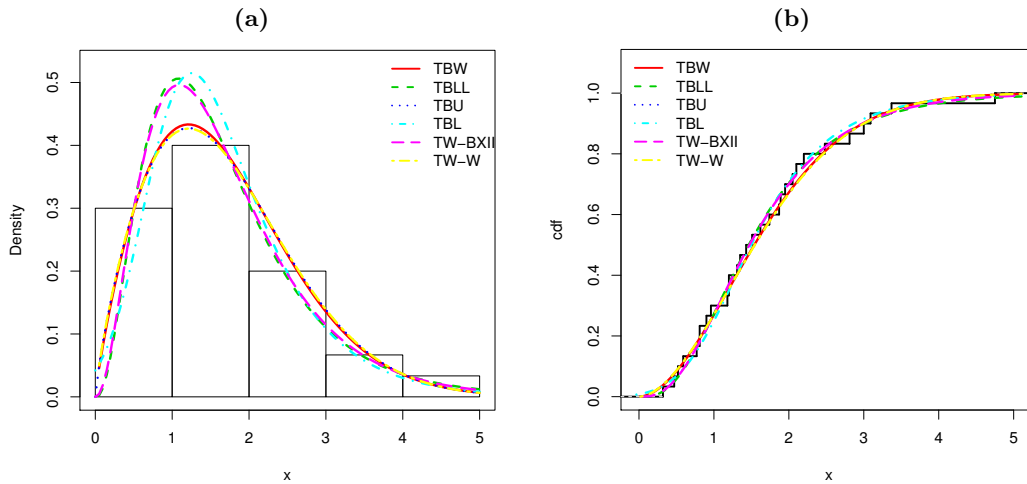


Figure 8: Estimated pdfs and cdfs for data set 2.

4. SIMULATION STUDY

In this section, the performance of the MLEs of the TBLL distribution parameters is discussed by means of Monte-Carlo simulation study. The following measures are used to evaluate the simulation results: Estimated bias, Root mean square error (RMSE) and coverage probability (CP). The simulation experiment was repeated $N=1,000$ times each with sample sizes $n = 20, 50, 100, 200, 300$ and 500 , where the samples are generated from the TBLL distribution, with $\theta = 4.5, c = 2.8, k = 0.8$, by using the inverse transform method. The MLEs of the parameters of TBLL distribution are obtained for each generated sample, $(\hat{\theta}, \hat{c}, \hat{k})$. The formulas for biases, RMSEs and CPs are given as follows.

Estimated bias of MLE $\hat{\Theta}$ of the parameter $\Theta = (\theta, c, k)$ is

$$\frac{1}{N} \sum_{i=1}^N (\hat{\Theta} - \Theta).$$

Root mean squared error (RMSE) of the MLE $\hat{\Theta}$ of the parameter $\Theta = (\theta, c, k)$ is

$$\sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{\Theta} - \Theta)^2}.$$

Coverage probability (CP) of 95% confidence intervals of the parameter $\Theta = (\theta, c, k)$ is the percentage of intervals that contain the true value of parameter Θ .

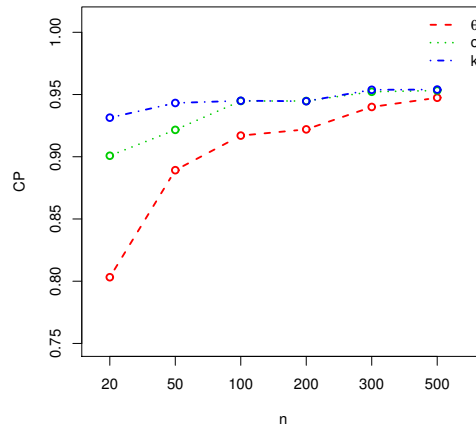


Figure 9: Estimated CPs for the selected parameters.

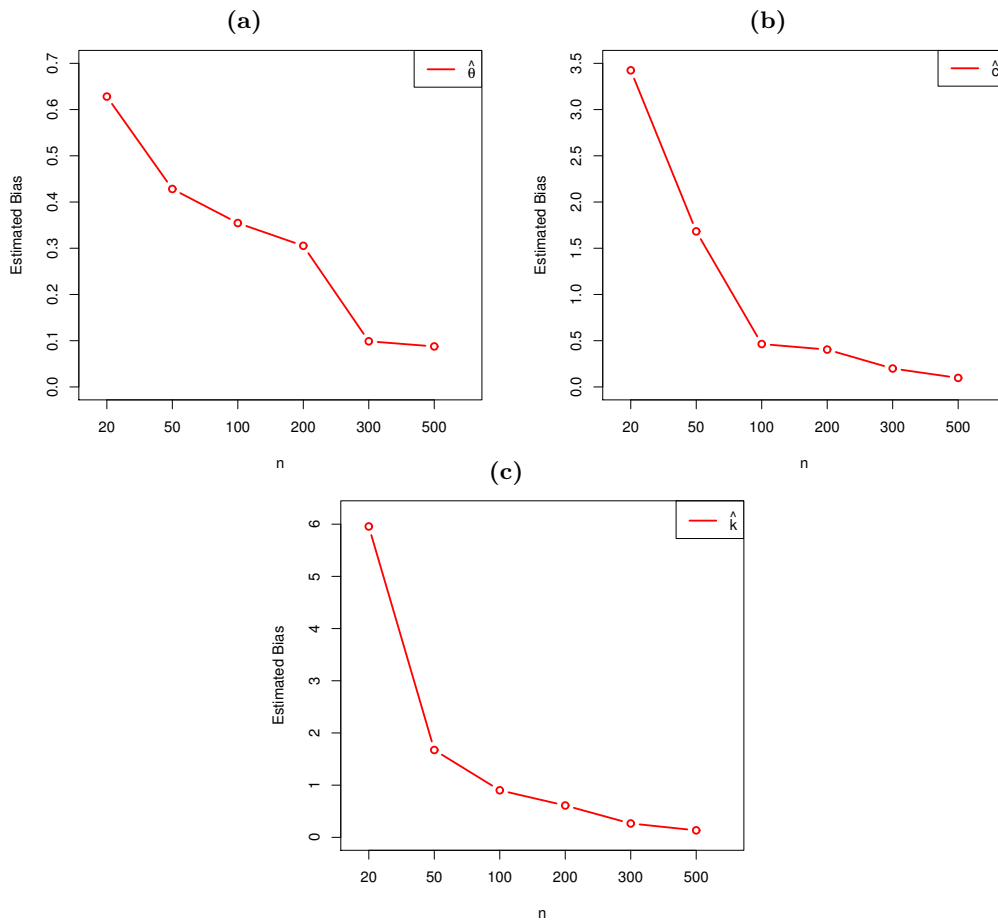


Figure 10: Estimated CPs for the selected parameters.

From Figures 9–11 we conclude that the estimated biases are positive for all parameters. The estimated biases decrease as the sample size n increases. Further, the estimated RMSEs are so closed to zero for large sample sizes. This result reveals the consistency property of the MLEs. The CP approaches to the nominal value (0.95) when the sample size increases.

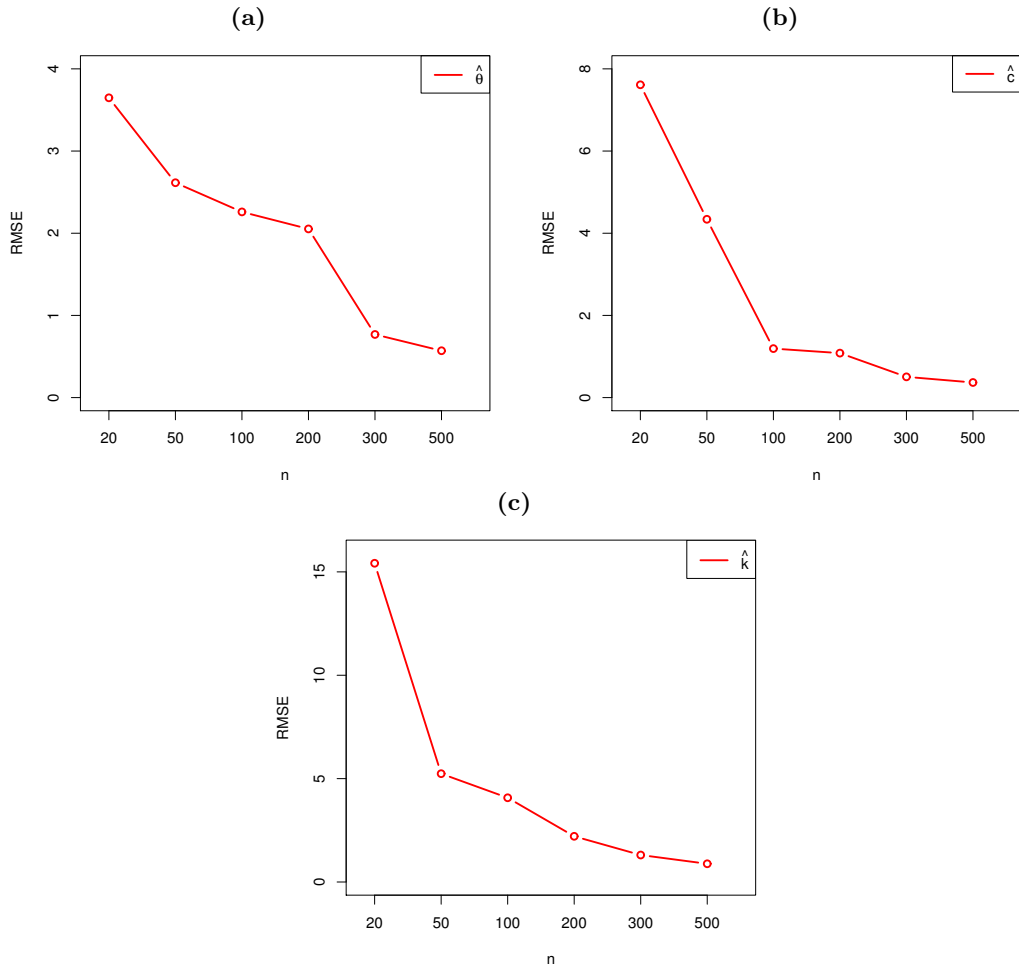


Figure 11: Estimated CPs for the selected parameters.

ACKNOWLEDGMENTS

The authors express their sincere thanks to the managing editor and the referee for careful reading of the article and providing valuable suggestions.

REFERENCES

- [1] ABAN, I.B.; MEERSCHAERT, M.M. and PANORSKA, A.K. (2006). Parameter estimation for the truncated Pareto distribution, *Journal of the American Statistical Association*, **101**, 270–277.
- [2] ALIZADEH, M.; CORDEIRO, GAUSS M.; NASCIMENTO, ABRAÃO D.C.; LIMA, MARIA DO CARMO S. and ORTEGA, EDWIN M.M. (2017). Odd-Burr generalized family of distributions with some applications, *Journal of Statistical Computation and Simulation*, **87**, 367–389.
- [3] ALIZADEH, M.; MEROVCI, F. and HAMEDANI, G.G. (2017). Generalized transmuted family of distributions: properties and applications, *Hacetepa Journal of Mathematics and Statistics*, **46**, 645–667.
- [4] AMINI, M.; MIRMOSTAFAEI, S.M.T.K. and AHMADI, J. (2014). Log-gamma-generated families of distributions, *Statistics*, **48**, 913–932.
- [5] BAKOUCH, H.S.; CHESNEAU, C. and KHAN, M.N. (2019). The extended odd family of probability distributions with practice to a submodel, *FILOMAT*, **33**, 3855–3867.
- [6] BURROUGHS, S.M. and TEBBENS, S.F. (2002). The upper-truncated power law applied to earthquake cumulative frequency-magnitude distributions, *Bulletin of the Seismological Society of America*, **92**, 2983–2993.
- [7] CORDEIRO, G.M. and DE CASTRO, M. (2011). A new family of generalized distributions, *Journal of Statistical Computation and Simulation*, **81**, 883–893.
- [8] EL-DEEB, A.M.H. (2015). *Weibull–Lomax distribution and its properties and applications*, A Thesis, Al-Azhar University – Gaza Deanship of Postgraduate Studies Faculty of Economics and Administrative Sciences Department of Statistics, page 113, unpublished thesis.
- [9] EUGENE, N.; LEE, C. and FAMOYE, F. (2002). Beta-normal distribution and its applications, *Communications in Statistics – Theory and Methods*, **31**, 497–512.
- [10] HINKLEY, D. (1977). On quick choice of power transformations, *Journal of the Royal Statistical, Series C*, **26**(1), 67–69.
- [11] JAMAL, F.; NASIR, M.A.; TAHIR, M.H. and MONTAZERI, N.H. (2017). The odd Burr-III family of distributions, *Journal of Statistics Applications and Probability*, **6**(1), 105–122.
- [12] NAJARZADEGAN, H.; ALAMATSAZ, M.H. and HAYATI, S. (2017). Truncated Weibull-G more flexible and more reliable than beta-G distribution, *International Journal of Statistics and Probability*, **6**(5), 1–17.
- [13] RISTIC, M.M. and BALAKRISHNAN, N. (2012). The gamma-exponentiated exponential distribution, *Journal of Statistical Computation and Simulation*, **82**, 1191–1206.
- [14] SABOOR, A.; BAKOUCH, H.S. and KHAN, M.N. (2016). Beta Sarhan–Zaindin modified Weibull distribution, *Applied Mathematical Modelling*, **40**, 6604–6621.
- [15] SARAN, J.; PUSHKARNA, N. and SEHGAL, S. (2019). Mid-truncated Burr XII distribution and its applications in order Statistics, *Statistics, Optimization and Information Computing*, **7**, 171–191.
- [16] SHAKED, M. and SHANTHIKUMAR, J.G. (1994). *Stochastic Orders and Their Applications*, Academic Press, New York.
- [17] ZANINETTI, L. and FERRARO, M. (2008). On the truncated Pareto distribution with application, *Central European Journal of Physics*, **6**, 1–6.
- [18] ZHANG, T. and XIE, M. (2011). On the upper truncated Weibull distribution and its reliability implications, *Reliability Engineering and System Safety*, **96**, 194–200.
- [19] ZOGRAFOS, K. and BALAKRISHNAN, N. (2009). On families of beta and generalized gamma-generated distributions and associated inference, *Statistical Methodology*, **6**, 342–362.

AN EFFICIENT MIXED RANDOMIZED RESPONSE MODEL FOR SENSITIVE CHARACTERISTIC IN SAMPLE SURVEY

Authors: AMOD KUMAR

– Mathematics and Computing, Indian Institute of Technology (Indian School of Mines),
Dhanbad-826004, India
amod.ism01@gmail.com

GAJENDRA K. VISHWAKARMA 

– Mathematics and Computing, Indian Institute of Technology (Indian School of Mines),
Dhanbad-826004, India
vishwagk@rediffmail.com

G.N. SINGH 

– Mathematics and Computing, Indian Institute of Technology (Indian School of Mines),
Dhanbad-826004, India
gnsingh_ism@yahoo.com

Received: April 2019

Revised: September 2019

Accepted: September 2019

Abstract:

- This paper proposes an efficient mixed randomized response (RR) model for estimating the proportion of individuals who possess to the sensitive attribute in the given population under both the conditions completely truthful reporting as well as less than completely truthful reporting and examined its properties. The proposed models are found to be dominant over Kim and Warde [13] model. It has also been extended for stratified random sampling. Numerical illustrations are presented to support the theoretical results.

Keywords:

- *randomized response technique; dichotomous; estimation of proportion; privacy; innocuous variable; sensitive characteristic.*

AMS Subject Classification:

- 62D05.

1. INTRODUCTION

In situations where potentially embarrassing or incriminating responses are sought, the randomized response technique (RRT) is effective in reducing non-sampling errors in sample surveys. In survey methodology, refusal to respond and lying are two major sources of non-sampling errors, as the stigma attached to certain practices (e.g. abortion and the use of illegal drugs) often leads to discrimination. Warner [28] did the pioneering work by suggesting a randomized response technique (RRT), which minimizes under reporting in survey data related to a socially undesirable or incriminating behaviour questions such as illegal earning or homosexuality among others. Warner [28] model requires the interviewee to give a “Yes” or “No” answers either to the sensitive question or to its negative depending on the outcome of a random device not reported to the interviewer. Further by introducing a choice of an unrelated question Greenberg *et al.* [7] modifying the Warner [28] randomized response model (RRM), the randomized response technique was further modified for different practical situation by Moors [17], Cochran [5], Fox and Tracy [6], Chaudhuri and Mukherjee [4], Hedayat and Sinha [8], Ryu *et al.* [19], Singh and Mangat [22], Tracy and Mangat [26], Tracy and Osahan [27], Singh [21], Singh and Tarray [23, 24, 25] and Kim and Warde [13] among others.

Kim and Warde [13] suggested a mixed randomized response model using simple random sampling with replacement which rectifies the privacy problem. Following the work of Kim and Warde [13], Amitava [1] and Hussain and Shabbir [10] suggested mixed randomized response technique (RRT) for complex survey designs and illustrated the superiority of their models over Kim and Warde [13] model.

Motivated with the above works, we have suggested a modified version of Kim and Warde [13] model and studied its properties in detail. We also present the less than completely truthful reporting counterpart of suggested model. It has been demonstrated that the suggested models perform better than the mixed randomized response model (RRM) of Kim and Warde [13]. We have also introduced the suggested model for stratified random sampling. The empirical studies are carried out; which showed dominance of proposed mixed randomized response models and stratified random sampling as well.

2. SUGGESTED MODEL

Let a sample of size n be selected from a finite population of size N using simple random sampling with replacement (SRSWR) scheme. Each respondent from the sample is instructed to answer the direct question “whether he/she is a member of the innocuous group?” If the answer to the initial direct question is “Yes” then he/she is instructed to go to the random device R_1 consisting of two statements:

- (i) “I am a member of the sensitive trait group”,
- (ii) “I am a member of the innocuous trait group”,

with probabilities P_1 and $(1 - P_1)$ respectively. If a respondent answers “No” to the direct

question, then the respondent is instructed to use the random device R_2 consisting of the statements on the first stage which is same as Mangat and Singh [16]:

- (i) “Do you possess the sensitive attribute A ”, with probability T ,
- (ii) “Go to the random device R_3 in the second stage”, with probability $(1 - T)$.

The respondents at the second stage are instructed to use the random device R_3 using three statements:

- (i) “I possess the sensitive attribute A ”,
- (ii) “Yes”,
- (iii) “No”,

with probabilities P , $(1 - P)/2$ and $(1 - P)/2$ respectively. When the outcome of random device R_3 is either (ii) or (iii), all the respondents, irrespective of whether they possess attribute A or not, are supposed to say “Yes” or “No” respectively. It is to be mentioned that the random device R_3 is due to Tracy and Osahan [27]. The survey procedures are performed under the assumption that both the sensitive and innocuous questions are unrelated and independent in a random device R_1 . To protect the respondents’ privacy, the respondents should not disclose to the interviewer the question they answered from either R_1 or R_2 or R_3 . Let n be the sample size confronted with a direct question and n_1 and $n_2 (= n - n_1)$ denote the number of “Yes” and “No” answers from the sample. Since all respondents using a random device R_1 already responded “Yes” from the initial direct question.

The probability ‘ Y ’ of getting “Yes” answers from the respondents using random device R_1 is given by

$$(2.1) \quad Y = P_1\pi_s + (1 - P_1)\pi_1,$$

where π_s is the proportion of “Yes” answer from the sensitive trait group and π_1 is the proportion of “Yes” so that $(\pi_1 = 1)$ answer from the innocuous question

$$(2.2) \quad Y = P_1\pi_s + (1 - P_1).$$

The probability ‘ Y^* ’ of getting “No” answers from the respondents using random device R_1 is given as

$$(2.3) \quad Y^* = 1 - [P_1\pi_s + (1 - P_1)].$$

Thus the maximum likelihood function is given by

$$(2.4) \quad L = \binom{n}{n_1} [P_1\pi_s + (1 - P_1)]^{n_1} [P_1(1 - \pi_s)]^{(n - n_1)}.$$

Taking log on the both sides of equation (2.4):

$$(2.5) \quad \log L = \log \binom{n}{n_1} + n_1 \log [P_1\pi_s + (1 - P_1)] + (n - n_1) \log [P_1(1 - \pi_s)].$$

Differentiating on both sides of equation (2.5) with respect to π_s and equating to zero, we have

$$(2.6) \quad P_1 \pi_s + (1 - P_1) = \frac{n_1}{n}.$$

This is maximum likelihood estimator of Y .

An unbiased estimator of π_s , in terms of the sample proportion of “Yes” responses \hat{Y} , becomes

$$(2.7) \quad \hat{\pi}_a = \frac{\hat{Y} - (1 - P_1)}{P_1},$$

where \hat{Y} is the sample proportion of “Yes” response, thus expected value of $\hat{\pi}_a$ is

$$(2.8) \quad E(\hat{\pi}_a) = \frac{E(\hat{Y}) - (1 - P_1)}{P_1} = \pi_s.$$

The variance of $\hat{\pi}_a$ is obtained as

$$(2.9) \quad V(\hat{\pi}_a) = \frac{1}{n_1} \left[\pi_s(1 - \pi_s) + \frac{(1 - \pi_s)(1 - P_1)}{P_1} \right].$$

The probability X of “Yes” answers from the respondents using random devices R_2 and R_3 is given as

$$(2.10) \quad X = T \pi_s + (1 - T) \left[P \pi_s + \frac{(1 - P)}{2} \right].$$

An unbiased estimator of π_s , in terms of the sample proportion of “Yes” responses \hat{X} , is given by

$$(2.11) \quad \hat{\pi}_b = \frac{\hat{X} - (1 - T) \frac{(1 - P)}{2}}{T + P(1 - T)}.$$

The variance of unbiased estimator $\hat{\pi}_b$ is obtained as

$$(2.12) \quad V(\hat{\pi}_b) = \frac{1}{n_2} \left[\pi_s(1 - \pi_s) + \frac{(1 - T)(1 - P)[2 - (1 - T)(1 - P)]}{4[T + P(1 - T)]^2} \right].$$

The estimator of π_s , in the terms of the sample proportion of “Yes” response $\hat{\pi}_a$ and $\hat{\pi}_b$, is

$$(2.13) \quad \hat{\pi}_{A1} = \left(\frac{n_1}{n} \right) \hat{\pi}_a + \left(\frac{n_2}{n} \right) \hat{\pi}_b, \quad \text{for } 0 < \frac{n_1}{n} < 1.$$

Since $\hat{\pi}_a$ and $\hat{\pi}_b$ are unbiased estimators, therefore the expected value of $\hat{\pi}_{A1}$ is

$$(2.14) \quad E(\hat{\pi}_{A1}) = \frac{n_1}{n} E(\hat{\pi}_a) + \frac{(n - n_1)}{n} E(\hat{\pi}_b) = \frac{n_1}{n} \pi_s + \frac{(n - n_1)}{n} \pi_s = \pi_s.$$

Thus, the proposed estimator $\hat{\pi}_{A1}$ is an unbiased estimator of π_s .

Since the random device R_1 and Tracy and Osahan [27] randomized response technique (consists of two random devices R_2 and R_3) used are independent. We derive the expression of variance of $\hat{\pi}_{A1}$ as

$$\begin{aligned}
 (2.15) \quad V(\hat{\pi}_{A1}) &= \frac{n_1^2}{n^2} V(\hat{\pi}_a) + \frac{n_2^2}{n^2} V(\hat{\pi}_b) \\
 &= \frac{n_1}{n^2} \left[\frac{(1 - \pi_s) [P_1 \pi_s + (1 - P_1)]}{P_1} \right] \\
 &\quad + \frac{n_2}{n^2} \left[\pi_s (1 - \pi_s) + \frac{(1 - T)(1 - P) [2 - (1 - T)(1 - P)]}{4 [T + P(1 - T)]^2} \right].
 \end{aligned}$$

Under the circumstances that the Warner [28] and Simmons *et al.* [20] method (known π_1) are equally confidential to respondents, Lanke [14] obtain a unique value of P as $P = 1/2 + P_1/[2P_1 + 4(1 - P_1)\pi_1]$, for every P_1 and every π_1 .

Since our proposed mixed model also use Simmons *et al.* [20] method when $\pi_1 = 1$, we may apply Lanke [14] technique in our proposed model. Thus we get

$$(2.16) \quad P = \frac{1}{(2 - P_1)}.$$

Putting $P = 1/(2 - P_1)$ in equation (2.12), we get

$$\begin{aligned}
 (2.17) \quad V(\hat{\pi}_b) &= \frac{1}{n_2} \left[\pi_s (1 - \pi_s) + \frac{(1 - T) \left(1 - \frac{1}{(1 - 2P_1)}\right) \left[2 - (1 - T) \left(1 - \frac{1}{(2 - P_1)}\right)\right]}{4 \left[T + \frac{1}{(2 - P_1)} (1 - T)\right]^2} \right] \\
 &= \left[\frac{\pi_s (1 - \pi_s)}{n_2} + \frac{(1 - T)(1 - P_1) [2(2 - P_1) - (1 - T)(1 - P_1)]}{4 n_2 [1 + T(1 - P_1)]^2} \right].
 \end{aligned}$$

Thus, we have the following theorem.

Theorem 2.1. *The variance of $\hat{\pi}_{A1}$ is given by*

$$\begin{aligned}
 (2.18) \quad V(\hat{\pi}_{A1}) &= \frac{\pi_s(1 - \pi_s)}{n} \\
 &\quad + \frac{1}{n} \left[\frac{\lambda(1 - \pi_s)(1 - P_1)}{P_1} + \frac{(1 - \lambda)(1 - T)(1 - P_1) [2(2 - P_1) - (1 - T)(1 - P_1)]}{4 [1 + T(1 - P_1)]^2} \right],
 \end{aligned}$$

for $n = n_1 + n_2$ and $\lambda = n_1/n$.

2.1. Efficiency comparison

In this section, the comparison of the proposed model under completely truthful reporting case has been made with Kim and Warde [13] model.

From Kim and Warde [13] model, we have

$$(2.19) \quad V(\hat{\pi}_{kw}) = \frac{\pi_s(1 - \pi_s)}{n} + \frac{(1 - P_1) [\lambda P_1(1 - \pi_s) + (1 - \lambda)]}{n P_1^2}.$$

The estimator $\hat{\pi}_{A1}$ is always more efficient than that of Kim and Warde [13] estimator $\hat{\pi}_{kw}$ if

$$V(\hat{\pi}_{kw}) > V(\hat{\pi}_{A1}),$$

which gives the conditions, when

$$\left[4(1 + T(1 - P_1))^2 - P_1^2(1 - T)(3 + T(1 - P_1) - P_1)\right] > 0.$$

To have a tangible idea about the performance of the proposed estimator $\hat{\pi}_{A1}$ over Kim and Warde [13] estimator $\hat{\pi}_{kw}$, we compute the percent relative efficiency $PRE(\hat{\pi}_{A1}, \hat{\pi}_{kw})$ for $\lambda = (0.7, 0.5, 0.3)$, $n = 1000$ and for different values of T , π_s , n_1 , n_2 and P_1 , and presented in Table 1:

$$(2.20) \quad PRE(\hat{\pi}_{A1}, \hat{\pi}_{kw}) = \frac{V(\hat{\pi}_{kw})}{V(\hat{\pi}_{A1})} \times 100.$$

Table 1: Percent relative efficiency of the proposed estimator $\hat{\pi}_{A1}$ with respect to Kim and Warde [13] estimator $\hat{\pi}_{kw}$.

π_s	$n = 1000$		λ	T	P_1								
	n_1	n_2			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	700	300	0.7	0.1	554.04	313.85	232.53	191.00	165.40	147.71	134.41	123.50	113.26
	500	500	0.5	0.5	1161.80	603.05	414.50	318.32	258.82	217.29	185.45	158.55	132.55
	300	700	0.3	0.9	2581.70	1278.40	838.29	613.08	472.76	373.61	296.28	230.14	167.51
0.2	700	300	0.7	0.1	601.54	331.57	240.35	193.90	165.40	145.84	131.31	119.75	109.77
	500	500	0.5	0.5	1266.60	638.56	427.09	319.64	253.65	208.18	174.11	146.69	122.86
	300	700	0.3	0.9	2794.00	1330.20	838.29	589.05	436.59	332.21	254.93	194.15	143.81
0.3	700	300	0.7	0.1	662.49	354.70	251.19	198.87	167.08	145.58	129.96	117.97	108.28
	500	500	0.5	0.5	1401.90	686.31	446.91	326.51	253.65	204.49	168.75	141.24	118.97
	300	700	0.3	0.9	3073.90	1410.90	858.93	584.52	421.02	312.88	236.26	179.24	135.16
0.4	700	300	0.7	0.1	743.37	385.71	266.30	206.54	170.71	146.89	129.96	117.37	107.67
	500	500	0.5	0.5	1567.20	739.28	466.33	331.77	252.32	200.28	163.76	136.86	116.27
	300	700	0.3	0.9	3454.00	1531.80	903.88	598.39	421.02	307.13	229.09	173.12	131.60
0.5	700	300	0.7	0.1	855.64	428.97	287.89	218.17	176.97	150.05	131.31	117.72	107.61
	500	500	0.5	0.5	1834.20	844.82	520.90	362.66	270.27	210.52	169.25	139.40	117.08
	300	700	0.3	0.9	3993.00	1713.50	982.27	634.07	436.59	312.88	230.23	172.49	130.82

It is observed from Table 1 and Figure 1 that:

- (a) For all the parametric combinations, the values of percent relative efficiencies are substantially exceeding 100, which indicate that the proposed estimator $\hat{\pi}_{A1}$ is uniformly better than Kim and Warde [13] estimator $\hat{\pi}_{kw}$.
- (b) It may also be seen that the values of the percent relative efficiencies decrease with the increasing values of P_1 . However, the values of the percent relative efficiencies are showing increasing trend with the decreasing values of λ when the values of P_1 are fixed.
- (c) From Figure 1 it may be observed that there is a large gain in efficiency by using the proposed estimator $\hat{\pi}_{A1}$ over Kim and Warde [13] estimator $\hat{\pi}_{kw}$, when the proportion of stigmatizing attribute is moderately large.

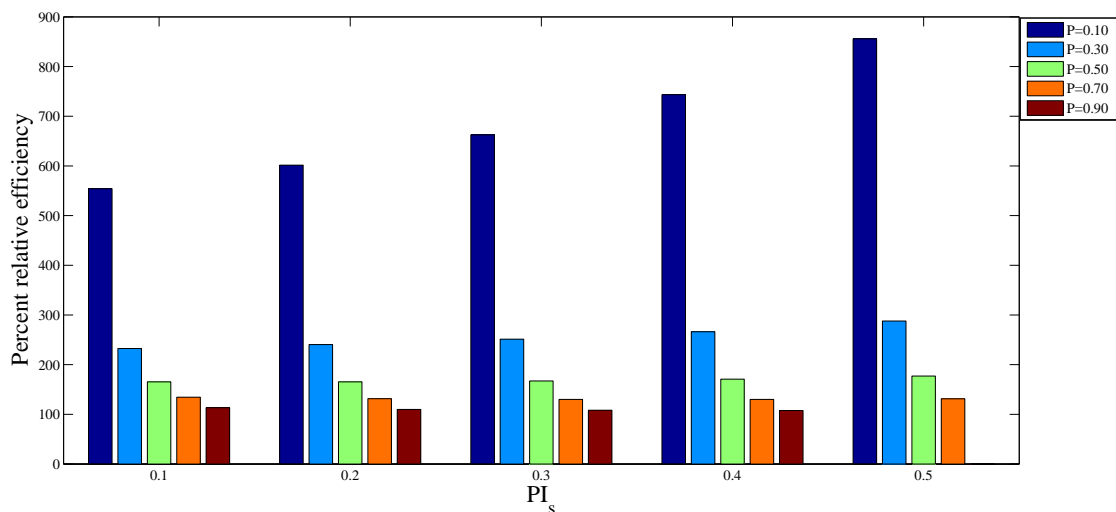


Figure 1: Percent relative efficiency of the proposed estimator $\hat{\pi}_{A1}$ with respect to Kim and Warde [13] estimator $\hat{\pi}_{kw}$ when $T = 0.1$ and $\lambda = 0.7$.

3. LESS THAN COMPLETELY TRUTHFUL REPORTING

Various authors including Mangat [15], Tracy and Osahan [27], Chang and Huang [2], Chang *et al.* [3], Kim and Warde [12], Kim and Elam [11], Nazuk and Shabbir [18] and cited therein has been consider the problem of “Less than completely truthful reporting” in RR technique. It is reasonably assumed that the persons who belong to sensitive trait group state truthful answers with probabilities T_1 , T_2 and T_3 in random devices R_1 , R_2 and R_3 respectively. The respondents in the non-sensitive group have no reason to tell a lie, they may lie for the sensitive group.

Since all respondents using a random device R_1 already responded “Yes” from the initial direct question, therefore $\pi_1 = 1$ in R_1 . Thus, the probability Y' of “Yes” answer for the random device R_1 is given by

$$(3.1) \quad Y' = P_1 \pi_s T_1 + (1 - P_1).$$

An estimator of π_s , in term of the sample proportion of “Yes” responses is given as

$$(3.2) \quad \hat{\pi}_{a(1)} = \frac{\hat{Y}' - (1 - P_1)}{P_1}.$$

Since each \hat{Y}' follows Binomial distribution $B(n_1, Y')$, therefore the estimator $\hat{\pi}_{a(1)}$ has the following bias and mean square error (MSE):

$$(3.3) \quad B(\hat{\pi}_{a(1)}) = \pi_s(T_1 - 1)$$

and

$$(3.4) \quad V(\hat{\pi}_{a(1)}) = \frac{Y'(1 - Y')}{n_1 P_1^2} = \frac{(1 - \pi_s T_1) [1 - P_1(1 - \pi_s T_1)]}{n_1 P_1}.$$

Thus, the MSE of $\hat{\pi}_{a(1)}$ is given by

$$(3.5) \quad \begin{aligned} \text{MSE}(\hat{\pi}_{a(1)}) &= V(\hat{\pi}_{a(1)}) + [B(\hat{\pi}_{a(1)})]^2 \\ &= \frac{(1 - \pi_s T_1) [1 - P_1(1 - \pi_s T_1)]}{n_1 P_1} + \pi_s^2 (T_1 - 1)^2. \end{aligned}$$

On the basis of the proposed procedure, the probability for the respondents who response “Yes” answer using random devices R_2 and R_3 is given by

$$(3.6) \quad X' = T \pi_s T_2 + (1 - T) \left[P \pi_s T_3 + \frac{(1 - P)}{2} \right].$$

By the method of moments, an estimator of population proportion π_s is obtained as

$$(3.7) \quad \hat{\pi}_{b(1)} = \frac{\hat{X}' - (1 - T) \frac{(1 - P)}{2}}{T + P(1 - T)}.$$

In random devices R_2 and R_3 , the same sensitive question is asked from the respondents who belong to rare sensitive group in the sample, so we take $T_2 = T_3$ in our case which is unlike as in case of Kim and Elam [11].

Since each \hat{X}' follows Binomial distribution $B(n_1, X')$, therefore the estimator $\hat{\pi}_{b(1)}$ has the following bias and MSE:

$$(3.8) \quad B(\hat{\pi}_{b(1)}) = \pi_s(T_2 - 1)$$

and

$$(3.9) \quad \begin{aligned} V(\hat{\pi}_{b(1)}) &= \frac{X'(1 - X')}{n_2 [T + P(1 - T)]^2} \\ &= \frac{\pi_s T_2 (1 - \pi_s T_2)}{n_2} + \frac{(1 - T) (1 - P) [2 - (1 - T) (1 - P)]}{4 n_2 [T + P(1 - T)]^2}. \end{aligned}$$

Therefore, the MSE of $\hat{\pi}_{b(1)}$ is given by

$$(3.10) \quad \begin{aligned} \text{MSE}(\hat{\pi}_{b(1)}) &= V(\hat{\pi}_{b(1)}) + [B(\hat{\pi}_{b(1)})]^2 \\ &= \frac{\pi_s T_2 (1 - \pi_s T_2)}{n_2} + \frac{(1 - T) (1 - P) [2 - (1 - T) (1 - P)]}{4 n_2 [T + P(1 - T)]^2} + \pi_s^2 (T_2 - 1)^2. \end{aligned}$$

Now, we propose the estimator for population proportion π_s in the terms of the sample proportion of “Yes” response $\hat{\pi}_{a(1)}$ and $\hat{\pi}_{b(1)}$ as

$$(3.11) \quad \hat{\pi}_A = \left(\frac{n_1}{n}\right) \hat{\pi}_{a(1)} + \left(\frac{n_2}{n}\right) \hat{\pi}_{b(1)} \quad \text{for } 0 < \frac{n_1}{n} < 1,$$

where $n_1 + n_2 = 1$.

Since both the estimators $\hat{\pi}_{a(1)}$ and $\hat{\pi}_{b(1)}$ are bias estimator of π_s , therefore the bias of $\hat{\pi}_A$ is given by

$$(3.12) \quad B(\hat{\pi}_A) = \pi_s \left[\left(\frac{n_1}{n}\right) (T_1 - 1) + \left(\frac{n_2}{n}\right) (T_2 - 1) \right],$$

and

$$(3.13) \quad \begin{aligned} \text{MSE}(\hat{\pi}_A) = & \frac{\lambda(1 - \pi_s T_1) [1 - P_1(1 - \pi_s T_1)]}{nP_1} \\ & + \frac{(1 - \lambda)}{n} \left[\pi_s T_2 (1 - \pi_s T_2) + \frac{(1 - T)(1 - P) [2 - (1 - T)(1 - P)]}{4[T + P(1 - T)]^2} \right] \\ & + \pi_s^2 [\lambda^2(T_1 - 1)^2 + (1 - \lambda)^2(T_2 - 1)^2]. \end{aligned}$$

Inserting Lanke [14] a unique value $P = 1/(2 - P_1)$ in equation (3.10), we get

$$(3.14) \quad \begin{aligned} \text{MSE}(\hat{\pi}_b(1)) = & \frac{\pi_s T_2 (1 - \pi_s T_2)}{n_2} \\ & + \frac{(1 - T)(1 - P_1) [2(2 - P_1) - (1 - T)(1 - P_1)]}{4n_2 [1 + T(1 - P_1)]^2} + \pi_s^2 (T_2 - 1)^2. \end{aligned}$$

Thus, we have the following theorem.

Theorem 3.1. *The MSE of $\hat{\pi}_A$ is given by*

$$(3.15) \quad \begin{aligned} \text{MSE}(\hat{\pi}_A) = & \frac{\pi_s [\lambda T_1 (1 - \pi_s T_1) + (1 - \lambda) T_2 (1 - \pi_s T_2)]}{n} \\ & + \frac{(1 - P_1)}{n} \left[\frac{\lambda(1 - \pi_s T_1)}{P_1} + \frac{(1 - \lambda)(1 - T) [2(2 - P_1) - (1 - T)(1 - P_1)]}{4 [1 + T(1 - P_1)]^2} \right] \\ & + \pi_s^2 [\lambda^2(T_1 - 1)^2 + (1 - \lambda)^2(T_2 - 1)^2], \end{aligned}$$

for $n = n_1 + n_2$ and $\lambda = n_1/n$.

3.1. Efficiency comparison

We compare the proposed model with Kim and Warde [13] model, under “Less than completely truthful reporting” situation.

The MSE of Kim and Warde [13] estimator $\hat{\pi}_{kw}$ under less than completely truthful reporting is given as

$$\begin{aligned}
 \text{MSE}(\hat{\pi}_{kw}) = & \frac{\pi_s [\lambda T_1(1 - \pi_s T_1) + (1 - \lambda) T_2(1 - \pi_s T_2)]}{n} \\
 (3.16) \quad & + \frac{(1 - P_1) [\lambda P_1(1 - \pi_s T_1) + (1 - \lambda)]}{nP_1^2} \\
 & + \pi_s^2 [\lambda^2 (T_1 - 1)^2 + (1 - \lambda)^2 (T_2 - 1)^2].
 \end{aligned}$$

The estimator $\hat{\pi}_A$ is always more efficient than that of Kim and Warde [13] estimator $\hat{\pi}_{kw}$ if

$$\text{MSE}(\hat{\pi}_{kw}) > \text{MSE}(\hat{\pi}_A),$$

which is true if

$$(3.17) \quad \left[\frac{(1 - T) [2(2 - P_1) - (1 - T)(1 - P_1)]}{4 [1 + T(1 - P_1)]^2} - \frac{1}{P_1^2} \right] > 0.$$

To have an idea about the magnitude of the percent relative efficiency of the proposed model in relation to Kim and Warde [13] model, we resort to an empirical investigation for $\lambda = (0.7, 0.5, 0.3)$, $n = 1000$, $T_1(T_2) = 0.7, 0.8, 0.9$ (0.6, 0.7, 0.8) and for different values of T , π_s , n_1 , n_2 and P_1 . The percent relative efficiency of the proposed estimator $\hat{\pi}_A$ with respect to Kim and Warde [13] estimator $\hat{\pi}_{kw}$ is defined as

$$(3.18) \quad \text{PRE}(\hat{\pi}_A, \hat{\pi}_{kw}) = \frac{\text{MSE}(\hat{\pi}_{kw})}{\text{MSE}(\hat{\pi}_A)} \times 100.$$

The following interpretations may be read out from Table 2 and Figure 2:

- (a) For all the parametric combinations, the values of percent relative efficiencies are substantially exceeding 100, which indicate that the proposed estimator $\hat{\pi}_A$ is uniformly better than Kim and Warde [13] estimator $\hat{\pi}_{kw}$.
- (b) Table 2 makes it visible that the values of percent relative efficiencies decrease with the increasing values of P_1 . Further, we observe that the percent relative efficiencies increase with the decreasing values of λ (and increasing values of T_1, T_2) when the values of P_1 are fixed.
- (c) It may also be seen that with the increase in the values of π_s there is the decreasing pattern in values of the percent relative efficiencies for fix values of P_1 .
- (d) From Figure 2 it is clear that there is less gain in the efficiency by using the proposed estimator $\hat{\pi}_A$ over Kim and Warde [13] estimator $\hat{\pi}_{kw}$, when the proportion of sensitive attribute is moderately large.

Table 2: Percent relative efficiency of the proposed estimator $\hat{\pi}_A$ with respect to Kim and Warde [13] estimator $\hat{\pi}_{kw}$ under the situation of “Less than completely truthful reporting”.

π_s	$n = 1000$		λ	T	T_1	T_2	P_1						
	n_1	n_2					0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	700	300	0.7	0.1	0.7	0.6	197.29	159.83	137.91	123.84	114.33	107.74	103.14
	500	500	0.5	0.5	0.8	0.7	347.62	257.62	203.92	168.40	143.33	124.85	110.82
	300	700	0.3	0.9	0.9	0.8	681.61	470.63	345.02	262.04	203.38	159.88	126.44
0.2	700	300	0.7	0.1	0.7	0.6	155.82	129.78	116.64	109.33	105.04	102.46	100.90
	500	500	0.5	0.5	0.8	0.7	253.81	185.33	149.67	129.12	116.52	108.51	103.33
	300	700	0.3	0.9	0.9	0.8	450.78	294.67	214.06	167.73	139.19	120.74	108.41
0.3	700	300	0.7	0.1	0.7	0.6	132.48	116.20	108.61	104.65	102.43	101.16	100.41
	500	500	0.5	0.5	0.8	0.7	194.00	148.49	126.73	115.02	108.23	104.11	101.57
	300	700	0.3	0.9	0.9	0.8	312.22	209.97	161.31	135.06	119.69	110.16	104.03
0.4	700	300	0.7	0.1	0.7	0.6	120.45	109.88	105.14	102.73	101.41	100.66	100.23
	500	500	0.5	0.5	0.8	0.7	160.79	130.25	116.28	108.98	104.85	102.40	100.90
	300	700	0.3	0.9	0.9	0.8	236.93	168.65	137.44	121.08	111.69	105.98	102.35
0.5	700	300	0.7	0.1	0.7	0.6	113.84	106.58	103.39	101.79	100.92	100.43	100.15
	500	500	0.5	0.5	0.8	0.7	141.78	120.40	110.84	105.93	103.18	101.56	100.58
	300	700	0.3	0.9	0.9	0.8	194.13	146.38	125.02	113.98	107.71	103.92	101.54

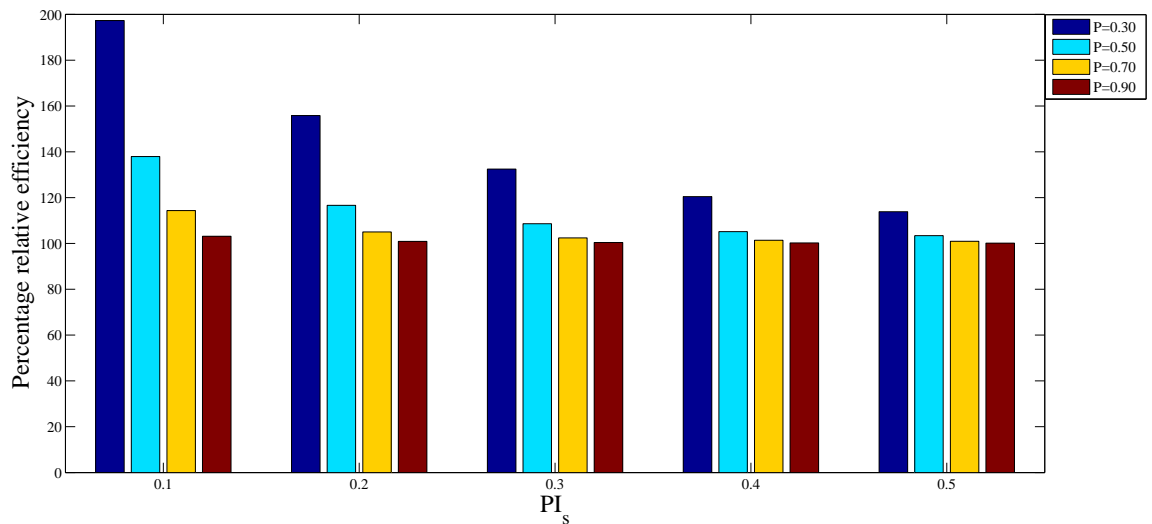


Figure 2: Percent relative efficiency of the proposed estimator $\hat{\pi}_A$ with respect to Kim and Warde [13] estimator $\hat{\pi}_{kw}$ under the condition of “Less than completely truthful reporting”, when $T = 0.1$ and $\lambda = 0.7$.

4. A MIX RANDOMIZED RESPONSE MODEL USING STRATIFICATION

4.1. A mixed Stratified randomized response (RR) model

Stratified random sampling is generally obtained by dividing the population into non-overlapping groups called strata and selecting a simple random sample from each stratum. The main advantage of the stratified random sampling is that the technique overcomes the limitation of the loss of individual characteristics of the respondents. A randomized response (RR) technique using stratified random sampling yields the group characteristics associated to each stratum estimator. Also, stratified random sampling protects a researcher from the possibility of obtaining a poor sample. Hong *et al.* [9] suggested a stratified RR technique using a proportional allocation. Kim and Warde [12] proposed a stratified randomized response model using an optimum allocation which is more efficient than that using a proportional allocation. Kim and Elam [11] suggested a two stage stratified Warner's RR model using optimal allocation. Further Kim and Warde [13] suggested a mixed stratified RR model.

In the proposed models, we assume that the population is partitioned into strata, and a sample is selected by using simple random sampling with replacement (SRSWR) scheme from each stratum. To get the full benefit from stratification, we assume that the number of units in each stratum is known. An individual respondent in a sample from each stratum is instructed to answer a direct question "I am a member of the innocuous trait group". Respondents reply the direct question by "Yes" or "No". If a respondent answers "Yes", then the respondent is instructed to go to the random device R_{k1} consisting of statements:

- (i) "I am a member of the sensitive trait group",
- (ii) "I am a member of the innocuous trait group",

with probabilities Q_k and $(1 - Q_k)$ respectively. If a respondent answers "No", then the respondent is instructed to use the random device R_{k2} consisting of two statements (see Mangat and Singh [16]):

- (i) "Do you possess the sensitive attribute A ", with probability T_k ,
- (ii) "Go to the third random device R_{k3} in the second stage", with probability $(1 - T_k)$.

The random device R_{k3} at the second stage consists of three statements:

- (i) "I possess the sensitive attribute A ",
- (ii) "Yes",
- (iii) "No",

with probabilities P_k , $(1 - P_k)/2$ and $(1 - P_k)/2$. When the outcome of random device R_{k3} is either (ii) or (iii), all the respondents, irrespective of whether they possess attribute A or not, are supposed to say "Yes" or "No" respectively. To protect the respondent's privacy, the respondents should not disclose to the interviewer the question they answered from either R_{k1} or R_{k2} or R_{k3} . Let m_k denote the number of units in the sample from stratum k and n as the total number of units in samples from all strata. Let m_{k1} be the number of people

responding “Yes” when respondents in a sample m_k were asked the direct question and m_{k2} be the number of people responding “No” when respondents in a sample m_k were asked the direct question, so that $n = \sum_{k=1}^r m_k = \sum_{k=1}^r (m_{k1} + m_{k2})$. Under the assumption that these “Yes” or “No” reports are made truthfully and Q_k and P_k are set by researcher. Thus, the probability Y_k of “Yes” answer from the respondents using the random device R_{k1} is given by

$$(4.1) \quad Y_k = Q_k \pi_{sk} + (1 - Q_k) \pi_{1k} \quad \text{for } k = 1, 2, \dots, r,$$

where π_{sk} is the proportion of respondents with the sensitive trait in stratum k , π_{1k} is the proportion of respondents with the innocuous trait group in stratum k .

Since the respondent performing a random device R_{k1} answered “Yes” to the direct question of the innocuous trait, if the respondent selects the same innocuous question from R_{k1} , then $\pi_{1k} = 1$ (see Kim and Warde [13]). Therefore, equation (4.1) becomes

$$(4.2) \quad Y_k = Q_k \pi_{sk} + (1 - Q_k) \quad \text{for } k = 1, 2, \dots, r.$$

An unbiased estimator of π_{sk} is given as

$$(4.3) \quad \hat{\pi}_{a1k} = \frac{\hat{Y}_k - (1 - Q_k)}{Q_k} \quad \text{for } k = 1, 2, \dots, r,$$

where \hat{Y}_k is the proportion of “Yes” answer in a sample in stratum k . Since each \hat{Y}_k follows Binomial distribution i.e. $\hat{Y}_k \sim B(m_{k1}, Y_k)$.

The variance of unbiased estimator $\hat{\pi}_{a1k}$ is given by

$$(4.4) \quad V(\hat{\pi}_{a1k}) = \frac{(1 - \pi_{sk}) [Q_k \pi_{sk} + (1 - Q_k)]}{m_{k1} Q_k}.$$

The probability X_k of “Yes” answer from the respondents using random devices R_{k2} and R_{k3} will be

$$(4.5) \quad X_k = T_k \pi_{sk} + (1 - T_k) \left[P_k \pi_{sk} + \frac{(1 - P_k)}{2} \right],$$

where π_{sk} is the proportion of respondents with the sensitive treat in stratum k .

An unbiased estimator of π_{sk} is given by

$$(4.6) \quad \hat{\pi}_{b1k} = \frac{\hat{X}_k - (1 - T_k) \frac{(1 - P_k)}{2}}{T_k + P_k (1 - T_k)},$$

where \hat{X}_k is the proportion of “Yes” responses in a sample from a stratum k . Since each \hat{X}_k follows Binomial distribution i.e. $\hat{X}_k \sim B(m_k, X_k)$. By using $m_k = m_{k1} + m_{k2}$ and $P_k = (2 - Q_k)^{-1}$ (see Lanke [14]), the variance of estimator $\hat{\pi}_{b1k}$ is given by

$$(4.7) \quad V(\hat{\pi}_{b1k}) = \left[\frac{\pi_{sk}(1 - \pi_{sk})}{m_{k2}} + \frac{(1 - T_k)(1 - Q_k) [2(2 - Q_k) - (1 - T_k)(1 - Q_k)]}{4 m_2 [1 + T_k(1 - Q_k)]^2} \right].$$

Now, we develop the unbiased estimator of π_{sk} , in terms of sample proportion of “Yes” responses \hat{Y}_k and \hat{X}_k , as

$$(4.8) \quad \hat{\pi}_{msk} = \left(\frac{m_{k1}}{m_k}\right) \hat{\pi}_{a1k} + \left(\frac{m_{k2}}{m_k}\right) \hat{\pi}_{b1k} \quad \text{for } 0 < \frac{m_{k1}}{m_k} < 1.$$

The variance of the estimator $\hat{\pi}_{msk}$ is given by

$$(4.9) \quad V(\hat{\pi}_{msk}) = \left[\frac{\pi_{sk}(1-\pi_{sk})}{m_k} + \frac{\lambda_k(1-\pi_{sk})(1-Q_k)}{m_k Q_k} + \frac{(1-\lambda_k)(1-T_k)(1-Q_k)[2(2-Q_k) - (1-T_k)(1-Q_k)]}{4m_k[1+T_k(1-Q_k)]^2} \right],$$

where $m_k = m_{k1} + m_{k2}$ and $\lambda_k = m_{k1}/m_k$.

Thus, the unbiased estimator of $\pi_s = \sum_{k=1}^r w_k \pi_{sk}$ is obtained as

$$(4.10) \quad \hat{\pi}_{Ak} = \sum_{k=1}^r w_k \hat{\pi}_{msk} = \sum_{k=1}^r w_k \left[\frac{m_{k1}}{m_k} \hat{\pi}_{a1k} + \frac{m_{k2}}{m_k} \hat{\pi}_{b1k} \right],$$

where N is the number of units in the whole population, N_k is the total number of units in stratum k , and $w_k = \frac{N_k}{N}$ for $k = 1, 2, \dots, r$ so that $w = \sum_{k=1}^r w_k = 1$. It can be shown that the proposed estimator $\hat{\pi}_{Ak}$ is unbiased for π_s . The variance of $\hat{\pi}_{Ak}$ is given by

$$(4.11) \quad V(\hat{\pi}_{Ak}) = \sum_{k=1}^r \frac{w_k^2}{m_k} \left[\pi_{sk}(1-\pi_{sk}) + \frac{\lambda_k(1-\pi_{sk})(1-Q_k)}{Q_k} + \frac{(1-\lambda_k)(1-T_k)(1-Q_k)[2(2-Q_k) - (1-T_k)(1-Q_k)]}{4[1+T_k(1-Q_k)]^2} \right].$$

Here, the requirement of doing the optimal allocation of a sample size n , we need to know $\lambda_k = m_{k1}/m_k$ and π_{sk} . In practice the information on $\lambda_k = m_{k1}/m_k$ and π_{sk} is usually unavailable. But if prior information about $\lambda_k = m_{k1}/m_k$ and π_{sk} is available from past experience, it will help to derive the following optimal allocation formula.

Theorem 4.1. *The optimum allocation of m to m_1, m_2, \dots, m_{r-1} and m_r to derive the minimum variance of the $\hat{\pi}_{Ak}$ subject to $n = \sum_{k=1}^r m_k$ is approximately given by*

$$(4.12) \quad \frac{m_k}{n} = \frac{A}{B},$$

where

$$A = w_k \left[\pi_{sk}(1-\pi_{sk}) + \frac{\lambda_k(1-\pi_{sk})(1-Q_k)}{Q_k} + \frac{(1-\lambda_k)(1-T_k)(1-Q_k)[2(2-Q_k) - (1-T_k)(1-Q_k)]}{4[1+T_k(1-Q_k)]^2} \right]^{\frac{1}{2}},$$

$$B = \sum_{k=1}^r w_k \left[\pi_{sk}(1-\pi_{sk}) + \frac{\lambda_k(1-\pi_{sk})(1-Q_k)}{Q_k} + \frac{(1-\lambda_k)(1-T_k)(1-Q_k)[2(2-Q_k) - (1-T_k)(1-Q_k)]}{4[1+T_k(1-Q_k)]^2} \right]^{\frac{1}{2}}.$$

Thus, the minimal variance of the estimator $\hat{\pi}_{Ak}$ is given by

$$\begin{aligned}
 (4.13) \quad V(\hat{\pi}_{Ak}) = & \frac{1}{n} \left[\sum_{k=1}^r w_k \left[\pi_{sk}(1 - \pi_{sk}) + \frac{\lambda_k(1 - \pi_{sk})(1 - Q_k)}{Q_k} \right. \right. \\
 & \left. \left. + \frac{(1 - \lambda_k)(1 - T_k)(1 - Q_k) [2(2 - Q_k) - (1 - T_k)(1 - Q_k)]}{4 [1 + T_k(1 - Q_k)]^2} \right]^{\frac{1}{2}} \right]^2,
 \end{aligned}$$

where $n = \sum_{k=1}^r m_k$, $m_k = m_{k1} + m_{k2}$ and $\lambda_k = m_{k1}/m_k$.

4.2. Efficiency comparison

To show the efficacious performance of the proposed stratified mixed randomized response model, we examine the efficiency comparison of the proposed estimator $\hat{\pi}_{Ak}$ over the proposed mixed randomized estimator $\hat{\pi}_{A1}$ and Kim and Warde [13] estimator $\hat{\pi}_{kw}$ respectively. The comparisons are given in the form of following theorems.

Theorem 4.2. *Suppose there are two strata (i.e. $k = 2$) in the population and $\lambda = m_{k1}/m_k$. The proposed stratified estimator $\hat{\pi}_{Ak}$ is always more efficient than that of usual proposed estimator $\hat{\pi}_{A1}$ where $P_1 = Q_1 = Q_2$, $\lambda = \lambda_1 = \lambda_2$ and $T = T_1 = T_2$.*

Proof: Under the assumption $k = 2$, $P_1 = Q_1 = Q_2$, $\lambda = \lambda_1 = \lambda_2$ and $T = T_1 = T_2$, the equation (4.13) can be rewritten as

$$\begin{aligned}
 (4.14) \quad V(\hat{\pi}_{Ak}) = & \frac{1}{n} \left[w_1 \left[\pi_{s1}(1 - \pi_{s1}) + \frac{\lambda(1 - \pi_{s1})(1 - P_1)}{P_1} \right. \right. \\
 & \left. \left. + \frac{(1 - \lambda)(1 - T)(1 - P_1) [2(2 - P_1) - (1 - T)(1 - P_1)]}{4 [1 + T(1 - P_1)]^2} \right]^{\frac{1}{2}} \right. \\
 & \left. + w_2 \left[\pi_{s2}(1 - \pi_{s2}) + \frac{\lambda(1 - \pi_{s2})(1 - P_1)}{P_1} \right. \right. \\
 & \left. \left. + \frac{(1 - \lambda)(1 - T)(1 - P_1) [2(2 - P_1) - (1 - T)(1 - P_1)]}{4 [1 + T(1 - P_1)]^2} \right]^{\frac{1}{2}} \right]^2.
 \end{aligned}$$

If we denote

$$\begin{aligned}
 a_1 &= \frac{(1 - \pi_{s1})(1 - P_1)}{P_1}, \\
 a_2 &= \frac{(1 - \pi_{s2})(1 - P_1)}{P_1}, \\
 b &= \frac{(1 - \lambda)(1 - T)(1 - P_1) [2(2 - P_1) - (1 - T)(1 - P_1)]}{4 [1 + T(1 - P_1)]^2},
 \end{aligned}$$

we can write equation (4.14) as

$$(4.15) \quad V(\hat{\pi}_{Ak}) = \frac{1}{n} \left[w_1 \left[\pi_{s1}(1 - \pi_{s1}) + \lambda a_1 + b \right]^{\frac{1}{2}} + w_2 \left[\pi_{s2}(1 - \pi_{s2}) + \lambda a_2 + b \right]^{\frac{1}{2}} \right]^2$$

From equation (2.18), we have

$$(4.16) \quad V(\hat{\pi}_{A1}) = \frac{1}{n} \left[(w_1 \pi_{s1} + w_2 \pi_{s2}) (1 - w_1 \pi_{s1} - w_2 \pi_{s2}) + \lambda (w_1 a_1 + w_2 a_2) + b \right].$$

Now, subtracting equation (4.15) from equation (4.16), we have

$$\begin{aligned} n[V(\hat{\pi}_{A1}) - V(\hat{\pi}_{Ak})] &= \left[(w_1 \pi_{s1} + w_2 \pi_{s2}) (1 - w_1 \pi_{s1} - w_2 \pi_{s2}) + \lambda (w_1 a_1 + w_2 a_2) + b \right] \\ &\quad - \left[w_1 \left[\pi_{s1}(1 - \pi_{s1}) + \lambda a_1 + b \right]^{\frac{1}{2}} + w_2 \left[\pi_{s2}(1 - \pi_{s2}) + \lambda a_2 + b \right]^{\frac{1}{2}} \right]^2 \\ &= w_1 \pi_{s1} + w_2 \pi_{s2} - 2 w_1 w_2 \pi_{s1} \pi_{s2} - w_1^2 \pi_{s1} - w_2^2 \pi_{s2} \\ &\quad - w_1^2 (\lambda a_1 + b) - w_2^2 (\lambda a_2 + b) + \lambda (w_1 a_1 + w_2 a_2) + b \\ &\quad - 2 w_1 w_2 \left[\pi_{s1}(1 - \pi_{s1}) + \lambda a_1 + b \right]^{\frac{1}{2}} \left[\pi_{s2}(1 - \pi_{s2}) + \lambda a_2 + b \right]^{\frac{1}{2}} \\ &= w_1 (\pi_{s1} + \lambda a_1) + w_2 (\pi_{s2} + \lambda a_2) \\ &\quad - w_1^2 (\pi_{s1} + \lambda a_1 + b) - w_2^2 (\pi_{s2} + \lambda a_2 + b) - 2 w_1 w_2 \pi_{s1} \pi_{s2} + b \\ &\quad - 2 w_1 w_2 \left[\pi_{s1}(1 - \pi_{s1}) + \lambda a_1 + b \right]^{\frac{1}{2}} \left[\pi_{s2}(1 - \pi_{s2}) + \lambda a_2 + b \right]^{\frac{1}{2}} \\ &> 0, \end{aligned}$$

which proves the theorem. \square

Theorem 4.3. Suppose there are two strata (i.e. $k = 2$) in the population and $\lambda = m_{k1}/m_k$. The proposed stratified estimator $\hat{\pi}_{Ak}$ is always more efficient than that of Kim and Warde [13] estimator $\hat{\pi}_{kw}$ where $P_1 = Q_1 = Q_2$, $\lambda = \lambda_1 = \lambda_2$ and $T = T_1 = T_2$.

Proof: Under the assumption $P_1 = Q_1 = Q_2$, $\lambda = \lambda_1 = \lambda_2$ and $T = T_1 = T_2$, the minimal variance of the Kim and Warde [13] estimator $\hat{\pi}_{kw}$ is given by

$$(4.17) \quad V(\hat{\pi}_{kw}) = \frac{1}{n} \left[w_1 (A_1 + b_1)^{\frac{1}{2}} + w_2 (A_2 + b_1)^{\frac{1}{2}} \right]^2,$$

where

$$\begin{aligned} A_1 &= \pi_{s1}(1 - \pi_{s1}) + \frac{\lambda(1 - P_1)(1 - \pi_{s1})}{P_1}, \\ A_2 &= \pi_{s2}(1 - \pi_{s2}) + \frac{\lambda(1 - P_1)(1 - \pi_{s2})}{P_1}, \\ b_1 &= \frac{(1 - \lambda)(1 - P_1)}{P_1^2}. \end{aligned}$$

Equation (4.15) can be rewritten as

$$(4.18) \quad V(\hat{\pi}_{Ak}) = \frac{1}{n} \left[w_1(A_1 + b)^{\frac{1}{2}} + w_2(A_2 + b)^{\frac{1}{2}} \right]^2.$$

From equations (4.17) and (4.18), we have

$$\begin{aligned} n \left[V(\hat{\pi}_{kw}) - V(\hat{\pi}_{Ak}) \right] &= \\ &= \left[w_1(A_1 + b_1)^{\frac{1}{2}} + w_2(A_2 + b_1)^{\frac{1}{2}} \right]^2 - \left[w_1(A_1 + b)^{\frac{1}{2}} + w_2(A_2 + b)^{\frac{1}{2}} \right]^2 \\ &= \left[w_1^2 b_1 + w_2^2 b_1 - w_1^2 b - w_2^2 b + 2 w_1 w_2 \left[(A_1 + b_1)^{\frac{1}{2}} (A_2 + b_1)^{\frac{1}{2}} - (A_1 + b)^{\frac{1}{2}} (A_2 + b)^{\frac{1}{2}} \right] \right] \\ &= \left[(b_1 - b) (w_1^2 + w_2^2) + 2 w_1 w_2 \left[(A_1 + b_1)^{\frac{1}{2}} (A_2 + b_1)^{\frac{1}{2}} - (A_1 + b)^{\frac{1}{2}} (A_2 + b)^{\frac{1}{2}} \right] \right] \\ &= (b_1 - b) \left[w_1^2 w_2^2 + 2 w_1 w_2 \frac{(A_1 + A_2 + b_1 + b)}{\left[(A_1 + b_1)^{\frac{1}{2}} (A_2 + b_1)^{\frac{1}{2}} - (A_1 + b)^{\frac{1}{2}} (A_2 + b)^{\frac{1}{2}} \right]} \right] \\ &> 0, \end{aligned}$$

since $(b_1 - b) > 0$.

Therefore, $n [V(\hat{\pi}_{kw}) - V(\hat{\pi}_{Ak})]$ is always positive. Thus the theorem is proved. □

We have shown the performance of proposed stratified estimator $\hat{\pi}_{Ak}$ over suggested mixed estimator $\hat{\pi}_{A1}$ and Kim and Warde [13] estimator $\hat{\pi}_{kw}$ in case of two strata (i.e. $k = 2$). Now, we calculate the percent relative efficiencies $\text{PRE}(\hat{\pi}_{Ak}, \hat{\pi}_{A1})$ and $\text{PRE}(\hat{\pi}_{Ak}, \hat{\pi}_{kw})$ for different values of T , π_s , n_1 , n_2 and P_1 , by using the following formulas:

$$(4.19) \quad \text{PRE}(\hat{\pi}_{Ak}, \hat{\pi}_{A1}) = \frac{V(\hat{\pi}_{A1})}{V(\hat{\pi}_{Ak})} \times 100$$

and

$$(4.20) \quad \text{PRE}(\hat{\pi}_{Ak}, \hat{\pi}_{kw}) = \frac{V(\hat{\pi}_{kw})}{V(\hat{\pi}_{Ak})} \times 100,$$

where

$$V(\hat{\pi}_{kw}) = \left[\sum_{k=1}^2 w_k \left[\frac{\pi_{sk}(1 - \pi_{sk})}{n} + \frac{(1 - Q_k) [\lambda_k Q_k (1 - \pi_{sk}) + (1 - \lambda_k)]}{n Q_k^2} \right]^{\frac{1}{2}} \right]^2.$$

We may observe from Tables 3–4:

- (a) For all the parametric combinations the values of percent relative efficiencies are substantially exceeding 100, which indicate that the proposed stratified estimator $\hat{\pi}_{Ak}$ is uniformly better than the proposed mixed estimator $\hat{\pi}_{A1}$ and Kim and Warde [13] estimator $\hat{\pi}_{kw}$ under optimum allocation condition.
- (b) It is also noted, from Table 3, the percent relative efficiencies increasing as the values of P_1 increases. Also the percent relative efficiencies almost increasing as the values of π_s increases for fixed values of λ and T .

- (c) From Table 4, we observe that with the increase in the values of P_1 there is a decreasing pattern in the values of percent relative efficiencies.
- (d) Figures 3–4 also show that there is a large gain in efficiencies by using the proposed stratified estimator $\hat{\pi}_{Ak}$ over the mixed estimator $\hat{\pi}_{A1}$ and Kim and Warde [13] stratified estimator, when the proportion of stigmatizing attribute is moderately large.

Table 3: Percent relative efficiency of the proposed stratified estimator $\hat{\pi}_{Ak}$ with respect to mixed estimator $\hat{\pi}_{A1}$.

π_{s1}	π_{s2}	π_s	w_1	w_2	λ	T	$P_1 = Q_1 = Q_2$						
							0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.08	0.13	0.1	0.6	0.4	0.2	0.1	100.031	100.041	100.051	100.063	100.076	100.090	100.107
						0.3	100.026	100.032	100.038	100.046	100.056	100.067	100.080
						0.6	100.023	100.027	100.031	100.036	100.042	100.049	100.058
						0.8	100.023	100.026	100.029	100.033	100.038	100.044	100.051
0.18	0.23	0.2	0.6	0.4	0.2	0.1	100.035	100.043	100.052	100.061	100.072	100.082	100.094
						0.3	100.031	100.037	100.043	100.050	100.058	100.067	100.078
						0.6	100.028	100.032	100.036	100.041	100.046	100.053	100.061
						0.8	100.028	100.031	100.035	100.039	100.043	100.049	100.055
0.28	0.33	0.3	0.6	0.4	0.2	0.1	100.040	100.047	100.055	100.063	100.071	100.080	100.089
						0.3	100.038	100.044	100.050	100.056	100.063	100.071	100.080
						0.6	100.035	100.039	100.043	100.047	100.053	100.059	100.066
						0.8	100.036	100.039	100.043	100.047	100.051	100.057	100.063
0.38	0.43	0.4	0.6	0.4	0.2	0.1	100.047	100.054	100.060	100.067	100.074	100.081	100.089
						0.3	100.049	100.054	100.060	100.066	100.072	100.079	100.087
						0.6	100.046	100.049	100.053	100.058	100.063	100.069	100.075
						0.8	100.047	100.051	100.055	100.059	100.064	100.069	100.075

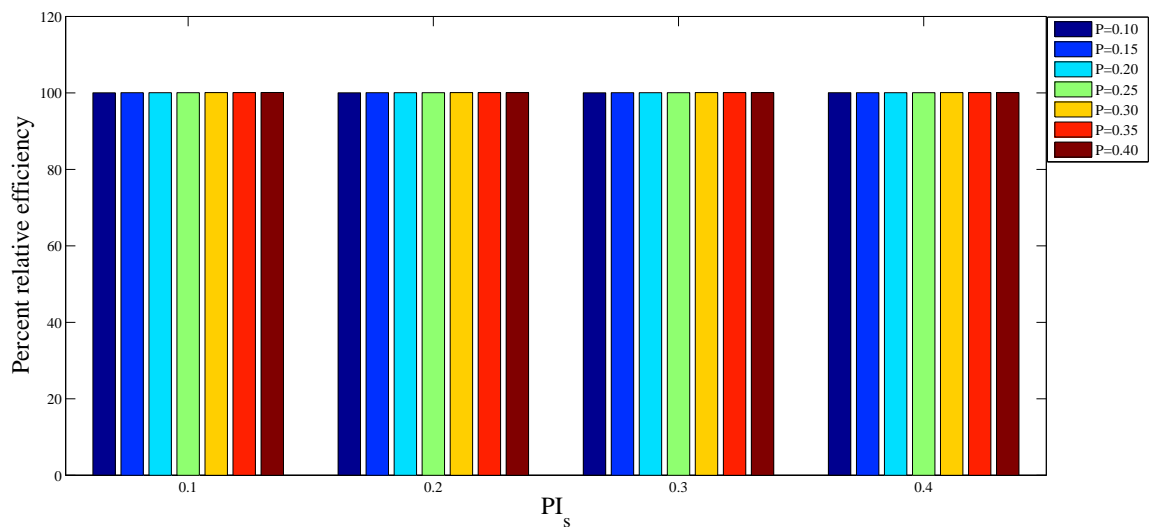


Figure 3: Percent relative efficiency of the proposed stratified estimator $\hat{\pi}_{Ak}$ with respect to mixed estimator $\hat{\pi}_{A1}$ when $T = 0.1$ and $\lambda = 0.2$.

Table 4: Percent relative efficiency of the proposed stratified estimator $\hat{\pi}_{Ak}$ with respect to Kim and Warde [13] stratified estimator $\hat{\pi}_{kw}$.

π_{s1}	π_{s2}	π_s	w_1	w_2	λ	T	$P_1 = Q_1 = Q_2$							
							0.10	0.15	0.20	0.25	0.30	0.35	0.40	
0.08	0.13	0.1	0.6	0.4	0.2	0.1	3481.8	2101.6	1442.9	1066.2	826.74	663.42	546.22	
						0.4	0.3	1631.1	1075.7	797.99	631.41	520.33	440.92	381.28
						0.6	0.1	794.72	546.74	422.30	347.29	297.01	260.87	233.57
						0.8	0.3	370.55	277.34	230.42	202.03	182.88	169.02	158.45
0.18	0.23	0.2	0.6	0.4	0.2	0.1	3667.2	2161.6	1454.6	1056.2	806.48	638.26	518.94	
						0.4	0.3	1768.8	1146.8	837.14	652.25	529.65	442.57	377.60
						0.6	0.1	864.48	586.05	446.58	362.71	306.65	266.47	236.23
						0.8	0.3	399.97	295.03	242.18	210.17	188.57	172.92	160.97
0.28	0.33	0.3	0.6	0.4	0.2	0.1	3914.3	2255.9	1490.7	1066.2	803.68	629.02	506.52	
						0.4	0.3	1946.4	1240.7	891.40	684.34	548.12	452.19	381.28
						0.6	0.1	953.86	636.79	478.46	383.62	320.52	275.53	241.87
						0.8	0.3	437.83	317.90	257.52	220.96	196.31	178.47	164.89
0.38	0.43	0.4	0.6	0.4	0.2	0.1	4245.8	2394.1	1555.4	1097.4	817.92	634.11	506.52	
						0.4	0.3	2182.3	1367.1	966.80	731.53	578.17	471.21	392.93
						0.6	0.1	1072.2	704.20	521.29	412.31	340.23	289.19	251.26
						0.8	0.3	488.32	348.50	278.17	235.65	207.04	186.39	170.73

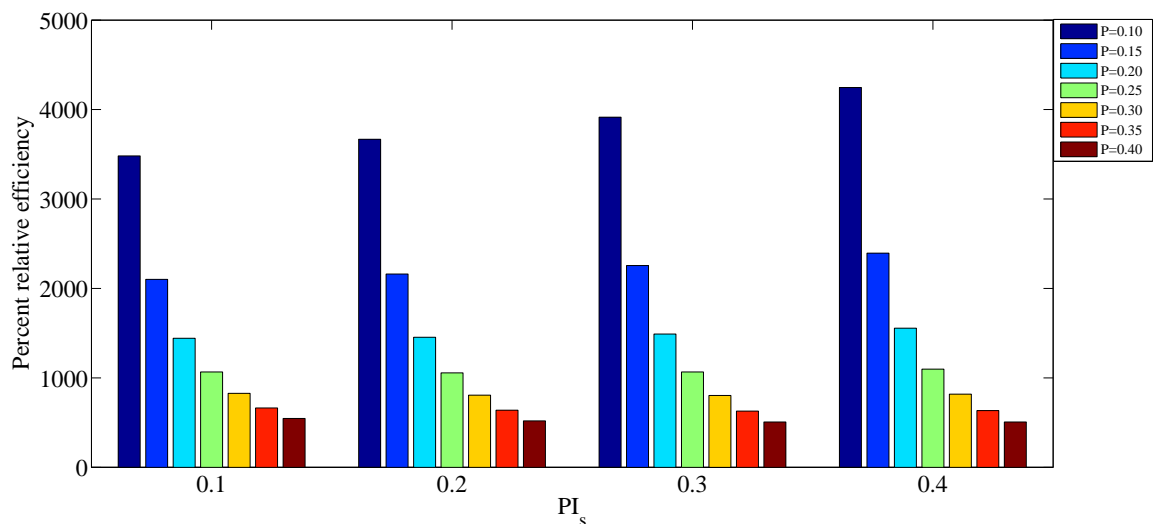


Figure 4: Percent relative efficiency of the proposed stratified estimator $\hat{\pi}_{Ak}$ with respect to Kim and Warde [13] stratified estimator $\hat{\pi}_{kw}$ when $T = 0.1$ and $\lambda = 0.2$.

5. CONCLUSIONS

In this paper, we have estimated the population proportion who possess to the sensitive attribute in the given population under both the situations of completely truthful reporting and less than completely truthful reporting as well as its stratified randomized response model. It has been shown that the proposed mixed randomized response models are better than the

Kim and Warde [13] mixed randomized response model with larger gain in efficiencies.

ACKNOWLEDGMENTS




Authors are heartily thankful to Editor-in-Chief and learned referee for their valuable comments which have made substantial improvement to bring the original manuscript to its present form. Authors are also grateful to the IIT(ISM), Dhanbad for providing the financial assistance and necessary infrastructure to carry out the present work.

REFERENCES

- [1] AMITAVA, S. (2005). Kim and Warde's mixed randomized response technique for complex surveys, *Journal of Modern Applied Statistical Methods*, **4**(2), 538–544.
- [2] CHANG, H.J. and HUANG, K.C. (2001). Estimation of proportion and sensitivity of a qualitative character, *Metrika*, **53**, 269–280.
- [3] CHANG, H.J.; WANG, C.L. and HUANG, K.C. (2004). Using randomized response to estimate the proportion and truthful reporting probability in a dichotomous finite population, *Journal of Applied Statistics*, **31**(5), 565–573.
- [4] CHAUDHURI, A. and MUKERJEE, R. (1998). *Randomized Response: Theory and Techniques*, Marcel-Dekker, New York, USA.
- [5] COCHRAN, W.G. (1977). *Sampling Technique*, 3rd Edition, New York, John Wiley and Sons, USA.
- [6] FOX, J.A. and TRACY, P.E. (1986). *Randomized Response: A Method of Sensitive Surveys*, Newbury Park, CA, SEGE Publications.
- [7] GREENBERG, B.; ABDUL-ELA, A.; SIMMONS, W.R. and HORVITZ, D.G. (1969). The unrelated question randomized response: theoretical framework, *Journal of American Statistical Association*, **64**, 529–539.
- [8] HEDAYAT, A.S. and SINHA, B.K. (1991). *Design and Inference in Finite Population Sampling*, New York, Wiley.
- [9] HONG, K.; YUM, J. and LEE, H. (1994). A stratified randomized response technique, *Korean Journal of Applied Statistics*, **7**, 141–147.
- [10] HUSSAIN, Z. and SHABBIR, J. (2007). Improvement of Kim and Warde's mixed randomized response technique for complex surveys, *InterStat*, July 003.
- [11] KIM, J.M. and ELAM, M.E. (2005). A two-stage stratified Warner's randomized response model using optimal allocation, *Metrika*, **61**, 1–7.
- [12] KIM, J.M. and WARDE, W.D. (2004). A stratified Warner randomized response model, *Journal of Statistical Planning and Inference*, **120**, 155–165.
- [13] KIM, J.M. and WARDE, W.D. (2005). A mixed randomized response model, *Journal of Statistical Planning and Inference*, **133**, 211–221.
- [14] LANKE, J. (1976). On the degree of protection in randomized interview internet, *International Statistical Review*, **44**, 80–83.

- [15] MANGAT, N.S. (1994). An improved randomized response strategy, *Journal of the Royal Statistical Society, B*, **56**(1), 93–95.
- [16] MANGAT, N.S. and SINGH, R. (1990). An alternative randomized procedure, *Biometrika*, **77**, 439–442.
- [17] MOORS, J.A. (1971). Optimization of the unrelated question randomized response model, *Journal of the American Statistical Association*, **66**, 627–929.
- [18] NAZUK, A. and SHABBIR, J. (2010). A new mixed randomized response model, *International Journal of Business and Social Science*, **1**(1), 186–190.
- [19] RYU, J.B.; HONG, K.H. and LEE, G.S. (1993). *Randomized Response Model*, Freedom Academy, Seoul, Korea.
- [20] SIMMONS, W.R.; HORVITZ, D.G. and SHAH, B.V. (1967). The unrelated question randomized response model, *Social Statistics Section of the American Statistical Association*, 65–72.
- [21] SINGH, S. (2003). *Advance Sampling Theory with Applications*, Kluwer Academic Publishers, Dordrecht, the Netherlands.
- [22] SINGH, R. and MANGAT, N.S. (1996). *Elements of Survey Sampling*, Kluwer Academic Publishers, Dordrecht, the Netherlands.
- [23] SINGH, H.P. and TARRAY, T.A. (2013a). An alternative to Kim and Warde’s mixed randomized response model, *Statistics and Operation Research Transactions*, **37**(2), 189–210.
- [24] SINGH, H.P. and TARRAY, T.A. (2013b). An improve mixed randomized response model, *Model Assisted Statistics and Applications*, **9**, 73–87.
- [25] SINGH, H.P. and TARRAY, T.A. (2013c). An alternative to Kim and Warde’s mixed randomized response technique, *Statistica*, **LXXIII**(3), 379–402.
- [26] TRACY, D.S. and MANGAT, N.S. (1996). Some developments in randomized response sampling during the last decade – a follow up of review by Chaudhuri and Mukherjee, *Journal of Applied Statistical Science*, **4**(2/3), 147–158.
- [27] TRACY, D.S. and OSAHAN, S.S. (1999). An improved randomized response technique, *Pakistan Journal of Statistics*, **15**, 1–6.
- [28] WARNER, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association*, **60**, 63–69.

A REGRESSION MODEL FOR POSITIVE DATA BASED ON THE SLASHED HALF-NORMAL DISTRIBUTION

- Authors: YOLANDA M. GÓMEZ 
– Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama,
Copiapó, Chile
yolanda.gomez@uda.cl
- DIEGO I. GALLARDO 
– Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama,
Copiapó, Chile
diego.gallardo@uda.cl
- MÁRIO DE CASTRO 
– Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo,
São Carlos, SP, Brazil
mcastro@icmc.usp.br

Received: September 2018

Revised: September 2019

Accepted: September 2019

Abstract:

- In this paper, we discuss several aspects about the slashed half-normal distribution. We reparameterize the model based on the mean and we perform comparisons with well-known regression models for positive data. Maximum likelihood estimation of the parameters is carried out through the *expectation-maximization* algorithm. Some properties of the estimators and two kinds of residuals are assessed in a simulation study. Two real datasets illustrated the proposed model as well as other three models for the sake of comparison.

Keywords:

- *EM algorithm; gamma distribution; half-normal distribution; slashed distribution.*

AMS Subject Classification:

- 62F86, 60E05.

1. INTRODUCTION

The half-normal (HN) distribution is a very important model in the study of skewed distributions. For instance, it is used in the stochastic representation of the skew-normal distribution in Azzalini [4, 5] and Henze [15]. Several papers in the literature have paid attention to the half-normal distribution. For instance, Chou and Liu [7] studied its properties and its uses in quality control. Pewsey [22, 23] studied asymptotic inference and maximum likelihood estimation for the general location-scale half-normal distribution. For analysis and applications from a Bayesian point of view, the reader is referred to Wiper *et al.* [32] and Khan and Islam [17]. Also, the `hnp` R package [20], generates half-normal plots with simulated envelopes using different diagnostics tools from a range of different fitted models. Even though the HN distribution accommodates only decreasing hazard rates, this distribution has been used to model positive data and is becoming an important model in reliability theory,. Some of the generalizations of this distribution can be found in Cooray and Ananda [8], Cordeiro *et al.* [9], Olmos *et al.* [21], Gómez and Bolfarine [13], Bourguignon *et al.* [6] and Asgharzadeh *et al.* [1], among others. Particularly, we focus on the extension proposed in Olmos *et al.* [21], named slash half-normal (SHN) distribution, where the goal is to increase the kurtosis with respect to its parent half-normal distribution, and hence be more useful for modeling positive datasets that may have a heavy right tail. In this work, we propose a reparameterization for this model based on the mean. We use this parameterization because it is convenient for proposing a regression model.

The article is organized as follows. In Section 2, we describe the reparametrized SHN regression model and compare it with some existing models. In Section 3, we describe parameter estimation by the maximum likelihood (ML) method using the *expectation-maximization (EM)* algorithm. Goodness of fit through residuals is discussed in Section 4. In Section 5, we carry out two simulation studies to assess the performance of the proposed estimators and the two kinds of residuals. In Section 6, we apply the proposed model to analyze two datasets on the diet of the hunter-gatherer and concentration of minerals in soil samples. Concluding remarks are given in Section 7.

2. THE PROPOSAL

In this section, we present the proposed reparameterization for the SHN model in terms of the mean. We also present three common distributions to accommodate positive data that also are reparametrized in terms of the mean: the gamma, Weibull and Birnbaum–Saunders models.

2.1. Reparametrized slashed half-normal model

The SHN model (Olmos *et al.* (2012) [21]) is built in the following way. If $X \sim \text{HN}(\sigma)$ ($\sigma > 0$) and $Z \sim \text{Beta}(\alpha, 1)$ are independent random variables, then

$$(2.1) \quad Y = \frac{X}{Z} \sim \text{SHN}(\sigma, \alpha),$$

where $\alpha > 0$ is a shape parameter that mainly controls the *right* tail of the distribution. Lower values of α ($0 < \alpha < 1$) lead to a heavier tail (see Figure 1 in Olmos *et al.* [21]). However, in practice we have found estimates for α greater than 1 (see the two examples in Olmos *et al.* [21] and our applications). For this reason, the potential advantages of the parameterization of the model in terms of the mean (mainly related to the interpretation of the coefficients in a regression model) justify the restriction $\alpha > 1$. Such kind of restriction is not uncommon in the literature. Without going further, the popular Student's *t* distribution has a finite mean if the degrees of freedom are greater than 1. We propose a reparameterization of the SHN model based on $\mu = \sqrt{2/\pi} \alpha \sigma / (\alpha - 1)$. The probability density function of the reparametrized SHN, henceforth RSHN(μ, α), is given by

$$(2.2) \quad f_{\text{RSHN}}(y; \mu, \alpha) = \alpha \sqrt{\frac{2^\alpha}{\pi}} \left[\sqrt{\frac{\pi}{2}} \frac{\mu(\alpha-1)}{\alpha} \right]^\alpha \Gamma\left(\frac{\alpha+1}{2}\right) y^{-(\alpha+1)} G\left[\frac{\alpha^2 y^2}{\pi \mu^2 (\alpha-1)^2}, \frac{\alpha+1}{2}\right],$$

for $y > 0$, where $\Gamma(\cdot)$ denotes the gamma function and $G(y, a) = \int_0^y u^{a-1} e^{-u} du / \Gamma(a)$ is the cumulative distribution function (cdf) of the gamma distribution with rate parameter equal to 1. Based on results in Olmos *et al.* [21], we have $\mathbb{E}(Y) = \mu$, for $\alpha > 1$,

$$\text{Var}(Y) = \frac{\mu^2}{2} \left[\pi - 2 + \frac{\pi}{\alpha(\alpha-2)} \right], \quad \text{for } \alpha > 2,$$

$$\sqrt{\nu_3} = \frac{\pi \sqrt{2(\alpha-2)} \left[\frac{4}{\pi} \alpha^2 (\alpha-2) (\alpha-3) - (\alpha-1)^2 (\alpha-4) (\alpha+1) \right]}{\sqrt{\alpha} (\alpha-3) \left[(\pi-2) \alpha (\alpha-2) + \pi \right]^{3/2}}, \quad \text{for } \alpha > 3,$$

and

$$\nu_4 = \frac{3\alpha(\alpha-2)^2(\alpha-3) \left[\pi^2(\alpha-1)^4 - 4\alpha^3(\alpha-4) \right] - 4\pi\alpha^2(\alpha-1)^2(\alpha-2)(\alpha-4)(\alpha^2-3\alpha+8)}{\alpha^2(\alpha-3)(\alpha-4) \left[(\pi-2)\alpha(\alpha-2) + \pi \right]^2},$$

for $\alpha > 4$, where $\sqrt{\nu_3}$ and ν_4 denote the skewness and kurtosis coefficients, respectively. Note that this parameterization is very convenient because the parameter μ is related only to the mean and the variance of the distribution.

2.2. Reparametrized gamma distribution

For $Y \sim \text{RG}(\mu, \phi)$ (the gamma model parametrized in terms of the mean), we have

$$\mathbb{E}(Y) = \mu, \quad \text{Var}(Y) = \frac{\mu^2}{\phi}, \quad \sqrt{\nu_3} = \frac{2}{\sqrt{\phi}} \quad \text{and} \quad \nu_4 = 3 + \frac{6}{\phi}.$$

The RSHN model is a competing distribution for the gamma distribution because the coefficient of variation (cv), skewness and kurtosis coefficients do not depend on μ in both models. Figure 1(a) shows the values of ϕ in the $\text{RG}(\mu, \phi)$ model and α in the $\text{RSHN}(\mu, \alpha)$ model that lead to the same values of cv. Figure 1(b) displays the kurtosis coefficient for those pairs (ϕ, α) corresponding to the same value of cv. It is clear that the gamma model is more flexible in the sense that it allows to obtain any positive value for the cv, whereas the RSHN distribution only supports values for cv greater than $[(\pi-2)/2]^{1/2} \approx 0.756$, i.e., greater than the cv of the half-normal distribution. However, there is a range of values of α such that,

for the same value of the cv, the RSHN distribution has a greater kurtosis coefficient than the gamma distribution. In short, in the RSHN model the variance is proportional to the square of the mean (similar to the gamma model), but the RSHN model has a greater kurtosis coefficient for a certain range of values of α .

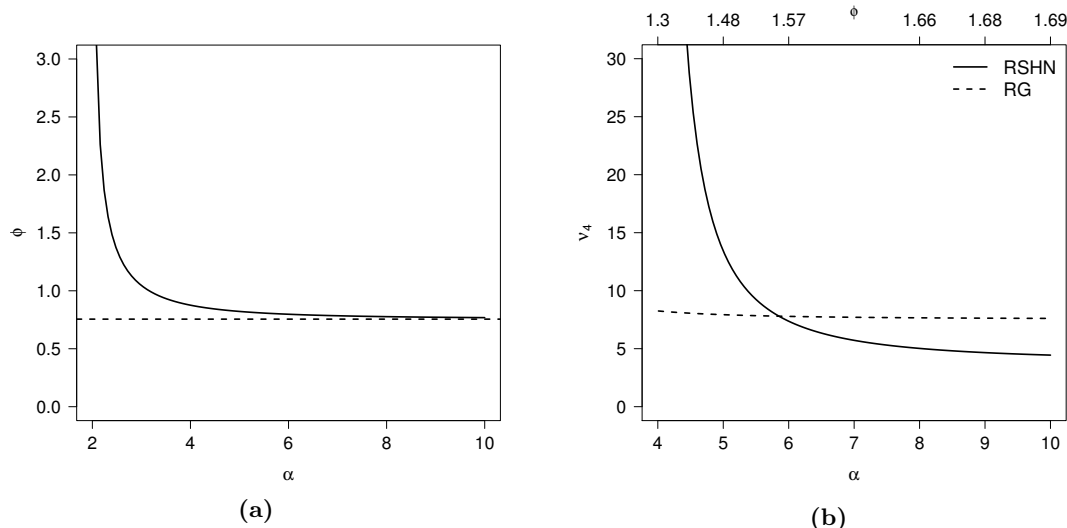


Figure 1: (a) Values for ϕ and α in the $RG(\mu, \phi)$ and $RSHN(\mu, \alpha)$ distributions that produce the same coefficient of variation and (b) their respective kurtosis coefficients.

2.3. Reparametrized Weibull and Birnbaum–Saunders distributions

The reparametrized form of the Weibull distribution with parameters $\mu > 0$ and $\delta > 0$ has probability density function

$$f_{RW}(y; \mu, \delta) = \frac{\phi}{\gamma} \left(\frac{y}{\gamma}\right)^{\delta-1} \exp\left[-\left(\frac{y}{\gamma}\right)^\delta\right], \quad \text{for } y > 0,$$

where $\gamma = \mu/\Gamma(1/\delta + 1)$, so that

$$\mathbb{E}(Y) = \mu \quad \text{and} \quad \text{Var}(Y) = \mu^2 \left\{ \frac{\Gamma(2/\delta + 1)}{[\Gamma(1/\delta + 1)]^2} - 1 \right\}.$$

We denote as $Y \sim RW(\mu, \delta)$.

In the same way, Santos-Neto *et al.* [31] also reparametrized the Birnbaum–Saunders distribution in terms of the mean. With parameters $\mu > 0$ and $\xi > 0$, the probability density function is given by

$$f_{RBS}(y; \mu, \xi) = \frac{\exp(\xi/2) \sqrt{\xi+1}}{4 \sqrt{\pi \mu} y^{3/2}} \left(y + \frac{\xi \mu}{\xi + \mu} \right) \exp \left\{ -\frac{\xi}{4} \left[\frac{y(\xi+1)}{\xi \mu} + \frac{\xi \mu}{y(\xi+1)} \right] \right\},$$

for $y > 0$, so that $\mathbb{E}(Y) = \mu$ and $\text{Var}(Y) = \mu^2(2\xi+5)/(\xi+1)^2$. We use the notation $Y \sim RBS(\mu, \xi)$. The RW and RG (Section 2.2) will be compared with the RSHN model fitted to real datasets in Section 6.

Remark 2.1. The RG and RW models are more flexible than the RBS and RSHN models in the sense that, for a given value of μ , they allow to obtain any positive value for the variance, whereas the RBS and RSHN models have some restrictions. However, even when all the models produce the same mean and variance, the skewness and kurtosis are not the same. Moreover, such terms do not depend on μ . Table 1 shows four models with the same mean and variance. However, the skewness and kurtosis coefficients are different.

Table 1: Examples of models with the same mean and variance.

Moment or coefficient	Model			
	RG($\mu, 1.333$)	RW($\mu, 1.158$)	RBS($\mu, 3.692$)	RSHN($\mu, 4.125$)
Mean	μ	μ	μ	μ
Variance	$0.75 \mu^2$	$0.75 \mu^2$	$0.75 \mu^2$	$0.75 \mu^2$
Skewness	1.732	1.390	12.662	1.791
Kurtosis	7.500	6.868	59.641	120.807

Remark 2.2. The mean and the variance of the RG, RW, RBS and RSHN models are μ and $\mu^2 w^2(\eta)$, where η represents ϕ, δ, ξ or α in each model, respectively, and $w(\cdot)$ is a positive function representing the coefficient of variation. This function is presented in Table 2. The computational implementation to model mean and dispersion parameters with a set of covariates linked to both components in RG and RW models is implemented in the `gamlss.dist` package in R (see Rigby and Stasinopoulos [28, 29]), while the RBS model is discussed in Santos-Neto *et al.* [30]. A similar scheme to model mean and dispersion might be considered for the RSHN distribution. However, we only consider a model for the mean parameter in this work.

Table 2: Summary for some models with quadratic variance function.

Model	RG(μ, ϕ)	RW(μ, δ)	RBS(μ, ξ)	RSHN(μ, α)
$w(\eta)$	$\frac{1}{\sqrt{\phi}}$	$\sqrt{\frac{\Gamma(2/\delta + 1)}{[\Gamma(1/\delta + 1)]^2}} - 1$	$\frac{\sqrt{(2\xi + 5)}}{\xi + 1}$	$\sqrt{\frac{1}{2} \left(\pi - 2 + \frac{\pi}{\alpha(\alpha - 2)} \right)}$

3. ESTIMATION

In this section, we discuss some details about the estimation procedure based on the ML method. We also consider an EM type algorithm to obtain a more stable estimation procedure. Henceforth, we consider a set of p observed covariates for each individual, say $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$. Since $\mu = \mathbb{E}(Y)$ is a positive parameter, we adopt the logarithmic link function $\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$, $i = 1, \dots, n$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients.

3.1. General context

In Olmos *et al.* [21], parameter estimation (without covariates) was carried out based on the direct maximization of the log-likelihood function using as initial values the method of moments estimates of the parameters. In our model, assuming the intercept is included, naive estimators for β_0 and α can be obtained ignoring the covariates, i.e., $\beta_1 = \dots = \beta_p = 0$. In this case, such estimators are given by

$$(3.1) \quad \widehat{\beta}_{0M} = \log(\overline{Y}) \quad \text{and} \quad \widehat{\alpha}_M = \frac{1}{2} + \frac{1}{2} \sqrt{1 + \frac{\pi}{2A_y - 2 + \pi}}, \quad \text{if } \overline{Y^2} > \frac{\pi}{2} \overline{Y}^2,$$

where $A_y = \overline{Y^2}/\overline{Y}^2$ and $\overline{Y^2}$ is the sample mean of the squared observations.

The log-likelihood function of $\boldsymbol{\psi} = (\boldsymbol{\beta}^\top, \alpha)^\top$ in a random sample with observations y_1, \dots, y_n is given by

$$(3.2) \quad \ell(\boldsymbol{\psi}) = c(\alpha) + \alpha \log(\mu) - (\alpha + 1) \sum_{i=1}^n \log(y_i) + \sum_{i=1}^n \log \left\{ G \left[\frac{\alpha^2 y^2}{\pi \mu^2 (\alpha - 1)^2}, \frac{\alpha + 1}{2} \right] \right\},$$

where $c(\alpha) = -n(\alpha - 1) \log(\alpha) - n\alpha \log(2)/2 + (\alpha - 1/2) \log(\pi) + \alpha \log(\alpha - 1) + n \log[\Gamma(\alpha/2 + 1/2)]$. However, direct maximization of (3.2) is not simple and may suffer from numerical instabilities. In Section 3.2, we propose a stable estimation procedure for this model based on the stochastic representation in (2.1). We develop in the sequel an EM algorithm (Dempster *et al.* [10]) for parameter estimation.

3.2. ECM and ECME algorithms

To facilitate the estimation process, we include latent variables Z_1, \dots, Z_n through the following hierarchical representation of the RSHN model:

$$Y_i | Z_i = z_i, \quad \mu_i \sim \text{HN} \left[\sqrt{\frac{\pi}{2}} \frac{\mu_i(\alpha - 1)}{\alpha z_i} \right] \quad \text{and} \quad Z_i \sim \text{Beta}(\alpha, 1).$$

Thus, the complete likelihood function for $\boldsymbol{\psi}$ is given by

$$L_c(\boldsymbol{\psi}) = \left(\sqrt{\frac{2}{\pi}} \frac{\alpha^2}{\alpha - 1} \right)^n \exp \left\{ - \sum_{i=1}^n [\log(\mu_i) - \alpha \log(z_i)] - \frac{\alpha^2}{\pi(\alpha - 1)^2} \sum_{i=1}^n \frac{y_i^2 Z_i^2}{\mu_i^2} \right\}.$$

Consequently, up to a constant, the complete log-likelihood function for $\boldsymbol{\psi}$ is

$$\ell_c(\boldsymbol{\psi}) = - \frac{\alpha^2}{\pi(\alpha - 1)^2} \sum_{i=1}^n \frac{y_i^2 z_i^2}{\mu_i^2} - \sum_{i=1}^n [\log(\mu_i) - \alpha \log(z_i)] + n[2 \log(\alpha) - \log(\alpha - 1)].$$

Let $\widehat{z}_i^2 = \mathbb{E}(Z_i^2 | \boldsymbol{\psi} = \widehat{\boldsymbol{\psi}})$, $\widehat{\log}(z_i) = \mathbb{E}(\log(Z_i) | \boldsymbol{\psi} = \widehat{\boldsymbol{\psi}})$ and $Q(\boldsymbol{\psi} | \widehat{\boldsymbol{\psi}}) = \mathbb{E}(\ell_c(\boldsymbol{\psi}) | \boldsymbol{\psi} = \widehat{\boldsymbol{\psi}})$. With these definitions,

$$Q(\boldsymbol{\psi} | \widehat{\boldsymbol{\psi}}) = - \frac{\alpha^2}{\pi(\alpha - 1)^2} \sum_{i=1}^n \frac{y_i^2 \widehat{z}_i^2}{\mu_i^2} - \sum_{i=1}^n [\log(\mu_i) - \alpha \widehat{\log}(z_i)] + n[2 \log(\alpha) - \log(\alpha - 1)].$$

In addition,

$$f(z_i | Y_i = y_i) \propto (z_i^2)^{\left(\frac{\alpha}{2}+1\right)-1} \exp\left[-\frac{\alpha^2 y_i^2 z_i^2}{\pi \mu_i^2 (\alpha - 1)^2}\right] I_{(0,1)}(z_i),$$

where $I_A(a) = 1$ if $a \in A$ and 0 otherwise. Define $W_i = Z_i^2$, $i = 1, \dots, n$. It is straightforward to show that

$$f(w_i | Y_i = y_i) \propto w_i^{\frac{\alpha+1}{2}-1} \exp\left[-\frac{\pi \mu_i^2 (\alpha - 1) y_i^2 w_i}{\alpha}\right] I_{(0,1)}(w_i),$$

so that

$$W_i | Y_i = y_i \sim \text{Gamma}\left[\frac{\alpha + 1}{2}, \frac{\pi \mu_i^2 (\alpha - 1) y_i^2}{\alpha}\right] I_{(0,1)},$$

i.e., the truncated gamma distribution on the $(0, 1)$ interval. Thus,

$$\widehat{z}_i^2 = \frac{\pi \mu_i (\alpha + 1) (\alpha - 1)^2 G\left[\frac{\alpha^2 y_i^2}{\pi \mu_i^2 (\alpha - 1)^2}, \frac{\alpha + 3}{2}\right]}{y_i^2 G\left[\frac{\alpha^2 y_i^2}{\pi \mu_i^2 (\alpha - 1)^2}, \frac{\alpha + 1}{2}\right]}.$$

However, a closed form expression for $\widehat{\log(z_i)}$ is not available, but it can be computed numerically noticing that $\mathbb{E}[\log(Z_i)] = \mathbb{E}[\log(W_i)]/2 = C_{i1}(\boldsymbol{\psi})/[2 C_{i0}(\boldsymbol{\psi})]$, where

$$(3.3) \quad C_{ij}(\boldsymbol{\psi}) = \int_0^1 [\log(w)]^j w^{\frac{\alpha+1}{2}-1} \exp\left[-\frac{\pi \mu_i^2 (\alpha - 1) y_i^2 w}{\alpha}\right] dw,$$

for $\alpha > 1$ and $j = 0, 1$. Note that if $W_i^* \sim \text{Gamma}(a_i, b_i)$, $a_i, b_i > 0$, then $\mathbb{E}[\log(W_i^*)] = \eta(a_i) - \log(b_i)$, with $\eta(\cdot)$ denoting the digamma function. For this reason, the convergence of $C_{i1}(\boldsymbol{\psi})$ is guaranteed because $C_{i1}(\boldsymbol{\psi}) < \mathbb{E}[\log(W_i^*)] < \infty$, taking a_i and b_i conveniently. Therefore, the k -th iteration of the ECM algorithm takes the form:

- **E step.** For $i = 1, \dots, n$, use $\widehat{\boldsymbol{\psi}}^{(k-1)}$, the estimate of $\boldsymbol{\psi}$ at the $(k - 1)$ -th iteration of the algorithm, to compute

$$\widehat{z}_i^{2(k)} = \frac{\pi \widehat{\mu}_i^{(k-1)} (\widehat{\alpha}^{(k-1)} + 1) (\widehat{\alpha}^{(k-1)} - 1)^2 G\left[\frac{\widehat{\alpha}^{2(k-1)} y_i^2}{\pi \widehat{\mu}_i^{2(k-1)} (\widehat{\alpha}^{(k-1)} - 1)^2}, \frac{\widehat{\alpha}^{(k-1)} + 3}{2}\right]}{y_i^2 G\left[\frac{\widehat{\alpha}^{2(k-1)} y_i^2}{\pi \widehat{\mu}_i^{2(k-1)} (\widehat{\alpha}^{(k-1)} - 1)^2}, \frac{\widehat{\alpha}^{(k-1)} + 1}{2}\right]}$$

and $\widehat{\log(z_i)}^{(k)} = C_{i1}(\widehat{\boldsymbol{\psi}}^{(k)})/[2 C_{i0}(\widehat{\boldsymbol{\psi}}^{(k)})]$, where $\widehat{\mu}_i^{(k-1)} = \exp(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^{(k-1)})$ and $C_{ij}(\boldsymbol{\psi})$, for $j = 0, 1$, is given in (3.3).

- **CM step I.** Given $\widehat{\alpha}^{(k-1)}$ and $\widehat{\mathbf{z}}^{2(k)} = (\widehat{z}_1^{2(k)}, \dots, \widehat{z}_n^{2(k)})^\top$, maximize the expression

$$-\frac{\widehat{\alpha}^{2(k-1)}}{\pi (\widehat{\alpha}^{(k-1)} - 1)^2} \sum_{i=1}^n \frac{y_i^2 \widehat{z}_i^{2(k)}}{\exp(2 \mathbf{x}_i^\top \boldsymbol{\beta})} - \sum_{i=1}^n \mathbf{x}_i^\top \boldsymbol{\beta}$$

with respect to $\boldsymbol{\beta}$ to obtain $\widehat{\boldsymbol{\beta}}^{(k)}$.

- **CM step II.** Given $\widehat{\boldsymbol{\beta}}^{(k)}$ and $\widehat{\log(\mathbf{z})}^{(k)} = (\widehat{\log(z_1)}^{(k)}, \dots, \widehat{\log(z_n)}^{(k)})^\top$, maximize the expression

$$-\frac{\alpha^2}{\pi (\alpha - 1)^2} \sum_{i=1}^n \frac{y_i^2 \widehat{z}_i^{2(k)}}{\widehat{\mu}_i^{2(k)}} + \alpha \sum_{i=1}^n \widehat{\log(z_i)}^{(k)} + n[2 \log(\alpha) - \log(\alpha - 1)]$$

with respect to α , subject to $\alpha > 1$, to obtain $\widehat{\alpha}^{(k)}$.

The maximization procedures in the CM steps can be performed using extant software, e.g., with the `optim` function in the R language [24]. The E and CM steps are repeatedly cycled until a suitable convergence rule is satisfied, e.g., the difference in successive values of the estimates given by the Euclidean norm $\|\boldsymbol{\psi}^{(k+1)} - \boldsymbol{\psi}^{(k)}\|$ is less than a tolerance value.

In practice, the implementation of the ECM algorithm in this form can be computationally expensive, mainly due to the computation of $\widehat{\log(z_i)}$, $i = 1, \dots, n$, in the E step. To avoid this problem and following the same idea used in [19], we can replace the CM step II by the following step:

- **CME step II.** Given $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(k)}$, update the estimate of α by maximizing the expression $\sum_{i=1}^n \log[f_{\text{RSHN}}(y_i; \widehat{\mu}_i^{(k)}, \alpha)]$ with respect to α , subject to $\alpha > 1$, where f_{RSHN} is presented in (2.2). In other words, α is updated based on the maximization of the observed log-likelihood function with $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(k)}$. This step involves a unidimensional maximization, which can be performed using, for instance, the Brent method available in the `optim` function in R.

Finally, the covariance matrix of $\widehat{\boldsymbol{\psi}}$ can be estimated based on the Hessian matrix of the observed log-likelihood function. The `numDeriv` R package [12] provides an accurate numerical approximation for this matrix. In Sections 5 and 6, this estimate of the covariance matrix of $\widehat{\boldsymbol{\psi}}$ is used to build approximate confidence intervals and to compute standard errors. Computational codes are available in *supplementary material*.

Remark 3.1. For the case without covariates, the CM step I is reduced to

$$\text{CM step I. Update } \mu \text{ as follows: } \widehat{\mu}^{(k)} = \frac{\widehat{\alpha}^{(k)}}{\widehat{\alpha}^{(k)} - 1} \left(\frac{2}{n\pi} \sum_{i=1}^n z_i^2 \widehat{y}_i^{2(k)} \right)^{1/2}.$$

Remark 3.2. In the RSHN regression model, when the intercept term is included in the model, an initial value to $\boldsymbol{\psi}$ can be obtained based on the moment estimators presented in (3.1). Such initial value can be considered as $\widehat{\boldsymbol{\psi}}^{(0)} = (\widehat{\beta}_{0M}, 0, \dots, 0, \widehat{\alpha}_M)$.

4. RESIDUAL DIAGNOSTICS FOR THE RSHN MODEL

In this section, we discuss some aspects related to the deviance and quantile residuals for the RSHN model.

4.1. Deviance residuals

Residual diagnostics for the RSHN model can be carried out using the deviance residuals defined as $r_{D_i} = \text{sign}(Y_i - \widehat{\mu}_i) \sqrt{2} [\ell(\widetilde{\mu}_i, \widehat{\alpha}) - \ell(\widehat{\mu}_i, \widehat{\alpha})]^{1/2}$, where $\ell(\cdot)$ denotes the log-likelihood function, $\widetilde{\mu}_i$ is the ML estimator of $\mu_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$ under the saturated model and $\widehat{\mu}_i$ is the ML estimator of μ_i under the working model (with $p < n$ regression coefficients).

For the RSHN regression model, with $\tilde{\mu}_i = Y_i$ and $\ell(\cdot)$ coming from (3.2), these residuals are given by

$$r_{D_i} = \text{sign}(Y_i - \hat{\mu}_i) \sqrt{2} \left(\hat{\alpha} \log(Y_i/\hat{\mu}_i) + \log \left\{ G \left[\frac{\hat{\alpha}^2}{\pi(\hat{\alpha} - 1)}, \frac{\hat{\alpha} + 1}{2} \right] \right\} - \log \left\{ G \left[\frac{Y_i^2 \hat{\alpha}^2}{\pi \hat{\mu}_i^2 (\hat{\alpha} - 1)}, \frac{\hat{\alpha} + 1}{2} \right] \right\} \right)^{1/2}, \quad \text{for } i = 1, \dots, n,$$

where $G(\cdot)$ is given in (2.2). If the model is correct, the approximate distribution of r_{D_i} , $i = 1, \dots, n$, is the standard normal. The normality of the residuals can be tested based on different tests such as the Shapiro–Wilk (SW), Anderson–Darling (AD) and Cramér–von Mises (CVM) tests [33]. Moreover, simulated envelopes (Atkinson [3]) are also useful to assess the fitting of the models.

4.2. Quantile residuals

A second alternative for residual analysis can be based on the normalized quantile residuals (Dunn and Smyth [11]). These residuals are defined as

$$r_{Q_i} = \Phi^{-1} [F(Y_i; \hat{\psi})], \quad i = 1, \dots, n,$$

where $F(\cdot; \psi)$ is the cdf of the response variable and $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard normal distribution. Except for the uncertainty due to estimation of the parameters, if the model is correct, r_{Q_i} , $i = 1, \dots, n$, constitute a random sample from the standard normal distribution. For the RSHN model, we have

$$r_{Q_i} = \hat{\alpha} \sqrt{\frac{2\hat{\alpha}}{\pi}} \left[\sqrt{\frac{\pi}{2}} \frac{\hat{\mu}_i (\hat{\alpha} - 1)}{\hat{\alpha}} \right]^{\hat{\alpha}} \Gamma \left(\frac{\hat{\alpha} + 1}{2} \right) \int_0^{Y_i - (\hat{\alpha} + 1)} u_i^{-\hat{\alpha} + 1} G \left[\frac{\hat{\alpha}^2 u_i^2}{\pi \hat{\mu}_i^2 (\hat{\alpha} - 1)^2}, \frac{\hat{\alpha} + 1}{2} \right] du_i,$$

where the integral can be computed numerically using, for instance, the `integrate` function in R.

5. SIMULATION STUDIES

In this section, we present two simulation studies. The first is devoted to assess the performance of the ML estimator for the RSHN model in finite samples when the model is well specified. The main goal of the second study is similar to the one in Leiva *et al.* [18], with the aim of assess the behavior of the deviance and normalized quantile residuals when the model is either well or misspecified.

5.1. Parameters recovery

We stress that in Olmos *et al.* [21], the authors did not carry out a simulation study, so that it is of interest to address this issue. To draw synthetic datasets from the RSHN model, we fix $\beta = (\beta_0, \beta_1, \beta_2)^\top$ (two covariates) and α at the true values in Table 3.

Table 3: Bias, average of the asymptotic standard error (SE), square root of the simulated mean squared error (RMSE) and coverage probability of the 95% asymptotic confidence intervals (CP) of the estimators under the RSHN regression model with 1,000 replications.

Parameter	True value	n = 50				n = 100				n = 200			
		Bias	SE	RMSE	CP	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP
α	2.5	1.962	1.987	1.659	0.906	1.655	1.311	1.178	0.913	1.001	0.926	0.879	0.931
β_0	0.5	-0.039	0.352	0.304	0.924	-0.028	0.287	0.231	0.928	-0.011	0.171	0.159	0.955
β_1	0.5	-0.001	0.272	0.245	0.933	-0.001	0.192	0.151	0.936	-0.001	0.125	0.111	0.947
β_2	0.05	-0.001	0.009	0.005	0.934	0.000	0.007	0.003	0.939	0.000	0.003	0.002	0.947
α	2.5	2.139	2.152	1.993	0.912	1.683	1.559	1.313	0.921	1.149	1.082	1.032	0.932
β_0	1.0	0.041	0.281	0.265	0.935	0.031	0.256	0.225	0.937	0.010	0.181	0.169	0.945
β_1	0.5	-0.031	0.223	0.201	0.936	-0.029	0.169	0.147	0.941	-0.024	0.127	0.119	0.941
β_2	0.05	-0.005	0.010	0.006	0.931	-0.004	0.007	0.004	0.937	-0.004	0.005	0.004	0.942
α	2.5	2.389	1.559	1.379	0.918	1.446	1.333	1.052	0.922	0.982	0.790	0.754	0.935
β_0	0.5	-0.089	0.369	0.311	0.912	-0.045	0.246	0.201	0.934	-0.021	0.171	0.152	0.952
β_1	0.5	0.031	0.249	0.219	0.924	0.005	0.178	0.152	0.931	0.003	0.130	0.111	0.947
β_2	0.025	0.001	0.010	0.005	0.926	0.000	0.008	0.003	0.931	0.000	0.003	0.002	0.939
α	2.5	2.424	1.587	1.401	0.914	1.452	1.156	1.038	0.921	0.951	0.891	0.858	0.941
β_0	1.0	-0.079	0.402	0.351	0.924	-0.059	0.271	0.217	0.943	-0.013	0.178	0.156	0.953
β_1	0.5	0.012	0.251	0.210	0.931	0.009	0.180	0.154	0.933	0.002	0.135	0.112	0.941
β_2	0.025	0.000	0.009	0.005	0.924	0.000	0.007	0.003	0.931	0.000	0.003	0.002	0.941
α	3.0	2.094	1.852	1.650	0.918	1.912	1.210	1.003	0.919	0.929	0.974	0.936	0.937
β_0	0.5	0.049	0.336	0.281	0.944	0.043	0.251	0.202	0.945	0.038	0.161	0.143	0.949
β_1	0.5	-0.005	0.242	0.200	0.942	-0.002	0.186	0.142	0.943	0.000	0.125	0.101	0.947
β_2	0.05	-0.001	0.009	0.004	0.922	-0.001	0.007	0.003	0.939	0.000	0.003	0.002	0.942
α	3.0	2.150	2.014	1.833	0.908	1.850	1.319	1.142	0.923	0.839	0.981	0.954	0.931
β_0	1.0	0.090	0.369	0.316	0.914	0.050	0.299	0.245	0.929	0.040	0.214	0.190	0.943
β_1	0.5	-0.049	0.241	0.206	0.932	-0.046	0.187	0.143	0.933	-0.038	0.111	0.097	0.944
β_2	0.05	-0.006	0.009	0.006	0.917	-0.005	0.008	0.005	0.933	-0.004	0.005	0.004	0.941
α	3.0	2.202	1.263	1.029	0.902	1.456	1.099	0.878	0.914	1.141	0.725	0.697	0.935
β_0	0.5	0.057	0.349	0.277	0.930	0.036	0.243	0.192	0.944	0.028	0.151	0.136	0.949
β_1	0.5	0.037	0.271	0.203	0.929	0.019	0.160	0.135	0.935	0.013	0.123	0.101	0.943
β_2	0.025	0.000	0.009	0.004	0.957	0.000	0.007	0.003	0.952	0.000	0.003	0.002	0.950
α	3.0	2.378	1.295	1.075	0.912	1.670	1.091	0.914	0.925	0.947	0.619	0.600	0.941
β_0	1.0	0.035	0.356	0.287	0.948	0.031	0.251	0.193	0.949	0.022	0.178	0.152	0.950
β_1	0.5	0.011	0.261	0.198	0.930	0.005	0.184	0.139	0.932	0.001	0.119	0.097	0.943
β_2	0.025	0.000	0.008	0.004	0.939	0.000	0.007	0.003	0.946	0.000	0.003	0.002	0.949
α	5.0	2.419	2.514	2.297	0.902	1.926	1.894	1.640	0.930	1.503	1.212	1.199	0.937
β_0	0.5	0.060	0.351	0.274	0.962	0.043	0.231	0.187	0.958	0.030	0.134	0.117	0.952
β_1	0.5	-0.007	0.246	0.177	0.959	-0.002	0.157	0.116	0.957	-0.001	0.099	0.086	0.956
β_2	0.05	-0.001	0.008	0.004	0.961	-0.001	0.007	0.003	0.960	0.000	0.003	0.002	0.957
α	5.0	2.134	2.152	1.958	0.904	1.069	1.419	1.275	0.912	0.825	0.974	0.951	0.939
β_0	1.0	0.082	0.362	0.270	0.910	0.078	0.266	0.213	0.934	0.044	0.200	0.181	0.947
β_1	0.5	-0.019	0.253	0.183	0.957	-0.015	0.184	0.136	0.954	-0.005	0.119	0.092	0.952
β_2	0.050	-0.005	0.009	0.005	0.902	-0.005	0.008	0.005	0.922	-0.004	0.005	0.004	0.939
α	5.0	1.354	2.055	1.728	0.902	1.029	1.462	1.284	0.919	0.899	1.034	0.995	0.932
β_0	0.5	0.025	0.314	0.246	0.959	0.018	0.233	0.179	0.954	0.013	0.145	0.126	0.953
β_1	0.5	0.014	0.256	0.186	0.958	0.008	0.176	0.128	0.957	0.008	0.099	0.084	0.944
β_2	0.025	0.000	0.008	0.004	0.958	0.000	0.007	0.003	0.956	0.000	0.003	0.002	0.952
α	5.0	1.768	2.263	1.928	0.922	1.483	1.500	1.396	0.930	1.156	1.127	1.091	0.938
β_0	1.0	-0.007	0.325	0.257	0.962	-0.005	0.221	0.177	0.955	-0.003	0.152	0.131	0.952
β_1	0.5	0.006	0.254	0.186	0.960	0.002	0.187	0.136	0.956	0.000	0.100	0.084	0.954
β_2	0.025	0.000	0.009	0.004	0.939	0.000	0.007	0.002	0.940	0.000	0.003	0.002	0.942

In practice, covariates may have any kind of association. Therefore, we assume that the values of one covariate depends on the other. In short, for $i = 1, \dots, n$, the steps to generate datasets are the following:

- Draw $x_{1i} \sim U(10, 90)$ (the uniform distribution).
- Draw $x_{2i} \sim \text{Bernoulli}(\theta_i)$, where $\theta_i = \exp(2 - 0.025 x_{1i}) / [1 + \exp(2 - 0.025 x_{1i})]$, i.e., $x_{2i} = 1$ with probability θ_i that varies between 0.438 and 0.852 depending on the value of x_{1i} .
- Compute $\mu_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$ and draw $W_i \sim \text{HN}(\sigma_i)$ independent from $Z_i \sim \text{Beta}(\alpha, 1)$, where $\sigma_i = \sqrt{2} \mu_i \alpha / [\sqrt{\pi} (\alpha - 1)]$.
- Compute $Y_i = W_i / Z_i$.

Once generated, the values of \mathbf{x}_i , $i = 1, \dots, n$, are kept fixed throughout the simulations. For each generated sample, we apply the scheme described in Section 3.2 to estimate $\boldsymbol{\beta}$ and α , while the standard errors of the estimates are computed from the Hessian matrix in Section 3.2. We report the average bias of the estimates (Bias), the average of the asymptotic standard error (SE), the square root of the simulated mean squared error (RMSE) and the coverage probability of the 95% asymptotic confidence intervals (CP).

We considered four different regression coefficients $\boldsymbol{\beta}$, namely, (0.5, 0.5, 0.05), (1.0, 0.5, 0.05), (0.5, 0.5, 0.025) and (1.0, 0.5, 0.025). Such values guarantee that the drawn values of y_i belong to the interval (1.649, 4.711) in all the cases. We also considered $\alpha \in \{2.5, 3.0, 5.0\}$ (that guarantees a finite value for the variance of y_i) and $n \in \{50, 100, 200\}$. The results presented in Table 3 were obtained from 1000 replications. Note that in all cases, the absolute value of bias and the RMSE decrease when n increases, suggesting that the estimators are consistent, and the coverage probabilities are close to the nominal value, as expected. Except for the estimator of α , we see that SE and RMSE get closer when the sample size increases, as expected from the asymptotic properties of the estimators. However, even for $n = 200$ the bias of $\hat{\alpha}$ is substantial. This result is in agreement with other slashed distributions in the literature (see, for instance, Astorga [2] and Reyes *et al.* [27, 26, 25]). This should not be a serious concern because in practice the most important inferences pertain to the mean of the response variable, which depends only on the regression coefficients vector $\boldsymbol{\beta}$. Additionally, since the coverage probability of the confidence interval for α ranges from 0.902 to 0.941, we see that the interval estimator behaves better than the point estimator.

5.2. Deviance and quantile residuals

In order to assess the performance of the distribution of the deviance and quantile residuals, we take samples drawn from the $\text{RG}(\mu_i, \phi=1)$ model (which also corresponds to the $\text{RW}(\mu_i, \delta=1)$ model) and $\text{RSHN}(\mu_i, \alpha=2.1)$ models, where $\mu_i = \exp(\beta_0 + \beta_1 x_i)$, and x_i was drawn from the $U(0, 10)$ distribution. For each sample, we fit the RSHN, RG, RW and RBS regression models and present the quantile-quantile (QQ) plots with simulated envelopes based on 1000 replicates for the deviance and quantile residuals. We consider three sample sizes: $n = 50$, $n = 100$ and $n = 200$. We also present the p -value for the SW, AD and CVM normality tests. Tables 4 and 5 show the QQ plots.

As expected, when the true model is the RG model, the QQ plots related to the RG and RW models present an approximately linear behavior and a good agreement with the standard normal distribution for the three sample sizes for both, deviance and quantile residuals.

Table 4: QQ plots with simulated envelopes for the deviance and quantile residuals when $RG(\mu_i, \phi = 1)$ is the true model.

Residual	n	Fitted model					
		RSHN	RG	RW	RBS		
Deviance	50						
		100					
			200				
	Quantile	50					
			100				
				200			

Moreover, the three normality tests do not reject the hypothesis of normality under the common significance levels. In counterpart, in this case the RSHN regression models yields unsatisfactory results and the normality assumption of the residuals is questionable.

Table 5: QQ plots with simulated envelopes for the deviance and quantile residuals when $RSHN(\mu_i, \alpha = 2.1)$ is the true model.

Residual	n	Fitted model					
		RSHN	RG	RW	RBS		
Deviance	50						
		100					
			200				
	Quantile	50					
			100				
				200			

When the true model is the RSHN model, as expected, the QQ plots for the deviance and quantile residuals of the RSHN model present a good agreement with the standard normal for all sample sizes. In addition, the deviance residuals for the RG and RW models only provides fair results when $n = 50$. This result suggest that the RG and RW regression models are very competitive in small sample sizes, even when the true model is not the RG model or the RW model. Finally, the deviance and quantile residuals of the RBS regression model are far away from the identity line in all the cases, suggesting poor results when the true model is the RG model or the RSHN model.

6. DATA ANALYSIS

In this section, the regression models formulated in Section 2 are applied in the analysis of two datasets.

6.1. Hunter-gatherer group dataset

In this section, the regression models formulated in Section 2 are applied in the analysis of a dataset described in Kelly [16]. The dataset is related to the macroecological relationship between the size of the homerange (measured in km^2) of a hunter-gatherer group (response variable) and the contribution (in percentage) of hunted foods to the diet. The dataset comprises 39 groups. The sample mean, median and standard deviation of the size of the homerange are 4004.4, 906.0 and 10728.1 km^2 , respectively, while the sample skewness and kurtosis coefficients are $\sqrt{\hat{\nu}_3} = 4.46$ and $\hat{\nu}_4 = 23.43$.

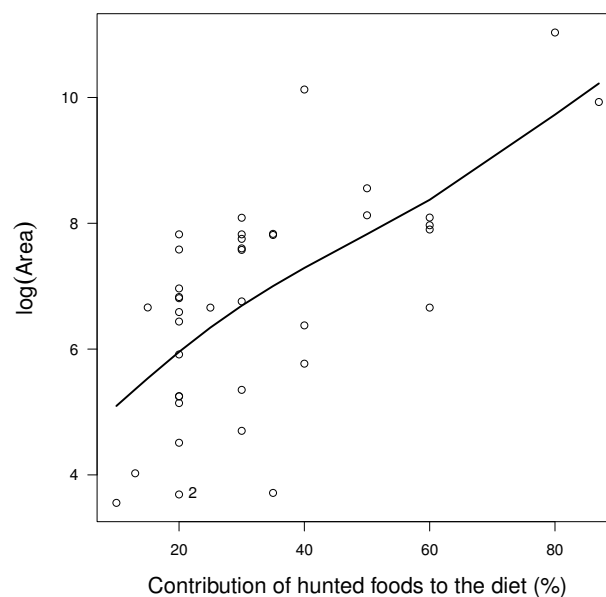


Figure 2: Scatterplot and smoothing spline of the homerange, in 1000 km^2 , and the contribution of hunted foods to the diet (observation 2 was perturbed in the analysis).

Figure 2 shows the scatterplot of the data and a smoothing spline, which indicates that the logarithmic link function is adequate. We fit the RG, RW, RBS and RSHN models, with results presented in Table 6. The deviance and quantile residuals plots with envelopes are presented in the upper panels in Figures 4 and 5. The lines in these plots represent the 2.5%, 50% and 97.5% quantile values of the residuals computed from 100 bootstrap samples generated from the models in Table 6. Note that, based on both residuals, all models seem appropriate for this dataset. Furthermore, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) values are similar for all models.

Table 6: Parameter estimates (standard errors) and information criteria for the RG, RW, RBS and RSHN regression models fitted to the hunter-gatherer group dataset.

Dataset	Parameter	Model			
		RG	RW	RBS	RSHN
Unperturbed	β_0	5.442 (0.504)	5.456 (0.436)	5.290 (0.478)	5.718 (0.136)
	β_1	0.063 (0.013)	0.062 (0.012)	0.067 (0.013)	0.059 (0.010)
	α	—	—	—	2.225 (1.541)
	ϕ	0.811 (0.159)	—	—	—
	δ	—	0.845 (0.100)	—	—
	ξ	—	—	0.805 (0.227)	—
	AIC	670.26	669.25	668.02	670.11
	BIC	675.25	674.24	673.01	675.10
Perturbed	β_0	6.588 (0.759)	6.345 (0.482)	6.407 (0.491)	6.332 (0.340)
	β_1	0.042 (0.020)	0.047 (0.013)	0.048 (0.014)	0.054 (0.013)
	α	—	—	—	1.517 (1.301)
	ϕ	0.602 (0.115)	—	—	—
	δ	—	0.695 (0.078)	—	—
	ξ	—	—	0.587 (0.227)	—
	AIC	698.80	693.57	691.66	688.20
	BIC	703.79	698.56	696.64	693.19

In order to illustrate the robustness of the RSHN model, we perturb the response variable of observation 2 in Figure 2 by adding two standard deviations (originally with an area of 4,000 km²). The lower panels in Figures 4 and 5 show the deviance and the quantile residuals plots for the models fitted to the perturbed data. Note that for both residuals, the SW, AD and CVM tests support that the residuals of the RSHN model come from the standard normal distribution for datasets without and with perturbation. This fact suggests that the RG, RW and RBS models do not yield a good fit for the perturbed dataset, differently from the RSHN model, which yields a good fit in both scenarios. Information criteria for the perturbed dataset in Table 6 also suggest that the best fit is achieved with the RSHN model. Due to the perturbation, estimates of the coefficient of the contribution of hunted foods to the diet (β_1) decrease 33.3%, 24.2% and 28.4% under the RG, RW and RBS models, respectively, whereas for the RSHN model the reduction amounts to 8.5%. Estimated means of the homerange for unperturbed and perturbed data are displayed in Figure 3. We stress that the ratio of the estimated area for unperturbed data to perturbed data is much more stable for the RSHN model, especially for large values of the contribution of hunted foods to the diet, as can be seen in Figure 3(c).

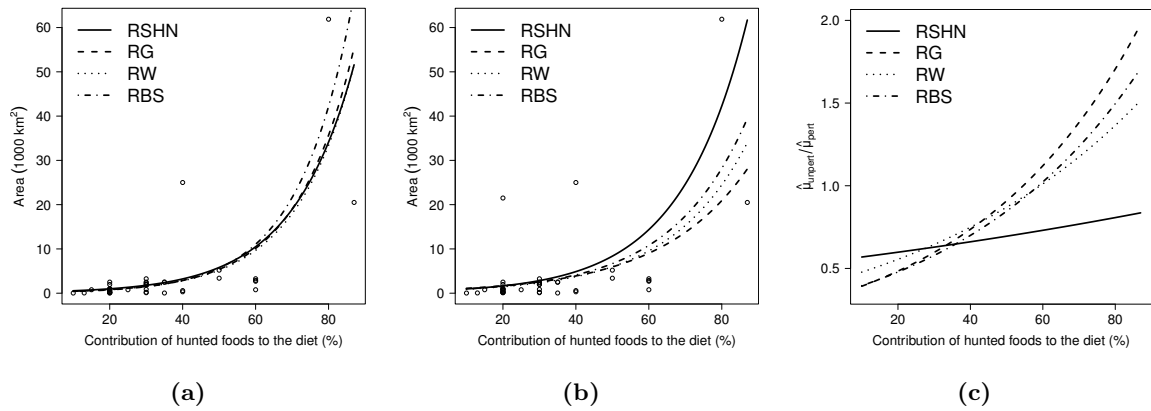


Figure 3: Scatterplot of the homerange and the contribution of hunted foods to the diet together with estimated means under different models for data (a) without perturbation and (b) with perturbation, and (c) ratio of the estimated area for unperturbed data ($\hat{\mu}_{unpert}$) to perturbed data ($\hat{\mu}_{pert}$).

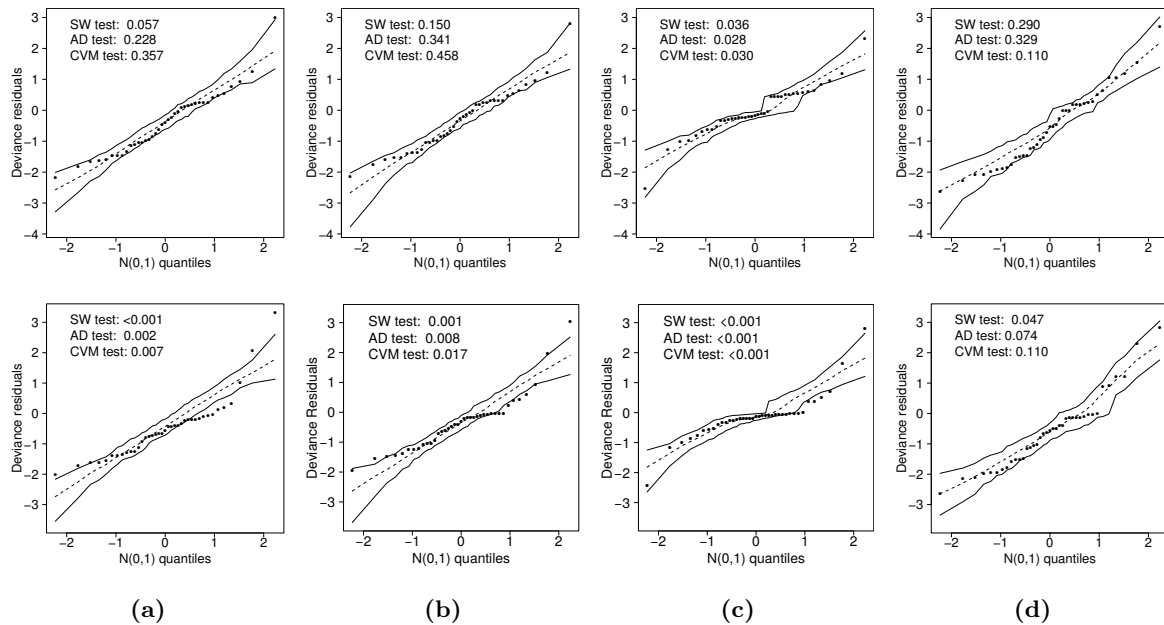


Figure 4: Deviance residual plots with simulated envelopes for the (a) RG, (b) RW, (c) RBS and (d) RSHN regression models fitted to the hunter-gatherer group dataset without perturbation (upper panel) and with perturbation (lower panel).

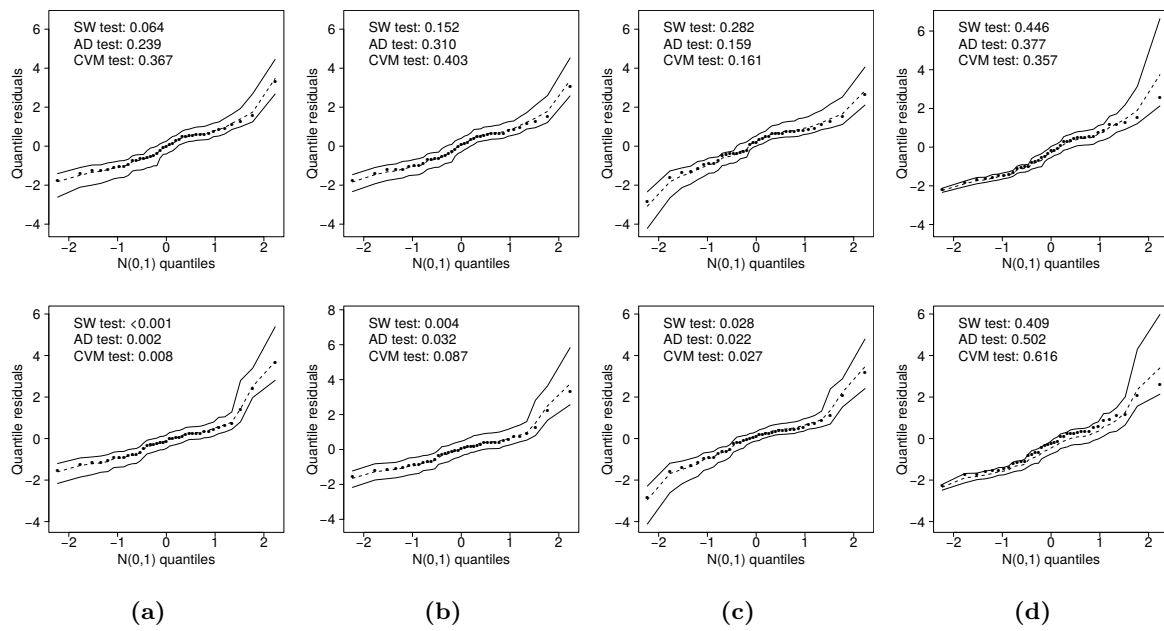


Figure 5: Quantile residual plots with simulated envelopes for the (a) RG, (b) RW, (c) RBS and (d) RSHN regression models fitted to the hunter-gatherer group dataset without perturbation (upper panel) and with perturbation (lower panel).

6.2. Minerals concentration dataset

This dataset is related to the concentration of some minerals in soil samples obtained at the Mining Department, University of Atacama, Chile. This dataset was previously analyzed in Gómez *et al.* [14] and Olmos *et al.* [21]. The measurements are related to nickel (Ni) and zinc (Zn) respectively. In our application, we consider to model jointly the positive measurements related to thorium (Th, $n = 71$), uranium (U, $n = 57$), vanadium (V, $n = 86$) and zinc (Zn, $n = 86$). The unit of measurement of the concentrations (response variable) is parts-per million (ppm). The dataset comprises 300 observations. The sample mean, median and standard deviation of the concentrations are 72.43, 29.00 and 110.06, respectively, while the sample skewness and kurtosis coefficients are $\sqrt{\hat{\nu}_3} = 4.37$ and $\hat{\nu}_4 = 35.87$. Note that the kurtosis is unusually greater than the kurtosis of the normal distribution. Given the high value of kurtosis, we consider appropriate to model this dataset with the RSHN model in Section 2, linking the covariates to the mean as $\mu_i = \exp(\beta_{Th} x_{iTh} + \beta_U x_{iU} + \beta_V x_{iV} + \beta_{Zn} x_{iZn})$, $i = 1, \dots, 300$, where x_{iTh} , x_{iU} , x_{iV} and x_{iZn} are indicator variables assuming the value 1 when the i -th observation corresponds to the referred mineral. We also compare the results with the RG, RW and RBS regression models. Results are presented in Table 7. Note that AIC and BIC attain the smallest values for the RSHN model. Figure 6 shows the histogram of thorium and zinc concentrations compared with the fitted density functions. Table 8 also presents the p -value for the univariate Kolmogorov–Smirnov (KS) test for comparison of empirical and fitted cdf’s from each mineral. Note that all p -values are greater than 5% for the RSHN model, suggesting a better fit for this model over the RG, RW and RBS models.

Table 7: Parameter estimates (standard errors) and information criteria for the RG, RW, RBS and RSHN regression models fitted to the minerals dataset.

Parameter	Model			
	RG	RW	RBS	RSHN
β_{Th}	2.871 (0.119)	2.866 (0.110)	3.005 (0.146)	2.989 (0.127)
β_U	2.436 (0.133)	2.434 (0.123)	2.508 (0.155)	2.581 (0.136)
β_V	4.896 (0.108)	4.892 (0.100)	4.646 (0.107)	5.071 (0.124)
β_{Zn}	4.572 (0.108)	4.589 (0.101)	4.555 (0.122)	4.458 (0.114)
α	—	—	—	2.871 (2.541)
ϕ	1.206 (0.088)	—	—	—
δ	—	1.080 (0.046)	—	—
ξ	—	—	1.147 (0.082)	—
AIC	2917.55	2920.67	2980.57	2906.97
BIC	2936.07	2939.18	2999.09	2925.49

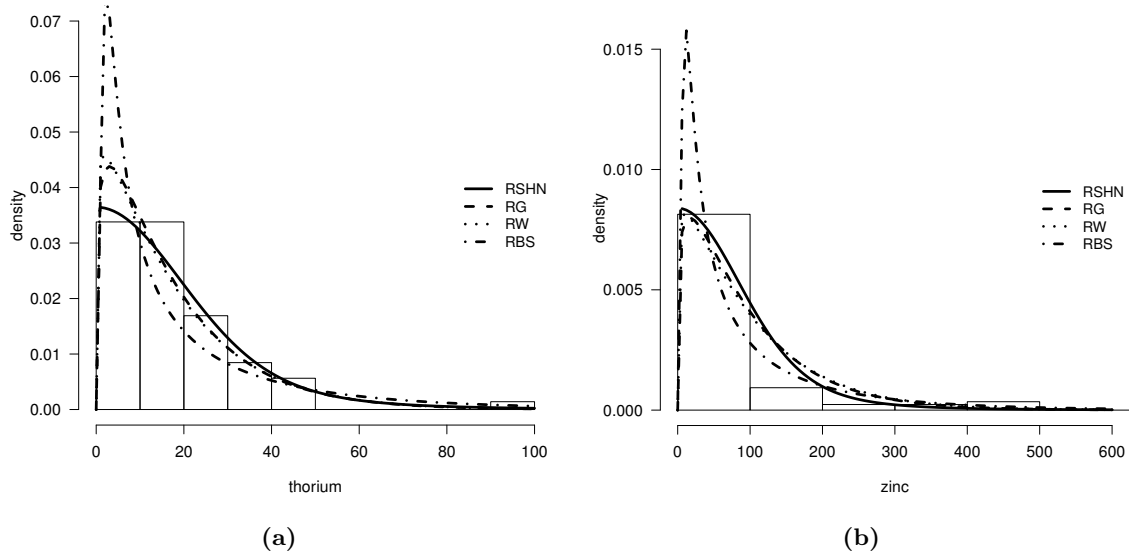


Figure 6: Histogram and fitted density functions for RSHN, RG, RW and RBS models in minerals dataset: (a) thorium and (b) zinc.

Table 8: p -values for the Kolmogorov–Smirnov goodness-of-fit test.

Mineral	RG	RW	RBS	RSHN
Th	0.269	0.195	0.009	0.580
U	0.947	0.955	0.119	0.535
V	0.105	0.112	<0.001	0.348
Zn	0.003	0.002	0.040	0.065

Besides the information criteria in Table 7, Figures 7 and 8 show the deviance and the quantile residuals plots for the fitted models. Note that for both residuals, the SW, AD and CVM tests support (at a 5% significance level) that only the residuals of the RSHN model come from the standard normal distribution. This fact suggests that the RG, RW and RBS models do not yield a good fit for this dataset.

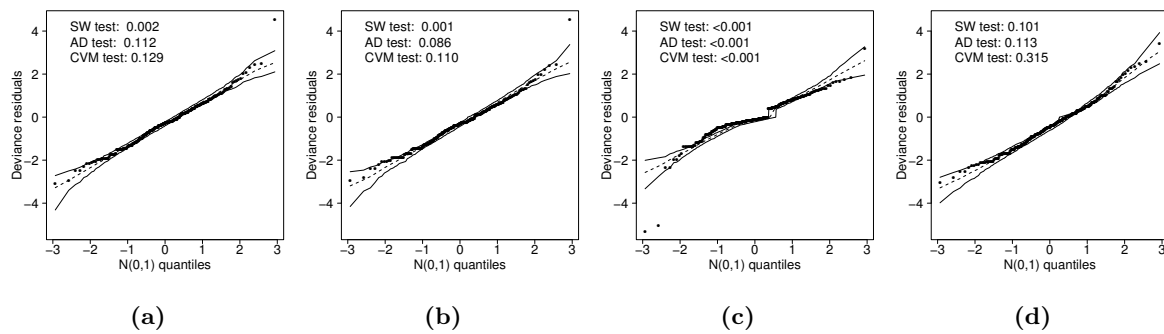


Figure 7: Deviance residual plots with simulated envelopes for the (a) RG, (b) RW, (c) RBS and (d) RSHN regression models fitted to the minerals dataset.

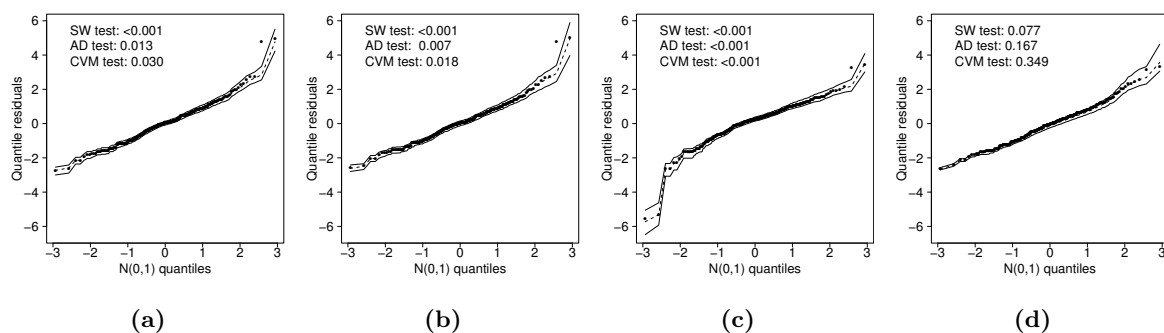


Figure 8: Quantile residual plots with simulated envelopes for the (a) RG, (b) RW, (c) RBS and (d) RSHN regression models fitted to the minerals dataset.

7. CONCLUSION

In this work, a reparameterization of the distribution proposed by Olmos *et al.* [21] based on the mean motivated us to propose a regression model for positive data. The proposed model is an alternative to some well-known models for positive response variables. Maximum likelihood estimates are computed with the EM algorithm. A simulation study was carried out to assess some properties of the proposed estimator. The analysis of two datasets illustrates the robustness of the model. Extensions of this work might include Bayesian inference, influence assessment and mixed models.

ACKNOWLEDGMENTS


We would like to thank the Editor, an AE and two anonymous reviewers for their time and valuable remarks. The work of the first author is partially supported by Proyecto DIUDA Programa de Inserción N.º 22367, Chile. The work of the second author is partially supported by grant FONDECYT 11160670, Chile. The work of the third author is partially supported by CNPq, Brazil.


REFERENCES

- [1] ASGHARZADEH, A.; NADARAJAH, S. and SHARAFI, F. (2018). Weibull Lindley distribution, *REVSTAT – Statistical Journal*, **16**, 87–113.
- [2] ASTORGA, J.M.; GÓMEZ, H.W. and BOLFARINE, H. (2017). Slashed generalized exponential distribution, *Communications in Statistics – Theory and Methods*, **46**, 2091–2102.
- [3] ATKINSON, A.C. (1985). *Plots, Transformations, and Regression*, Clarendon, Oxford.
- [4] AZZALINI, A. (1985). A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, **12**, 171–178.
- [5] AZZALINI, A. (1986). Further results on a class of distributions which includes the normal ones, *Statistics*, **12**, 199–208.
- [6] BOURGUIGNON, M.; LEÃO, J.; LEIVA, V. and SANTOS-NETO, M. (2017). The transmuted Birnbaum–Saunders distribution, *REVSTAT – Statistical Journal*, **15**, 601–628.
- [7] CHOU, C.Y. and LIU, H.R. (1998). Properties of the half-normal distribution and its application to quality control, *Journal of Industrial Technology*, **14**, 4–7.
- [8] COORAY, K. and ANANDA, M.M.A. (2008). A generalization of the half-normal distribution with applications to lifetime data, *Communications in Statistics – Theory and Methods*, **10**, 195–224.
- [9] CORDEIRO, G.M.; PESCIM, R.R. and ORTEGA, E.M.M. (2012). The Kumaraswamy generalized half-normal distribution for skewed positive data, *Journal of Data Science*, **10**, 195–224.
- [10] DEMPSTER, A.P.; LAIRD, N.M. and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.
- [11] DUNN, P.K. and SMYTH, G.K. (1996). Randomized quantile residuals, *Journal of Computational and Graphical Statistics*, **5**, 236–244.
- [12] GILBERT, P. and VARADHAN, R. (2016). numDeriv: Accurate Numerical Derivatives, *R package version 2016.8-1*, <http://CRAN.R-project.org/package=numDeriv>.
- [13] GÓMEZ, Y.M. and BOLFARINE, H. (2015). Likelihood-based inference for power half-normal distribution, *Journal of Statistical Theory and Applications*, **14**, 383–398.
- [14] GÓMEZ, H.W.; VENEGAS, O. and BOLFARINE, H. (2006). Skew-symmetric distributions generated by the distribution function of the normal distribution, *Envirometrics*, **18**, 395–407.
- [15] HENZE, N. (1986). A probabilistic representation of the skew-normal distribution, *Scandinavian Journal of Statistics*, **13**, 271–275.
- [16] KELLY, R.L. (2013). *The Lifeways of Hunter-Gatherers: The Foraging Spectrum*, 2nd ed., Cambridge, New York.

- [17] KHAN, M.A. and ISLAM, H.N. (2012). Bayesian analysis of system availability with half-normal life time, *The Mathematical Scientist*, **9**, 203–209.
- [18] LEIVA, V.; FERREIRA, M.; GOMES, M.I. and LILLO, C. (2016). Extreme value Birnbaum–Saunders regression models applied to environmental data, *Stochastic Environmental Research and Risk Assessment*, **30**, 1045–1058.
- [19] LIN, T.T.; LEE, J.C. and HSIEH, W.J. (2007). Robust mixture modeling using the skew t distribution, *Statistics and Computing*, **17**, 81–92.
- [20] MORAL, R.A.; HINDE, J. and DEMÉTRIO, C.G.B. (2017). Half-normal plots and overdispersed models in R: the hnp package, *Journal of Statistical Software*, **81**, 1–23.
- [21] OLMOS, N.M.; VARELA, H.; GÓMEZ, H.W. and BOLFARINE, H. (2012). An extension of the half-normal distribution, *Statistical Papers*, **53**, 875–886.
- [22] PEWSEY, A. (2002). Large-sample inference for the general half-normal distribution, *Communications in Statistics – Theory and Methods*, **31**, 1045–1054.
- [23] PEWSEY, A. (2004). Improved likelihood based inference for the general half-normal distribution, *Communications in Statistics – Theory and Methods*, **33**, 197–204.
- [24] R CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org/>.
- [25] REYES, J.; IRIARTE, Y.; JODRÁ, P. and GÓMEZ, H.W. (2019). The slash Lindley–Weibull distribution, *Methodology and Computing in Applied Probability*, **21**, 235–251.
- [26] REYES, J.; VENEGAS, O. and GÓMEZ, H.W. (2017a). Modified slash Lindley distribution, *Journal of Probability and Statistics*, Article ID 6303462, 1–9.
- [27] REYES, J.; VILCA, F.; GALLARDO, D.I. and GÓMEZ, H.W. (2017b). Modified slash Birnbaum–Saunders distribution, *Haceteppe Journal of Mathematics and Statistics*, **46**, 969–984.
- [28] RIGBY, R.A. and STASINOPOULOS, D.M. (2005). Generalized additive models for location, scale and shape (with discussion), *Applied Statistics*, **54**, 507–554.
- [29] RIGBY, R.A. and STASINOPOULOS, D.M. (2019). gamlss.dist: distributions for generalized additive models for location scale and shape, *R package version 5.1-3*, <https://CRAN.R-project.org/package=gamlss.dist>.
- [30] SANTOS-NETO, M.; CYSNEIROS, F.J.A.; LEIVA, V. and BARROS, M. (2014). On a reparameterized Birnbaum–Saunders distribution and its moments, estimation and applications, *REVSTAT – Statistical Journal*, **12**, 247–272.
- [31] SANTOS-NETO, M.; CYSNEIROS, F.J.; LEIVA, V. and BARROS, M. (2016). Reparameterized Birnbaum–Saunders regression models with varying precision, *Electronic Journal of Statistics*, **10**, 2825–2855.
- [32] WIPER, M.P.; GIRÓN, F.J. and PEWSEY, A. (2008). Objective Bayesian inference for the half-normal and half- t distributions, *Communications in Statistics – Theory and Methods*, **37**, 3165–3185.
- [33] YAZICI, B. and YOCALAN, S. (2007). A comparison of various tests of normality, *Journal of Statistical Computation and Simulation*, **77**, 175–183.

COMPARISON OF THE LIKELIHOOD RATIOS OF TWO DIAGNOSTIC TESTS SUBJECT TO A PAIRED DESIGN: CONFIDENCE INTERVALS AND SAMPLE SIZE

Authors: JOSÉ ANTONIO ROLDÁN-NOFUENTES 
– Statistics (Biostatistics), School of Medicine, University of Granada,
Granada, Spain
jaroldan@ugr.es

SAAD BOUH SIDATY-REGAD 
– Public Health and Epidemiology, School of Medicine, University of Nouakchott,
Mauritania

Received: November 2018

Revised: June 2019

Accepted: September 2019

Abstract:

- Positive and negative likelihood ratios are parameters which are used to assess and compare the effectiveness of binary diagnostic tests. Both parameters only depend on the sensitivity and specificity of the diagnostic test and are equivalent to a relative risk. This article studies the comparison of the likelihood ratios of two binary diagnostic tests subject to a paired design through confidence intervals. Six approximate confidence intervals are presented for the ratio of the likelihood ratios, and simulation experiments are carried out to study the coverage probabilities and the average lengths of the intervals considered, and some general rules of application are proposed. A method is also proposed to determine the sample size necessary to estimate the ratio between the likelihood ratios with a determined precision. The results were applied to two real examples.

Keywords:

- *binary diagnostic test; likelihood ratios; sample size.*

AMS Subject Classification:

- 62P10, 6207.

1. INTRODUCTION

A diagnostic test is a medical test that is applied to an individual in order to determine the presence or absence of a disease. When the result of a diagnostic test is positive or negative, the diagnostic test is called a binary diagnostic test (BDT). A stress test for the diagnosis of coronary disease is an example of BDT. The effectiveness of a BDT is measured in terms of two fundamental parameters: sensitivity and specificity. The sensitivity (Se) is the probability of the BDT being positive when the individual has the disease, and the specificity (Sp) is the probability of the BDT being negative when the individual does not have it. The Se and the Sp of a BDT are estimated in relation to a gold standard (GS), which is a medical test which objectively determines whether or not an individual has the disease or not. An angiography for coronary disease is an example of GS. Other parameters that are used to assess the effectiveness of a BDT are the likelihood ratios (LRs) ([10, 17]). When the BDT is positive, the likelihood ratio, called the positive likelihood ratio (LR^+), is the ratio between the probability of correctly classifying an individual with the disease and the probability of incorrectly classifying an individual who does not have it. When the BDT is negative, the likelihood ratio, called the negative likelihood ratio (LR^-), is the ratio between the probability of incorrectly classifying an individual who has the disease and the probability of correctly classifying an individual who does not have it. The LRs only depend on the sensitivity and the specificity of the BDT and do not depend on the disease prevalence, and therefore the LRs are superior parameters of the accuracy of a BDT ([10, 17]). The comparison of the parameters of two BDTs has been the subject of numerous studies in Statistical literature. When the two BDTs and the GS are applied to all of the individuals in a random sample sized n (paired design), the comparison of the two sensitivities (specificities) is made by applying a comparison test of two paired binomial proportions. Subject to this same sample design, the comparison of the LRs of two BDTs is more complex. Leisenring and Pepe [6] studied the estimation of the LRs of a BDT through a regression model. Pepe [10] adapted this model to compare the LRs of two BDTs, for which in the regression model a variable dummy is considered to compare a BDT in relation to another. Moreover, Pepe [10] proposed a confidence interval for the ratio of the two positive (negative) LRs estimating the variance of the ratios subject to the null hypothesis of equality of the two LRs. Section 3.1 summarizes the method of Pepe [10]. Biggerstaff [1] proposed a graphical method to compare the LRs of two (or more) BDTs. Nevertheless, this method is not inferential and can only be applied to the estimators. Roldán-Nofuentes and Luna [12] studied hypothesis tests to compare the LRs individually and simultaneously, and they also studied the same problem for the case of ordinal diagnostic tests. The hypothesis tests proposed by Roldán-Nofuentes and Luna [12] are based on the logarithmic transformation of the ratio of the positive (negative) LRs, and therefore by inverting the test statistics of the individual tests, confidence intervals are obtained for the ratio of the two LRs (in Section 3.2 we summarize this method). Dolgun *et al.* [3] extended the method of Leisenring and Pepe [6] to compare the LRs simultaneously. Comparing the sensitivities (specificities) of two BDTs, we compare the intrinsic accuracy of both BDTs, and we determined which BDT is more accurate for an individual who has the disease (which BDT has the greatest sensitivity) or for an individual who does not have the disease (which BDT has the greatest specificity). Comparing the positive (negative) LRs of two BDTs it is possible to quantify with which BDT it is more likely to obtain a positive (negative) result for the BDT for an individual who has the disease than for an individual who does not.

In this manuscript we study the comparison of the LRs of two BDTs through confidence intervals (CIs), making the following contributions: a) four intervals to compare the LRs, and b) a method to calculate the sample size to compare the LRs through CIs. Section 2 presents the LRs and their properties. Section 3 presents the CIs studied by Pepe [10], by Roldán-Nofuentes and Luna [12], and four new CIs are proposed: a Wald type interval, an interval based on the Fieller method, a Bootstrap interval based on the bias-corrected interval, and a Bayesian interval based on non-informative beta distributions and on the application of the Monte Carlo method. In Section 4, simulation experiments are carried out to study the coverage probabilities and the average lengths of the CIs presented in Section 3. Section 5 presents a method to calculate the sample size to compare the LRs through CIs. In Section 6, the results are applied to two real examples, and in Section 7 the results obtained are discussed.

2. LIKELIHOOD RATIOS

Let us consider a BDT that is assessed in relation to a GS. Let T be the variable that models the result of the BDT: $T = 1$ when the BDT is positive and $T = 0$ when it is negative. Let D be the variable that models the result of the GS: $D = 1$ when the individual has the disease and $D = 0$ when this is not the case. Let $\pi = P(D=1)$ be the disease prevalence in the population studied, and $\bar{\pi} = 1 - \pi$. The positive LR ([10, 17]) is defined as

$$(2.1) \quad LR^+ = \frac{P(T=1 \mid D=1)}{P(T=1 \mid D=0)} = \frac{Se}{1 - Sp},$$

and the negative LR as

$$(2.2) \quad LR^- = \frac{P(T=0 \mid D=1)}{P(T=0 \mid D=0)} = \frac{1 - Se}{Sp}.$$

The LRs vary between 0 and infinity, and have the following properties:

- a) If the BDT and the GS are independent then $LR^+ = LR^- = 1$.
- b) If the BDT correctly classifies all of the individuals then $LR^+ = \infty$ and $LR^- = 0$.
- c) If $LR^+ > 1$ then a positive result of the BDT is more probable for an individual who has the disease than for an individual who does not.
- d) If $LR^- < 1$ then a negative result of the BDT is more probable for an individual who does not have the disease than for an individual who does.
- e) The LRs quantify the increase in knowledge of the presence of the disease through the application of the BDT. Before applying the BDT, the odds of an individual having the disease are

$$\text{pre-test odds} = \frac{\pi}{1 - \pi},$$

where π is the disease prevalence. After applying the BDT, the odds are

$$\text{post-test odds} = \frac{P(D=1 \mid T=i)}{P(D=0 \mid T=i)}, \quad i = 0, 1.$$

The LRs relate the pre-test odds and the post-test odds:

$$\begin{aligned} \text{post-test odds } (T=1) &= LR^+ \times \text{pre-test odds,} \\ \text{post-test odds } (T=0) &= LR^- \times \text{pre-test odds.} \end{aligned}$$

Therefore, the likelihood ratios quantify the change in the odds of the disease obtained by knowledge of the application of the BDT.

We then study the comparison of the LRs of two BDTs subject to a paired design through CIs.

3. CONFIDENCE INTERVALS

Let us consider two BDTs that are assessed in relation to the same GS. Let T_h be the variable that models the result of the h -th BDT, with $h = 1, 2$, defined in a similar way to the variable T given in Section 2. Let Se_h and Sp_h be the sensitivity and the specificity of the h -th BDT, and LR_h^+ and LR_h^- the positive and negative likelihood ratios respectively. Table 1 shows the frequencies and the theoretical probabilities obtained when comparing two BDTs in relation to a GS subject to a paired design. In the observed frequencies given in Table 1, the only value set by the researcher is the sample size n .

Table 1: Frequencies and probabilities subject to a paired design.

Frequencies					
	$T_1 = 1$		$T_1 = 0$		Total
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	
$D = 1$	s_{11}	s_{10}	s_{01}	s_{00}	s
$D = 0$	r_{11}	r_{10}	r_{01}	r_{00}	r
Total	$s_{11} + r_{11}$	$s_{10} + r_{10}$	$s_{01} + r_{01}$	$s_{00} + r_{00}$	n

Probabilities					
	$T_1 = 1$		$T_1 = 0$		Total
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	
$D = 1$	p_{11}	p_{10}	p_{01}	p_{00}	π
$D = 0$	q_{11}	q_{10}	q_{01}	q_{00}	$\bar{\pi}$
Total	$p_{11} + q_{11}$	$p_{10} + q_{10}$	$p_{01} + q_{01}$	$p_{00} + q_{00}$	1

Applying the model of conditional dependence of Vacek [14], the theoretical probabilities are expressed as

$$(3.1) \quad p_{ij} = \pi \left[Se_1^i (1 - Se_1)^{1-i} Se_2^j (1 - Se_2)^{1-j} + \delta_{ij} \varepsilon_1 \right]$$

and

$$(3.2) \quad q_{ij} = \bar{\pi} \left[Sp_1^{1-i} (1 - Sp_1)^i Sp_2^{1-j} (1 - Sp_2)^j + \delta_{ij} \varepsilon_0 \right],$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = -1$ if $i \neq j$, with $i, j = 0, 1$, and verifying that $\pi = \sum_{ij} p_{ij}$ and $\bar{\pi} = \sum_{ij} q_{ij}$. The parameters ε_1 and ε_0 are the dependence factors between the two BDTs when $D = 1$ and when $D = 0$ respectively, verifying that

$$0 \leq \varepsilon_1 \leq \text{Min}\{Se_1(1 - Se_2), Se_2(1 - Se_1)\}$$

and

$$0 \leq \varepsilon_0 \leq \text{Min}\{Sp_1(1 - Sp_2), Sp_2(1 - Sp_1)\}.$$

If $\varepsilon_1 = \varepsilon_0 = 0$ then the two BDTs are conditionally independent on the disease, which is not normally a realistic one. In practice, the BDTs are conditionally dependent on the disease, so that $\varepsilon_1 > 0$ and/or $\varepsilon_0 > 0$. The frequencies of Table 1 are the product of a multinomial distribution whose vector of probabilities is $\psi = (p_{11}, p_{10}, p_{01}, p_{00}, q_{11}, q_{10}, q_{01}, q_{00})^T$. The maximum likelihood estimators of these probabilities are $\hat{p}_{ij} = s_{ij}/n$ and $\hat{q}_{ij} = r_{ij}/n$, those of π and $\bar{\pi}$ are $\hat{\pi} = s/n$ and $\hat{\bar{\pi}} = r/n$, and the variance-covariance matrix of $\hat{\psi}$ is $\Sigma_{\hat{\psi}} = \{\text{diag}(\psi) - \psi\psi^T\}/n$.

In terms of the probabilities of the vector ψ , the sensitivity and the specificity of each BDT are written as $Se_1 = (p_{10} + p_{11})/\pi$, $Sp_1 = (q_{00} + q_{01})/\bar{\pi}$, $Se_2 = (p_{01} + p_{11})/\pi$ and $Sp_2 = (q_{00} + q_{10})/\bar{\pi}$. The estimators of the sensitivities and the specificities are $\widehat{Se}_1 = \frac{s_{11} + s_{10}}{s}$, $\widehat{Se}_2 = \frac{s_{11} + s_{01}}{s}$, $\widehat{Sp}_1 = \frac{r_{01} + r_{00}}{r}$ and $\widehat{Sp}_2 = \frac{r_{10} + r_{00}}{r}$, and those of the dependence factors are $\widehat{\varepsilon}_1 = \frac{\hat{p}_{11}}{\hat{\pi}} - \widehat{Se}_1\widehat{Se}_2 = \frac{s_{11}s_{00} - s_{10}s_{01}}{s}$ and $\widehat{\varepsilon}_0 = \frac{\hat{q}_{00}}{\hat{\bar{\pi}}} - \widehat{Sp}_1\widehat{Sp}_2 = \frac{r_{11}r_{00} - r_{10}r_{01}}{r}$. Applying the delta method, it holds that the variances-covariances of \widehat{Se}_h and \widehat{Sp}_h are

$$(3.3) \quad \begin{aligned} \text{Var}(\widehat{Se}_h) &\approx \frac{Se_h(1 - Se_h)}{n\pi}, & \text{Var}(\widehat{Sp}_h) &\approx \frac{Sp_h(1 - Sp_h)}{n\bar{\pi}}, \\ \text{Cov}(\widehat{Se}_1, \widehat{Se}_2) &\approx \frac{\varepsilon_1}{n\pi}, & \text{Cov}(\widehat{Sp}_1, \widehat{Sp}_2) &\approx \frac{\varepsilon_0}{n\bar{\pi}}. \end{aligned}$$

The rest of the covariances are zero. Regarding the LR_s, applying the delta method again, their variances-covariances (the proof can be seen in Appendix A) are

$$(3.4) \quad \begin{aligned} \text{Var}(\widehat{LR}_h^+) &\approx \frac{Se_h^2 \text{Var}(\widehat{Sp}_h) + (1 - Sp_h)^2 \text{Var}(\widehat{Se}_h)}{(1 - Sp_h)^4}, \\ \text{Var}(\widehat{LR}_h^-) &\approx \frac{(1 - Se_h)^2 \text{Var}(Sp_h) + Sp_h^2 \text{Var}(\widehat{Se}_h)}{Sp_h^4}, \\ \text{Cov}(\widehat{LR}_1^+, \widehat{LR}_1^+) &\approx \frac{Se_1 Se_2 \text{Cov}(\widehat{Sp}_1, \widehat{Sp}_2) + (1 - Sp_1)(1 - Sp_2) \text{Cov}(\widehat{Se}_1, \widehat{Se}_2)}{(1 - Sp_1)^2 (1 - Sp_2)^2}, \\ \text{Cov}(\widehat{LR}_1^-, \widehat{LR}_1^-) &\approx \frac{(1 - Se_1)(1 - Se_2) \text{Cov}(\widehat{Sp}_1, \widehat{Sp}_2) + Sp_1 Sp_2 \text{Cov}(\widehat{Se}_1, \widehat{Se}_2)}{Sp_1^2 Sp_2^2}. \end{aligned}$$

Substituting in the previous expressions the parameters with their estimators, we obtain the expressions of the estimators of the variances-covariances. Pepe [10] studied the comparison of the LR_s considering the ratio between them, i.e. $\omega^+ = LR_1^+/LR_2^+$ and $\omega^- = LR_1^-/LR_2^-$. Roldán-Nofuentes and Luna [12] considered the Napierian logarithm of ω . In this study, we are going to follow the same criteria as Pepe, and therefore we are going to compare the LR_s through CIs for ω^+ and ω^- . From here onwards, we are going to consider that LR_h is LR_h^+ or LR_h^- , and that ω is ω^+ or ω^- , depending on whether we compare the positive LR_s or the negative LR_s. If the CI for ω contains the value one, then we do not reject the equality of

the LRs of both BDTs; in the opposite case, the LR of a BDT is significantly higher than that of the other BDT. Applying the delta method (see Appendix A), the variance of $\widehat{\omega}$ is

$$(3.5) \quad \text{Var}(\widehat{\omega}) \approx \omega^2 \left[\frac{\text{Var}(\widehat{LR}_1)}{LR_1^2} + \frac{\text{Var}(\widehat{LR}_2)}{LR_2^2} - \frac{2 \text{Cov}(\widehat{LR}_1, \widehat{LR}_2)}{LR_1 LR_2} \right].$$

Then six CIs are presented for each ratio ω^+ and ω^- . The first interval was proposed by Pepe [10], the second is deduced from the study by Roldán-Nofuentes and Luna [12], and the rest of the intervals are contributions made by this manuscript.

3.1. Regression model

Leisenring and Pepe [6] studied the estimation of the LRs of a BDT in presence of covariates through a regression model. For the positive LR, the regression model with p covariates is $\ln(LR^+(X_1)) = \beta_0 + \sum_{i=1}^p \beta_i X_{1p}$, where β_i are the parameters of the model and $X_1 = (X_{11}, \dots, X_{1p})$ is the matrix of covariates. This model can be used to compare two BDTs ([10]), i.e. $\ln[LR^+(X_T)] = \beta_0 + \beta_1 X_T$, where X_T is a variable dummy to compare a BDT in relation to another. The regression model to compare the two negative LRs is $\ln[LR^-(X_T)] = \alpha_0 + \alpha_1 X_T$. In these models, the ratio ω^+ is estimated as $e^{\widehat{\beta}_1}$ and the ratio ω^- as $e^{\widehat{\alpha}_1}$. The confidence interval for ω^+ is

$$(3.6) \quad \widehat{\omega}^+ \times \exp \left\{ \pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}_0[\ln(\widehat{\omega}^+)]} \right\},$$

where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ -th percentile of the standard normal distribution and

$$\widehat{\text{Var}}_0[\ln(\widehat{\omega}^+)] \approx \frac{1 - \widehat{Se}_1}{s \widehat{Se}_1} + \frac{\widehat{Sp}_1}{r(1 - \widehat{Sp}_1)} + \frac{1 - \widehat{Se}_2}{s \widehat{Se}_2} + \frac{\widehat{Sp}_2}{r(1 - \widehat{Sp}_2)}$$

is the estimated variance of $\widehat{\omega}^+$ subject to the null hypothesis $H_0: LR_1^+ = LR_2^+$. The confidence interval for ω^- is similar to the previous one, where

$$\widehat{\text{Var}}_0[\ln(\widehat{\omega}^-)] \approx \frac{\widehat{Se}_1}{s(1 - \widehat{Se}_1)} + \frac{1 - \widehat{Sp}_1}{r \widehat{Sp}_1} + \frac{\widehat{Se}_1}{s(1 - \widehat{Se}_1)} + \frac{1 - \widehat{Sp}_1}{r \widehat{Sp}_1}.$$

The book by Pepe [10] discusses the confidence interval obtained from the regression model.

3.2. Logarithmic interval

Roldán-Nofuentes and Luna [12] studied a hypothesis test to compare the positive (negative) LRs of two BDTs subject to a paired design. These hypothesis tests are based on the transformation of the Napierian logarithm of the ratio between the two positive (negative) LRs, i.e., $H_0: \ln(\omega) = 0$ vs $H_1: \ln(\omega) \neq 0$, where ω is $\omega^+ = LR_1^+/LR_2^+$ or $\omega^- = LR_1^-/LR_2^-$, and the test statistic is

$$(3.7) \quad \frac{\ln(\widehat{\omega})}{\sqrt{\widehat{\text{Var}}[\ln(\widehat{\omega})]}} \rightarrow N(0,1),$$

where $\widehat{\text{Var}}[\ln(\widehat{\omega})]$ is an unrestricted estimator of the variance and is calculated applying the delta method (see Appendix A), i.e.

$$(3.8) \quad \text{Var}[\ln(\widehat{\omega})] \approx \frac{\text{Var}(\widehat{LR}_1)}{LR_1^2} + \frac{\text{Var}(\widehat{LR}_2)}{LR_2^2} - \frac{2 \text{Cov}(\widehat{LR}_1, \widehat{LR}_2)}{LR_1 LR_2},$$

and substituting in this expression each parameter with its estimator. Inverting the test statistic (3.7), it holds that the CI for $\ln(\omega)$ is $\ln(\widehat{\omega}) \pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\ln(\widehat{\omega})]}$. Finally, the logarithmic CI for ω is

$$(3.9) \quad \widehat{\omega} \times \exp \left\{ \pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\ln(\widehat{\omega})]} \right\}.$$

Roldán-Nofuentes and Luna studied the size (and the power) of the test $H_0: \ln(\omega) = 0$ through simulation experiments. As the logarithmic interval (3.9) is obtained by inverting the test statistic (3.7), the coverage probability of this interval is equal to 1 minus the type I error obtained in the simulations carried out by Roldán-Nofuentes and Luna, and therefore the results are equivalent.

3.3. Wald CI

The Wald interval ([15]) is a classic interval for a parameter. Assuming the asymptotic normality of $\widehat{\omega}$, i.e. $\widehat{\omega} \xrightarrow[n \rightarrow \infty]{} N[\omega, \text{Var}(\omega)]$, the Wald CI for ω is

$$(3.10) \quad \widehat{\omega} \left[1 \pm z_{1-\alpha/2} \sqrt{\frac{\widehat{\text{Var}}(\widehat{LR}_1)}{\widehat{LR}_1^2} + \frac{\widehat{\text{Var}}(\widehat{LR}_2)}{\widehat{LR}_2^2} - \frac{2 \widehat{\text{Cov}}(\widehat{LR}_1, \widehat{LR}_2)}{\widehat{LR}_1 \widehat{LR}_2}} \right].$$

3.4. Fieller CI

The Fieller method ([5]) is a classic method used to calculate a CI for the ratio of two parameters, and requires us to assume that the estimators are distributed according to a bivariate normal distribution. Therefore, assuming the bivariate normality, i.e. $(\widehat{LR}_1, \widehat{LR}_2)^\top \xrightarrow[n \rightarrow \infty]{} N[(LR_1, LR_2)^\top, \Sigma_{LR}]$, where

$$\Sigma_{LR} = \begin{pmatrix} \text{Var}(LR_1) & \text{Cov}(LR_1, LR_2) \\ \text{Cov}(LR_1, LR_2) & \text{Var}(LR_2) \end{pmatrix},$$

and, applying the Fieller method, it is verified that

$$\widehat{LR}_1 - \omega \widehat{LR}_2 \xrightarrow[n \rightarrow \infty]{} N\left(0, \text{Var}(LR_1) - 2\omega \text{Cov}(LR_1, LR_2) + \omega^2 \text{Var}(LR_2)\right).$$

The Fieller CI is obtained by searching for the set of values for ω that satisfy the inequality

$$\frac{(\widehat{LR}_1 - \omega \widehat{LR}_2)^2}{\widehat{\text{Var}}(\widehat{LR}_1) - 2\omega \widehat{\text{Cov}}(\widehat{LR}_1, \widehat{LR}_2) + \omega^2 \widehat{\text{Var}}(\widehat{LR}_2)} < z_{1-\alpha/2}^2.$$

Solving this inequation, the Fieller CI for ω is

$$(3.11) \quad \frac{\widehat{LR}_1 \widehat{LR}_2 - \widehat{\sigma}_{12} z_{1-\alpha/2}^2 \pm \sqrt{\left(\widehat{LR}_1 \widehat{LR}_2 - \widehat{\sigma}_{12} z_{1-\alpha/2}^2\right)^2 - \left(\widehat{LR}_1^2 - \widehat{\sigma}_{11} z_{1-\alpha/2}^2\right)\left(\widehat{LR}_2^2 - \widehat{\sigma}_{22} z_{1-\alpha/2}^2\right)}}{\left(\widehat{LR}_2^2 - \widehat{\sigma}_{22} z_{1-\alpha/2}^2\right)},$$

where $\widehat{\sigma}_{ii} = \widehat{\text{Var}}(\widehat{LR}_i)$ and $\widehat{\sigma}_{12} = \widehat{\text{Cov}}(\widehat{LR}_1, \widehat{LR}_2)$. This interval is valid when $\left(\widehat{LR}_1 \widehat{LR}_2 - \widehat{\sigma}_{12} z_{1-\alpha/2}^2\right)^2 > \left(\widehat{LR}_1^2 - \widehat{\sigma}_{11} z_{1-\alpha/2}^2\right)\left(\widehat{LR}_2^2 - \widehat{\sigma}_{22} z_{1-\alpha/2}^2\right)$ and $\widehat{LR}_2^2 - \widehat{\sigma}_{22} z_{1-\alpha/2}^2 \neq 0$.

3.5. Bootstrap CI

The Bootstrap method is one which is widely used for the estimation of parameters. The Bootstrap CI is calculated generating B random samples with replacement from the sample sized n , and then a CI is calculated. For the interval, we considered the bias-corrected Bootstrap CI ([4]). For each one of the B samples with replacement, we calculate the estimators of the LRs and of ω , i.e. \widehat{LR}_{1Bi} , \widehat{LR}_{2Bi} and $\widehat{\omega}_{Bi}$, with $i = 1, \dots, B$. The parameter ω is estimated as the average of the B Bootstrap estimations, i.e.

$$\widehat{\omega}_B = \frac{1}{B} \sum_{i=1}^B \widehat{\omega}_{Bi}.$$

Let $A = \#(\widehat{\omega}_{Bi} < \widehat{\omega})$ be the number of samples in which the Bootstrap estimator $\widehat{\omega}_{Bi}$ is lower than the maximum likelihood estimator $\widehat{\omega}$. Let $\widehat{z}_0 = \Phi^{-1}(A/B)$, where $\Phi^{-1}(\cdot)$ is the inverse function of the standard normal cumulative distribution function. Let $q_1 = \Phi(2\widehat{z}_0 - z_{1-\alpha/2})$ and $q_2 = \Phi(2\widehat{z}_0 + z_{1-\alpha/2})$, then the bias-corrected Bootstrap CI is

$$(3.12) \quad \left(\widehat{\omega}_B^{(q_1)}, \widehat{\omega}_B^{(q_2)}\right),$$

where $\widehat{\omega}_B^{(q)}$ is the q -th quantile of the distribution of the B Bootstrap estimations of ω . The bias-corrected bootstrap CI is consistent, as it verifies ([13]) that $P[\sqrt{n}(\widehat{\omega}_n - \omega) \leq x] - P_B[\sqrt{n}(\widehat{\omega}_{B,n} - \widehat{\omega}_n) \leq x]$ converges in probability to zero when the sample size is very large ($n \rightarrow \infty$) for every value x , where P_B is the bootstrap distribution and $\widehat{\omega}_{B,n}$ is the upper (lower) limit of the bootstrap CI.

3.6. Bayesian CI

The previous CIs are all frequentists, the problem can also be addressed from a Bayesian perspective. Conditioning on $D = 1$, i.e. on the individuals who have the disease, it is verified that $s_{11} + s_{10} \rightarrow B(s, Se_1)$ and that $s_{11} + s_{01} \rightarrow B(s, Se_2)$. Conditioning on $D = 0$ it is verified that $r_{01} + r_{00} \rightarrow B(r, Sp_1)$ and that $r_{10} + r_{00} \rightarrow B(r, Sp_2)$. Considering the distribution of the BDT 1, the estimators of its sensitivity and specificity are $\widehat{Se}_1 = \frac{s_{11} + s_{10}}{s}$ and $\widehat{Sp}_1 = \frac{r_{01} + r_{00}}{r}$, which are estimators of binomial proportions. In a similar way, the estimators $\widehat{Se}_2 = \frac{s_{11} + s_{01}}{s}$

and $\widehat{Sp}_2 = \frac{r_{10}+r_{00}}{r}$ are also estimators of binomial proportions. Therefore, for these estimators, conjugate beta prior distributions are proposed, i.e.

$$(3.13) \quad \widehat{Se}_h \rightarrow \text{Beta}(\alpha_{Se_h}, \beta_{Se_h}) \quad \text{and} \quad \widehat{Sp}_h \rightarrow \text{Beta}(\alpha_{Sp_h}, \beta_{Sp_h}),$$

with $h = 1, 2$. Let $\mathbf{n} = (s_{11}, s_{10}, s_{01}, s_{00}, r_{11}, r_{10}, r_{01}, r_{00})$ be the vector of observed frequencies, then the posteriori distributions for the estimators of the sensitivity and the specificity of the BDT 1 are

$$(3.14) \quad \widehat{Se}_1 | \mathbf{n} \rightarrow \text{Beta}(s_{11} + s_{10} + \alpha_{Se_1}, s_{01} + s_{00} + \beta_{Se_1})$$

and

$$(3.15) \quad \widehat{Sp}_1 | \mathbf{n} \rightarrow \text{Beta}(r_{01} + r_{00} + \alpha_{Sp_1}, r_{11} + r_{10} + \beta_{Sp_1}).$$

In a similar way, the posteriori distributions for the estimators of the sensitivity and the specificity of the BDT 2 are

$$(3.16) \quad \widehat{Se}_2 | \mathbf{n} \rightarrow \text{Beta}(s_{11} + s_{01} + \alpha_{Se_2}, s_{10} + s_{00} + \beta_{Se_2})$$

and

$$(3.17) \quad \widehat{Sp}_2 | \mathbf{n} \rightarrow \text{Beta}(r_{10} + r_{00} + \alpha_{Sp_2}, r_{11} + r_{01} + \beta_{Sp_2}).$$

Once all the distributions have been defined, the posteriori distribution for the LR_s of each BDT, and for ω^+ and ω^- , can be approximated by applying the Monte Carlo method ([2]). This method consists of generating M random values of the posteriori distributions given in equations (3.13) to (3.17). In each interaction the generated values of sensitivities (\widehat{Se}_{hi}) and specificities (\widehat{Sp}_{hi}) are plugged in the equations $\widehat{LR}_{hi}^+ = \frac{\widehat{Se}_{hi}}{1-\widehat{Sp}_{hi}}$ and $\widehat{LR}_{hi}^- = \frac{1-\widehat{Se}_{hi}}{\widehat{Sp}_{hi}}$, and from these each ratio $\widehat{\omega}_i$ is calculated. As an estimator of each ratio, the average of the M Bayesian estimations is calculated, i.e. $\widehat{\omega}_{Ba} = \frac{1}{M} \sum_{i=1}^M \widehat{\omega}_i$. Finally, from the M values $\widehat{\omega}_i$ a CI based on the quantiles is calculated, i.e. the $100 \times (1 - \alpha)\%$ CI for ω is

$$(3.18) \quad \left(\widehat{\omega}_{Ba}^{(\alpha/2)}, \widehat{\omega}_{Ba}^{(1-\alpha/2)} \right),$$

where $\widehat{\omega}_{Ba}^{(q)}$ is the q -th quantile of the distribution of the M Bayesian estimations $\widehat{\omega}_i$.

All of the CIs presented are for $\omega = LR_1/LR_2$. If we want to calculate the CI for LR_2/LR_1 ($= \omega' = 1/\omega$), the regression, logarithmic, Fieller, Bootstrap and Bayesian intervals are obtained by calculating the inverse of each boundary of the corresponding interval for ω . Nevertheless, the Wald CI for ω' is obtained from the Wald CI for ω dividing each boundary by $\widehat{\omega}^2$, i.e. if (L_ω, U_ω) is the Wald CI for ω then the Wald CI for $\omega' = 1/\omega$ is $(L_\omega/\widehat{\omega}^2, U_\omega/\widehat{\omega}^2)$.

4. SIMULATION EXPERIMENTS

Monte Carlo simulation experiments were carried out to study the coverage probability (CP) and the average length (AL) of each one of the CIs presented in the Section 3. For this purpose, $N = 10,000$ random samples of multinomial distributions with sizes $n = \{50, 100, 200, 300, 400, 500, 1000\}$ were generated, and their probabilities were calculated from

equations (3.1) and (3.2). As sensitivity and specificity of each BDT, the values $Se_h, Sp_h = \{0.70, 0.75, \dots, 0.90, 0.95\}$ were taken, which are realistic values in clinical practice, and the LRs were calculated with the equations $LR_h^+ = Se_h/(1 - Sp_h)$ and $LR_h^- = (1 - Se_h)/Sp_h$ with $h = 1, 2$. For the disease prevalence, $\pi = \{10\%, 25\%, 50\%\}$ was considered, and for the dependence factors ε_1 and ε_0 intermediate values (50% of the maximum value of each ε_i) and high values (80% of the maximum value of each ε_i) were taken, i.e.

$$\varepsilon_1 = k \times \text{Min} \left\{ Se_1(1 - Se_2), Se_2(1 - Se_1) \right\}$$

and

$$\varepsilon_0 = k \times \text{Min} \left\{ Sp_1(1 - Sp_2), Sp_2(1 - Sp_1) \right\},$$

where $k = \{0.50, 0.80\}$. Once the value of the parameters in each scenario was set, the probabilities of each multinomial distribution were calculated by substituting the value of the parameters in equations (3.1) and (3.2).

For the Bootstrap interval, for each one of the N random samples generated, $B = 2,000$ replacement samples were generated in turn, and from the B replacement samples the bias-corrected bootstrap CI was calculated through the method described in Section 3.5.

Regarding the Bayesian CI, for the estimators of the two sensitivities and of the two specificities, the Beta(1,1) distribution was considered as prior distribution. The choice of this distribution is justified by the fact that it is a non-informative distribution, which is flat for every possible value of the sensitivities and the specificities, and it has a minimum impact on the posteriori distributions. Moreover, for each one of the N generated random samples, $M = 10,000$ random samples were generated in turn, and from the M samples the Bayesian CI was been calculated by applying the method described in Section 3.6.

The simulation experiments were designed so that in every random sample generated, it is possible to estimate all the parameters and their variances-covariances. Therefore, if a parameter could not be estimated in a sample (for example, $\widehat{Se}_h = 0$) then that sample was discarded and another one was generated in its place. This problem mainly occurred in the samples with $n = 50$. In each one of the scenarios considered (values set for $Se_h, Sp_h, \pi, \varepsilon_1$ and ε_0) the CP and the AL were calculated for each one of the six CIs for ω^+ and ω^- . The CP of each CI was calculated as the quotient between the number of intervals that contained the parameter (ω^+ or ω^- , depending on the case) and the number of samples generated N , and the AL was calculated adding the length of the N intervals and dividing this number by N . As the confidence level we took 95%.

The comparison of the asymptotic behaviour of the CIs was made following the criterion based on whether the CI “fails” or “does not fail” for a confidence of 95%. This criterion, which has been used by other authors [11, 7, 8, 9], establishes that a CI fails (or does not fail) if its coverage probability is $\leq 93\%$ ($> 93\%$). The selection of the CI with the best asymptotic behaviour was made through the following steps:

- 1) Choose the CIs with the fewest failures;
- 2) Choose the CIs which are the most accurate, i.e. those with least AL, and among these those which have a CP closest to 95%.

This method is justified in Appendix B.

4.1. Positive LRs

Tables 2 and 3 show some of the results obtained for the intervals of ω^+ , considering two different scenarios of sensitivities and specificities. In these tables, failures are indicated in bold type. From the results of the experiments, the following conclusions are reached:

- a) Regression CI. The CI obtained applying the regression method does not fail, and it has a CP of 100% or very close to this value. In general terms, its AL is larger than that of the rest of the intervals.
- b) Logarithmic CI. The logarithmic CI does not fail. In very general terms, when the sample size is small ($n = 50$) or moderate ($n = 100$) its CP is 100% or very near to this value. When the sample size is large ($n = 200 - 400$) or very large ($n \geq 500$) its CP fluctuates around 95%. The AL of this interval is lower than that of the interval calculated through regression.
- c) Wald CI. When $\omega^+ \neq 1$, this interval may fail if $n \leq 100$ and the prevalence is moderate ($\pi = 25\%$) or large ($\pi = 50\%$), whereas if $n \geq 200$ the interval does not fail. When $\omega^+ = 1$ the interval does not fail. In situations in which the Wald CI does not fail, its CP and AL are very similar to those of the logarithmic CI.
- d) Fieller CI. The Fieller CI does not fail. In general terms, its CP is 100% or very close to this value when $n \leq 100$. When $n \geq 200$ its CP behaves in a very similar way to the CP of the logarithmic and Wald intervals (and the ALs are very similar). Therefore, when $n \geq 200$, the behaviour of the Fieller CI is very similar to the logarithmic and Wald intervals.
- e) Bootstrap CI. In very general terms, when $n \leq 100$ this interval may fail if $\omega^+ \neq 1$ or its CP is equal (or very near) to 100% if $\omega^+ = 1$. When $n \geq 200$, the Bootstrap CI does not fail, its CP fluctuates around 95% and its AL is very similar to that of the logarithmic, Wald and Fieller intervals. Therefore, when $n \geq 200$ the Bootstrap interval has an asymptotic behaviour which is very similar to that of logarithmic, Wald and Fieller intervals.
- f) Bayesian CI. The Bayesian CI does not fail and has a CP and an AL which are very similar to those of the interval obtained by regression. The CP and the AL of the Bayesian interval are almost always higher than those of the logarithmic, Wald, Fieller and Bootstrap intervals.

Table 2: Coverage probabilities (%) and average lengths of the CIs for the ratio of the two positive LR_s (I).

$LR_1^+ = 9.5, LR_2^+ = 4.5, LR_1^- = 0.056, LR_2^- = 0.125, \omega^+ = 2.111, \omega^- = 0.444,$ $Se_1 = 0.95, Sp_1 = 0.90, Se_2 = 0.90, Sp_2 = 0.80$												
n	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
$\pi = 10\%, \varepsilon_1 = 0.0225, \varepsilon_0 = 0.0400$												
50	99.95	7.06	99.40	5.72	97.20	4.53	100	8.93	98.30	3.69	99.90	5.90
100	99.25	5.73	97.90	4.75	97.40	4.16	99.80	5.64	98.50	3.09	99.10	5.42
200	99.40	3.04	96.85	2.49	96.60	2.38	97.90	2.61	96.90	2.52	99.30	3.04
300	98.90	2.26	96.15	1.86	96.10	1.81	96.85	1.90	95.60	1.89	99.00	2.27
400	99.10	1.86	95.90	1.53	95.85	1.50	96.10	1.55	95.80	1.55	99.15	1.86
500	98.50	1.61	95.55	1.33	95.45	1.31	95.90	1.35	95.05	1.34	98.35	1.62
1000	98.20	1.07	95.45	0.89	95.30	0.88	95.65	0.89	95.35	0.90	98.20	1.08
$\pi = 10\%, \varepsilon_1 = 0.0360, \varepsilon_0 = 0.0640$												
50	99.95	6.54	99.10	4.78	95.50	3.93	99.95	7.74	91.80	2.72	99.95	5.48
100	99.90	5.15	98.60	3.76	96.55	3.39	99.45	4.57	95.60	2.51	99.90	4.91
200	99.60	2.93	96.90	2.09	96.00	2.01	98.15	2.19	96.35	1.95	99.55	2.93
300	99.65	2.21	96.30	1.57	95.90	1.53	97.25	1.61	96.00	1.53	99.60	2.22
400	99.80	1.82	95.90	1.30	95.95	1.28	97.10	1.32	96.30	1.28	99.85	1.83
500	99.75	1.59	95.80	1.13	95.75	1.12	96.35	1.15	95.65	1.13	99.80	1.60
1000	99.55	1.07	95.45	0.76	95.35	0.76	95.70	0.77	95.50	0.76	99.60	1.08
$\pi = 25\%, \varepsilon_1 = 0.0225, \varepsilon_0 = 0.0400$												
50	99.85	6.04	97.80	4.89	91.30	3.95	99.90	6.38	93.60	3.28	99.65	5.49
100	99.50	5.19	97.90	4.28	95.05	3.74	99.40	4.52	97.45	2.72	99.35	4.90
200	98.45	2.96	95.60	2.44	94.75	2.32	97.30	2.50	95.90	2.62	98.40	2.91
300	98.45	2.28	95.45	1.88	95.25	1.83	97.05	1.91	94.95	2.03	98.30	2.25
400	99.00	1.91	96.10	1.59	95.95	1.55	96.65	1.60	95.60	1.68	98.85	1.90
500	98.55	1.65	95.60	1.37	95.25	1.35	96.15	1.38	95.55	1.43	98.55	1.65
1000	98.30	1.14	95.15	0.95	94.90	0.94	95.30	0.95	94.65	0.97	98.35	1.14
$\pi = 25\%, \varepsilon_1 = 0.0360, \varepsilon_0 = 0.0640$												
50	100	5.77	96.80	4.21	91.50	3.50	99.65	5.56	83.55	2.31	100	5.25
100	99.85	4.45	95.40	3.19	91.85	2.88	97.15	3.45	89.15	2.20	99.80	4.25
200	99.60	2.85	96.15	2.02	94.00	1.95	96.40	2.08	94.85	1.93	99.60	2.80
300	99.40	2.23	94.15	1.59	94.10	1.55	95.15	1.62	94.10	1.60	99.40	2.21
400	99.55	1.87	94.95	1.32	94.85	1.30	95.15	1.34	94.65	1.35	99.50	1.85
500	99.15	1.66	94.85	1.18	94.75	1.16	95.70	1.19	95.05	1.21	99.15	1.65
1000	99.50	1.14	95.00	0.81	95.15	0.81	95.70	0.82	94.90	0.83	99.30	1.14
$\pi = 50\%, \varepsilon_1 = 0.0225, \varepsilon_0 = 0.0400$												
50	99.75	5.98	96.75	4.88	89.35	3.97	99.75	6.11	86.45	4.31	99.55	5.39
100	99.60	5.91	96.35	4.87	92.20	3.80	98.90	5.22	94.45	3.81	99.40	5.38
200	98.85	3.78	95.90	3.13	94.15	2.89	97.70	3.21	96.85	3.10	98.70	3.65
300	98.50	2.87	95.00	2.38	94.70	2.26	96.40	2.41	95.40	2.61	98.30	2.82
400	98.50	2.40	95.35	1.99	95.05	1.92	96.80	2.02	94.65	2.20	98.25	2.37
500	98.35	2.08	95.80	1.72	95.45	1.68	95.25	1.74	95.25	1.88	98.20	2.06
1000	97.50	1.41	94.55	1.17	94.80	1.15	95.50	1.17	93.80	1.22	97.60	1.40
$\pi = 50\%, \varepsilon_1 = 0.0360, \varepsilon_0 = 0.0640$												
50	99.90	5.47	94.15	4.03	88.70	3.28	99.20	5.26	67.35	2.89	99.80	4.97
100	99.85	5.20	93.80	3.80	91.40	3.22	96.65	4.24	78.55	2.43	99.75	4.79
200	99.70	3.45	93.75	2.47	93.65	2.32	93.70	2.56	89.75	2.15	99.45	3.34
300	99.55	2.72	94.65	1.93	94.45	1.86	94.65	1.98	94.10	1.90	99.55	2.67
400	99.65	2.33	95.15	1.66	94.90	1.62	95.45	1.69	95.35	1.69	99.65	2.31
500	99.45	2.06	95.55	1.46	95.15	1.43	95.25	1.48	96.00	1.51	99.20	2.04
1000	99.20	1.40	94.75	1.00	94.80	0.99	94.85	1.00	94.80	1.03	99.25	1.40

Table 3: Coverage probabilities (%) and average lengths of the CIs for the ratio of the two positive LRs (II).

$LR_1^+ = 6, LR_2^+ = 6, LR_1^- = 0.118, LR_2^- = 0.118, \omega^+ = 1, \omega^- = 1,$ $Se_1 = 0.90, Sp_1 = 0.85, Se_2 = 0.90, Sp_2 = 0.85$												
n	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
$\pi = 10\%, \varepsilon_1 = 0.0450, \varepsilon_0 = 0.0638$												
50	99.95	3.61	99.50	2.51	99.85	2.18	100	4.67	100	1.96	99.95	3.16
100	99.80	2.38	97.75	1.65	97.90	1.52	98.85	2.37	98.60	1.51	99.75	2.33
200	99.65	1.33	96.40	0.92	96.90	0.89	97.65	1.02	97.00	0.91	99.60	1.35
300	99.65	1.00	96.25	0.70	96.45	0.68	97.90	0.74	96.75	0.69	99.70	1.01
400	99.65	0.84	95.60	0.58	96.00	0.58	96.95	0.61	96.10	0.58	99.65	0.84
500	99.50	0.72	95.30	0.51	95.70	0.50	96.35	0.52	95.70	0.51	99.60	0.73
1000	99.25	0.48	94.65	0.34	94.30	0.34	95.15	0.35	94.80	0.34	99.25	0.49
$\pi = 10\%, \varepsilon_1 = 0.0720, \varepsilon_0 = 0.1020$												
50	100	3.18	100	1.79	99.90	1.62	100	3.65	100	1.43	100	2.77
100	100	2.19	99.85	1.11	99.75	1.06	100	1.58	99.95	0.99	100	2.15
200	100	1.28	98.15	0.60	98.20	0.59	98.75	0.67	98.55	0.57	100	1.29
300	100	0.98	97.05	0.45	97.15	0.45	97.45	0.48	97.95	0.43	100	0.98
400	100	0.82	96.85	0.37	96.90	0.37	97.05	0.39	97.15	0.37	100	0.82
500	100	0.71	96.30	0.33	96.40	0.32	96.80	0.34	96.65	0.32	100	0.72
1000	100	0.49	95.80	0.22	95.80	0.22	96.15	0.22	96.32	0.22	100	0.49
$\pi = 25\%, \varepsilon_1 = 0.0450, \varepsilon_0 = 0.0638$												
50	99.90	3.24	99.35	2.25	99.55	1.97	100	3.58	99.95	1.81	99.85	3.06
100	99.65	2.05	96.95	1.39	96.95	1.30	100	1.78	99.15	1.38	99.75	2.00
200	99.30	1.24	95.00	0.86	94.85	0.84	98.45	0.94	95.00	0.90	99.15	1.23
300	99.70	0.97	94.45	0.68	94.10	0.66	97.35	0.71	94.20	0.70	99.65	0.96
400	99.45	0.82	95.55	0.57	94.85	0.57	97.10	0.60	95.05	0.59	99.35	0.82
500	99.45	0.73	94.70	0.51	94.15	0.50	96.15	0.53	94.25	0.52	99.40	0.72
1000	99.60	0.51	95.45	0.36	95.25	0.36	95.85	0.36	95.15	0.36	99.50	0.51
$\pi = 25\%, \varepsilon_1 = 0.0720, \varepsilon_0 = 0.1020$												
50	100	2.80	100	1.49	99.85	1.38	100	2.51	100	1.27	100	2.66
100	100	1.93	99.30	0.89	99.25	0.86	100	1.15	100	0.82	100	1.89
200	100	1.21	96.95	0.53	96.50	0.53	98.70	0.59	98.30	0.53	100	1.20
300	100	0.96	95.85	0.42	95.65	0.42	96.75	0.45	97.65	0.42	100	0.95
400	100	0.82	95.35	0.36	94.95	0.36	96.30	0.38	96.35	0.37	100	0.82
500	100	0.73	95.25	0.32	95.25	0.32	95.90	0.33	95.80	0.33	100	0.73
1000	100	0.50	95.25	0.22	95.25	0.22	95.70	0.23	95.40	0.23	100	0.50
$\pi = 50\%, \varepsilon_1 = 0.0450, \varepsilon_0 = 0.0638$												
50	99.95	3.27	99.95	2.27	99.60	1.97	100	3.54	100	1.67	99.95	3.06
100	100	2.51	98.90	1.69	97.65	1.52	100	2.39	99.85	1.50	99.85	2.39
200	99.55	1.54	95.60	1.06	94.30	1.01	98.80	1.22	96.45	1.12	99.35	1.51
300	99.35	1.20	96.00	0.83	95.10	0.81	97.70	0.90	95.65	0.86	99.25	1.19
400	99.55	1.02	95.40	0.71	95.40	0.69	96.10	0.75	95.55	0.74	99.50	1.01
500	99.55	0.89	95.20	0.62	94.75	0.61	96.20	0.65	94.15	0.64	99.50	0.89
1000	99.55	0.61	94.40	0.43	94.75	0.43	95.75	0.44	94.25	0.44	99.50	0.61
$\pi = 50\%, \varepsilon_1 = 0.0720, \varepsilon_0 = 0.1020$												
50	100	2.81	100	1.50	99.95	1.38	100	2.58	100	1.24	100	2.66
100	100	2.25	99.90	1.05	99.70	1.00	100	1.51	100	0.94	100	2.16
200	100	1.49	99.20	0.66	98.45	0.65	99.95	0.77	99.95	0.64	100	1.47
300	100	1.17	97.70	0.51	97.05	0.50	99.50	0.56	99.45	0.51	100	1.16
400	100	1.00	96.50	0.43	96.40	0.43	98.55	0.46	97.95	0.44	100	0.99
500	100	0.89	95.75	0.39	95.35	0.38	97.55	0.40	96.80	0.39	100	0.88
1000	99.95	0.61	95.55	0.27	95.25	0.27	96.65	0.28	95.60	0.27	99.95	0.61

4.2. Negative LRs

Tables 4 and 5 show some of the results obtained for ω^- considering the same scenarios as for ω^+ . Failures are indicated in bold type. From the results, the following conclusions are obtained:

- a) Regression CI. This interval has an asymptotic behaviour which is very similar to that of the same interval for ω^+ .
- b) Logarithmic CI. In general terms, this interval can fail when $\omega^+ \neq 1$ and the dependence factors are high, whatever the sample size may be. This interval does not fail when $\omega^+ = 1$, and its CP is 100% or very near to this value when $n \leq 100$, and even with $n \geq 200$ if the prevalence is small. When this interval does not fail, its AL is lower than that of the interval obtained through regression.
- c) Wald CI. The Wald CI does not fail, and its CP is 100% (or very near) when $n \leq 100$, and its CP fluctuates around 95% when $n \geq 200$. The AL of the Wald CI is slightly lower than that of the logarithmic CI (when this does not fail), and its CP shows better fluctuations around 95% than that of the logarithmic interval.
- d) Fieller CI. This interval does not show any failures. In very general terms, the Fieller CI has a very similar CP to that of the Wald CI when $\omega^+ \neq 1$. When $\omega^+ = 1$, the CP of the Fieller CI is 100% (or near) when $n \leq 100$, and fluctuates around 95% if $n \geq 200$. Its AL is greater than that of the Wald CI, especially when $n \leq 500$.
- e) Bootstrap CI. This interval has many failures when $\omega^+ \neq 1$, especially when the prevalence is small or moderate, and regardless of the sample size. When $\omega^+ = 1$, the interval does not fail, and its CP is greater than that of the Wald CI or the logarithmic CI, especially when the prevalence is small or moderate. Regarding the Fieller CI, the CP of the Bootstrap interval is very similar to that of the Fieller interval, and its AL is slightly lower than that of the Fieller CI, especially for $n \leq 500$.
- f) Bayesian CI. The same as for ω^+ , the Bayesian CI for ω^- does not fail and has a CP and an AL which are very similar to those of the interval obtained through regression. The same as for ω^+ , the CP and the AL of the Bayesian interval are higher than those of the logarithmic, Wald, Fieller and Bootstrap intervals.

Table 4: Coverage probabilities (%) and average lengths of the CIs for the ratio of the two negative LR_s (I).

$LR_1^+ = 9.5, LR_2^+ = 4.5, LR_1^- = 0.056, LR_2^- = 0.125, \omega^+ = 2.111, \omega^- = 0.444,$ $Se_1 = 0.95, Sp_1 = 0.90, Se_2 = 0.90, Sp_2 = 0.80$												
n	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
$\pi = 10\%, \varepsilon_1 = 0.0225, \varepsilon_0 = 0.0400$												
50	99.95	2.07	97.30	1.65	96.85	1.27	99.50	2.71	14.80	1.79	99.75	1.90
100	99.90	2.02	96.60	1.59	96.05	1.17	99.60	2.49	35.50	1.83	99.85	1.81
200	99.95	1.99	96.15	1.42	95.90	1.09	99.55	2.40	53.70	1.68	99.85	1.79
300	99.85	1.81	95.45	1.30	95.15	1.03	99.05	1.99	75.95	1.59	99.75	1.65
400	99.85	1.67	96.55	1.23	95.55	0.97	99.10	1.75	86.05	1.55	99.75	1.54
500	99.80	1.62	96.95	1.20	95.95	0.96	98.80	1.70	88.80	1.48	99.60	1.50
1000	99.55	1.22	96.90	0.93	95.90	0.81	97.85	1.16	95.80	1.05	99.45	1.16
$\pi = 10\%, \varepsilon_1 = 0.0360, \varepsilon_0 = 0.0640$												
50	100	2.18	92.60	1.63	99.95	1.31	99.50	2.83	5.50	1.70	100	2.01
100	100	2.11	90.85	1.53	98.90	1.19	99.25	2.57	17.25	1.66	100	1.96
200	100	2.16	91.15	1.38	98.35	1.12	99.25	2.48	33.00	1.53	100	1.91
300	99.95	1.94	90.20	1.21	97.60	1.01	98.10	2.02	54.45	1.43	99.90	1.78
400	99.95	1.76	92.40	1.13	97.10	0.95	97.65	1.64	65.25	1.39	99.90	1.63
500	99.90	1.68	92.80	1.09	96.10	0.91	97.85	1.55	70.45	1.35	99.85	1.56
1000	99.90	1.22	93.40	0.79	95.60	0.71	97.45	0.97	84.65	0.93	99.80	1.16
$\pi = 25\%, \varepsilon_1 = 0.0225, \varepsilon_0 = 0.0400$												
50	100	2.06	97.80	1.56	96.35	1.18	99.30	2.66	34.05	1.86	99.75	1.87
100	100	1.87	96.20	1.34	95.95	1.04	99.65	2.13	64.85	1.67	99.80	1.70
200	99.65	1.64	96.00	1.22	95.80	0.98	98.00	1.77	89.30	1.50	99.60	1.52
300	99.50	1.44	95.95	1.07	95.60	0.90	97.40	1.46	93.15	1.28	99.45	1.35
400	99.10	1.21	95.75	0.93	95.40	0.81	96.55	1.16	95.35	1.05	98.90	1.15
500	99.50	1.06	95.55	0.82	95.45	0.73	96.00	0.97	95.60	0.89	99.20	1.01
1000	98.60	0.65	95.20	0.52	95.15	0.50	94.65	0.55	95.55	0.52	98.45	0.64
$\pi = 25\%, \varepsilon_1 = 0.0360, \varepsilon_0 = 0.0640$												
50	100	2.13	91.90	1.48	99.90	1.19	99.30	2.60	18.35	1.71	99.95	1.95
100	100	2.07	90.35	1.29	99.00	1.08	98.45	2.31	37.80	1.53	99.95	1.89
200	99.85	1.71	91.65	1.09	96.55	0.92	97.40	1.58	67.35	1.35	99.80	1.59
300	99.85	1.48	92.25	0.95	96.35	0.82	97.15	1.28	77.20	1.14	99.75	1.39
400	99.85	1.26	91.90	0.81	95.90	0.72	96.85	1.02	82.05	0.94	99.85	1.20
500	99.85	1.06	92.70	0.69	95.70	0.63	96.35	0.80	87.20	0.77	99.65	1.02
1000	99.50	0.65	94.45	0.43	95.35	0.42	96.20	0.45	94.40	0.44	99.55	0.64
$\pi = 50\%, \varepsilon_1 = 0.0225, \varepsilon_0 = 0.0400$												
50	99.90	1.82	97.65	1.35	99.90	1.07	99.60	2.13	71.70	1.76	99.85	1.69
100	99.85	1.67	96.35	1.23	99.30	0.98	99.05	1.82	84.60	1.56	99.80	1.55
200	99.70	1.23	97.10	0.94	96.95	0.81	98.00	1.19	96.05	1.07	99.60	1.17
300	98.75	0.92	96.25	0.73	94.40	0.66	95.60	0.81	97.25	0.76	98.50	0.89
400	98.55	0.75	95.45	0.60	94.45	0.56	95.25	0.64	96.80	0.61	98.60	0.73
500	98.15	0.66	94.35	0.53	94.40	0.50	94.10	0.55	95.05	0.53	97.80	0.65
1000	98.65	0.44	95.20	0.35	95.20	0.35	94.80	0.36	94.35	0.36	98.45	0.43
$\pi = 50\%, \varepsilon_1 = 0.0360, \varepsilon_0 = 0.0640$												
50	100	1.90	92.35	1.25	99.30	1.04	98.35	2.01	47.90	1.60	99.95	1.77
100	100	1.74	92.05	1.11	97.80	0.93	97.80	1.63	60.20	1.43	99.95	1.62
200	100	1.26	93.55	0.82	96.30	0.73	97.45	1.02	81.85	0.97	99.90	1.20
300	99.65	0.94	94.70	0.62	95.15	0.58	96.65	0.70	90.15	0.67	99.50	0.91
400	99.70	0.77	94.55	0.51	95.30	0.48	95.95	0.54	93.10	0.52	99.50	0.75
500	99.75	0.65	95.30	0.44	95.20	0.42	95.85	0.46	94.80	0.44	99.55	0.64
1000	99.65	0.43	95.75	0.30	94.80	0.29	95.40	0.30	96.30	0.29	99.55	0.43

Table 5: Coverage probabilities (%) and average lengths of the CIs for the ratio of the two negative LRs (II).

$LR_1^+ = 6, LR_2^+ = 6, LR_1^- = 0.118, LR_2^- = 0.118, \omega^+ = 1, \omega^- = 1,$ $Se_1 = 0.90, Sp_1 = 0.85, Se_2 = 0.90, Sp_2 = 0.85$												
n	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
$\pi = 10\%, \varepsilon_1 = 0.0450, \varepsilon_0 = 0.0638$												
50	100	2.55	100	1.84	99.50	1.55	100	3.35	100	1.72	100	2.34
100	100	2.54	100	1.74	98.85	1.44	99.95	3.04	100	1.65	100	2.33
200	100	2.52	100	1.58	95.90	1.36	99.90	3.01	100	1.56	100	2.32
300	100	2.48	100	1.52	93.85	1.34	99.60	2.70	100	1.52	100	2.31
400	100	2.39	99.65	1.51	93.15	1.32	99.20	2.53	100	1.51	99.90	2.26
500	100	2.35	99.65	1.43	94.35	1.31	99.05	2.45	100	1.50	100	2.25
1000	99.85	1.98	97.15	1.33	93.95	1.24	96.85	1.86	98.70	1.38	99.85	1.91
$\pi = 10\%, \varepsilon_1 = 0.0720, \varepsilon_0 = 0.1020$												
50	100	2.73	100	1.76	99.90	1.56	100	3.45	100	1.39	100	2.51
100	100	2.68	100	1.56	99.80	1.40	100	2.84	100	1.34	100	2.49
200	100	2.65	100	1.42	99.80	1.30	100	2.78	100	1.23	100	2.40
300	100	2.61	100	1.28	98.60	1.19	99.95	2.62	100	1.12	100	2.33
400	100	2.52	100	1.19	97.70	1.11	98.90	2.05	100	1.07	100	2.27
500	100	2.42	100	1.13	97.10	1.07	97.95	1.85	100	1.03	100	2.18
1000	100	1.91	99.80	0.85	96.80	0.82	97.15	1.16	100	0.80	100	1.85
$\pi = 25\%, \varepsilon_1 = 0.0450, \varepsilon_0 = 0.0638$												
50	100	2.56	100	1.72	98.20	1.46	100	3.23	100	1.67	100	2.40
100	100	2.51	100	1.53	95.45	1.35	99.85	2.91	100	1.55	99.95	2.35
200	99.95	2.40	99.50	1.50	93.90	1.31	98.90	2.57	99.95	1.53	99.90	2.20
300	99.85	2.26	98.55	1.48	94.65	1.25	98.00	2.35	99.75	1.47	99.80	2.15
400	99.70	1.98	96.95	1.33	93.05	1.19	96.20	1.85	98.20	1.37	99.55	1.92
500	99.55	1.74	95.20	1.18	92.35	1.10	94.55	1.50	96.40	1.24	99.40	1.70
1000	99.25	1.15	94.80	0.79	94.25	0.75	94.40	0.86	94.15	0.84	99.20	1.13
$\pi = 25\%, \varepsilon_1 = 0.0720, \varepsilon_0 = 0.1020$												
50	100	2.86	100	1.57	99.80	1.41	100	3.11	100	1.40	100	2.65
100	100	2.81	100	1.37	98.90	1.26	100	2.95	100	1.22	100	2.54
200	100	2.45	100	1.14	97.75	1.07	100	1.86	100	1.04	100	2.30
300	100	2.22	99.90	1.01	97.20	0.97	99.80	1.54	100	0.93	100	2.10
400	100	1.92	96.95	0.86	96.80	0.83	98.55	1.17	99.95	0.80	100	1.85
500	100	1.69	96.55	0.74	96.45	0.72	98.15	0.94	99.90	0.71	100	1.65
1000	100	1.13	96.05	0.49	95.95	0.48	96.50	0.53	98.45	0.49	100	1.10
$\pi = 50\%, \varepsilon_1 = 0.0450, \varepsilon_0 = 0.0638$												
50	100	2.45	100	1.59	95.50	1.40	99.90	2.80	100	1.65	99.95	2.31
100	99.95	2.42	99.25	1.56	94.70	1.35	99.00	2.69	99.95	1.55	99.90	2.25
200	99.80	2.01	96.90	1.34	93.30	1.25	96.20	1.89	98.85	1.37	99.70	1.94
300	99.65	1.57	96.75	1.08	94.30	1.05	96.30	1.29	97.20	1.15	99.60	1.54
400	99.70	1.32	95.40	0.91	94.65	0.88	95.20	1.02	95.20	0.97	99.70	1.30
500	99.70	1.17	95.10	0.81	94.90	0.78	94.70	0.88	94.20	0.85	99.65	1.15
1000	99.40	0.78	95.20	0.54	94.60	0.54	95.05	0.56	94.75	0.56	99.35	0.77
$\pi = 50\%, \varepsilon_1 = 0.0720, \varepsilon_0 = 0.1020$												
50	100	2.67	99.95	1.36	99.50	1.26	99.95	2.64	100	1.30	100	2.51
100	100	2.49	100	1.16	98.45	1.09	100	1.95	100	1.09	100	2.36
200	100	1.94	99.55	0.86	97.30	0.83	99.40	1.18	100	0.81	100	1.88
300	100	1.55	98.80	0.67	97.00	0.66	98.55	0.80	99.75	0.65	100	1.51
400	100	1.30	96.95	0.56	96.90	0.55	97.80	0.63	99.60	0.55	100	1.28
500	100	1.14	96.25	0.50	96.25	0.49	96.05	0.54	98.20	0.50	100	1.13
1000	100	0.78	95.35	0.34	95.10	0.34	95.35	0.35	95.30	0.35	100	0.77

4.3. Rules of application

Considering the asymptotic behaviour of each one of the CIs studied, it is possible to give some general rules of application for the CIs studied. These rules of application are for the different scenarios considered in the simulation experiments, scenarios that correspond to realistic values of prevalence, sensitivities and specificities in clinical practice. Based on the sample size, which in practice is the only parameter set by the researcher, the rules are the following:

- a) For the ratio ω^+ , use the logarithmic CI, whatever the sample size may be, although when $n \geq 200$ we can also use the Wald, the Fieller and the Bootstrap intervals.
- b) For the ratio ω^- , use the Wald CI, whatever the sample size may be.

5. SAMPLE SIZE

An important question when comparing two parameters is the calculation of the sample size necessary to compare the parameters with a determined error and power. In the context of the comparison of the LRs, Roldán-Nofuentes and Luna [12] proposed a method to calculate the sample size to solve the hypothesis test $H_0: \ln(\omega) = 0$ vs $H_1: \ln(\omega) \neq 0$. We then study the same problem but from the perspective of the CIs. Therefore, we study the problem of calculating the sample size necessary to estimate the ratio between the two LRs with a precision δ and a confidence $100(1 - \alpha)\%$. As in the previous sections, we consider that ω is ω^+ or ω^- . Let us first consider the Wald CI, which can be applied both to estimate ω^+ (with $n \geq 200$) and ω^- (for any sample size). Based on the asymptotic normality of the estimator of ω , it is verified that

$$\hat{\omega} \in \omega \pm z_{1-\alpha/2} \sqrt{\text{Var}(\hat{\omega})},$$

i.e. the probability of obtaining an estimator $\hat{\omega}$ is in this interval with a probability $100(1-\alpha)\%$. Let us consider that $LR_2 > LR_1$ and, therefore, that $\omega < 1$ (the Wald interval will be lower than one) and let δ be the precision set by the researcher. As it has been assumed that $\omega < 1$, then δ must be lower than one, and if we want to have a high level of precision then δ must be a small value. The sample size n is calculated from the expression

$$(5.1) \quad \delta = z_{1-\alpha/2} \omega \sqrt{\frac{\text{Var}(\widehat{LR}_1)}{LR_1^2} + \frac{\text{Var}(\widehat{LR}_2)}{LR_2^2} - \frac{2 \text{Cov}(\widehat{LR}_1, \widehat{LR}_2)}{LR_1 LR_2}}.$$

This equation is obtained from the Wald CI (equation (3.10)). Substituting the variances and the covariance with their respective expressions given in equations (3.4) and clearing n we obtain the expression of the sample size to estimate ω with a precision δ and a confidence $100(1 - \alpha)\%$. For ω^+ the equation of the sample size is

$$(5.2) \quad n = \left(\frac{z_{1-\alpha/2} \omega^+}{\delta}\right)^2 \left[\sum_{h=1}^2 \left(\frac{1 - Se_h}{\pi Se_h} + \frac{Sp_h}{\bar{\pi}(1 - Sp_h)} \right) - \frac{2 \varepsilon_1}{\pi Se_1 Se_2} - \frac{2 \varepsilon_0}{\bar{\pi}(1 - Sp_1)(1 - Sp_2)} \right],$$

and for ω^- is

$$(5.3) \quad n = \left(\frac{z_{1-\alpha/2} \omega^-}{\delta}\right)^2 \left[\sum_{h=1}^2 \left(\frac{Se_h}{\pi(1 - Se_h)} + \frac{1 - Sp_h}{\bar{\pi} Sp_h} \right) - \frac{2 \varepsilon_1}{\pi(1 - Se_1)(1 - Se_2)} - \frac{2 \varepsilon_0}{\bar{\pi} Sp_1 Sp_2} \right].$$

If it is considered that $\omega > 1$ (and consequently the Wald CI is higher than one) the BDTs can always be permuted and ω will then be lower than one. Another alternative consists of setting a value for a precision δ' , in a similar way to the previous situation when $\omega < 1$, and then apply equation (5.2) or (5.3) considering $\delta = \hat{\omega}^2 \delta'$. As is explained at the end of Section 3, this is due to the fact that if (L_ω, U_ω) is the Wald CI for $\omega = LR_1/LR_2 < 1$ then the Wald CI for $\omega' = 1/\omega = LR_2/LR_1$ is $(\frac{L_\omega}{\hat{\omega}^2}, \frac{U_\omega}{\hat{\omega}^2})$. It is easy to check that the calculated value of the sample size n is the same both if $\omega < 1$ (with precision δ) and if $\omega > 1$ (with precision $\delta = \hat{\omega}^2 \delta'$).

In order to be able to apply the previous equations, it is necessary to know the sensitivities, the specificities (and therefore the LRs, ω^+ and ω^-), the dependence factors between the two BDTs (ε_i) and the prevalence (π). In practice, these values can be estimated from a pilot sample or can be obtained from another similar study. Therefore, the method to calculate the sample size requires us to know some estimations of the accuracy (Se and Sp) of each BDT, of the dependence factors between the BDTs and of the disease prevalence, obtained for example from a pilot study or from other previous studies. The method to calculate the size of the sample consists of the following steps:

- Step 1.** Take a pilot sample sized n_0 (in general terms, $n_0 \geq 200$ if ω^+ is estimated to then be able to calculate the Wald CI), and with this sample we calculate \widehat{Se}_h , \widehat{Sp}_h (and therefore \widehat{LR}_h , $\hat{\omega}^+$ and $\hat{\omega}^-$), $\hat{\varepsilon}_i$ and $\hat{\pi}$. The Wald CI for ω is then calculated, and if this interval has a precision δ , i.e. $z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\omega})} \leq \delta$, then the required precision has been reached; if not, go to the following step.
- Step 2.** Based on the estimations obtained in Step 1, calculate the sample size n applying equation (5.2) or (5.3).
- Step 3.** Take the sample of n individuals (add $n - n_0$ individuals to the initial pilot sample), and from this new sample we calculate \widehat{Se}_h , \widehat{Sp}_h , $\hat{\varepsilon}_i$, $\hat{\pi}$ and the Wald CI. If the Wald CI has a precision δ , then the set precision has been achieved; if not, consider the new sample to be a pilot sample ($n_0 = n$) and go back to Step 1.

This proposed procedure to calculate the sample size is iterative, and therefore it does not guarantee that with the sample size calculated we can then estimate the parameter ω with the required precision. Moreover, if the researcher sets a precision δ^+ to estimate ω^+ and also sets a precision δ^- to estimate ω^- , once both sample sizes have been calculated through the previous method, the researcher must take a sample size of at least the maximum of the two sample sizes, to thus guarantee the precision in both estimations. In general, the calculation of the sample size makes sense when the confidence interval for ω does not contain the value one, since in this situation (the interval contains the value one) the equality of both LRs is not rejected and it does not make sense to determine how much larger one LR is compared to the other. Nevertheless, if the pilot sample is small (for example to estimate ω^-) and the Wald CI for ω^- contains the value 1, it may be useful to calculate the sample size to estimate the ω^- . In this situation, the Wald CI for ω^- will be very wide (as the pilot sample is small) and may contain the value 1 even if LR_1^- and LR_2^- are different.

The calculation of the sample size depends on the estimations obtained from an initial pilot sample. In order to study the effect that this sample has on the calculation of the sample size, simulation experiments were carried out which were similar to those carried out in Section 4. From the values of the parameters, we calculated the sample size n applying equation (5.2)

or (5.3) depending on the case, taking a precision equal to 0.10, and we then generated $N = 10,000$ random samples of multinomial distributions sized n . In each one of the N random samples, we calculated the sample size n'_i from the estimators calculated with the random sample, and then calculated the average sample size $\bar{n} = \sum n'_i / N$ and the relative bias $RB(n') = (\bar{n} - n) / n$. Table 6 shows the results obtained for the scenarios considered in Tables 2 and 4 ($\omega \neq 1$). From the results, it holds that the dependence factors ε_i have an important effect on the calculation of the sample size, the sample size is smaller when the dependence factors are larger. Moreover, the increase in the prevalence means an increase (decrease) in the sample size to estimate ω^+ (ω^-). The relative biases obtained are very small, and therefore the sample sizes calculated from equations (5.2) and (5.3) are robust. Consequently, the initial pilot sample does not have an important effect on the determination of the sample size to estimate ω .

Table 6: Sample size to estimate ω .

$LR_1^+ = 9.5, LR_2^+ = 4.5, LR_1^- = 0.056, LR_2^- = 0.125,$ $\omega^+ = 2.111, \omega^- = 0.444,$ $Se_1 = 0.95, Sp_1 = 0.90, Se_2 = 0.90, Sp_2 = 0.80$			
Sample size for ω^+			
	$\pi = 10\%$	$\pi = 25\%$	$\pi = 50\%$
$\varepsilon_1 = 0.0225, \varepsilon_0 = 0.0400$			
Sample size	958	1073	1571
Average sample size	981	1084	1597
Relative bias (%)	2.40	1.03	1.66
$\varepsilon_1 = 0.0360, \varepsilon_0 = 0.0640$			
Sample size	701	786	1152
Average sample size	734	796	1160
Relative bias (%)	4.71	1.27	0.69
Sample size for ω^-			
	$\pi = 10\%$	$\pi = 25\%$	$\pi = 50\%$
$\varepsilon_1 = 0.0225, \varepsilon_0 = 0.0400$			
Sample size	14439	5793	2922
Average sample size	14715	5896	2966
Relative bias (%)	1.91	1.78	1.51
$\varepsilon_1 = 0.0360, \varepsilon_0 = 0.0640$			
Sample size	10336	4147	2092
Average sample size	10482	4186	2118
Relative bias (%)	1.41	0.94	1.24

If the initial pilot sample has a small or moderate size, then in order to estimate ω^+ we use the logarithmic CI. In this situation, the process is similar to the previous one, and the sample size is calculated from the equation $\ln(\delta) = z_{1-\alpha/2} \sqrt{\text{Var}[\ln(\hat{\omega}^+)]}$, where the expression of $\text{Var}[\ln(\hat{\omega}^+)]$ is given in equation (3.8). Following a similar process to the previous one, it holds that

$$(5.4) \quad n = \left(\frac{z_{1-\alpha/2}}{\ln(\delta)} \right)^2 \left[\sum_{h=1}^2 \left(\frac{1 - Se_h}{\pi Se_h} + \frac{Sp_h}{\bar{\pi}(1 - Sp_h)} \right) - \frac{2\varepsilon_1}{\pi Se_1 Se_2} - \frac{2\varepsilon_0}{\bar{\pi}(1 - Sp_1)(1 - Sp_2)} \right].$$

6. APPLICATIONS

The results obtained were applied to two real examples: a study of the diagnosis of coronary disease and another study of the diagnosis of colorectal cancer.

6.1. Diagnosis of coronary disease

The results obtained were applied to the study of Weiner *et al.* [16] on the diagnosis of coronary disease, which is a widely used study to illustrate statistical methods for the estimation and comparison of parameters of BDTs. Weiner *et al.* studied the diagnosis of coronary artery disease using as diagnostic tests the exercise test and the resting EKG, and the coronary arteriography as a GS. Table 7 shows the frequencies obtained by applying the three medical tests to a sample of 1,465 males, where T_1 models the result of the exercise test, T_2 models the result of the resting EKG and D the result of the GS. Table 7 also shows the estimations of the LRs (ω) and their standard errors (SE), as well as the CIs for ω^+ and ω^- .

Table 7: Diagnosis of coronary disease.

Frequencies					
	$T_1 = 1$		$T_1 = 0$		Total
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	
$D = 1$	224	591	32	176	1023
$D = 0$	35	80	41	286	442
Total	259	671	73	462	1465

Results				
	$\widehat{Se} \pm SE$	$\widehat{Sp} \pm SE$	$\widehat{LR}^+ \pm SE$	$\widehat{LR}^- \pm SE$
Exercise test	0.797 ± 0.013	0.740 ± 0.021	3.065 ± 0.250	0.274 ± 0.019
Resting EKG	0.250 ± 0.014	0.828 ± 0.018	1.453 ± 0.171	0.906 ± 0.026
\widehat{p}	$\widehat{\epsilon}_1$	$\widehat{\epsilon}_0$	$\widehat{\omega}^+ \pm SE$	$\widehat{\omega}^- \pm SE$
0.698	0.020	0.034	2.109 ± 0.273	0.302 ± 0.021

CIs for ω^+		
Regression CI	Logarithmic CI	Wald CI
(1.589, 2.786)	(1.632, 2.713)	(1.569, 2.639)
Fieller CI	Bootstrap CI	Bayesian CI
(1.647, 2.765)	(1.501, 2.612)	(1.668, 2.567)

CIs for ω^-		
Regression CI	Logarithmic CI	Wald CI
(0.263, 0.351)	(0.265, 0.348)	(0.262, 0.345)
Fieller CI	Bootstrap CI	Bayesian CI
(0.262, 0.346)	(0.280, 0.348)	(0.264, 0.343)

For ω^+ , from any of the six CIs (all of them are greater than one) it holds that the positive LR of the exercise test is significantly larger than the positive LR of the resting EKG, i.e. a positive result in the exercise test is more indicative of the presence of the disease than a positive result in the resting EKG. Interpreting the results of the logarithmic CI, the positive LR of the exercise test is (with a confidence of 95%) a value between 1.632 and 2.713 times larger than the positive LR of the resting EKG.

Regarding ω^- , all of the CIs intervals (all are less than one) we reject the equality of the two negative LRs, and it holds that a negative result for the resting EKG is more indicative of the absence of the disease than a negative result of the exercise test. Interpreting the Wald CI, the negative LR of the resting EKG is (with a confidence of 95%) a value between 2.872 ($= 0.262/0.302^2$) and 3.783 ($= 0.345/0.302^2$) times larger than the negative LR of the exercise test.

Moreover, in order to illustrate the method to calculate the sample size, we are going to consider that the researcher wants to estimate ω^+ with a precision equal to 0.10, which can be considered to be a high precision. The Wald CI for ω^+ is (1.569, 2.639), and therefore multiplying this interval by $1/(\hat{\omega}^+)^2 = 1/2.109^2$ it holds that the 95% Wald CI for $\omega'^+ = LR_2^+/LR_1^+$ is (0.353, 0.593), and the precision is 0.12. As 0.12 is higher than 0.10, it is necessary to increase the sample size to estimate ω^+ with the required precision. Setting the confidence at 95% and taking $\delta = (\hat{\omega}^+)^2 \delta' = 2.109^2 \times 0.10 \approx 0.445$, applying equation (5.2) it holds that $n = 2,146$. Consequently, it is necessary to add 681 new individuals to the initial sample of 1,465 individuals, and once the data are obtained it is necessary to check that the required precision has been achieved.

6.2. Diagnosis of colorectal cancer

The results obtained were applied to a study of the diagnosis of colorectal cancer, using as diagnostic tests Fecal Occult Blood Testing (FOBT) and Fecal Immunochemical Testing (FIT), and the biopsy as the GS. Table 8 shows the results obtained by applying the three tests to a sample of 168 adult men with suspicious symptoms of the disease, where the variable T_1 models the result of the FOBT, T_2 models the result of the FIT and D models the result of the biopsy. This data came from a study carried out at the University Hospital of Granada (Spain). Table 8 also shows the estimations of the LRs, their standard errors and the confidence intervals for ω^+ and ω^- . Applying the rule given in Section 4.3, as $n = 168 < 200$ the logarithmic CI for ω^+ must be used in addition to the Wald CI for ω^- . For ω^+ , the logarithmic CI contains the value one, and therefore we do not reject the equality of both positive LRs. Regarding ω^- , the Wald CI does not contain the value one, and therefore we reject the equality of both negative LRs. Thus, a negative result for the FOBT is more indicative of the presence of colorectal cancer than a negative result for the FIT. The negative LR of the FOBT is (with a confidence of 95%) a value between 1.321 and 3.183 times larger than the negative LR of the FIT. The Wald CI for $1/\omega^-$ is (0.260, 0.628), calculated as $(1.321/2.252^2, 3.183/2.252^2)$.

In order to illustrate in this example the method of sample size calculation, let us suppose that the researchers want to estimate $1/\omega^-$ with a precision equal to 0.10, or in other words, to estimate ω^- with a precision of $0.10 \times (\widehat{\omega}^-)^2 = 0.10 \times 2.252^2 \approx 0.50$. As with the sample of 168 individuals the precision obtained with the Wald CI for ω^- is $0.931 > 0.50$, or rather a precision equal to 0.184 (> 0.10) with the Wald CI for $1/\omega^-$, then it is necessary to calculate the sample size. Considering the sample of 168 individuals to be a pilot sample, applying equation (5.3) it holds that $n = 561$. Therefore, 561 individuals are needed (we have to add 393 to the sample of 168) in order to estimate ω^- ($1/\omega^-$) with a precision equal to 0.50 (0.10) with a confidence of 95%.

Table 8: Diagnosis of colorectal cancer.

Frequencies					
	$T_1 = 1$		$T_1 = 0$		Total
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	
$D = 1$	68	1	18	13	100
$D = 0$	4	2	1	61	68
Total	72	3	19	74	168

Results				
	$\widehat{Se} \pm SE$	$\widehat{Sp} \pm SE$	$\widehat{LR}^+ \pm SE$	$\widehat{LR}^- \pm SE$
FOBT	0.690 ± 0.046	0.912 ± 0.034	7.841 ± 3.093	0.340 ± 0.052
FIT	0.860 ± 0.035	0.926 ± 0.032	11.622 ± 5.057	0.151 ± 0.038
\widehat{p}	$\widehat{\epsilon}_1$	$\widehat{\epsilon}_0$	$\widehat{\omega}^+ \pm SE$	$\widehat{\omega}^- \pm SE$
0.595	0.087	0.052	0.675 ± 0.215	2.252 ± 0.475

CIs for ω^+		
Regression CI	Logarithmic CI	Wald CI
(0.212, 2.108)	(0.356, 1.255)	(0.254, 1.096)
Fieller CI	Bootstrap CI	Bayesian CI
(0.278, 2.277)	(0.281, 1.283)	(0.222, 2.057)

CIs for ω^-		
Regression CI	Logarithmic CI	Wald CI
(1.265, 4.001)	(1.488, 3.403)	(1.321, 3.183)
Fieller CI	Bootstrap CI	Bayesian CI
(1.556, 3.894)	(1.553, 3.778)	(1.281, 4.006)

7. DISCUSSION

The LRs are parameters that are used to assess and compare the effectiveness of BDTs, and only depend on the accuracy (sensitivity and specificity) of the BDT. The comparison of the positive (negative) LRs of two BDTs subject to a paired design is a topic which has not been widely studied in Statistical literature and consists of the comparison of two relative risks subject to the same type of design. The previous studies ([10, 6, 12, 3]) focused mainly on the study of hypothesis tests to compare the positive (negative) LRs of the two BDTs. The comparison of the positive (negative) LRs through CIs has been the object of the very little research, and the studies that have been published by Pepe [10] and Roldán-Nofuentes and Luna [12] have focused on proposing CIs without dealing with this question in more depth. In this article, we extend the scope of these previous studies, proposing four new intervals: three of which are frequentist (Wald, Fieller and Bootstrap) and one which is Bayesian. The Wald and Fieller intervals are based on the asymptotic normality of the ratio of the LRs, and the Bootstrap interval is based on the fact that the bootstrap estimator of the ratio of the LRs can be transformed to a normal distribution. Regarding the Bayesian interval, this was obtained by applying the Monte Carlo method considering a priori non-informative distributions. The importance of the study of the CIs for the ratio of the positive (negative) LRs does not only lie in the fact that these CIs allow us to compare the two positive (negative) LRs, but also that it allows us to determine (when the equality of both CIs is rejected) how much bigger one CI than the other, which means an advantage over the hypothesis tests.

The comparison of the asymptotic behaviour of the six CIs was studied through simulation experiments. The results of these experiments has shown that, in the scenarios considered, in order to estimate the ratio $\omega^+ = LR_1^+ / LR_2^+$, in general terms, the intervals with the best behaviour are the logarithmic (for all the sample sizes), the Wald, Fieller and Bootstrap intervals (these last three for large or very large samples); whereas in order to estimate $\omega^- = LR_1^- / LR_2^-$ the interval with the best behaviour is the Wald interval (for all of the samples sizes). The use of different CIs for ω^+ and for ω^- may be due to the convergence to the normal distribution of the estimators. For an informative BDT, i.e. for a BDT whose Youden index is higher than 0 ($Y = Se + Sp - 1 > 0$), it must be verified that $LR^+ > 1$ and that $LR^- < 1$. Then, considering that the two BDTs are informative (as should be the case in clinical practice), ω^+ is the ratio between two values greater than 1 and ω^- is the ratio between two values lower than 1. For ω^+ , $\ln \hat{\omega}^+$ converges better to the normal distribution than $\hat{\omega}^+$ for $n < 200$, but when $n \geq 200$ both ($\hat{\omega}^+$ and $\ln \hat{\omega}^+$) has a good approximation to the normal distribution. The Wald CI for ω^- has a better asymptotic behaviour than the logarithmic CI for ω^- , which must be due to the fact that $\hat{\omega}^-$ converges more quickly to the normal distribution (even with large samples) than $\ln \hat{\omega}^-$.

An important question when comparing parameters of two BDTs is the calculation of the sample size necessary to compare the parameters based on certain specifications. When a hypothesis test is carried out, the sample size is calculated based on an error α , a power θ and a difference (or ratio) to be detected among the parameters. Roldán-Nofuentes and Luna [12] proposed a method to calculate sample size to solve the hypothesis test ($H_0: \ln \omega = 0$) of equality of the positive (negative) LRs. This article proposes, as a complement to the study of the CIs, a method to determine the sample size necessary to estimate the ratio between the LRs with a previously set precision. This is a topic that has never been studied and,

therefore, represents a contribution to Statistical literature on the subject analysed in this article. The method, which is based on the Wald (logarithmic) CI, requires knowledge of the estimations of the sensitivities, specificities, dependence factors and disease prevalence. These estimations can be obtained from a pilot sample or another similar study and, therefore, as it depends on the pilot sample selected. Therefore, the method does not guarantee that with the calculated sample size the parameter ω can be estimated with the set precision, and it is necessary to check this precision.

The intervals studied in this article can also be applied when the sample design is case-control. In this type of design, the two BDTs are applied to all of the individuals in two random samples, one of n_1 individuals with the disease and another one of n_2 individuals without the disease. If this type of sampling is used, two multinomial distributions are involved, one for the case sample, whose probabilities are $p_{ij} = Se_1^i(1 - Se_1)^{1-i} Se_2^j(1 - Se_2)^{1-j} + \delta_{ij} \varepsilon_1$ with $\sum p_{ij} = 1$, and the other for the control sample, whose probabilities are $q_{ij} = Sp_1^{1-i}(1 - Sp_1)^i Sp_2^{1-j}(1 - Sp_2)^j + \delta_{ij} \varepsilon_0$ with $\sum q_{ij} = 1$. Here, the variances-covariances of the sensitivities and specificities, obtained applying the delta method, are

$$\begin{aligned} \text{Var}(\widehat{Se}_h) &\approx \frac{Se_h(1 - Se_h)}{n_1}, & \text{Var}(\widehat{Sp}_h) &\approx \frac{Sp_h(1 - Sp_h)}{n_2}, \\ \text{Cov}(\widehat{Se}_1, \widehat{Se}_2) &\approx \frac{\varepsilon_1}{n_1}, & \text{Cov}(\widehat{Sp}_1, \widehat{Sp}_2) &\approx \frac{\varepsilon_0}{n_2}. \end{aligned}$$

The equations of the estimators and of the variances-covariances given in the regression, logarithmic, Wald and Fieller intervals are valid substituting s with n_1 and r with n_2 . Regarding the Bootstrap interval, it is necessary to generate B samples with replacement from the case sample and another B samples with replacement from the control sample, and the process is the same as the one described in Section 3.5. Regarding the Bayesian interval, the process is similar substituting s with n_1 and r with n_2 .

The methodology used in this article, both to obtain the CIs and to calculate the sample size, can be used to compare other parameters of BDTs, e.g. the odds ratios. The odds ratio of a BDT is defined as $OR = SeSp / [(1 - Se)(1 - Sp)]$ and is a measure of the association between the BDT and the GS. It is easy to check that the ratio of the odds ratios of two BDTs is $LR_1^+ LR_2^- / (LR_1^- LR_2^+)$, and therefore from this expression it is possible to deduce CIs similar to those given in Section 3 and can also be applied to the same procedure as in Section 5 to determine the sample size necessary to compare the odds ratios of two BDTs through a CI.

In this manuscript we studied the comparison of the LRs of two binary diagnostic tests. When the diagnostic test is quantitative, its accuracy is measured by the area under the ROC curve. The LRs are related to the equation of the ROC curve. Thus, for a single quantitative diagnostic test, for each one of the cut off points c of the estimated ROC curve a value for \widehat{Se} and a value $1 - \widehat{Sp}$ are obtained, and therefore a value for \widehat{LR}^+ (and another one for \widehat{LR}^-). For \widehat{LR}^+ , its numerator \widehat{Se} is the “y” coordinate of the estimated ROC curve, and the denominator $1 - \widehat{Sp}$ is the “x” coordinate of the estimated ROC curve. The estimator of LR for an interval (c_1, c_2) of test values corresponds to the slope of the line segment between c_1 and c_2 on the estimated ROC curve. In the case of two quantitative diagnostic test, for each cut off point of each estimated ROC curve, we obtain a value for $\widehat{\omega}^+$ and another one for $\widehat{\omega}^-$, and therefore it is possible to calculate the CIs studied in Section 3.

A. APPENDIX

The asymptotic variances-covariances of all of the parameters were obtained applying the delta method. Let $\boldsymbol{\theta} = (Se_1, Sp_1, Se_2, Sp_2)^\top$ be a vector whose components are the sensitivities and the specificities, let $\mathbf{LR} = (LR_1^+, LR_2^+, LR_1^-, LR_2^-)^\top$ be a vector whose components are the positive LRs and the negative LRs, and $\boldsymbol{\omega} = (\omega^+, \omega^-)^\top$. The matrix of asymptotic variances-covariances of $\hat{\boldsymbol{\theta}}$ is

$$\Sigma_{\hat{\boldsymbol{\theta}}} = \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \right) \Sigma_{\hat{\boldsymbol{\psi}}} \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \right)^\top.$$

Regarding the LRs, the matrix of asymptotic variances-covariances of $\widehat{\mathbf{LR}}$ is

$$\Sigma_{\widehat{\mathbf{LR}}} = \left(\frac{\partial \mathbf{LR}}{\partial \boldsymbol{\theta}} \right) \Sigma_{\hat{\boldsymbol{\theta}}} \left(\frac{\partial \mathbf{LR}}{\partial \boldsymbol{\theta}} \right)^\top.$$

Finally, the matrix of asymptotic variances-covariances of $\hat{\boldsymbol{\omega}}$ is

$$\Sigma_{\hat{\boldsymbol{\omega}}} = \left(\frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{\theta}} \right) \Sigma_{\hat{\boldsymbol{\theta}}} \left(\frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{\theta}} \right)^\top.$$

The matrix of asymptotic variances-covariances of $\ln(\hat{\boldsymbol{\omega}})$ is calculate in a similar way, i.e.

$$\Sigma_{\ln(\hat{\boldsymbol{\omega}})} = \left(\frac{\partial \ln(\boldsymbol{\omega})}{\partial \boldsymbol{\theta}} \right) \Sigma_{\hat{\boldsymbol{\theta}}} \left(\frac{\partial \ln(\boldsymbol{\omega})}{\partial \boldsymbol{\theta}} \right)^\top.$$

Performing the algebraic operations in each one of the previous expressions and substituting each parameter with its estimator, we obtain the asymptotic variances-covariances given in the equations (3.3), (3.4), (3.5) and (3.8) respectively.

B. APPENDIX

The selection of the CI with the best asymptotic behaviour was made through the following steps:

- 1) Choose the CIs with the least failures ($CP > 93\%$);
- 2) Choose the CIs which are the most precise (lowest AL) and among those which have a CP closest to 95%.

The first step in this method establishes that the CI does not fail when $CP > 93\%$. The confidence level was set at 95%, i.e. $\gamma = 1 - \alpha = 0.95$ was set as the nominal confidence and, therefore, a nominal error $\alpha = 5\%$. Let γ^* be the calculated CP, then $\Delta\alpha = \gamma^* - \gamma = \alpha - \alpha^*$, where α^* is the type I error. Furthermore, the hypothesis test to check the equality of the two LRs is $H_0: LR_1 = LR_2$ vs $H_1: LR_1 \neq LR_2$, which is equivalent to checking $H_0: \omega = 1$ vs $H_0: \omega \neq 1$. In Step 1, a CI fails if $CP \leq 93\%$, i.e. if $\Delta\alpha \leq -2$. In this situation, the type I error of the hypothesis test is $\geq 7\%$, and therefore it is a very liberal hypothesis test and can give false significances. If $\Delta\alpha > 2\%$, i.e. $CP > 97\%$, then the hypothesis test is very conservative (its type I error is very small, $< 3\%$), but does not give false significances. Therefore, the choice of the CI is linked to the decisions of the hypothesis test, and it is preferable to choose a conservative test rather than a very liberal one (as there will be no false significances due to the fact that its type I error is lower than the nominal one).

ACKNOWLEDGMENTS

This research was supported by the Spanish Ministry of Economy, Grant Number MTM2016-76938-P. We thank the two referees, the Associate Editor and the Editor (M. Ivette Gomes) of REVSTAT for their helpful comments that improved the quality of the paper.

REFERENCES

- [1] BIGGERSTAFF, B.J. (2000). Comparing diagnostic tests: a simple graphic using likelihood ratios, *Statistics in Medicine*, **19**, 649–663.
- [2] BOOS, D.D. and STEFANSKI, L.A. (2013). *Essential Statistical Inference. Theory and Method*, Springer, New York.
- [3] DOLGUN, N.A.; GOZUKARA, H. and KARAAGAOGLU, E. (2012). Comparing diagnostic tests: test of hypothesis for likelihood ratios, *Journal of Statistical Computation and Simulation*, **82**, 369–381.
- [4] EFRON, B. and TIBSHIRANI, R.J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.
- [5] FIELLER, E.C. (1940). The biological standardization of insulin, *Journal of the Royal Statistical Society*, **7**, 1–64.
- [6] LEISENRING, W. and PEPE, M.S. (1998). Regression modelling of diagnostic likelihood ratios for the evaluation of medical diagnostic tests, *Biometrics*, **54**, 444–452.
- [7] MARTÍN-ANDRÉS, A. and ÁLVAREZ-HERNÁNDEZ, M. (2014). Two-tailed asymptotic inferences for a proportion, *Journal of Applied Statistics*, **41**, 1516–1529.
- [8] MARTÍN-ANDRÉS, A. and ÁLVAREZ-HERNÁNDEZ, M. (2014). Two-tailed approximate confidence intervals for the ratio of proportions, *Statistics and Computing*, **24**, 65–75.
- [9] MONTERO-ALONSO, M.A. and ROLDÁN-NOFUENTES, J.A. (2019). Approximate confidence intervals for the likelihood ratios of a binary diagnostic test in the presence of partial disease verification, *Journal of Biopharmaceutical Statistics*, **29**, 56–81.
- [10] PEPE, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, New York.
- [11] PRICE, R.M. and BONETT, D.G. (2004). An improved confidence interval for a linear function of binomial proportions, *Computational Statistics and Data Analysis*, **45**, 449–456.
- [12] ROLDÁN-NOFUENTES, J.A. and LUNA DEL CASTILLO, J.D. (2007). Comparison of the likelihood ratios of two binary diagnostic tests in paired designs, *Statistics in Medicine*, **26**, 4179–4201.
- [13] SHAO, J. and TU, D. (1995). *The Jackknife and Bootstrap*, Springer, New York.
- [14] VACEK, P.M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests, *Biometrics*, **41**, 959–968.
- [15] WALD, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large, *Transactions of the American Mathematical Society*, **5**, 426–482.
- [16] WEINER, D.A.; RYAN, T.J.; MCCABE, C.H.; KENNEDY, J.W.; SCHLOSS, M.; TRISTANI, F.; CHAITMAN, B.R. and FISHER, L.D. (1979). Correlations among history of angina, ST-segment and prevalence of coronary artery disease in the coronary artery surgery study (CASS), *New England Journal of Medicine*, **301**, 230–235.
- [17] ZHOU, X.H.; OBUCHOWSKI, N.A. and MCCLISH, D.K. (2011). *Statistical Methods in Diagnostic Medicine*, Second Edition, Wiley, New York.

REVSTAT – STATISTICAL JOURNAL

Background

Statistics Portugal (INE, I.P.), well aware of how vital a statistical culture is in understanding most phenomena in the present-day world, and of its responsibility in disseminating statistical knowledge, started the publication of a scientific statistical journal called *Revista de Estatística*. The original language used in this publication was Portuguese and the idea behind it was to publish it, three times a year, containing original research results, and application studies, namely in the economic, social and demographic fields.

In 1998 it was decided that the publication should also include papers in English. This step was taken to achieve a broader dissemination, and to encourage foreign contributors to submit their work for publication.

At the time, the Editorial Board was mainly comprised of Portuguese university professors. It is now comprised of international university faculties and this has been the first step aimed at changing the character of *Revista de Estatística* from a national to an international scientific journal.

We have also initiated a policy of publishing special volumes that may be thematic highlighting areas of interest or associated with scientific events in Statistics. For example, in 2001, a special issue of *Revista de Estatística* was published containing three volumes of extended abstracts of the invited contributed papers presented at the 23rd European Meeting of Statisticians.

In 2003, the name of the Journal has been changed to REVSTAT - STATISTICAL JOURNAL, now fully published in English, with a prestigious international editorial board, aiming to become a reference scientific journal that promotes the dissemination of relevant research results in Statistics.

The editorial policy of REVSTAT Statistical Journal is mainly placed on the originality and importance of the research.

All articles consistent with REVSTAT aims and scope will undergo scientific evaluation by at least two reviewers, one from the Editorial Board and another external.

The only working language allowed is English.

Abstract and Indexing Services

The REVSTAT is covered by the following abstracting/indexing services:

- Current Index to Statistics
- Google Scholar
- Mathematical Reviews® (MathSciNet®)
- Science Citation Index Expanded
- Zentralblatt für Mathematic
- Scimago Journal & Country Rank
- Scopus

Instructions to Authors

Articles must be written in English and will be submitted according to the following guidelines:

The corresponding author sends the manuscript in PDF format to the Executive Editor (revstat@ine.pt) with the Subject "New Submission to REVSTAT"; a MS#REVSTAT reference will be assigned later.

Optionally, in a mail cover letter, authors are welcome to suggest one of the Editors or Associate Editors, whose opinion may be considered suitable to be taken into account.

The submitted manuscript should be original and not have been previously published nor about to be published elsewhere in any form or language, avoiding concerns about self-plagiarism'.

Content published in this journal is peer-reviewed (Single Blind).

All research articles will be refereed by at least two researchers, including one from the Editorial Board unless the submitted manuscript is judged unsuitable for REVSTAT or does not contain substantial methodological novelty, in which case is desk rejected.

Manuscripts should be typed only in black, in double-spacing, with a left margin of at least 3 cm, with numbered lines, and with less than 25 pages. Figures (minimum of 300dpi) will be reproduced online in colours, if produced this way; however, authors should take into account that the printed version is always in black and grey tones.

The first page should include the name, ORCID iD (optional), Institution, country, and mail-address of the author(s) and a summary of fewer than one hundred words, followed by a maximum of six keywords and the AMS 2000 subject classification.

Authors are encouraged to submit articles using LaTeX, in the REVSTAT style, which is available at the LaTeX2e MACROS webpage.

References about the format and other useful information on the submission are available in the LaTeX2e Templates page.

Acknowledgments of people, grants or funds should be placed in a short section before the References title page. Note that religious beliefs, ethnic background, citizenship and political orientations of the author(s) are not allowed in the text.

Supplementary files (in REVSTAT style) may be published online along with an article, containing data, programming code, extra figures, or extra proofs, etc; however, REVSTAT is not responsible for any supporting information supplied by the author(s).

Any contact with REVSTAT must always contain the assigned REVSTAT reference number.

Accepted papers

Authors of accepted papers are requested to provide the LaTeX files to the Secretary of the REVSTAT revstat@ine.pt. The authors should also mention if figure files were included, and submit electronic figures separately in .gif, .jpg, .png or .pdf format. Figures must be a minimum of 300dpi.

Copyright and reprints

Upon acceptance of an article, the author(s) will be asked to transfer copyright of the article to the publisher, Statistics Portugal, in order to ensure the widest possible dissemination of information, namely through the Statistics Portugal website (<http://www.ine.pt>).

After assigning copyright, authors may use their own material in other publications provided that REVSTAT is acknowledged as the original place of publication. The Executive Editor of the Journal must be notified in writing in advance.

Editorial Board

Editor-in-Chief

Isabel Fraga Alves, University of Lisbon, Portugal

Co-Editor

Giovani L. Silva, University of Lisbon, Portugal

Associate Editors

Marília Antunes, University of Lisbon, Portugal

Barry Arnold, University of California, USA

Narayanaswamy Balakrishnan, McMaster University, Canada

Jan Beirlant, Katholieke Universiteit Leuven, Belgium

Graciela Boente, University of Buenos Aires, Argentina

Paula Brito, University of Porto, Portugal

Valérie Chavez-Demoulin, University of Lausanne, Switzerland

David Conesa, University of Valencia, Spain

Charmaine Dean, University of Waterloo, Canada

Fernanda Figueiredo, University of Porto, Portugal

Jorge Milhazes Freitas, University of Porto, Portugal

Alan Gelfand, Duke University, USA

Stéphane Girard, Inria Grenoble Rhône-Alpes, France

Marie Kratz, ESSEC Business School, France

Victor Leiva, Pontificia Universidad Católica de Valparaíso, Chile

Artur Lemonte, Federal University of Rio Grande do Norte, Brazil

Shuangzhe Liu, University of Canberra, Australia

Maria Nazaré Mendes-Lopes, University of Coimbra, Portugal

Fernando Moura, Federal University of Rio de Janeiro, Brazil

John Nolan, American University, USA

Paulo Eduardo Oliveira, University of Coimbra, Portugal

Pedro Oliveira, University of Porto, Portugal

Carlos Daniel Paulino (2019-2021), University of Lisbon, Portugal

Arthur Pewsey, University of Extremadura, Spain

Gilbert Saporta, Conservatoire National des Arts et Métiers, France

Alexandra M. Schmidt, McGill University, Canada

Julio Singer, University of Sao Paulo, Brazil

Manuel Scotto, University of Lisbon, Portugal

Lisete Sousa, University of Lisbon, Portugal

Milan Stehlík, University of Valparaíso, Chile and LIT-JK University Linz, Austria

María Dolores Ugarte, Public University of Navarre, Spain

Executive Editor

José A. Pinto Martins, Statistics Portugal

Secretariat

José Cordeiro, Statistics Portugal

Olga Bessa Mendes, Statistics Portugal