ARTICLE TEMPLATE

An Alternative Discrete Analogue of the Half-Logistic Distribution Based on Minimization of a Distance between Cumulative Distribution Functions

Alessandro Barbiero[a] and Asmerilda Hitaj[b]

[a]Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, via Conservatorio 7, 20122 Milan (Italy); [b]Department of Economics, Università degli Studi dell'Insubria, via Monte Generoso 71, 21100 Varese (Italy)

ARTICLE HISTORY
Compiled December 26, 2023

ABSTRACT
A discrete version of the continuous half-logistic distribution is introduced, which is based on the minimization of the Cramér distance between the corresponding continuous and step-wise cumulative distribution functions. The expression of the probability mass function is derived in an analytic form, and some properties of the distribution - mainly related to moments and reliability concepts - are discussed. As for sample estimation, three different techniques are suggested, whose theoretical and empirical features are examined also through a Monte Carlo simulation study, comprising several parameter and sample size combinations. A comparison is also made between the proposed distribution and a discrete version already proposed in the literature, based on a different rationale, and a main difference is highlighted. A count regression model is suggested where the response variable follows the discrete half-logistic distribution and artificial and real data are used to illustrate its use. Finally, the performance of the proposed distribution over other classical models is discussed based on a real data set.

KEYWORDS
count distribution; Cramér distance; cumulative distribution function; logistic distribution; survival data

1. Introduction

The half-logistic distribution is a continuous random distribution supported on $\mathbb{R}^+$ obtained by folding the logistic distribution, which is defined on $\mathbb{R}$, about the origin (Balakrishnan, 1985). Thus, if Y is a random variable (rv) that follows the logistic distribution with parameter $\theta > 0$, with cumulative distribution function (cdf) $F_Y(y) = \frac{1}{1+e^{-\theta y}}$ and probability density function (pdf) $f_Y(y) = \frac{\theta e^{-\theta y}}{(1+e^{-\theta y})^2}$, the rv $X = |Y|$ follows the the half-logistic distribution with the same parameter θ; the expression of its pdf is

$$f(x) = \frac{2\theta e^{-\theta x}}{(1+e^{-\theta x})^2}, \quad x \in \mathbb{R}^+, \theta \in \mathbb{R}^+; \tag{1}$$

CONTACT Alessandro Barbiero. Email: alessandro.barbiero@unimi.it

whereas for its cdf it is

$$F(x) = \frac{2}{1+e^{-\theta x}} - 1 = \frac{2e^{\theta x}}{1+e^{\theta x}} - 1 = \frac{e^{\theta x}-1}{e^{\theta x}+1}, \quad x \in \mathbb{R}^+. \tag{2}$$

The expectation is $\mu = \log 4/\theta$.

Barbiero and Hitaj (2020) introduced a discrete analogue of the half-logistic distribution defined through (1) or (2), by imposing the matching of the survival function $P(X \geq x)$ at each integer value of the support, i.e., defining the probability mass function (pmf) as $p(i) = p_i = F(i+1) - F(i)$, $i = 0, 1, 2, \dots$. The probabilities of this discrete analogue of the half-logistic distribution are then:

$$p_i = 2\left[1+e^{\theta i}\right]^{-1} - 2\left[1+e^{\theta(i+1)}\right]^{-1} = 2e^{-\theta i}/[1+e^{-\theta i}] - 2e^{-\theta(i+1)}/[1+e^{-\theta(i+1)}], \tag{3}$$

for $x = 0, 1, 2, \dots$. The p_i's are a decreasing sequence for any possible value of θ; therefore, the mode is always 0.

In this paper, we introduce and discuss an alternative discrete version of the continuous half-logistic distribution by following a different approach, based on the minimization of a discrepancy measure between the continuous cdf of the parent distribution and the step-wise cdf of the discrete counterpart (Barbiero & Hitaj, 2021). The distance chosen is the Cramér distance, defined as

$$d(F,G) = \int_{\mathbb{R}} |F(x) - G(x)|^2 \mathrm{d}x, \tag{4}$$

where F and G are the continuous and step-wise cdf of the continuous random distribution and of its discrete version, respectively. The paper is structured as follows: In the next section, we provide the general solution to the problem stated above and then derive in particular the “optimal” discrete counterpart of the half-logistic distribution, by providing the analytic expression of its pmf. Some properties of the proposed distribution are discussed. The third section is devoted to sample estimation; the maximum likelihood method, the method of moment and the method of proportion are discussed and empirically compared through a Monte Carlo simulation study. The fourth section suggests a regression model where the response variable follows the discrete half-logistic distribution. The fifth and final section presents an application to a real dataset, on which the proposed discrete distribution is fitted.

2. Definition of an alternative discrete version of the half-logistic distribution

If G is a stepwise cdf, supported on the non-negative integers $i \in \{0, 1, 2, \dots\}$, which can be seen as a discrete version of a continuous cdf F, supported on the positive half-line, letting $Q_i = G(i)$, the Cramér distance (4) can be rewritten as

$$d(F,G) = \sum_{i=0}^{\infty} \int_i^{i+1} |F(x) - Q_i|^2 \mathrm{d}x.$$

By minimizing the function above with respect to the Q_i's, we obtain the "optimal" values as (Barbiero & Hitaj, 2021)

$$Q_i = \int_i^{i+1} F(x)\mathrm{d}x. \tag{5}$$

In fact, the first-order derivative of $d(F,G)$ with respect to Q_i is

$$\frac{\mathrm{d}d(F,G)}{\mathrm{d}Q_i} = -2\int_i^{i+1} F(x)\mathrm{d}x + 2Q_i,$$

and setting it equal to zero leads to the expression for Q_i showed above. The second-order derivative of $d(F,G)$ with respect to Q_i is $\frac{\mathrm{d}^2 d(F,G)}{\mathrm{d}Q_i^2} = 2 > 0$, which confirms that the Q_i's in (5) constitute an absolute minimum. Lingering over their expression, it can be deduced that the optimal Q_i's represent a series of non-decreasing values bounded between 0 and 1, tending to 1 as i tends to ∞. The corresponding probabilities p_i are obtained by difference as $p_i = Q_i - Q_{i-1}$, if $i \geq 1$, whereas $p_0 = Q_0$. Automatically, the p_i constitute a valid series of probabilities, since $p_i \geq 0$ for any $i \geq 0$ and $\sum_{i=0}^{\infty} p_i = Q_0 + \sum_{i=1}^{\infty}(Q_i - Q_{i-1}) = \lim_{i\to\infty} Q_i = 1$.

The optimal discrete analogue of the half-logistic distribution has then cumulative probabilities given by (Barbiero & Hitaj, 2023)

$$\begin{aligned} Q_i &= \int_i^{i+1} \left(1 - \frac{2}{1+e^{\theta x}}\right)\mathrm{d}x = 1 - 2\int_i^{i+1} 1 - \frac{e^{\theta x}}{1+e^{\theta x}}\mathrm{d}x = 1 - 2\left[x - \frac{\log(1+e^{\theta x})}{\theta}\right]_i^{i+1} \\ &= 1 - 2 + \left[\frac{2\log(1+e^{\theta x})}{\theta}\right]_i^{i+1} = \frac{2}{\theta}\log\frac{1+e^{\theta(i+1)}}{1+e^{\theta i}} - 1, \end{aligned} \tag{6}$$

for $i = 0, 1, 2, \ldots$, so that the probabilities are

$$\begin{cases} p_0 = Q_0 = \dfrac{2}{\theta}\log\dfrac{1+e^{\theta}}{2} - 1 \\ p_i = Q_i - Q_{i-1} = \dfrac{2}{\theta}\log\dfrac{(1+e^{\theta(i+1)})(1+e^{\theta(i-1)})}{(1+e^{\theta i})^2}, \quad i = 1, 2, \ldots. \end{cases} \tag{7}$$

It can be proved that $p_0 < p_1$ if and only if θ is smaller than a certain value θ^*. In fact, $p_0 < p_1$ if and only if $\frac{2}{\theta}\log\frac{1+e^{\theta}}{2} - 1 < \frac{2}{\theta}\log\frac{2(1+e^{2\theta})}{(1+e^{\theta})^2}$, which is equivalent to $(1+e^{\theta})^3 < 4e^{\theta/2}(1+e^{2\theta})$ or, by setting $\omega = e^{\theta}$, $(1+\omega)^3 < 4w^{0.5}(1+\omega^2)$, which is satisfied for any $\omega < \omega^* = 8.35241$, i.e., for any $\theta < \theta^* = 2.12255$. Conversely, for any $\theta > \theta^*$, $p_0 > p_1$, whereas if $\theta = \theta^*$, it follows $p_0 = p_1$.

It can be also proved that $p_i > p_{i+1}$ for any $i \geq 1$. In fact, $p_i - p_{i+1} = \frac{2}{\theta}\log\frac{(1+e^{\theta(i+1)})^3(1+e^{\theta(i-1)})}{(1+e^{\theta i})^3(1+e^{\theta(i+2)})}$, or $p_i - p_{i+1} = \frac{2}{\log\omega}\log\frac{(1+\omega^{i+1})^3(1+\omega^{i-1})}{(1+\omega^i)^3(1+\omega^{i+2})}$. But $\frac{(1+\omega^{i+1})^3(1+\omega^{i-1})}{(1+\omega^i)^3(1+\omega^{i+2})} > 1$ for any positive integer i and for any $\omega > 1$; in fact, both the numerator and the denominator of the above expression are positive and, after some algebraic steps, one derives that $(1+\omega^{i+1})^3(1+\omega^{i-1}) - (1+\omega^i)^3(1+\omega^{i+2}) = (\omega-1)^3\omega^{i-1}(\omega^{2i+1}-1)$, which is clearly greater than 0 for any $\omega > 1$ and $i \geq 1$. So we have proved the assertion.

As a direct consequence of the two results above, we conclude that the proposed alternative discrete counterpart of the half-logistic distribution is unimodal with mode equal to 1 if $\theta < \theta^*$, with mode equal to 0 if $\theta > \theta^*$; it is bimodal with modes at 0 and 1 if $\theta = \theta^*$. This is a very relevant difference with respect to the model (3). Figure 1 displays, for the integers 0 to 10, the probabilities for the two models when $\theta = 1/2$. Figure 2 displays for different values of θ the probabilities of the proposed distribution, from which one can note the different value(s) of the mode.

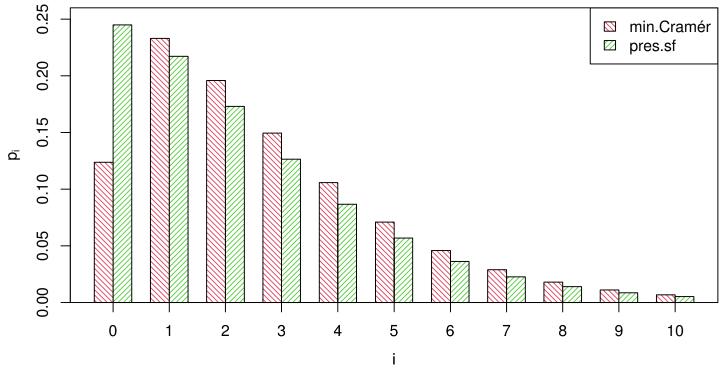


Figure 1. Pmf of the proposed discrete counterpart, based on (7) and derived by minimizing the Cramér distance between cdf, and of the discrete counterpart proposed by Barbiero and Hitaj (2020), (3), based on the preservation of the survival function; $\theta = 1/2$.

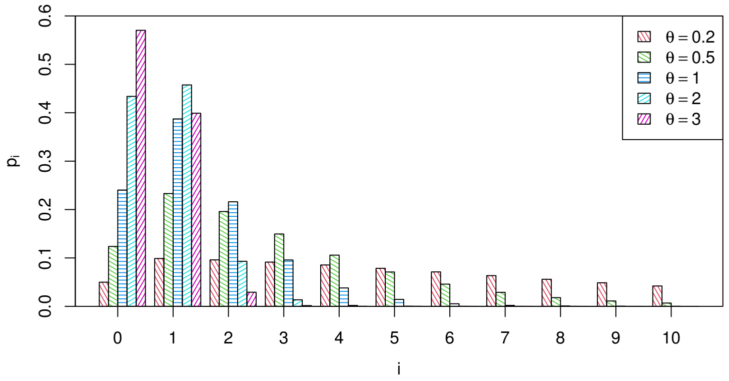


Figure 2. Pmf of the proposed discrete counterpart for different values of θ. Note that when $\theta = 3$ ($> \theta^* = 2.12255$) the distribution has a unique mode at 0; for all the other values of θ (which are $< \theta^*$) the mode is 1.

The quantile of level u, x_u, $0 < u < 1$, can be obtained by computing the generalized inverse of the cumulative probabilities (6). From $u = \frac{2}{\theta} \log \frac{1+e^{\theta(i+1)}}{1+e^{\theta i}}$ we obtain $e^{\theta(1+u)/2} = \frac{1+e^{\theta}e^{\theta i}}{1+e^{\theta i}}$ and $e^{\theta i}(e^{\theta} - e^{\theta(1+u)/2}) = e^{\theta(1+u)/2} - 1$, so that

$$x_u = \left\lceil \frac{1}{\theta} \log \frac{e^{\theta(1+u)/2} - 1}{e^{\theta} - e^{\theta(1+u)/2}} \right\rceil, \tag{8}$$

where $\lceil \cdot \rceil$ denotes the ceiling function.

The naïve failure rate function, which for a discrete rv is defined as $r_i = p_i/P(\tilde{X} \geq$

$i) = p_i/(1-Q_{i-1})$, for the discrete half-logistic distribution is thus equal to

$$r_i = \frac{\log \frac{(1+e^{\theta(i+1)})(1+e^{\theta(i-1)})}{(1+e^{\theta i})^2}}{\left(\theta - \log \frac{1+e^{\theta i}}{1+e^{\theta(i-1)}}\right)},$$

for $i = 1, 2, \ldots$ (with the assumption that $Q_{-1} = 0$ so that $r_0 = p_0$). It is always bounded between 0 and 1. It can be numerically proved that the failure rate function is strictly increasing with i and that it tends asymptotically to the value $r_\infty = \frac{e^\theta - 1}{e^\theta}$.

It can be easily shown that the expectation of the alternative discrete half-logistic coincides with that of the parent continuous distribution. This is a general property holding for the discrete counterparts of positive rvs obtained by minimizing the Cramér distance (4) (Chakraborti, Jardim, & Epprecht, 2019). In fact, denoting the continuous rv and its optimal counterpart by X and $\tilde{X}$, respectively, and recalling an alternative formulation of the expected value for positive rvs, one shows that

$$\mathbb{E}(\tilde{X}) = \sum_{i=0}^{\infty}(1-Q_i) = \sum_{i=0}^{\infty}\left(1 - \int_i^{i+1} F(x)\mathrm{d}x\right) = \int_0^\infty (1-F(x))\mathrm{d}x = \mathbb{E}(X).$$

The second raw moment can be computed as (Chakraborti et al., 2019)

$$\begin{aligned}\mathbb{E}(\tilde{X}^2) &= 2\sum_{i=0}^{\infty} i(1-Q_i) + \mathbb{E}(\tilde{X}) = 2\sum_{i=0}^{\infty} i\left(1 - \frac{2}{\theta}\log\frac{1+e^{\theta(i+1)}}{1+e^{\theta i}} + 1\right) + \log 4/\theta \\ &= 4\sum_{i=0}^{\infty} i\left(1 - \frac{1}{\theta}\log\frac{1+e^{\theta(i+1)}}{1+e^{\theta i}}\right) + \log 4/\theta\end{aligned}$$

and then

$$\mathrm{Var}(\tilde{X}) = 4\sum_{i=0}^{\infty} i\left(1 - \frac{1}{\theta}\log\frac{1+e^{\theta(i+1)}}{1+e^{\theta i}}\right) + \log 4/\theta - (\log 4/\theta)^2.$$

The values of the variance as a function of the parameter θ can be recovered numerically, as well as these of the dispersion index $\mathrm{DI}(\tilde{X}) = \mathrm{Var}(\tilde{X})/\mathbb{E}(\tilde{X})$. Figure 3 graphs the dispersion index as a function of the parameter θ. The function is first decreasing, attaining the value 1 (corresponding to equi-dispersion) at $\theta_1 = 1.14426$ and then reaching its minimum value $\mathrm{DI}_{\min}$ at $\theta_{\min} = 2.970796$; then it is strictly increasing and asymptotically tends to 1. It should be noted that for $\theta \to \infty$ the proposed discrete half-logistic distribution converges to a degenerate rv taking the value 0 with probability 1.

In a similar manner, one can numerically reconstruct the values of the customary indexes of skewness (normalized third central moment: $\mathbb{E}(\tilde{X} - \mathbb{E}(\tilde{X}))^3/\mathrm{Var}^{1.5}(\tilde{X})$), and kurtosis (normalized fourth central moment: $\mathbb{E}(\tilde{X} - \mathbb{E}(\tilde{X}))^4/\mathrm{Var}^2(\tilde{X})$) as functions of θ; the corresponding graphs are depicted in Figure 4. We can notice that the proposed distribution is always positively skewed; it is leptokurtic (kurtosis greater than 3) for values of θ external to the interval $(\theta_a, \theta_b) = (2.966018, 6.352766)$; it is platykurtic for values of θ internal to the same interval. For θ tending to $+\infty$ both skewness and kurtosis diverge to $+\infty$ (but this is not an interesting case; the rv converges

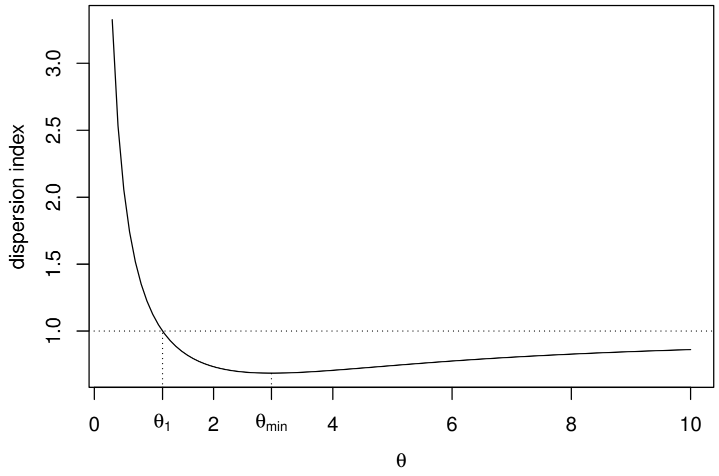
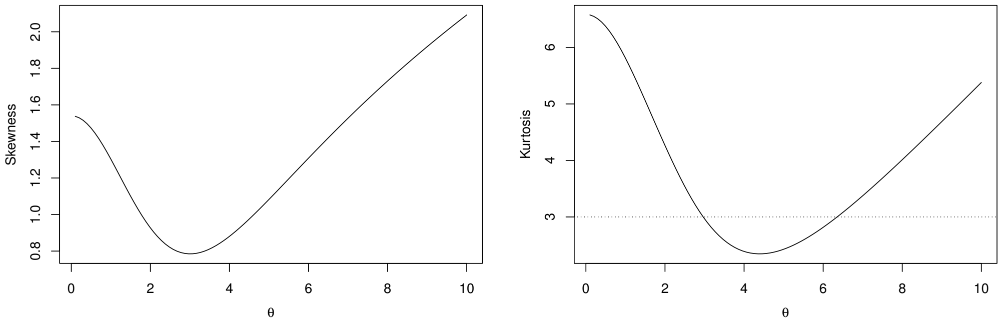


Figure 3. Plot of the Dispersion Index as a function of θ.

to a degenerate rv, as mentioned before). For θ tending to 0^+, the skewness of the distribution tends to about 1.540, whereas the kurtosis tends to about 6.584.



(a) Skewness.

(b) Kurtosis.

Figure 4. Skewness and kurtosis for the proposed discrete half-logistic model.

The Zero-Modification (ZM) index is defined as $ZM = 1 + \log(P(\tilde{X}) = 0)/\mathbb{E}(\tilde{X})$. A positive value of ZM indicates zero-inflation, conversely, a negative value of ZM indicates zero-deflation; a zero value for ZM denotes neither inflation nor deflation for the zero probability. For the proposed model, it takes the expression $ZM = 1 + \frac{\theta}{\log 4} \log \left[\frac{2}{\theta} \log \frac{1+e^{\theta}}{2} - 1\right]$. Figure 5 displays the values of the ZM index as a function of θ. The function, starting from the limiting value 1 for $\theta \to 0^+$ is first decreasing, attaining the value 0 at $\theta_0 = 0.9251821$ and then the minimum value -0.2195279 at $\theta_{\text{min}} = 2.59023$. Then, it becomes strictly increasing and tends asymptotically to zero for $\theta \to \infty$. The shape of the ZM index function is very similar to that of the DI presented in Figure 3.

2.0.1. Infinite divisibility

It can be shown that the proposed parametric distribution is not infinitely divisible. In fact, it is well-known that a necessary condition for a discrete distribution to be infinitely divisible is that $p_1^2 \leq 2p_0p_2$ (Steutel & Van Harn, 2003). However, if we let

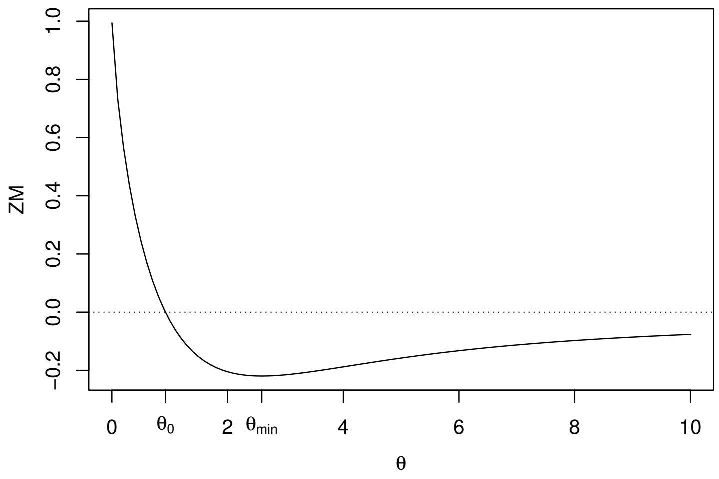


Figure 5. Plot of the Zero-Modified index as a function of θ.

$\theta = 1$, we obtain $p_1 = 0.3871036$, $p_0 = 0.240229$, and $p_2 = 0.215986$ and the above inequality is not satisfied; this implies that the discrete half-logistic is not an infinitely divisible parametric family.

3. Parameter estimation

Given an iid sample $(x_1, x_2, \ldots, x_n)$, which we assume coming from the alternative discrete half-logistic distribution (7), the unknown parameter θ can be estimated by resorting to one of the following methods.

3.1. Maximum likelihood method

The maximum likelihood estimate $\hat{\theta}_{ML}$ of θ is the value maximizing the log-likelihood function $\ell(\theta; x_1, \ldots, x_n) = \sum_{j=1}^{n} \log p_{x_j}(\theta)$. The expression of the log-likelihood function is

$$\ell(\theta; x_1, \ldots, x_n) = n_0 \log\left(\frac{2}{\theta}\log\frac{1+e^{\theta}}{2} - 1\right) + \sum_{i=1}^{x_{(n)}} n_i \log\left[\frac{2}{\theta}\log\frac{(1+e^{\theta(i+1)})(1+e^{\theta(i-1)})}{(1+e^{\theta i})^2}\right]$$

where n_i, $i = 0, 1, \ldots, x_{(n)}$, is the sample absolute frequency of the value i; $x_{(n)} = \max\{x_j; j = 1, 2, \ldots, n\}$. Due to the complicated expression of ℓ, it is not possible to derive a closed-form expression of $\hat{\theta}_{ML}$, but any standard optimization routine, such Newton-Raphson based methods, can be used in order to recover it numerically.

In order to obtain interval estimates for θ, one can consider constructing $100 \times (1-\alpha)\%$ log-likelihood based confidence intervals (CIs) (see, e.g., Bolker & R Development Core Team, 2022) or, more easily, use large-sample approximations for the $100 \times (1-\alpha)$ two-sided symmetric CIs: $\hat{\theta} \mp z_{1-\alpha/2}\sqrt{\hat{J}_n(\hat{\theta}_{ML})^{-1/2}}$, where $\hat{J}_n(\hat{\theta}_{ML}) = -\left[\frac{\mathrm{d}^2\ell(\theta; x_1, \ldots, x_n)}{\mathrm{d}\theta^2}\right]_{\theta=\hat{\theta}_{ML}}$ is the Fisher observed information computed at the MLE of θ.

3.2. Moment method

We have already seen that the expectation of the proposed discrete model is $\mathbb{E}(\tilde{X}) = \log 4/\theta$. Then, by equating this expectation to the sample mean $\bar{x} = \sum_{i=1}^{n} x_i/n$, we derive the moment estimate as $\hat{\theta}_M = \log 4/\bar{x}$.

3.3. Method of proportion

The method of proportion, suitable for discrete distributions, was probably first conceived in Khan, Khalique, and Abouammoh (1989), where it was applied to the type I discrete Weibull distribution. We have already seen that the mode of the proposed discrete model is 1 for any $\theta < \theta^* = 2.12255$ and its probability, according to (7), is $p_1 = \frac{2}{\theta} \log \frac{2(1+e^{2\theta})}{(1+e^{\theta})^2}$. By equating this probability to the sample relative frequency of 1s, which we denote by $\hat{p}_1$, we obtain the following non-linear equation in θ: $2(1 + e^{2\theta}) = e^{\theta \hat{p}_1/2}(1 + e^{\theta})^2$, which, by setting $\omega = e^{\theta}$ $(\omega > 1)$, becomes $\omega^{2+\hat{p}_1/2} - 2\omega^2 + 2\omega^{1+\hat{p}_1/2} + \omega^{\hat{p}_1/2} - 2 = 0$, which can be solved numerically, and possibly provides a (unique?) feasible root $\omega_P^{(1)}$ and a corresponding estimate $\hat{\theta}_P^{(1)} = \log \omega_P^{(1)}$. If we consider the function $g_1(\omega) = \omega^{2+\hat{p}_1/2} - 2\omega^2 + 2\omega^{1+\hat{p}_1/2} + \omega^{\hat{p}_1/2} - 2$, we have that $g_1(1) = 0$ and $\lim_{\omega\to\infty} g_1(\omega) = \infty$. Moreover, $g_1'(\omega) = (2 + \hat{p}_1/2)\omega^{1+\hat{p}_1/2} - 4\omega + 2(1 + \hat{p}_1/2)\omega^{\hat{p}_1/2} + \hat{p}_1/2\omega^{\hat{p}_1/2-1}$, so that $g_1'(1) = 2\hat{p}_1 > 0$ (g is strictly increasing at 1). Now, g_1, for ω greater than 1, can have no real roots (in this case, g_1 is strictly increasing over the whole interval $(1, +\infty)$), can have two real roots, or (this is a limit case) a unique root, which occurs for $\hat{p}_1 \approx 0.4602$: in this case the unique solution is $\omega = 1.803$ $(\theta = 0.5895)$. The former two situations are depicted in Figure 6, which correspond to the cases $\hat{p}_1 = 0.5$ and $\hat{p}_1 = 0.45$. In the first case, there are no feasible real roots ω larger than 1, and thus the method of proportion fails in recovering a valid estimate for θ. In the second case, there are two roots for ω both larger than 1, leading to two possible valid estimates for θ, namely 1.469882 and 2.203663.

Alternatively, one can consider matching the probability of 0 and the corresponding sample relative frequency, $\hat{p}_0$. After simple algebraic steps, one obtains the following equation in ω, $2\omega^{(1+\hat{p}_0)/2} - \omega - 1 = 0$, which yields a root $\omega_P^{(0)}$ and the corresponding estimate $\hat{\theta}_P^{(0)} = \log \omega_P^{(0)}$. Does the solution $\omega_P^{(0)}$ (and then $\hat{\theta}_P^{(0)}$) always exist? Being $0 \le \hat{p}_0 \le 1$, we can state that the function $g_0(\omega) = 2\omega^{(1+\hat{p}_0)/2} - \omega - 1$ is continuous, it is equal to 0 at 1 for any feasible value of $\hat{p}_0$, it tends to $-\infty$ when ω tends to ∞ and its derivative $g_0'(\omega) = (1 + \hat{p}_0)\omega^{(\hat{p}_0-1)/2} - 1$ is positive for $1 < \omega < \omega^*$ and is equal to 0 at $\omega^* = \left(\frac{1}{\hat{p}_0+1}\right)^{\frac{2}{\hat{p}_0-1}} > 1$, with $g_0(\omega^*) > 0$. Thus ω^* is an absolute maximum point and the function takes the value 0 at some point $\omega_P > \omega^*$; ω_P and θ_P are unique. Due to this result, it is preferable to base the method of proportion on the matching of the zero probability with the sample relative frequency of zeros.

3.4. Monte Carlo simulation study

In this section, we have estimated, using $B = 50000$ Monte Carlo simulations, the average bias and the root mean squared error of the three estimators (ML, MM, MP - the latter based on the sample proportion of zeros) of the parameter θ of the discrete half-logistic distribution; as well as the coverage probability and the coverage length of ML-based confidence intervals. To run the simulation plan, we have considered $\theta = 0.05; 0.1; 0.2; 0.5; 1; 2$ and sample sizes $n = 25; 50; 100$. The inverse-transform method

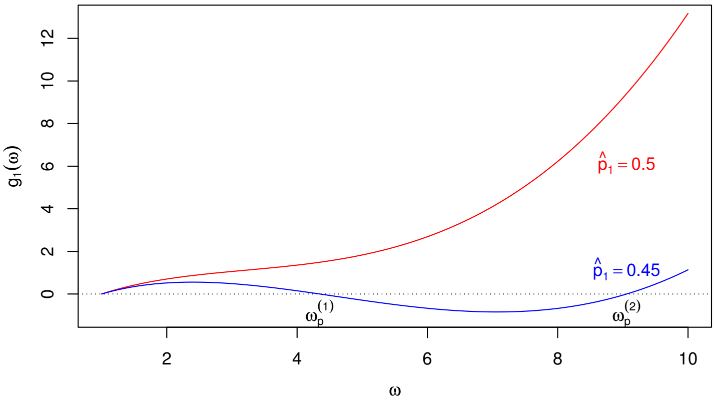


Figure 6. Graph of the function $g_1(\omega)$ involved in the method of proportion based on the matching of probability and sample relative frequency of 1. Two values of $\hat{p}_1$ are considered.

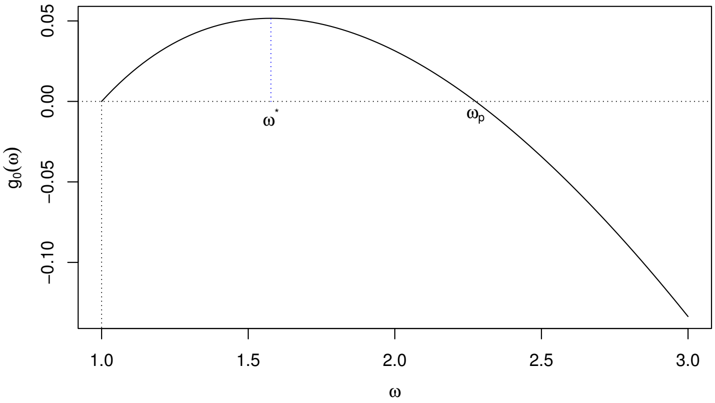


Figure 7. Example of graph of the function $g_0(\omega)$ involved in the method of proportion based on the matching of probability and sample relative frequency of 0.

for discrete distributions (Rubinstein & Kroese, 2016) was implemented to generate the pseudo-random samples, based on the expression for the quantile function given in (8). The quantities of interest were estimated by the following expression:

$$\begin{aligned}
\text{bias}(\hat{\theta}) &= \frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_i - \theta)\\
\text{rmse}(\hat{\theta}) &= \sqrt{\frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_i - \theta)^2}\\
\text{CL}_\theta(n) &= \frac{1}{B}\sum_{i=1}^{B}\theta_{U,i} - \theta_{L,i}\\
\text{CP}_\theta(n) &= \frac{1}{B}\sum_{i=1}^{B}\mathbb{1}\{\theta_{L,i} \le \theta \le \theta_{U,i}\}
\end{aligned}$$

where $\hat{\theta}_i$ is the estimate computed on the i-th sample; $\theta_{U,i}$ and $\theta_{L,i}$ are the upper and lower bounds, respectively, of the 95% log-likelihood-based confidence interval for θ built upon the i-th sample, and $\mathbb{1}(\cdot)$ denotes the indicator function.

The values of the average bias and average root mean squared error along with those of the coverage probabilities and the coverage lengths are reported in Table 1.

Focusing on the ML, we note that its average bias is always positive. For a fixed value of θ, it decreases with n and for a fixed n it increases with θ (actually, also the relative bias increases with θ). The rmse is, as expected, a decreasing function of n, for an assigned value of θ. For a fixed n, the rmse increases with θ (the relative rmse is just slightly increasing with θ).

As for the MM, we easily note that, if compared to the ML, for any scenario it is characterized by a positive but (slightly) smaller bias but also shows a (slightly) larger rmse.

The MP estimator shows much larger values of rmse than the other two estimators, especially when θ is smaller: in those cases, in fact, the number of zeros contained in the sample is closer to zero and then the MP turns out to be inefficient, since it exploits less sample information. For large values of θ (when the sample typically contains a significant quote of zeros), the bias of the MP estimator can be smaller than that of the two competing estimators, but the rmse is still larger.

As for the coverage probabilities, $\text{CP}_\theta(n)$ is always very close to the nominal 95%, even for the smallest sample size here examined ($n = 25$); some discrepancies (with either over or under-coverage) can be detected for $\theta = 2$. The average length $\text{CL}_\theta(n)$, as expected, decreases with n for a fixed value of θ; it increases with θ for a fixed value of n.

4. Regression

The fact that the expectation of the discrete half-logistic distribution is given by $\mu = \log 4/\theta$ prompts its use for count regression models. A reparameterization is

Table 1. Monte Carlo simulation results

n	θ	ML	MM	MP	ML	MM	MP	log-likelihood CIs	
			bias			rmse		CP	CL
25	0.05	0.0015	0.0014	0.1355	0.0091	0.0091	0.1500	0.9495	0.0338
	0.1	0.0031	0.0029	0.1137	0.0182	0.0183	0.1494	0.9491	0.0677
	0.2	0.0061	0.0058	0.0794	0.0364	0.0367	0.1670	0.9493	0.1356
	0.5	0.0159	0.0150	0.0247	0.0925	0.0933	0.2652	0.9489	0.3434
	1	0.0358	0.0333	0.0222	0.1949	0.1981	0.4001	0.9495	0.7195
	2	0.1055	0.0915	0.0744	0.4763	0.4875	0.6689	0.9447	1.7375
50	0.05	0.0007	0.0007	0.0570	0.0061	0.0062	0.0746	0.9498	0.0235
	0.1	0.0015	0.0014	0.0398	0.0123	0.0124	0.0839	0.9499	0.0471
	0.2	0.0029	0.0028	0.0174	0.0246	0.0248	0.1156	0.9503	0.0943
	0.5	0.0075	0.0071	0.0032	0.0623	0.0630	0.1926	0.9492	0.2384
	1	0.0170	0.0160	0.0103	0.1298	0.1327	0.2789	0.9507	0.4956
	2	0.0483	0.0427	0.0350	0.3024	0.3145	0.4477	0.9486	1.1445
100	0.05	0.0004	0.0003	0.0198	0.0043	0.0043	0.0421	0.9497	0.0165
	0.1	0.0007	0.0007	0.0092	0.0086	0.0086	0.0586	0.9500	0.0330
	0.2	0.0015	0.0014	0.0015	0.0171	0.0173	0.0865	0.9500	0.0662
	0.5	0.0038	0.0036	0.0016	0.0433	0.0439	0.1368	0.9498	0.1671
	1	0.0085	0.0081	0.0056	0.0899	0.0922	0.1958	0.9500	0.3463
	2	0.0235	0.0210	0.0171	0.2049	0.2144	0.3093	0.9507	0.7869

suggested, according to which the pmf can be defined as

$$\begin{cases} p_0 = Q_0 = 2\sigma \log \dfrac{1+e^{1/\sigma}}{2} - 1 \\ p_i = Q_i - Q_{i-1} = 2\sigma \log \dfrac{(1+e^{(i+1)/\sigma})(1+e^{(i-1)/\sigma})}{(1+e^{i/\sigma})^2}, \end{cases} \tag{9}$$

where the new parameter $\sigma > 0$ corresponds to $1/\theta$, and then $\mu = \sigma \log 4$. Then σ, by using a logarithmic link function, can be modelled as

$$\log \sigma = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

or, equivalently,

$$\sigma = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k), \tag{10}$$

where $x_1, \dots, x_k$ are the k regressors with coefficients $\beta_1, \dots, \beta_k$ and β_0 is the intercept. We recall that the appropriate choice of the link function is based on the domain of the parameter: here, since σ can take on only positive values, the log-link function is suitable for this regression model.

Using the re-parameterized pmf of the discrete half-logistic, given in (9), the log-likelihood function of the discrete half-logistic regression model, based on a sample $\boldsymbol{y} = (y_1, \dots, y_n)$, is

$$\ell(\sigma) = \sum_{i=1}^{n} \log p_{y_i}(\sigma_i)$$

with $\sigma_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$, being $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)^T$ the vector of parameters and $\boldsymbol{x}_i = (1, x_{i1}, \dots, x_{ik})^T$ the vector of covariates for the i-th observation. The parameter vector $\boldsymbol{\beta}$ can be estimated by maximizing $\ell(\sigma)$ with respect to $\boldsymbol{\beta}$, a task which can be performed numerically resorting to some appropriate optimization routine (Bolker & R Development Core Team, 2022).

After fitting a count regression model, it is essential to consider a diagnostics analysis to investigate its appropriateness (see, e.g., Feng, Li, & Sadeghpour, 2020, for a review of diagnostic tools for count regression). Given that the response is discrete, it is advised performing a residual analysis on the basis of the randomised quantile residuals, as developed by Dunn and Smyth (1996) and used in many other studies (e.g., Klakattawi, Vinciotti, and Yu 2018). In particular, we let $e_i = \Phi^{-1}(u_i)$, $i = 1, \dots, n$, where $\Phi(\cdot)$ is the quantile function of the standard normal distribution and u_i is a uniform random variable on the interval $(a_i, b_i] = (\lim_{y \to y_i^-} F(y; \hat{\sigma}_i), F(y_i; \hat{\sigma}_i)] \approx [F(y_i - 1; \hat{\sigma}_i), F(y_i; \hat{\sigma}_i)]$. These residuals follow the standard normal distribution, apart from the sampling variability in $\hat{\sigma}_i$. Hence, the validity of a discrete half-logistic regression model can be assessed using goodness-of-fit investigations of the normality of the residuals, such as Q-Q plots and normality tests.

4.1. A simulation study

We illustrate the count regression model above by a simple simulation study. We considered two independent covariates, $X_1 \sim N(0, 1)$ and $X_2 \sim \text{Uniform}(0, 10)$. We assumed the regression parameters to take values $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2) = (0.5, 0.75, -0.25)$. Then, we sampled n values (x_{1i}, x_{2i}), $i = 1, \dots, n$, of the covariates and drew the corresponding responses y_i from a discrete half-logistic distribution whose parameter σ_i is calculated as in Equation (10).

Estimation of β_0, β_1 and β_2 is carried out by resorting to ML estimation, implemented by the function `mle2` within the `R` package `bbmle`, using the Nelder-Mead optimization routine. Table 2 reports the estimates of the parameters, averaged over $1,000$ Monte Carlo runs, together with the average bias and the root mean squared error (rmse); the average length of the 95% log-likelihood based confidence intervals for σ is also reported.

When the results given in Table 2 are examined, one concludes that the estimated biases are near the zero for all sample sizes and decreasing with n in absolute value at least for β_0 (for the other two parameters, it doesn't occur, but this is plausibly due to the sampling error); the rmse's decrease with n, as one should have expected; analogously for the average lengths of the confidence intervals. These results empirically show that the MLEs of the parameters of the discrete half-logistic regression model are asymptotically unbiased and consistent.

4.2. Data set 1: badhealth

We consider a dataset coming from the German Health Survey and available in the `COUNT` package (Hilbe, 2016), with the name of `badhealth`, under the `R` programming language (R Core Team, 2023). The response variable is the number of visits to doctors during 1998. Two predictors are considered: an indicator variable representing patients claiming to be in bad health (1) or not (0), and the age of the patient. The response variable ranges from 0 to 40 visits and has a sample mean of 2.3532 and variance of 11.9818, suggesting overdispersion relative to Poisson regression.

The summary results of the discrete half-logistic regression (reported in Table 3) if compared to the findings in Klakattawi et al. (2018), related to Poisson, discrete Weibull and negative binomial regressions, highlight how the proposed regression

Table 2. Monte Carlo simulation results for a discrete half-logistic regression model

n	parameter	mle	bias	rmse	length
25	β_0	0.4329	-0.0671	0.4517	1.7364
	β_1	0.7668	0.0168	0.3034	1.111
	β_2	-0.2590	-0.0090	0.0981	0.3690
50	β_0	0.4832	-0.0168	0.2843	1.1282
	β_1	0.7483	-0.0017	0.1842	0.7155
	β_2	-0.2558	-0.0058	0.0616	0.2437
100	β_0	0.4849	-0.0151	0.1964	0.7748
	β_1	0.7468	-0.0032	0.1190	0.4828
	β_2	-0.2502	-0.0002	0.0411	0.1649
200	β_0	0.4962	-0.0038	0.1422	0.5372
	β_1	0.7472	-0.0028	0.0910	0.3341
	β_2	-0.2509	-0.0009	0.0306	0.1144

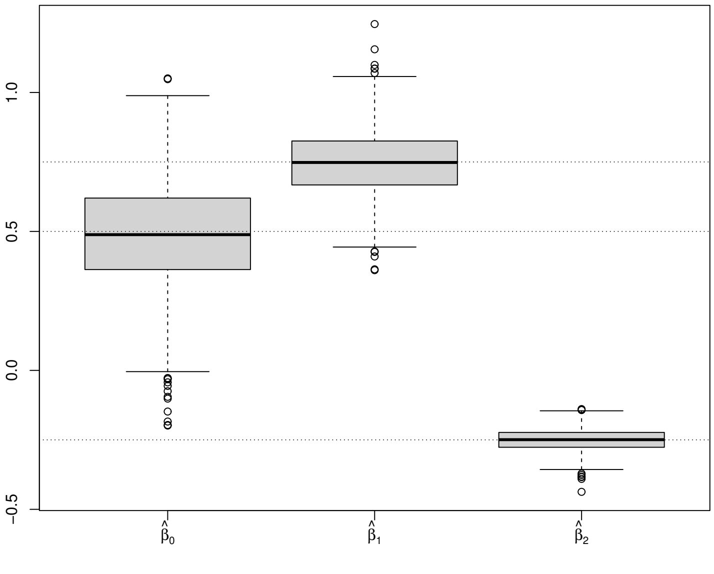


Figure 8. Monte Carlo distribution of the MLEs of the coefficients for the simulated regression model with the response variable following the discrete half-logistic distribution; $n = 100$.

Table 3. Summary results from the regression model fitted to the the bad health data ($n = 1127$, $p = 3$)

Parameter	Estimate (s.e.)
intercept (β_0)	0.1002 (0.0920)
sex (β_1)	1.0576 (0.0864)
age (β_1)	0.0083 (0.0024)
Deviance $= -2\ell_{\max}$	4727.62
AIC $= 2p - 2\ell_{\max}$	4733.62
BIC $= p \ln n - 2\ell_{\max}$	4748.70

Table 4. Summary statistics for the variables of strikes data set ($n = 108$)

Variables	Min	Median	Max
strikes (response)	0	5	18
output (explanatory)	−0.140	0.000	0.086

model, though far superior to Poisson regression, provides worse results than the latter two, which are, however, two-parameter distributions and thus more flexible. The AIC statistic for the discrete-half logistic model equals 4733.62, whereas for the Poisson regression model it is 5638.552, for the Negative Binomial regression model 4475.285 and for the discrete Weibull regression model 4474.973.

4.3. Data set 2: StrikeNb

The data set is available in the `Ecdat` package under the name `StrikeNb` and is originally reported in Croissant and Graves (2022). The data set `StrikeNb` reports the contract strikes in the U.S. manufacturing industries from 1968 to 1976 ($n = 108$). The number of strikes (`strikes`) and lockouts by economic activity (`output`) are defined as the response and explanatory variable, respectively. Table 4 shows the descriptive summary of these variables. The percentage of zeros in observed response variable is 4.63. Also, this data set indicates an over-dispersion problem with the index of dispersion 2.685.

The regression model which relates the number of strikes to the unique covariate is based on the equation $\sigma_i = \exp(\beta_0 + \beta_1 \texttt{output})$. We show the estimated parameters and summarize the fitted model in Table 5. We note that fitting the response values without taking into account of the covariate, would lead to an estimate of σ equal to $\hat{\sigma} = 3.720414$, which, by the way, is very close to the mean of the predicted $\hat{\sigma}_i$ by the regression model ($\bar{\hat{\sigma}}_i = 3.715634$); the deviance is 569.5 for the null model and 564.4545 for the model with the covariate, whose *AIC* is 568.4545. The Poisson regression model provides a deviance equal to 626.565 and an *AIC* equal to 630.565; the negative binomial model a deviance equal to 563.682 and an *AIC* equal to 569.682 (see, for example, Jornsatian & Bodhisuwan, 2022).

Figure 9 displays the observed and theoretical frequencies for the strikes' counts. Figure 10 displays the QQ-plot of the randomised quantile residuals versus the standard normal quantiles, from which one can conclude that the distribution of the randomised residuals matches the normal distribution.

Table 5. Summary results from the regression model fitted to the the strikes data

Parameter	Mean (s.e.)	95% CI
intercept (β_0)	1.308(0.079)	$(1.157, 1.469)$
output (β_1)	3.430(1.503)	$(0.446, 6.349)$
Deviance $= -2\ell_{\max}$	564.45	
AIC $= 2p - 2\ell_{\max}$	568.45	
BIC $= p \ln n - 2\ell_{\max}$	573.82	

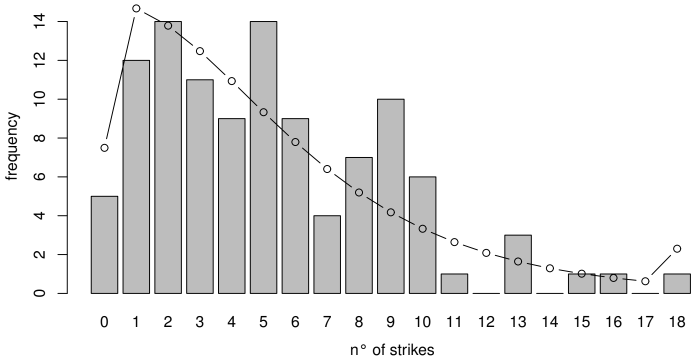


Figure 9. Observed and fitted frequencies for the strikes data set under the regression model with the discrete half-logistic distribution.

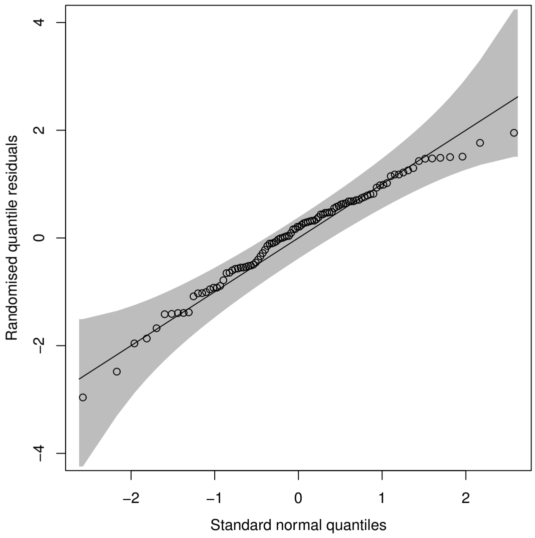


Figure 10. QQ-plot of the randomised quantile residuals versus the standard normal quantiles (strikes data set).

Table 6. Distribution of number of outbreaks of strikes, from Ridout and Besbeas (2004)

count	observed frequency	theoretical frequency
0	46	51.39
1	76	69.69
2	24	25.61
3	9	7.01
≥ 4	1	2.30
total	156	156

5. Real data illustration

We consider the dataset presented in Ridout and Besbeas (2004) and reported here in Table 6. For these data, the sample mean is $\bar{x} = 0.9936$ and the sample variance $s_X^2 = 0.7419$ (so the data result under-dispersed). Fitting the data through the alternative half-logistic might be plausible, since there is a mode at 1. The MLE of θ results equal to 1.424 (which, by the way, is a parameter value inducing under-dispersion, recall Figure 3) and using this estimate we reconstructed the theoretical frequencies, which are displayed in the last column of Table 6. Pooling the last two counts (3 and 4), we calculated the usual chi-squared statistics, $X^2 = \sum_{i=0}^{3}(n_i - \hat{n}_i)^2/\hat{n}_i$, where n_i and $\hat{n}_i$ are the observed and theoretical frequencies of the count i; its value is 1.2896. Under the null hypothesis that the data come from the proposed distribution, the X^2 statistic asymptotically tends to be distributed as a chi-squared with 2 degrees of freedom; the approximate p-value of the chi-squared test is therefore 0.5247 and indicates a more than satisfactory fit of the model. The maximum value of the log-likelihood function is -188.104; the AIC value is 378.208. All these results, if compared to those of the statistical models analyzed in Chakraborty and Gupta (2015), highlight that the proposed alternative discrete half-logistic distribution has a superior goodness-of-fit.

For the sake of completeness, let us also consider the two other estimation methods discussed in the third section. The moment method provides $\hat{\theta}_M = 1.3952$. The method of proportion, if based on the the matching between sample frequency and probability of 1, is not able to provide a feasible estimate of θ; in fact, the value $\hat{p}_1 = n_1/n = 0.4872$ cannot be attained by p_1, which is bounded from above by 0.4602. If we base the method of proportion on the matching of frequency and probability of 0, being $\hat{p}_0 = 0.2949$, we obtain $\hat{\theta}_P^{(0)} = 1.2539$, which is slightly smaller than the estimates derived through the moment and maximum likelihood methods.

6. Final remarks

We introduced and discussed the main properties and inferential issues of a discrete analogue of the continuous half-logistic distribution, focusing in particular on the shape of its pmf, on its moments, and on sample estimation, also suggesting a count regression model where the response variable follows this new distribution. The discussion on these theoretical features is supported by applications on real data sets. A possible extension or generalization of this discrete distribution can be conceived, in order to make it more flexible and apt to catch the features present in real data, by introducing an additional parameter. The cdf of this extension can be then defined by resorting to the exponentiation of the one-parameter cdf (Lee, Famoye, & Alzaatreh, 2013), letting the cumulative probabilities be equal to

$$Q_i = \left[\frac{2}{\theta}\log\frac{(1+e^{\theta(i+1)})}{(1+e^{\theta i})} - 1\right]^{\alpha}, \tag{11}$$

with $\alpha > 0$ being the additional parameter. Introducing α allows the pmf corresponding to (11) to exhibit different shapes and in particular allows the mode to take on values different from 0 and 1 as occurs with the simple discrete half-logistic distribution; see Figure 11, where all the 12 combinations (θ, α) originated from $\theta \in \{1, 1/2, 1/4\}$ and $\alpha \in \{1/2, 1, 2, 3\}$ are considered. A deeper analysis of the count distribution in (11) can be the object of future research.

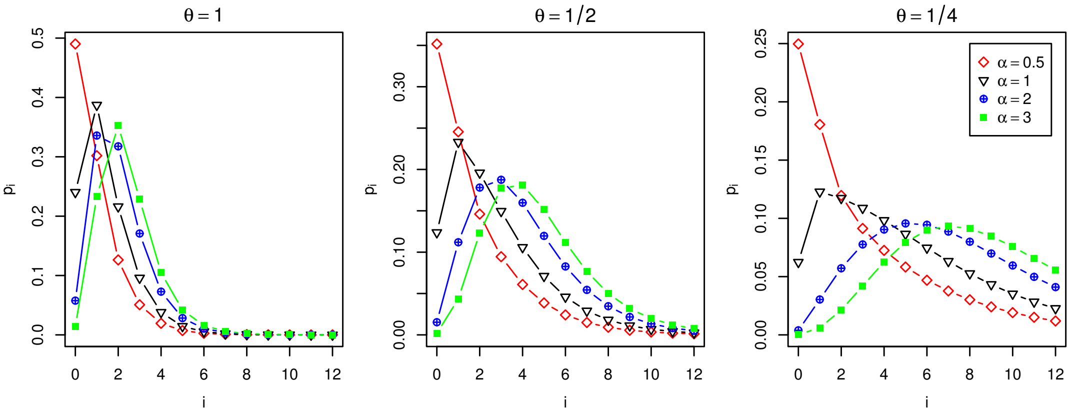


Figure 11. Pmf of a possible generalized discrete half-logistic rv, based on Eq. (11).

Acknowledgments

The authors are thankful to the two anonymous referees for their comments, which helped to improve the manuscript's quality.

The corresponding author acknowledges "PSR Linea 2" 2021 funding from the Università degli Studi di Milano.

References

Balakrishnan, N. (1985). Order statistics from the half logistic distribution. *Journal of Statistical Computation and Simulation*, *20*(4), 287–309.

Barbiero, A., & Hitaj, A. (2020). A discrete analogue of the half-logistic distribution. In *2020 International Conference on Decision Aid Sciences and Application (DASA)* (pp. 64–67).

Barbiero, A., & Hitaj, A. (2021). A new method for building a discrete analogue to a continuous random variable based on minimization of a distance between distribution functions. In *2021 International Conference on Data Analytics for Business and Industry (ICDABI)* (pp. 338–341).

Barbiero, A., & Hitaj, A. (2023). An alternative discrete analogue of the half-logistic distribution. *Proceedings of International Mathematical Sciences*, *5*(2), 14–18.

Bolker, B., & R Development Core Team. (2022). bbmle: Tools for general maximum likelihood estimation [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=bbmle` (R package version 1.0.25)

Chakraborti, S., Jardim, F., & Epprecht, E. (2019). Higher-order moments using the survival function: The alternative expectation formula. *The American Statistician*, *73*(2), 191-194.

Chakraborty, S., & Gupta, R. D. (2015). Exponentiated geometric distribution: another generalization of geometric distribution. *Communications in Statistics-Theory and Methods*, *44*(6), 1143–1157.

Croissant, Y., & Graves, S. (2022). Ecdat: Data sets for econometrics [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=Ecdat` (R package version 0.4-2)

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236–244.

Feng, C., Li, L., & Sadeghpour, A. (2020). A comparison of residual diagnosis tools for diagnosing regression models for count data. *BMC Medical Research Methodology*, *20*(1), 1–21.

Hilbe, J. M. (2016). Count: Functions, data and code for count data [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=COUNT` (R package version 1.3.4)

Jornsatian, C., & Bodhisuwan, W. (2022). Bayesian inference for negative binomial—beta exponential distribution and its regression model. *Lobachevskii Journal of Mathematics*, *43*(9), 2501–2514.

Khan, M. A., Khalique, A., & Abouammoh, A. (1989). On estimating parameters in a discrete Weibull distribution. *IEEE Transactions on Reliability*, *38*(3), 348–350.

Klakattawi, H. S., Vinciotti, V., & Yu, K. (2018). A simple and adaptive dispersion regression model for count data. *Entropy*, *20*(2), 142.

Lee, C., Famoye, F., & Alzaatreh, A. Y. (2013). Methods for generating families of univariate continuous distributions in the recent decades. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(3), 219–238.

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ridout, M. S., & Besbeas, P. (2004). An empirical model for underdispersed count data. *Statistical Modelling*, *4*(1), 77–89.

Rubinstein, R. Y., & Kroese, D. P. (2016). *Simulation and the Monte Carlo method.* John Wiley & Sons.

Steutel, F. W., & Van Harn, K. (2003). *Infinite divisibility of probability distributions on the real line.* CRC Press.