The Truncated Inflated Beta Distribution
Communications in Statistics - Theory and Methods

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ista20

The Truncated Inflated Beta Distribution

Gustavo H. A. Pereira, Denise A. Botter & Mônica C. Sandoval

Department of Statistics, University of São Paulo, São Paulo, Brazil


To cite this article: Gustavo H. A. Pereira, Denise A. Botter & Mônica C. Sandoval (2012): The Truncated Inflated Beta Distribution, Communications in Statistics - Theory and Methods, 41:5, 907-919

To link to this article: http://dx.doi.org/10.1080/03610926.2010.530370

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
The Truncated Inflated Beta Distribution

GUSTAVO H. A. PEREIRA, DENISE A. BOTTER, AND MÔNICA C. SANDOVAL

Department of Statistics, University of São Paulo, São Paulo, Brazil

The study of proportions is a common topic in many fields of study. The standard beta distribution or the inflated beta distribution may be a reasonable choice to fit a proportion in most situations. However, they do not fit well variables that do not assume values in the open interval \( (0, c) \), \( 0 < c < 1 \). For these variables, the authors introduce the truncated inflated beta distribution (TBEINF). This proposed distribution is a mixture of the beta distribution bounded in the open interval \( (c, 1) \) and the trinomial distribution. The authors present the moments of the distribution, its scoring vector, and Fisher information matrix, and discuss estimation of its parameters. The properties of the suggested estimators are studied using Monte Carlo simulation. In addition, the authors present an application of the TBEINF distribution for unemployment insurance data.

Keywords Beta distribution; Empirical distribution function; Inflated distributions; Maximum likelihood estimator; Proportions; Truncated inflated beta distribution.

Mathematics Subject Classification Primary 62F10; Secondary 62P20.

1. Introduction

Rates, proportions, and ratios are commonly studied in many subject areas. These variables are often measured in the open interval \( (0, 1) \). In such cases, the standard beta may be a convenient probability distribution to fit the variable. It is very flexible and its density can assume different symmetric or skewed shapes. Regression models when response variable has standard beta distribution were introduced by Paolino (2001), Kieschnick and McCullough (2003), and Ferrari and Cribari-Neto (2004). Recent contributions in this area were made by Ospina et al. (2006), Smithson and Verkuilen (2006), Simas et al. (2010), Espinheira et al. (2008a,b), Venezuela (2008), and Ferrari and Pinheiro (2010).

Sometimes, a proportion may assume the values of zero or one with positive probability. In such cases, the standard beta is not a reasonable choice because it is an absolutely continuous distribution in the open interval \( (0, 1) \). A possible solution to study these proportions is to use an inflated distribution (Tu, 2002). Such classes

Received November 11, 2009; Accepted October 4, 2010
Address correspondence to Gustavo Pereira, Statistics, University of São Paulo, Rua do Matão, 1010, Butantã, São Paulo, 05508-090, Brazil; E-mail: ghapereira@terra.com.br

907
of distributions are a mixture of two others. A degenerate or discrete distribution and another known probability distribution. Therefore, these classes may be useful when a variable assumes one or more values more frequently than expected by a known probability distribution. Inflated beta distributions were introduced by Hoff (2007), Cook et al. (2008), and Ospina and Ferrari (2010), whose work considered the standard beta distribution inflated at one, at zero, and at zero and one, respectively. For other distributions, some studies in this area include Aitchison (1955), Feuerverger (1979), Lambert (1992), Ridout et al. (1998), Vieira et al. (2000), Hall (2000), Greene (1994), and Heller et al. (2006).

The standard beta distribution or the inflated beta distribution may be a reasonable choice to fit proportions in most situations. However, they do not fit well variables that cannot assume values in the open interval \((0, c)\), \(0 < c < 1\). Variables related to a kind of double-bounded payment amount when studied as a proportion of the maximum payment amount have this feature. In Brazil, for example, the amount of unemployment insurance benefit is a function of previous wages of a claimant and it is double-bounded. Therefore, the variable obtained by the ratio between the amount of the unemployment insurance benefit and the maximum allowable benefit can assume the value of zero (for an unemployed person who is not eligible for benefits) and any real number in the closed interval \((c, 1]\). This variable has positive probability at points zero, \(c\), and one because many unemployed people receive the minimum amount, while many others receive the maximum. In many countries, the amount of public retirement benefits and the amount that a worker pays monthly to the government in order to receive retirement benefits in the future when studied as a proportion of its maximum allowable amount have these features as well.

Another variable with the same features can be found in the credit market. A credit card holder receives a monthly statement indicating the total amount owed and the minimum payment. He or she can choose to pay any amount between the minimum due and the total balance. Hence, the ratio between the payment amount and the total amount owed (proportional payment amount [PPA]) has the same features as the variables discussed previously. It has positive probability at points zero, \(c\), and one because many credit card holders do not have enough money to pay anything, and many others can pay only the minimum due, whereas there are many others who pay the entire amount owed to avoid interest charges. The study of the PPA in a financial institution such as a credit card company is useful because the distribution of the variable is directly related to credit card profitability.

In this work, we introduce the truncated inflated beta distribution (TBEINF) for variables that assume values at zero, at one, and at a known value \(c\) with positive probability, and at any real number in the open interval \((c, 1]\). The remainder of the article is organized as follows. Section 2 introduces TBEINF, its properties, and estimation of parameters by conditional moments and maximum likelihood. Section 3 extends the results of the preceding section when there are more inflation points. In the following section, Monte Carlo simulation is used to study the performance of the proposed estimators. Section 5 fits TBEINF distribution to real data. Concluding remarks are given in Sec. 6.

2. The Truncated Inflated Beta Distribution

The beta distribution bounded in the open interval \((a, b)\), \(a\) and \(b\) known, is usually parameterized using two shape parameters (Johnson et al., 1995,
Chapter 25). We use an alternative parameterization indexed by the mean ($\mu$, $a < \mu < b$) and a precision parameter ($\phi > 0$; Ferrari and Cribari-Neto, 2004). In this parameterization, the beta distribution has density function

$$f(y; \mu, \phi, a, b) = \frac{\Gamma(\phi)(y-a)\left(\frac{y-a}{b-a}\right)^{\phi-1}(b-y)\left(\frac{b-y}{b-a}\right)^{\phi-1}}{\Gamma\left(\frac{a-c}{b-a}\right) \Gamma\left(\frac{b-c}{b-a}\right) (b-a)^{\phi-1}}, \quad y \in (a, b),$$

(1)

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

The beta distribution, for its flexibility, is usually a good choice to fit a double-bounded variable. As discussed in Sec. 1, there are proportions that assume values at zero, at one, and at a known value $c$ with positive probability, and at any real number in the open interval $(c, 1)$. Hence, to fit these variables, it is reasonable to introduce a distribution that is a mixture of the beta distribution and a discrete distribution. The proposed distribution is a mixture of the beta distribution bounded in the open interval $(c, 1)$ and the trinomial distribution. We call it TBEINF and its probability density function is given by

$$f(y; z, \gamma_0, \gamma_1, \mu, \phi, c) = \begin{cases} 
\frac{z}{y} & \text{if } y = 0 \\
\frac{z}{y} & \text{if } y = 1 \\
z(1-\gamma_0 - \gamma_1) & \text{if } y = c \\
(1-z)f(y; \mu, \phi, c, 1) & \text{if } y \in (c, 1),
\end{cases}$$

(2)

where $z$ is the mixture parameter ($P(Y \in \{0, 1, c\})$); $\gamma_0$ and $\gamma_1$ are the multinomial parameters ($P(Y = 0 \mid Y \in \{0, 1, c\})$ and $P(Y = 1 \mid Y \in \{0, 1, c\})$, respectively). $Y$ is a random variable with TBEINF distribution, and $f(y; \mu, \phi, c, 1)$ is the probability density function of a beta($\mu, \phi, c, 1$). In this case, the parameter space of $z$ and $\gamma_0$ is the open interval $(0, 1)$, $c < \mu < 1$, $\phi > 0$, and $0 < \gamma_1 < 1 - \gamma_0$. We will use a simpler parameterization obtained defining $\delta_0 = \gamma_0$, $\delta_1 = \gamma_1$, and $\delta_c = z(1-\gamma_0 - \gamma_1)$. In this parameterization, $\delta_0 \in (0, 1)$, $\mu \in (c, 1)$, $\phi > 0$, $0 < \delta_1 < 1 - \delta_0$, $0 < \delta_c < 1 - \delta_0 - \delta_1$, and the probability density function of the TBEINF distribution is given by

$$f(y; \delta_0, \delta_1, \delta_c, \mu, \phi, c) = \begin{cases} 
\delta_0 & \text{if } y = 0 \\
\delta_1 & \text{if } y = 1 \\
\delta_c & \text{if } y = c \\
(1-\delta_0 - \delta_1 - \delta_c)f(y; \mu, \phi, c, 1) & \text{if } y \in (c, 1),
\end{cases}$$

(3)

where $\delta_0$, $\delta_1$, and $\delta_c$ are $P(Y = 0)$, $P(Y = 1)$, and $P(Y = c)$, respectively; $\mu = E(Y \mid Y \in (c, 1))$; $\phi$ is a precision parameter; and $f(y; \mu, \phi, c, 1)$, as defined in (1). In both parameterizations, there are restrictions in the parameter space. However, as will be seen later, this is not an issue because the maximum likelihood estimators of the restricted parameters have a closed form and always satisfy the restrictions. The TBEINF distribution tends to the BEINF distribution introduced by Ospina and Ferrari (2010) as $c$ and $\delta_c$ tend to 0.

It is possible to propose a mixture of the simplex distribution and the trinomial distribution to fit these proportions. However, in this article we focus on the TBEINF because, in general, the beta distribution seems to fit bounded variables better than the simplex distribution (Kieschnick and McCullough, 2003; Miyashiro, 2008).
The \( r \)th raw moment and variance of the TBEINF distribution are given by

\[
E(Y^r) = c_1^{r} \delta_c + \delta_1 + (1 - x) \mu_r, \quad r = 1, 2, \ldots
\]

\[
\text{Var}(Y) = x V_1 + (1 - x) V_2 + (1 - x) \left[ \frac{c_1^{\delta_c} + \delta_1}{\sqrt{2}} - \sqrt{2} \mu \right]^2.
\]

where \( x = \delta_0 + \delta_1 + \delta_c \), \( V_1 = \text{Var}(Y | I_{[0,c]}) = 1 = (c_1^2 \delta_c(\delta_0 - \delta_c) + \delta_1(\delta_0 - \delta_1) - 2c_2(\delta_c, \delta_0)/x^2 \), \( V_2 = \text{Var}(Y | I_{[0,c,1]} = 0) = [(\mu - c)(1 - \mu)]/(\phi + 1), \) and \( \mu \) is the \( r \)th raw moment of the beta distribution bounded in the interval \((c, 1)\) given by \( \mu_0 = 1, \mu_1 = \mu, \) and \( \mu = [(1 - c) \Gamma([1 + r - i]/\phi) + \sum_{i=1}^{n-1} (1 - i)(n - i)] \Gamma([r + i]/\phi), \) \( p = [(\mu - c)/(1 - c)] \phi \). The results (4) can be proved using the random variables \( E(Y^r | I_{[0,c,1]}(Y)) \) and \( \text{Var}(Y | I_{[0,c,1]}(Y)) \).

Equation (3) can be written as

\[
f(y; \eta) = \exp[y^T T(y) - A(\eta)] h(y),
\]

where \( \eta = (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5)^T, \) \( T(y) = (t_1(y), t_2(y), t_3(y), t_4(y), t_5(y))^T, \) \( \theta = (\delta_0, \delta_1, \delta_c, \mu, \phi)^T, \) \( \eta_1 = \log(\delta_0/C(\theta), \eta_2 = \log(\delta_1/C(\theta), \eta_3 = \log(\delta_c/C(\theta), \eta_4 = [(\mu - c)/(1 - c)] \phi, \eta_5 = [(1 - \mu)/(1 - c)] \phi, \text{ and } \phi C(\theta) = [(1 - \mu_0 - \delta_1 - \delta_c) \Gamma(\phi)]/[\Gamma([1 - c] \phi) [\Gamma([1 - \mu] \phi)]/(1 - c)] \phi]. \) \( T(y) = I_{[0,c]}(y), t_1(y) = I_{[0,c]}(y), t_2(y) = I_{[1]}(y), t_3(y) = I_{[c]}(y), t_4(y) = \log(y - c) \) if \( y \in (c, 1) \) and 0 if \( y \in [0, c, 1], t_5(y) = \log(1 - y) \) if \( y \in (c, 1) \) and 0 if \( y \in [0, c, 1], A(\eta) = -\log[\Gamma(\eta_4 + \eta_5)]/[\Gamma(\eta_4) \Gamma(\eta_5)](1 - c)^{\mu + \eta_5 - 1} + (\phi \eta_1 \phi \eta_5 + \phi \eta_5) \Gamma(\eta_4 + \eta_5)], h(y) = 1/((c - 1)(1 - y)) \) if \( y \in (c, 1), 1 \) if \( y \in [0, c, 1], A(\eta_1) = -\log[\Gamma(\eta_4 + \eta_5)]/[\Gamma(\eta_4) \Gamma(\eta_5)](1 - c)^{\mu + \eta_5 - 1} + (\phi \eta_1 \phi \eta_5 + \phi \eta_5) \Gamma(\eta_4 + \eta_5)], h(y) = 1/((c - 1)(1 - y)) \) if \( y \in (c, 1), 1 \) if \( y \in [0, c, 1], A(\eta_2) = -\log[\Gamma(\eta_4 + \eta_5)]/[\Gamma(\eta_4) \Gamma(\eta_5)](1 - c)^{\mu + \eta_5 - 1} + (\phi \eta_1 \phi \eta_5 + \phi \eta_5) \Gamma(\eta_4 + \eta_5)], h(y) = 1/((c - 1)(1 - y)) \) if \( y \in (c, 1), 1 \) if \( y \in [0, c, 1], A(\eta_3) = -\log[\Gamma(\eta_4 + \eta_5)]/[\Gamma(\eta_4) \Gamma(\eta_5)](1 - c)^{\mu + \eta_5 - 1} + (\phi \eta_1 \phi \eta_5 + \phi \eta_5) \Gamma(\eta_4 + \eta_5)], h(y) = 1/((c - 1)(1 - y)) \) if \( y \in (c, 1), 1 \) if \( y \in [0, c, 1], A(\eta_4) = -\log[\Gamma(\eta_4 + \eta_5)]/[\Gamma(\eta_4) \Gamma(\eta_5)](1 - c)^{\mu + \eta_5 - 1} + (\phi \eta_1 \phi \eta_5 + \phi \eta_5) \Gamma(\eta_4 + \eta_5)], h(y) = 1/((c - 1)(1 - y)) \) if \( y \in (c, 1), 1 \) if \( y \in [0, c, 1], A(\eta_5) = -\log[\Gamma(\eta_4 + \eta_5)]/[\Gamma(\eta_4) \Gamma(\eta_5)](1 - c)^{\mu + \eta_5 - 1} + (\phi \eta_1 \phi \eta_5 + \phi \eta_5) \Gamma(\eta_4 + \eta_5)], h(y) = 1/((c - 1)(1 - y)) \) if \( y \in (c, 1), 1 \) if \( y \in [0, c, 1], \) in addition, neither the \( \text{is} \) nor the \( \text{is} \) satisfy a linear constraint, and parameter space contains a five-dimensional rectangle. Therefore, TBEINF is an exponential family distribution of full rank (Lehmann and Casella, 1998, Sec. 1.5). As a consequence, given a sample \( y_1, \ldots, y_n \) of the TBEINF distribution, \( (T_1, T_2, T_3, T_4, T_5) = \sum_{i=1}^{n} (t_1(y_i), t_2(y_i), t_3(y_i), t_4(y_i), t_5(y_i)) \) is a complete sufficient statistic.

Using (3) and given the sample \( y_1, \ldots, y_n \), the log-likelihood function can be written as

\[
l(\theta) = l_1(\theta) + l_2(\theta),
\]

where

\[
l_1(\theta) = T_1 \log(\delta_0) + T_2 \log(\delta_1) + T_3 \log(\delta_c) + (n - T_1 - T_2 - T_3) \log(1 - \delta_0 - \delta_1 - \delta_c)
\]

and

\[
l_2(\theta) = (n - T_1 - T_2 - T_3) \log \left[ \frac{\Gamma(\phi)(1 - c)^{\phi} \Gamma([1 - \phi]/\phi)}{\Gamma([\phi \mu - \phi]/\phi) \Gamma([\phi \mu \phi]/\phi)} \right]
\]

\[
+ \left[ \left( \frac{\phi \mu - c}{1 - c} \phi - 1 \right) T_4 + \left( \frac{1 - \mu}{1 - c} \phi - 1 \right) T_5 \right]
\]

One can notice that \( l_1(\theta) \) depends only on parameters \( \delta_0, \delta_1, \) and \( \delta_c \), and \( l_2(\theta) \) depends only on \( \mu \) and \( \phi \). Therefore, \( (\delta_0, \delta_1, \delta_c) \) and \( (\mu, \phi) \) are separable parameters and estimation of \( (\delta_0, \delta_1, \delta_c) \) can take place separately from \( (\mu, \phi) \) and vice versa.
(Pace and Salvan, 1997, p. 128). The score function obtained differentiating (6) with respect to each of the parameters is given by \( (U_{\theta_0}, U_{\theta_1}, U_\mu, U_\phi) \), where \( U_{\theta_0} = T_1/\delta_0 - (n - T_1 - T_2 - T_3)/(1 - z), \) \( U_{\theta_1} = T_2/\delta_1 - (n - T_1 - T_2 - T_3)/(1 - z), \) \( U_\mu = T_3/\delta_2 - (n - T_1 - T_2 - T_3)/(1 - z), \) \( U_\phi = \phi/(1 - c) \{(n - T_1 - T_2 - T_3)\psi(1 - \mu)/(1 - c) - \psi(\phi(\mu - c)/(1 - c)) + T_4 - T_3\}. \) \( U_\phi = (n - T_1 - T_2 - T_3)\psi(\phi - [(\mu - c)/(1 - c)] - [(1 - \mu)/(1 - c)]\phi(1 - \mu)/(1 - c) - \log(1 - c)) + [(\mu - c)/(1 - c)]T_4 + [(1 - \mu)/(1 - c)]T_5, \) and \( \psi(\cdot) \) is the digamma function. Solving \((U_{\theta_0}, U_{\theta_1}, U_\mu) = (0, 0, 0)\), we find the maximum likelihood estimators, \( \hat{\theta}_0 = T_1/n, \) \( \hat{\theta}_1 = T_2/n \) and \( \hat{\delta}_2 = T_3/n. \) They are unbiased and a function of a complete sufficient statistic; hence, \( \hat{\theta}_0, \hat{\delta}_1 \) and \( \hat{\delta}_2 \) are uniform minimum variance unbiased estimators (UMVUE) of \( \theta_0, \delta_1, \) and \( \delta_2, \) respectively. The variance of \( \hat{\delta}_k, k = 0, 1, c \) is \( \delta_k(1 - \delta_k)/n. \) The equation system \((U_\mu, U_\phi) = (0, 0)\) does not have an algebraic solution. Therefore, the maximum likelihood estimators of \( \mu \) and \( \phi \) must be obtained using numerical methods. In Secs. 4 and 5, we will use the BFGS quasi-Newton method with analytical derivatives (Nocedal and Wright, 2006, Sec. 6.1).

One can obtain closed-form estimators of \( \mu \) and \( \phi \) based on conditional moments of \( Y \) given that \( Y \in (c, 1) \). The distribution of \( Y \mid Y \in (c, 1) \) is beta\( (\mu, \phi, c, 1) \); hence, \( E[Y \mid Y \in (c, 1)] = \mu \) and \( \text{Var}[Y \mid Y \in (c, 1)] = (\mu - c)/(1 + \phi). \) Solving the system of equations \([\mu, (\mu - c)/(1 + \phi)] = [\bar{y}, s^2], \) where \( \bar{y} = \sum_{y \in c, 1} y/[(n - T_1 - T_2 - T_3)] \) and \( s^2 = \sum_{y \in c, 1} [(y - \bar{y})^2/(n - T_1 - T_2 - T_3)] \), we find the estimators \( \hat{\mu} = \bar{y} \) and \( \hat{\phi} = (\bar{y} - c)/(1 - \bar{y})s^2 - 1. \)

The Fisher information matrix, \( K(\partial) \), is obtained using this, in an exponential family distribution of full rank, \( E[t_k(Y)] = \hat{c}A(\eta)/\hat{c} \eta_k \) (Lehmann and Casella, 1998, p. 27). After some algebra, we find that

\[
K(\partial) = \begin{pmatrix}
  k_{\delta_0 \delta_0} & k_{\delta_0 \delta_1} & k_{\delta_0 \mu} & 0 & 0 \\
  k_{\delta_1 \delta_0} & k_{\delta_1 \delta_1} & k_{\delta_1 \mu} & 0 & 0 \\
  k_{\mu \delta_0} & k_{\mu \delta_1} & k_{\mu \mu} & 0 & 0 \\
  0 & 0 & 0 & k_{\mu \phi} & k_{\phi \phi} \\
  0 & 0 & 0 & k_{\phi \mu} & k_{\phi \phi}
\end{pmatrix},
\]

where \( k_{\delta_0 \delta_0} = (1 - \delta_1 - \delta_2)/[\delta_0(1 - z)], \) \( k_{\delta_0 \delta_1} = (1 - \delta_0 - \delta_2)/[\delta_1(1 - z)], \) \( k_{\mu \delta_0} = (1 - \delta_0 - \delta_2)/[\delta_0(1 - z)], \) \( k_{\mu \phi} = k_{\delta_0 \phi} = (1 - z)^{-1}, \) \( k_{\mu \mu} = [\phi(1 - c)]^2(1 - z) \psi(\phi(\mu - c)/(1 - c))^2(1 - z) + \psi(\phi(\mu - c)/(1 - c)) + \psi(\phi(\mu - c)/(1 - c))^2, \) \( k_{\phi \phi} = (1 - z)[[(\mu - c)/(1 - c)] - \psi(\phi(\mu - c)/(1 - c)) - (1 - \mu)\psi(\phi(1 - \mu)/(1 - c)). \)

The asymptotic properties of the maximum likelihood estimator, \( \sqrt{n}(\hat{\theta} - \theta) \rightarrow D N(0, K(\partial)^{-1}) \) (regularity conditions are satisfied because TBEINF distribution belongs to the exponential family of full rank). Hence, an asymptotic \((100 \times \gamma)\% \) confidence interval for the \( k \)th parameter of the TBEINF distribution is given by

\[
\text{IC}(\theta_k; \gamma) = \hat{\theta}_k \pm z_{(1-\gamma)/2}(k_{\hat{\theta}_k \hat{\theta}_k}/n)^{1/2},
\]
where \( z_{(1+\gamma)/2} \) is the \((1+\gamma)/2\) quantile of the \(N(0, 1)\) and \( \hat{k}^{0, \delta} \) is the \((k, k)\) term of \( K(\hat{\theta})^{-1} \) evaluated at \( \hat{\theta} \). In addition, an asymptotic \((100 \times \gamma)\%\) confidence region for the five parameters of the TBEINF distribution can be written (Johnson and Wichern, 2007, p. 221) as

\[
n(\hat{\theta} - \theta)^\top K(\hat{\theta})(\hat{\theta} - \theta) \leq [5(n - 1)/(n - 5)]F_{5,n-5;5},
\]

where \( F_{5,n-5;5} \) is the \( \gamma \) quantile of the of the \( F \) distribution with 5 and \( n - 5 \) degrees of freedom, and \( \hat{K}(\hat{\theta}) \) is \( K(\theta) \) evaluated at \( \hat{\theta} \). The confidence intervals and the confidence region can be used to test hypotheses about the parameters. The asymptotic distribution of the maximum likelihood estimator of any differentiable function of \( \theta \) can be obtained using the delta method.

3. Variables with More Inflation Points

The results presented in Sec. 2 can be extended easily for cases in which there are more inflation points. Suppose that a variable \( X \) assumes values at zero, at one, at a known value \( c \), and at \( v \) (integer greater than zero) other additional known values in the interval \((0, c)\) with positive probability and at any real number in the open interval \((c, 1)\). If admitting that the distribution of the variable \( X \) is a mixture of the beta distribution bounded in the open interval \((c, 1)\) and the multinomial distribution, then its probability density function, in a parameterization similar to (3), can be written as

\[
f(x; \delta_0, \delta_1, \delta_c, \lambda_1, \ldots, \lambda_v, \mu, \phi) = \begin{cases} 
\delta_0 & \text{if } x = 0 \\
\delta_1 & \text{if } x = 1 \\
\delta_c & \text{if } x = c \\
\lambda_i & \text{if } x = s_i, i = 1, 2, \ldots, v \\
(1 - \tau)f(x; \mu, \phi, c, 1) & \text{if } x \in (c, 1),
\end{cases}
\]

where \( s_i \) are the known values in the interval \((0, c)\) in which \( P(X = s_i) > 0 \), \( \lambda_i = P(X = s_i) \), and \( \tau = \delta_0 + \delta_1 + \delta_c + \sum_{i=1}^{v} \lambda_i \).

The moments of \( X \), estimation of its parameters and the Fisher information matrix can be obtained as in Sec. 2. The results obtained, in general, have more terms, but they are very similar to those presented in the previous section. The \( r \)th raw moment of \( X \), for example, is given by \( E(X^r) = c^r \delta_1 + \delta_c + \sum_{i=1}^{v} s_i^r \lambda_i + (1 - \tau)\mu_r \). The maximum likelihood estimator of \( \lambda_i \) is \( \hat{\lambda}_i = \sum_{j=1}^{n} I_{(s_i)}(y_j)/n \) and the estimators of \( \delta_0, \delta_1, \delta_c, \mu, \phi \) do not change. The elements of the Fisher information matrix relative to \( \hat{\lambda}_i \) can be written as \( k_{\hat{\lambda}_i, \hat{\lambda}_j} = (1 - \tau + \lambda_i)/[\hat{\lambda}_j(1 - \tau)] \)\( k_{\hat{\lambda}_i, \hat{\lambda}_j} = k_{\hat{\lambda}_j, \hat{\lambda}_i} = (1 - \tau)^{-1} \) and \( k_{\hat{\lambda}_i, \hat{\mu}} = k_{\hat{\lambda}_i, \phi} = 0 \). The results are almost the same if the values \( s_i \) are not restricted to the interval \((0, c)\).

4. Simulation Studies and BEINF and TBEINF Comparison

The conditional moments (CM) estimators and maximum likelihood estimators of \( \mu, \phi, E(\gamma) \), and \( \text{Var}(\gamma) \) in the TBEINF distribution were compared through two Monte Carlo simulation studies. The maximum likelihood estimators of \( \delta_0, \delta_1, \) and \( \delta_c \) are not
Simulation results for the TBEINF distribution for some values of $n$: $\mu = 0.4$, $\phi = 2$, $\delta_0 = 0.3$, $\delta_1 = 0.1$, $\delta_c = 0.2$, $c = 0.2$, $E(Y) = 0.3$, $\text{Var}(Y) = 0.098$

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\mu$ Bias</th>
<th>$\mu$ RMSE</th>
<th>$\phi$ Bias</th>
<th>$\phi$ RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>-0.0003</td>
<td>-0.0035</td>
<td>0.0742</td>
<td>0.0727</td>
</tr>
<tr>
<td>50</td>
<td>0.0000</td>
<td>-0.0015</td>
<td>0.0457</td>
<td>0.0441</td>
</tr>
<tr>
<td>100</td>
<td>0.0002</td>
<td>-0.0007</td>
<td>0.0318</td>
<td>0.0307</td>
</tr>
<tr>
<td>500</td>
<td>0.0000</td>
<td>-0.0002</td>
<td>0.0143</td>
<td>0.0137</td>
</tr>
<tr>
<td>1000</td>
<td>0.0000</td>
<td>-0.0001</td>
<td>0.0100</td>
<td>0.0096</td>
</tr>
<tr>
<td>10000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0032</td>
<td>0.0030</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n$</th>
<th>$E(Y)$ Bias</th>
<th>$E(Y)$ RMSE</th>
<th>$\text{Var}(Y)$ Bias</th>
<th>$\text{Var}(Y)$ RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.0000</td>
<td>-0.0012</td>
<td>0.0710</td>
<td>0.0707</td>
</tr>
<tr>
<td>50</td>
<td>-0.0001</td>
<td>-0.0007</td>
<td>0.0443</td>
<td>0.0440</td>
</tr>
<tr>
<td>100</td>
<td>0.0000</td>
<td>-0.0003</td>
<td>0.0313</td>
<td>0.0311</td>
</tr>
<tr>
<td>500</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0141</td>
<td>0.0139</td>
</tr>
<tr>
<td>1000</td>
<td>-0.0001</td>
<td>-0.0001</td>
<td>0.0099</td>
<td>0.0098</td>
</tr>
<tr>
<td>10000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0032</td>
<td>0.0031</td>
</tr>
</tbody>
</table>

considered because they are UMVUE. For each case, we used 10,000 replications and obtained the estimated bias and the estimated root mean square error (RMSE) of the estimators. The two simulation studies described subsequently and the application of Sec. 5 were performed using the Ox language (Doornik, 2007).

The first simulation study was performed for the TBEINF distribution with $\delta_0 = 0.3$, $\delta_1 = 0.1$, $\delta_c = 0.2$, $\mu = 0.4$, $\phi = 2$, $c = 0.2$, and some values of $n$. Table 1 summarizes the simulation results. The estimators $\hat{\mu}$, $E(Y)$, $\text{Var}(Y)$, and $\hat{\phi}$ are slightly biased especially in small sample sizes, while the estimators $\tilde{\mu}$ and $\tilde{E}(Y)$ seem to be unbiased. However, analyzing the RMSE, the performance of the maximum likelihood estimators for these parameters appears to be slightly better than the CM estimators. In summary, the CM and maximum likelihood estimators of $\mu$, $E(Y)$ and $\text{Var}(Y)$ perform well even in small sample sizes. On the other hand, the estimators $\hat{\phi}$ and $\tilde{\phi}$ considerably overestimate $\phi$ for small sample sizes and $\hat{\phi}$ performs better than $\tilde{\phi}$. It is interesting to note that $\hat{\phi}$ is very biased, but $\text{Var}(Y)$, which is a function of $\phi$, is only slightly biased.

In the second simulation study, we fixed $c = 0.2$ and $n = 100$ and performed simulations for some values of the parameters of the TBEINF distribution. Table 2 presents the simulation results. The parameter space of $\phi$ does not have an upper bound and, in the table, we compare the performance of the estimators for two very different values of $\phi$. For this reason, to better study the estimators of $\phi$, for this parameter, the table presents the relative bias (bias/$\phi$) and the relative RMSE.
Table 2
Simulation results for the TBEINF distribution for some values of $\mu$, $\phi$, $\delta_0$, $\delta_1$, $\delta_2$: $n = 100$, $c = 0.2$ ($\delta$ sm.: $\delta_0 = 0.075$, $\delta_1 = 0.025$, $\delta_2 = 0.050$, $\delta$ la.: $\delta_0 = 0.3$, $\delta_1 = 0.1$, $\delta_2 = 0.2$)

<table>
<thead>
<tr>
<th>Parameters values</th>
<th>$\hat{\mu}$</th>
<th>$\bar{\mu}$</th>
<th>RMSE</th>
<th>$\hat{\phi}$</th>
<th>$\bar{\phi}$</th>
<th>RMSE</th>
<th>Relat. Bias (%)</th>
<th>Relat. RMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta$ sm., $\mu = 0.4$, $\phi = 2$</td>
<td>0.0000</td>
<td>$-0.0003$</td>
<td>0.0219</td>
<td>2.0210</td>
<td>3.4907</td>
<td>3.6775</td>
<td>18.9642</td>
<td>16.3079</td>
</tr>
<tr>
<td>$\delta$ sm., $\mu = 0.4$, $\phi = 100$</td>
<td>0.0000</td>
<td>$0.0000$</td>
<td>0.0219</td>
<td>2.0210</td>
<td>3.7220</td>
<td>3.6756</td>
<td>16.8567</td>
<td>16.7500</td>
</tr>
<tr>
<td>$\delta$ sm., $\mu = 0.6$, $\phi = 2$</td>
<td>0.0001</td>
<td>$0.0000$</td>
<td>0.0249</td>
<td>2.0239</td>
<td>2.5823</td>
<td>2.9657</td>
<td>15.6590</td>
<td>14.0930</td>
</tr>
<tr>
<td>$\delta$ sm., $\mu = 0.6$, $\phi = 100$</td>
<td>0.0000</td>
<td>$0.0000$</td>
<td>0.0044</td>
<td>2.0044</td>
<td>3.7919</td>
<td>3.8267</td>
<td>16.7634</td>
<td>16.7709</td>
</tr>
<tr>
<td>$\delta$ la., $\mu = 0.4$, $\phi = 2$</td>
<td>0.0002</td>
<td>$-0.0007$</td>
<td>0.0318</td>
<td>0.0307</td>
<td>8.6767</td>
<td>8.3245</td>
<td>31.3073</td>
<td>26.4972</td>
</tr>
<tr>
<td>$\delta$ la., $\mu = 0.4$, $\phi = 100$</td>
<td>$-0.0001$</td>
<td>$-0.0001$</td>
<td>0.0055</td>
<td>0.0055</td>
<td>8.2501</td>
<td>8.2402</td>
<td>27.7933</td>
<td>27.6028</td>
</tr>
<tr>
<td>$\delta$ la., $\mu = 0.6$, $\phi = 2$</td>
<td>$-0.0006$</td>
<td>$-0.0007$</td>
<td>0.0367</td>
<td>0.0354</td>
<td>6.5076</td>
<td>7.4273</td>
<td>25.3738</td>
<td>23.6132</td>
</tr>
<tr>
<td>$\delta$ la., $\mu = 0.6$, $\phi = 100$</td>
<td>0.0000</td>
<td>$0.0000$</td>
<td>0.0063</td>
<td>0.0063</td>
<td>8.4845</td>
<td>8.5580</td>
<td>27.5290</td>
<td>27.5513</td>
</tr>
</tbody>
</table>

(RMSE/$\phi$). For all values of the parameters studied, the estimators of $\mu$, $E(Y)$, and $\text{Var}(Y)$ seem to be unbiased or only slightly biased and $\hat{\phi}$ and $\bar{\phi}$ overestimate $\phi$. In addition, the performances of all estimators are better when the $\delta$ parameters are small than when they are large. This result was expected because the estimators of $\mu$ and $\phi$, in practice (see Sec. 2), only use the observations in which $Y_i \in (c, 1)$. As the values of $\delta$ parameters increase, the number of observations used to estimate $\mu$ and $\phi$ decreases and, in consequence, the performances of the estimators worsen. The value of $\phi$ seems to considerably affect the finite-sample behavior of the estimators of $\mu$ and $\phi$. The performances of $\hat{\mu}$ and $\bar{\mu}$ are much better when $\phi = 100$ than when $\phi = 2$. This is reasonable because as $\phi$ increases the variance of $Y$ decreases. Moreover, the performance of $\hat{\phi}$ and $\bar{\phi}$ is almost the same when $\phi = 100$, but the first performs better when $\phi = 2$. On the other hand, in relation to the performance of the estimators, the value of $\mu$ seems to be less important than the values of $\phi$ and $\delta$ parameters. In summary, the studied estimators of $\mu$, $E(Y)$, and $\text{Var}(Y)$ perform well, and $\hat{\phi}$ and $\bar{\phi}$ are good estimators of $\phi$ only when sample size is large and, in general, the maximum likelihood estimators seem to be slightly better than the CM estimators.

Figure 1 compares the cumulative distribution function (cdf) of the TBEINF with $\delta_0 = 0.3$, $\delta_1 = 0.1$, $\mu = 0.4$, $\phi = 2$, and some values of $c$ and $\delta_2$, and the cdf
Figure 1. The TBEINF distribution function for some values of $c$ and $\delta_c$ ($\mu = 0.4$, $\phi = 2$, $\delta_0 = 0.3$, $\delta_1 = 0.1$) and the BEINF distribution function with the same mean, variance, $\delta_0$, and $\delta_1$. Of the BEINF with the same mean, the same variance, and the same probability of assuming the values 0 and 1. It can be noticed that for $c = 0.01$ and $\delta_c = 0.01$, both distributions have almost the same cdf. However, as $c$ and $\delta_c$ increase, the difference between the TBEINF cdf and the BEINF cdf increases rapidly. The value of $c$ seems to cause greater difference between the cdfs than the value of $\delta_c$. This difference, for example, is higher when $c = 0.3$ and $\delta_c = 0.01$ than when $c = 0.01$ and $\delta_c = 0.3$. The graphs suggest that if someone does not notice that $P(Y \in (0, c)) = 0$ and fits the BEINF distribution, they will make a considerable mistake. The error would be insubstantial only if $c$ and $\delta_c$ were very small.

5. Application
In Brazil, the amount of the unemployment insurance benefit paid out to a claimant is a function of the previous wages of a given person, and it is bounded in the interval $[465.00; 870.01]$ (currency: Reais, year 2009). As mentioned in Sec. 1, many unemployed individuals receive the minimum amount, while many others receive the maximum. In addition, there are unemployed workers who are not eligible for benefits. Therefore, if studying the amount of the unemployment insurance benefits as a proportion of its maximum allowable amount, the TBEINF distribution may be a reasonable choice to fit the variable.

We considered a sample of 2,322 unemployed individuals from the state of Bahia, Brazil, who received their first unemployment insurance payment in May/2009. Data were supplied by Ministério do Trabalho e Emprego do Brasil. We studied the variable defined as the ratio of the amount of the third unemployment insurance payment to the maximum amount of the unemployment insurance benefit (proportional amount of the unemployment insurance benefit [PAUIB]). In the sample, the values zero, one, and $c = 0.5345$ were assumed by 99, 198, and 1,210 people, respectively. The sample mean and the sample standard deviation for the observations in the interval $(c, 1)$ were 0.7196 and 0.1268, respectively. Using the results given in Sec. 2, it is possible to obtain point (maximum likelihood) and
Table 3

Point (maximum likelihood) and interval estimates (95% confidence level) of the parameters of the TBEINF distribution to the proportional amount of the unemployment insurance benefits in Bahia, Brazil

<table>
<thead>
<tr>
<th></th>
<th>( \delta_0 )</th>
<th>( \delta_1 )</th>
<th>( \delta_c )</th>
<th>( \mu )</th>
<th>( \phi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum likelihood estimates</td>
<td>0.0426</td>
<td>0.0853</td>
<td>0.5211</td>
<td>0.7226</td>
<td>2.2250</td>
</tr>
<tr>
<td>Lower limit</td>
<td>0.0344</td>
<td>0.0739</td>
<td>0.5008</td>
<td>0.7142</td>
<td>2.0393</td>
</tr>
<tr>
<td>Upper limit</td>
<td>0.0509</td>
<td>0.0966</td>
<td>0.5414</td>
<td>0.7310</td>
<td>2.4107</td>
</tr>
</tbody>
</table>

interval estimates of the parameters (Table 3). Employing Eq. (9), a confidence region for the five parameters was also calculated, which is not shown here because the equation has many terms.

The confidence interval and the confidence region were based on the asymptotic properties of the maximum likelihood estimator. In order to study whether they are good enough in a sample size of 2,322 individuals, a Monte Carlo simulation study was performed. We conducted the simulation for a variable with TBEINF distribution with parameters equal to those estimated, and we used 10,000 replications. Table 4 shows the coverage probability estimates. It can be noticed that all estimates are close to the theoretical confidence levels. The table suggests that the intervals and region given in (8) and (9) will perform well if the sample size is as large as 2,322. The TBEINF distribution was proposed to fit some variables in econometrics. In this area, sample size is usually not an issue and it is often larger than 2,322. For this reason, in practice, results (8) and (9) are accurate enough to estimate the parameters and test hypotheses about them.

Figure 2 presents, in the left plot, the empirical distribution function of PAUIB and the TBEINF distribution function with parameters given by their maximum likelihood estimators. It can be noticed that both distribution functions are very close to each other. However, as the TBEINF distribution has a parameter for each of the discrete values, both curves are exactly equal in the interval \([0,c]\). Therefore, it is more convenient to study a conditional distribution function \(P(PAUIB \leq x | c < PAUIB < 1)\). The right plot of Fig. 2 shows the empirical and the TBEINF conditional distribution function. In the entire interval \((0.5345, 1.0000)\), both curves are not distant to each other. As PAUIB clearly has positive probability to assume the values zero, one, and \(c\), and can assume any value in the interval \((c, 1)\), it will have TBEINF distribution if it has beta distribution in the interval \((c, 1)\). To

Table 4

Coverage probability estimates of the confidence intervals and confidence region for the parameters of the TBEINF distribution \(n = 2,322\) and parameters values equal to estimates presented in Table 3

<table>
<thead>
<tr>
<th>Confidence level</th>
<th>( \delta_0 )</th>
<th>( \delta_1 )</th>
<th>( \delta_c )</th>
<th>( \mu )</th>
<th>( \phi )</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90</td>
<td>0.8940</td>
<td>0.8962</td>
<td>0.9026</td>
<td>0.8971</td>
<td>0.9062</td>
<td>0.8983</td>
</tr>
<tr>
<td>0.95</td>
<td>0.9449</td>
<td>0.9472</td>
<td>0.9501</td>
<td>0.9533</td>
<td>0.9507</td>
<td>0.9503</td>
</tr>
<tr>
<td>0.99</td>
<td>0.9899</td>
<td>0.9880</td>
<td>0.9919</td>
<td>0.9899</td>
<td>0.9893</td>
<td>0.9906</td>
</tr>
</tbody>
</table>
study this condition, we performed an Anderson-Darling goodness-of-fit test with half-sample method (Stephens, 1986, p. 169). The hypothesis that PAUIB has beta distribution in the interval \((c, 1)\) was not rejected \((p > .10)\). Figure 2 and the result of the goodness-of-fit test suggest that it is reasonable to fit the TBEINF distribution for the variable PAUIB.

6. Conclusion

In this work, we introduced the TBEINF distribution that is discrete at points 0, 1, and \(c\), and continuous in the open interval \((c, 1)\). We presented certain properties of the distribution, performed simulation studies, and showed an application for unemployment insurance data.

We mentioned four different variables that might be fitted by the TBEINF distribution. There might be other variables with the same features. The proposed distribution may be useful to governmental and financial institutions. In countries where the amount of the unemployment insurance benefit is bounded, for example, governmental organizations can use the TBEINF distribution, its estimated parameters, and a forecast of the number of the unemployed people to perform a simulation study. Given an estimate of the number of the unemployed people in the near future, such studies could estimate the conditional distribution of unemployment insurance expenses. Therefore, the proposed distribution may be useful to budget unemployment benefit expenses. In financial institutions, it could be helpful for developing regression models with TBEINF response variable for the PPA. For each customer, the estimate of \(\delta_0\) is a credit risk measure of that individual. On the other hand, the estimates of \(\delta_0, \delta_1, \delta_c,\) and \(\mu\) are directly related to customer profitability. Hence, a regression model with TBEINF response variable may help financial institutions manage credit card risks and rewards.

The examples mentioned previously and in Sec. 1 and the results obtained in Secs. 4 and 5 suggest that the TBEINF distribution can be useful for studying proportions that cannot assume values in the interval \((0, c)\).
Acknowledgments

Insurance unemployment data were kindly provided by Gabrielle Souza and Márcio Borges of Ministério do Trabalho e Emprego do Brasil. We thank two anonymous referees and professor Gizelton Alencar for helpful comments and suggestions on an earlier draft. This research was partially supported by CAPES and CNPq.

References


