SORT 44 (2) July-December 2020, 265-284 DOI: 10.2436/20.8080.02.102

Discrete generalized half-normal distribution and its applications in quantile regression

Diego I. Gallardo[1], Emilio Gómez-Déniz[2] and Héctor W. Gómez[3]

Abstract

A new discrete two-parameter distribution is introduced by discretizing a generalized half-normal distribution. The model is useful for fitting overdispersed as well as underdispersed data. The failure function can be decreasing, bathtub shaped or increasing. A reparameterization of the distribution is introduced for use in a regression model based on the median. The behaviour of the maximum likelihood estimates is studied numerically, showing good performance in finite samples. Three real data set applications reveal that the new model can provide a better explanation than some other competitors.

MSC: 62E10, 62F10, 62P05.

Keywords: Discretizing, generalized half-normal distribution, failure function, health, quantile regression, stochastic order.

1 Introduction

Kemp (2008) introduced a discrete version of the half-normal distribution which, by analogy with the continuous half-normal distribution, is the maximum entropy distribution with specified mean and variance and support on $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Another way of introducing a discrete version of a continuous model is by discretizing it as follows: if $S_Y(x)$ denotes the survival function of a continuous random variable Y with domain in the positive line, the probability mass function (PMF) of its analogue discrete random variable, X, is given by

$$P(X = k) = p_k = S_Y(k) - S_Y(k+1), \quad k \in \mathbb{N}_0. \tag{1}$$

[1] Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó, Chile, e-mail: diego.gallardo@uda.cl

[2] Department of Quantitative Methods in Economics and TiDES Institute, University of Las Palmas de Gran Canaria, e-mail: emilio.gomez-deniz@ulpgc.es

[3] Departamento de Matemáticas, Facultad de Ciencias Básicas, Universidad de Antofagasta, Antofagasta, Chile, e-mail: hector.gomez@uantof.cl

Received: November 2018
Accepted: June 2020

A classical example is geometric distribution, which can be derived by applying the above discretizing procedure to the negative exponential distribution. Other examples can be found in Nakagawa and Osaki (1975), which obtained the discrete Weibull distribution, Krishna and Singh (2009), the discrete Burr distribution, Gómez-Déniz and Calderín (2011), the discrete Lindley distribution, among many others. This method was also applied by Gómez-Déniz, Vázquez-Polo and García-García (2014) to obtain a discrete version for a generalization of the half-normal distribution based on a skew version of the normal distribution. The resulting discrete distribution differs from that studied in Kemp (2008). The reader can consult the work of Chakraborty (2015) in which different methods and classification are exposed in the discretization procedure of a continuous random variable.

The generalization of the half-normal distribution used in Gómez-Déniz et al. (2014) is based on the idea in Marshal and Olkin (1997). Other generalizations of the half-normal distribution have been proposed in the statistical literature. Here we consider the one in Cooray and Ananda (2008), whose derivation follows from considerations of the relationship between static fatigue crack extension and the failure time of a certain specimen. Its survival function is given by

$$S_Y(x;\sigma,\beta) = 2\Phi\left(-\left(\frac{x}{\sigma}\right)^{\beta}\right), \quad x \geq 0, \tag{2}$$

for some $\sigma,\beta > 0$, where $\Phi(\cdot)$ stands for the cumulative distribution function (CDF) of the standard normal distribution. If a positive random variable Y has survival function (2) we will say that it has a generalized half-normal (GHN) distribution and it will be denoted as $Y \sim GHN(\sigma,\beta)$. The associated discrete version X obtained by applying (1), which will be called the discrete generalized half-normal (DGHN) distribution, has PMF

$$P(X=k;\sigma,\beta) = p(k;\sigma,\beta) = 2\left\{\Phi_\psi\left((k+1)^\beta\right) - \Phi_\psi\left(k^\beta\right)\right\}, \quad x \in \mathbb{N}_0 \tag{3}$$

for some $\sigma,\beta > 0$, where $\psi = \sigma^\beta$ and $\Phi_\sigma(x) = \Phi(x/\sigma)$. If a random variable X taking values on $\mathbb{N}_0$ has PMF (3), we write $X \sim DGHN(\sigma,\beta)$. The new model is different from the one studied in Kemp (2008); for $\beta = 1$ it coincides with that introduced in Gómez-Déniz et al. (2014); for other parameter values, the resulting models are rather different. Figure 1 displays the PMF of X for several parameter. Looking at this figure we see that quite different shapes can be obtained by varying the parameter values.

The discretization of a continuous variable in order to obtain a discrete distribution has been developed with great enthusiasm in recent decades. The simple idea is to start from a continuous random variable that follows a certain probability distribution and for which the distribution function (survival) has a closed form expression. Except for a few occasions (the discretization of the exponential distribution that gives rise to the geometric discrete distribution and the discretization of the Lindley distribution (Gómez-Déniz and Calderín, 2011), the mean and any other superior moment are not obtained in a closed manner.

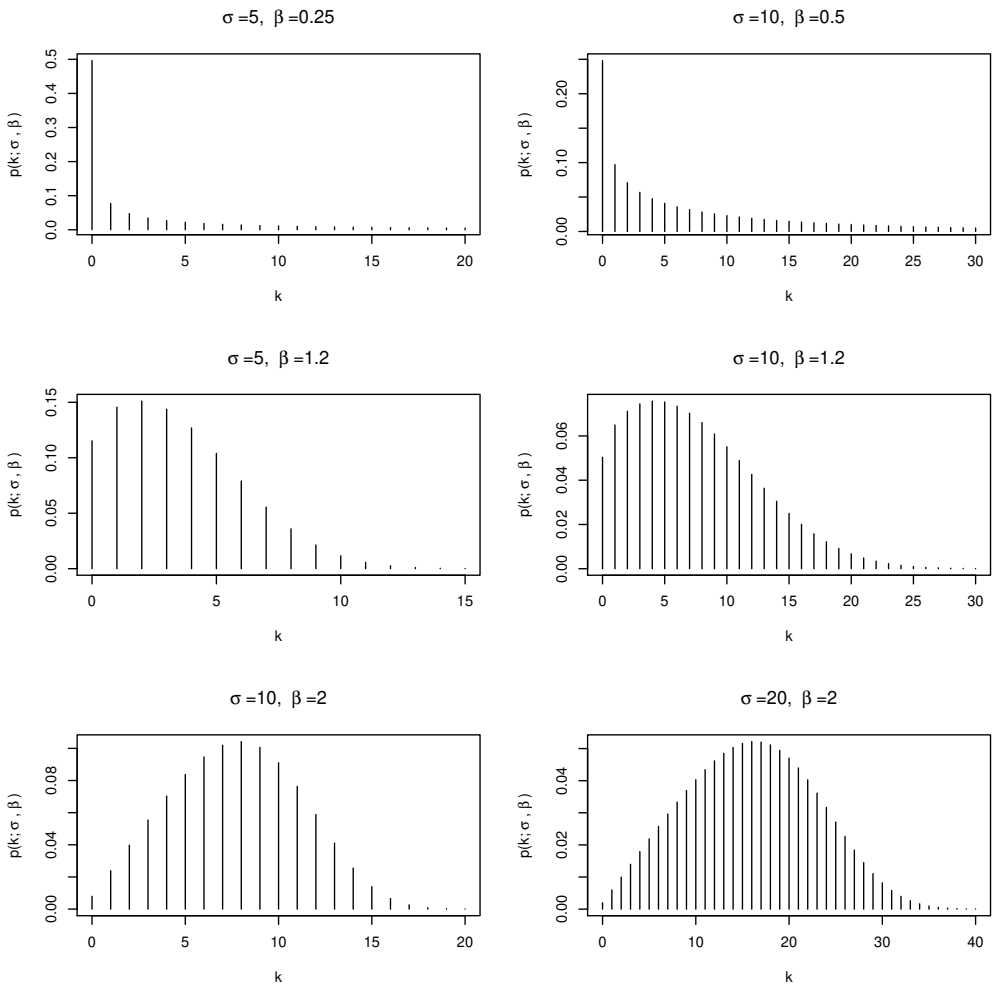

Figure 1: *Some examples of probability mass functions of the DGHN distribution for different values of the parameters β and σ.*

This is a great disadvantage for a researcher who wishes to carry out more in-depth studies on the variable that he wishes to study. For example, a regression study, i.e. explaining the effect that a series of factors can have on the dependent variable, is impossible to perform by ordinary methods.

However, the fact that the distribution function has a closed form makes it easier to calculate the quantile function and therefore to obtain the median. In this case, the initial probability function can be reparametrized as a function of certain parameters, one of which is precisely this quantile, the median. This procedure allows regression analysis to be carried out in a similar way to that traditionally used when trying to explain the mean of the response variable as a function of covariates, which is impossible for the distribution studied here. We therefore propose this line of action in the present work:

we will study the factors that affect the median of the distribution initially verifying that the reparameterization on the median provides a good fit of the data analysed.

The paper is organized as follows. Section 2 gives the expression of some functions associated with the model: the CDF, the survival function and the quantile function; it also explains how to generate random values from the new law, and studies some properties of the model such as unimodality and the fact that its members can be ordered stochastically. Graphical representations show that the family is quite flexible in several senses: it can used to model overdispersed and underdispersed data; it is also seen that the failure function can be decreasing, bathtub shaped or increasing. Section 3 deals with the point estimation of the two parameters. We offer a method of getting a starting point for the optimization problem involved by means of maximum likelihood (ML) estimates. The performance of the ML estimators is studied numerically and shows good behaviour. Finally, Section 4 considers three real data sets. The data are fitted both to the model presented in this paper and to other competitors. The proposed family provides a much better explanation than the other distributions, showing the practical usefulness of the new distribution.

2 Some properties of the discrete generalized half-normal distribution

Let $X \sim DGHN(\sigma,\beta)$, from (3) it readily follows that

$$\frac{p_k}{p_{k-1}} = \frac{\Phi_\psi\left((k+1)^\beta\right) - \Phi_\psi\left(k^\beta\right)}{\Phi_\psi\left(k^\beta\right) - \Phi_\psi\left((k-1)^\beta\right)}, \quad k = 1,2,\ldots,$$

where $p_0 = 1 - 2\Phi_\psi(1)$.

Let $X \sim DGHN(\sigma,\beta)$, from (3) it readily follows that the CDF of X is given by

$$F(k;\sigma,\beta) = 2\Phi_\psi((k+1)^\beta) - 1, \quad k \in \mathbb{N}_0,$$

the survival function of X is

$$S(k;\sigma,\beta) = 2\Phi_\psi(-(k+1)^\beta), \quad k \in \mathbb{N}_0,$$

and the quantile function is given by

$$Q(u;\sigma,\beta) = \left[\sigma\left\{\Phi^{-1}\left(\frac{1+u}{2}\right)\right\}^{1/\beta} - 1\right], \quad u \in (0,1),$$

where $[\cdot]$ denotes the integer part. As a special case, the median is

$$Q(0.5;\sigma,\beta) = \left[\sigma\left\{\Phi^{-1}\left(\frac{3}{4}\right)\right\}^{1/\beta} - 1\right] \approx \left[\sigma\,(0.6745)^{1/\beta} - 1\right]. \tag{4}$$

Because the DGHN distribution is a discrete version of the GHN model, random values can be generated from this distribution as follows:

(i) Generate $u \sim \mathcal{U}(0,1)$.
(ii) Compute $t = \sigma\left(-\Phi^{-1}(u/2)\right)^{1/\beta}$.
(iii) Do $X = [t]$.

2.1 Moments

The moments of X are given by

$$\begin{aligned} E(X^r) &= 2\sum_{k\geq 0} k^r \left\{\Phi_\psi\left((k+1)^\beta\right) - \Phi_\psi\left(k^\beta\right)\right\} \\ &= 2\sum_{k\geq 0}\{(k+1)^r - k^r\}\Phi_\psi\left(-(k+1)^\beta\right). \end{aligned} \tag{5}$$

As $[Y]^r \leq Y^r$, for $r \geq 1$, it follows directly that $E(X^r) < \infty$, $\forall r \in \mathbb{N}$.

In practice, many count data sets exhibit overdispersion and, although less frequently, also underdispersion. Figure 2 shows the value of the quotient $D = Var(X)/E(X)$ when

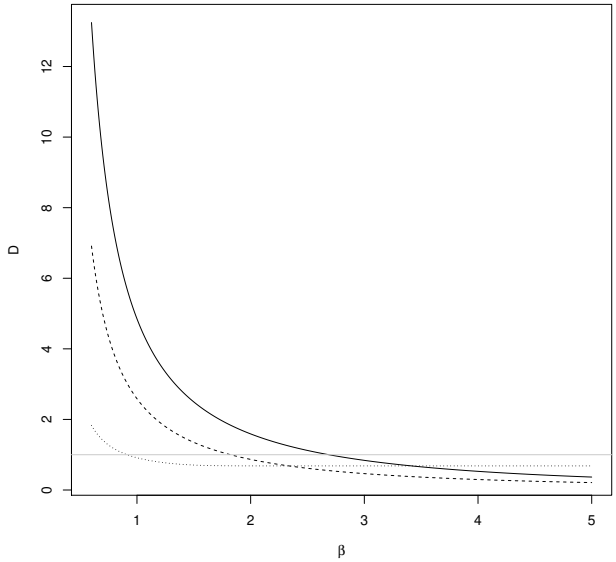

Figure 2: $D = Var(X)/E(X)$ *for* $\sigma = 1$ *(dotted),* $\sigma = 5$ *(dashed) and* $\sigma = 10$ *(solid), the horizontal line* $D = 1$ *is in grey.*

$X \sim DGHN(\sigma,\beta)$ for $\sigma = 1,5,10$ as a function of β. Looking at this figure it can be seen that for each σ the value of D can be greater than, equal to or less than 1 as the value of β increases. In this sense, the new model is quite flexible.

2.2 Mode

Looking at Figure 1 we see that in all cases the PMF is unimodal. Next we show that this is the case for all members in the family. Moreover, we will prove that for $0 < \beta < 1$ the PMF is decreasing. With this aim, we first give a preliminary lemma.

Lemma 1 *If $Y \sim GHN(\sigma,\beta)$ with probability density function $f(x;\sigma,\beta)$, then, as a function of x,*

(a) $f(x;\sigma,\beta)$ is strictly decreasing, if $0 < \beta < 1$, $\forall\sigma > 0$.

(b) $f(x;\sigma,\beta)$ is (strictly) log-concave , if $\beta \geq 1$, $\forall\sigma > 0$.

Proof (a) If $0 < \beta < 1$ then $f(x;\sigma,\beta)$ is proportional to the product of two strictly decreasing functions: $f_1(x) = x^{\beta-1}$ and $f_2(x) = \exp(-0.5x^{2\beta}/\sigma^{2\beta})$; thus it is a strictly decreasing function.

(b) Routine calculations show that $\frac{\partial^2}{\partial x^2} f(x;\sigma,\beta) = -\frac{\beta-1}{x^2} - \frac{\beta(2\beta-1)}{\sigma^{2\beta}} x^{2(\beta-1)}$, which is strictly negative, thus implying the result. ■

Now, we state the following proposition related to the DGHN model.

Proposition 1 *Let $X \sim DGHN(\sigma,\beta)$.*

(a) If $0 < \beta < 1$ and $\sigma > 0$, then $p(k;\sigma,\beta) > p(k+1;\sigma,\beta)$, $\forall k \in \mathbb{N}_0$.

(b) If $\beta \geq 1$ and $\sigma > 0$, then $p(k;\sigma,\beta)^2 \geq p(k-1;\sigma,\beta)p(k+1;\sigma,\beta)$, $\forall k \in \mathbb{N}_0$.

We study separately the two cases: $0 < \beta < 1$ and $\beta \geq 1$.

Proof (a) It is a direct consequence of Lemma 1 (a).

(b) Note that $P(X = k;\sigma,\beta)$ in equation (3) can be written as $P(X = k;\sigma,\beta) = \int_k^{k+1} f(x;\sigma,\beta)dx$. Then, for $\beta \geq 1$, it is a direct consequence of Theorem 2.8. in Dharmadhikari and Joag-Dev (1988) taking $g(x) = f(x;\sigma,\beta)$ (which is log-concave by lemma 1 part b), $\mathcal{B}_n = (0,\infty)$ and $B = (k,k+1) \subseteq \mathcal{B}_n$, that the DGHN distribution is log-concave; the result is immediate. ■

As an immediate consequence of Proposition 1 we state the following.

Corollary 1 *Let $X \sim DGHN(\sigma,\beta)$. X is unimodal. If $0<\beta<1$ the unique mode is attained at $x=0$.*

As commented in Keilson and Gerber (1971), unimodality guarantees that the distribution has all moments, and that the convolution of p_k with any unimodal discrete distribution is also unimodal and log-concave.

2.3 The failure rate function

The failure (or hazard) rate function for the probability function under consideration is given by

$$h(k;\sigma,\beta)=\frac{\Phi_\psi(-k^\beta)}{\Phi_\psi(-(k+1)^\beta)}-1, \quad k\in\mathbb{N}_0.$$

Theorem 9.6 in Dharmadhikari and Joag-Dev (1988) showed that if a random variable is log-concave then it has an increasing failure rate (IFR). Furthermore, Lariviere and Porteus (2001) introduced the concept of generalized failure rate function, defined as $g(k;\sigma,\beta)=k\,h(k;\sigma,\beta)$ for $k\in\mathbb{N}_0$, and showed that the distributions with increasing generalized failure rate (IGFR) have useful applications in operations management (see also Lariviere 2006). It is clear that if a random variable is IFR then it is also IGFR. Accordingly, by the log-concavity of the distribution discussed in Section 2.2, the following result can be established for the discrete generalized half-normal distribution.

Corollary 2 (i) *If $\beta\geq 1$ then the $DGHN(\sigma,\beta)$ distribution is* IFR *and* IGFR .

Figure 3 displays the failure rate function for several parameter values. Looking at this figure, it can be seen that the model is useful for fitting a wide range of shapes: decreasing, bathtub and increasing. Figure 4 shows the different patterns of the failure rate function (IFR, Bathtub and DFR) accordingly to the values of σ and β. We highlight that for $0<\beta\leq 1/2$ the model seems to be DFR, whereas for $1/2<\beta<1$ the behaviour of the failure rate also depends on σ.

The next proposition shows the limit of the failure rate for $k\to+\infty$.

Proposition 2 *Let $X\sim DGHN(\sigma,\beta)$. Therefore, the failure rate satisfies*

$$\lim_{k\to\infty} h(k;\sigma,\beta)=\begin{cases} 0 & \text{if } 0<\beta<1/2,\\ \exp\left(\frac{1}{2\sigma}\right)-1 & \text{if } \beta=1/2,\\ \infty & \text{if } \beta>1/2.\end{cases}$$

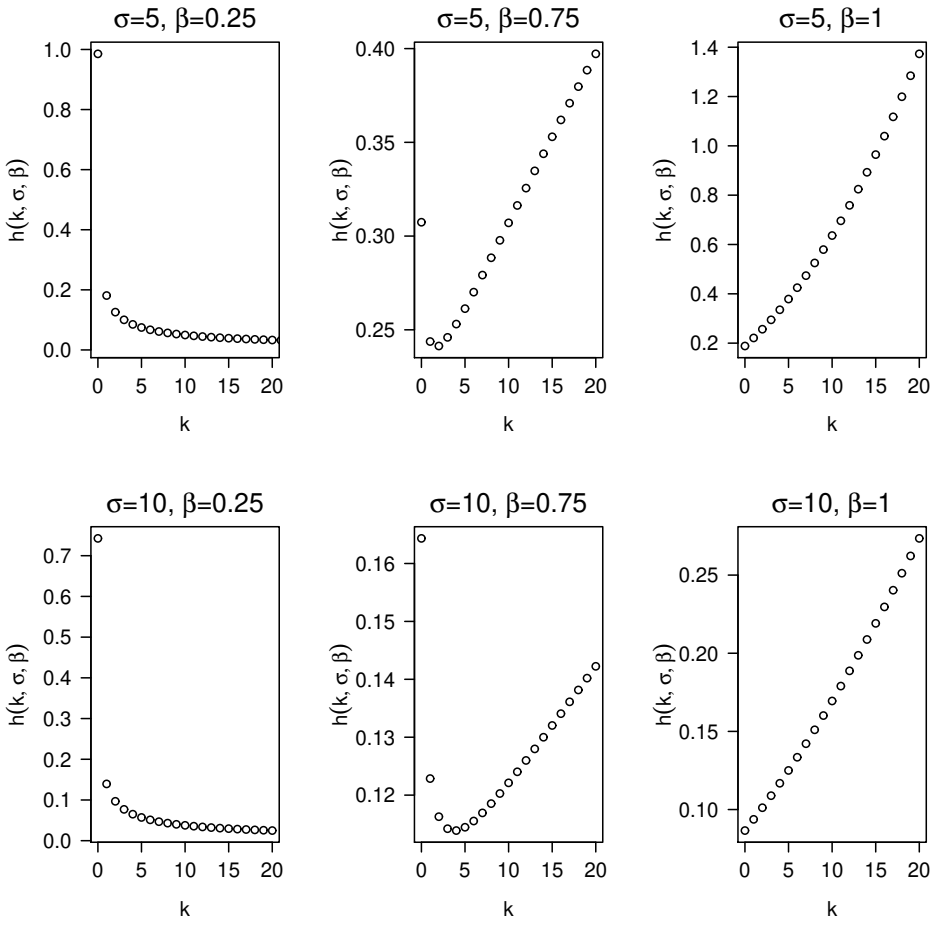

Figure 3: *Failure rate function for several parameter values.*

Proof Using the L'Hôpital rule and the continuity of the limit, we have

$$\lim_{k\to\infty} h(k;\sigma,\beta) = \lim_{k\to\infty}\left(\frac{k}{1+k}\right)^{\beta-1} \exp\left\{-\frac{1}{2\sigma^{2\beta}}\lim_{k\to\infty}\frac{\left[1-\left(1+\frac{1}{k}\right)^{2\beta}\right]}{k^{-2\beta}}\right\} - 1$$

Applying the L'Hôpital rule again in the second limit, we have

$$\begin{aligned}\lim_{k\to\infty} h(k;\sigma,\beta) &= \exp\left\{\frac{1}{2\sigma^{2\beta}}\lim_{k\to\infty}\frac{\left(1+\frac{1}{k}\right)^{2\beta-1}}{k^{-2\beta+1}}\right\} - 1 \\ &= \exp\left\{\frac{1}{2\sigma^{2\beta}}\lim_{k\to\infty}(1+k)^{2\beta-1}\right\} - 1.\end{aligned}$$

The result is obtained separating the cases $0 < \beta < 1/2$, $\beta = 1/2$ and $\beta > 1/2$. ■

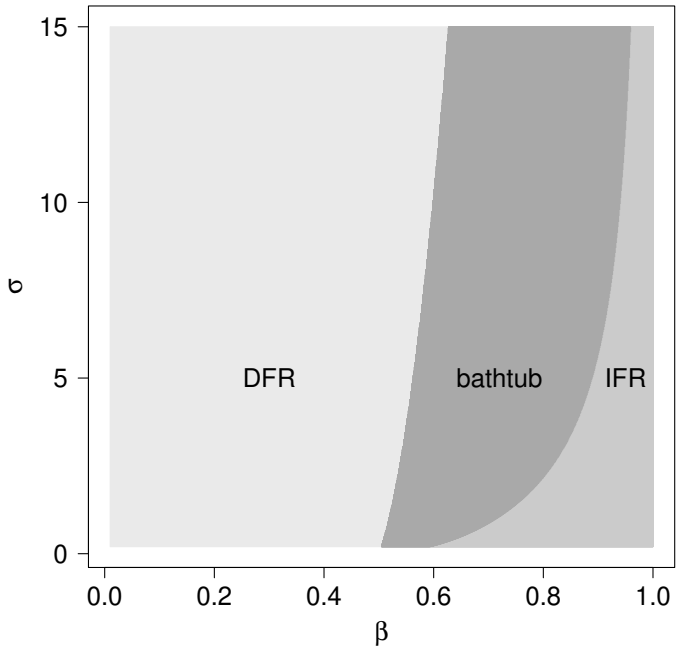

Figure 4: *Shapes for the failure rate of DGHN*(σ,β) *for* $0<\beta<1$.

2.4 Stochastic orderings

This subsection shows that the members of the new model can be stochastically ordered according to the parameter values. With this aim, we first recall the following definition:

Definition 1 *Let X_1 and X_2 be two random variables with distribution functions F_1 and F_2, respectively. Then X_1 is said to be stochastically smaller than X_2, denoted by $X_1 \leq_{st} X_2$, if $F_1(x) \geq F_2(x)$ for all x.*

The DGHN family can be ordered in the following way.

Proposition 3 *(a) Let $X_1 \sim DGHN(\sigma,\beta_1)$ and $X_2 \sim DGHN(\sigma,\beta_2)$, for some σ,β_1, $\beta_1>0$. Then, $X_2 \leq_{st} X_1$ if and only if $\beta_1 \geq \beta_2$.*

(b) Let $X_1 \sim DGHN(\sigma_1,\beta)$ and $X_2 \sim DGHN(\sigma_2,\beta)$, for some $\sigma_1,\sigma_2,\beta>0$. Then, $X_2 \leq_{st} X_1$ if and only if $\sigma_1 \geq \sigma_2$.

Proof (a) Let $\psi_i=\sigma^{\beta_i}$, $i=1,2$. We have $X_2 \leq_{st} X_1$ if and only if $P(X_2 \geq x) \leq P(X_1 \geq x)$ for all $x \in \mathbb{N}_0$ if and only if $2\Phi_{\psi_2}(-(x+1)^{\beta_2}) \leq 2\Phi_{\psi_1}(-(x+1)^{\beta_1})$ for all $x \in \mathbb{N}_0$ if and only if $\beta_1 \geq \beta_2$.

(b) The result can be shown using a similar argument to (a). ■

The following corollary is a consequence of Proposition 3.

Corollary 3 *(i) If* $X_1 \sim DGHN(\sigma,\beta_1)$ *and* $X_2 \sim DGHN(\sigma,\beta_2)$*, with* $\beta_1 \geq \beta_2$*, then* $E(X_2^r) \leq E(X_1^r)$*, for all* $r > 0$*.*

(ii) If $X_1 \sim DGHN(\sigma_1,\beta)$ *and* $X_2 \sim DGHN(\sigma_2,\beta)$*, with* $\sigma_1 \geq \sigma_2$*, then* $E(X_2^r) \leq E(X_1^r)$*, for all* $r > 0$*.*

3 Point estimation

3.1 Without covariates

Let $X_1,\ldots,X_n$ be independent and identically distributed (IID) from $X \sim DGHN(\sigma,\beta)$, and let the observed values be denoted by $x_1,\ldots,x_n$. The log-likelihood function for (σ,β) is

$$\ell(\sigma,\beta) = n\log(2) + \sum_{i=1}^{n} \log\left\{\Phi_\psi\left((x_i+1)^\beta\right) - \Phi_\psi\left(x_i^\beta\right)\right\}. \tag{6}$$

The derivatives of the log-likelihood function are

$$\frac{\partial}{\partial\sigma}\ell(\sigma,\beta) = -\frac{\beta}{\sigma}\sum_{i=1}^{n} \frac{\phi\left(\frac{(x_i+1)^\beta}{\sigma^\beta}\right)\frac{(x_i+1)^\beta}{\sigma^\beta} - \phi\left(\frac{x_i^\beta}{\sigma^\beta}\right)\frac{x_i^\beta}{\sigma^\beta}}{\Phi_\psi\left((x_i+1)^\beta\right) - \Phi_\psi\left(x_i^\beta\right)}, \tag{7}$$

$$\frac{\partial}{\partial\beta}\ell(\sigma,\beta) = \frac{1}{\sigma^\beta}\sum_{i=1}^{n} \frac{\phi\left(\frac{(x_i+1)^\beta}{\sigma^\beta}\right)(x_i+1)^\beta\log\left(\frac{x_i+1}{\sigma}\right) - \phi\left(\frac{x_i^\beta}{\sigma^\beta}\right)x_i^\beta\log\left(\frac{x_i}{\sigma}\right)}{\Phi_\psi\left((x_i+1)^\beta\right) - \Phi_\psi\left(x_i^\beta\right)}. \tag{8}$$

The ML estimates of the parameters satisfy the system that results from equating to 0 in equations (7) and (8). Nevertheless, since this system does not have an explicit solution, in order to obtain the ML estimates it is preferable to maximize function (6). This can be carried out, for example, by using the BFGS algorithm available in the `optim` function of the R programming language (R Core Team, 2016). The BFGS algorithm requires a starting point, which must be inside the feasible region. The estimators obtained from equating any two observed frequencies to their theoretical values can be used as the starting point. For example, if $\widehat{p}_i$ denotes the observed frequency of the value i, for $i = 0,1$ (the zero-frequency and the one-frequency method), the system is

$$\widehat{p}_0 = 2\Phi_\psi(1) - 1 \qquad \text{and} \qquad \widehat{p}_1 = 2\left\{\Phi_\psi\left(2^\beta\right) - \Phi_\psi(1)\right\}.$$

The solutions for ψ and β obtained from the above equations are

$$\tilde{\psi} = \left[\Phi^{-1}\left(\frac{1+\widehat{p}_0}{2}\right)\right]^{-1} \qquad \text{and} \qquad \tilde{\beta} = \frac{\log\tilde{\psi} + \log\Phi^{-1}\left(\widehat{p}_1/2 + \Phi\left(1/\tilde{\psi}\right)\right)}{\log 2}.$$

Therefore, the solution for σ is $\tilde{\sigma} = \tilde{\psi}^{1/\tilde{\beta}}$.

In order to assess numerically the performance of the ML estimates, a simulation study was carried out. Below we describe the study and summarize the results obtained. For several values of the parameters ($\beta = 0.8$, 1.0, 1.3 and $\sigma = 1$, 5) and sample sizes ($n = 30$, 50, 100) 1000 random samples were generated. In each case, the ML estimates of β and σ were computed, as well as their standard error based on the hessian matrix of the model. Table 1 reports the bias, the root of the mean squared error ($\sqrt{MSE}$) and the coverage probability (CP) of the 95% level interval obtained from the asymptotic normality of the ML estimates. As expected, the bias and the $\sqrt{MSE}$ decrease as the sample size increases. Also as expected, the closeness of the CP to its nominal value increases as the sample size increases.In all cases the empirical coverages is quite close to 0.95.

Table 1: *Results for the ML estimates in the DGHN model.*

			$n=30$			$n=50$			$n=100$		
β	σ		bias	$\sqrt{MSE}$	CP	bias	$\sqrt{MSE}$	CP	bias	$\sqrt{MSE}$	CP
0.8	1	$\widehat{\beta}$	0.157	0.443	0.970	0.067	0.249	0.962	0.030	0.125	0.954
		$\widehat{\sigma}$	0.006	0.211	0.955	0.001	0.166	0.952	0.003	0.117	0.950
	5	$\widehat{\beta}$	0.040	0.150	0.955	0.027	0.113	0.950	0.013	0.075	0.952
		$\widehat{\sigma}$	-0.007	0.901	0.926	-0.007	0.695	0.933	-0.003	0.490	0.940
1	1	$\widehat{\beta}$	0.304	0.679	0.970	0.156	0.468	0.971	0.047	0.210	0.961
		$\widehat{\sigma}$	-0.002	0.166	0.969	0.001	0.131	0.963	-0.002	0.094	0.953
	5	$\widehat{\beta}$	0.055	0.190	0.949	0.030	0.137	0.952	0.015	0.092	0.954
		$\widehat{\sigma}$	-0.017	0.708	0.931	-0.005	0.550	0.937	-0.009	0.387	0.944
1.3	1	$\widehat{\beta}$	0.648	0.948	0.975	0.520	0.868	0.975	0.266	0.620	0.975
		$\widehat{\sigma}$	-0.002	0.118	0.980	-0.001	0.094	0.980	0.001	0.070	0.957
	5	$\widehat{\beta}$	0.071	0.237	0.958	0.039	0.175	0.948	0.021	0.116	0.957
		$\widehat{\sigma}$	-0.023	0.549	0.926	-0.020	0.427	0.932	-0.006	0.299	0.947

3.2 Estimation in a DGHN regression model

Unfortunately, the mean of the DGHN has a complicated form (see equation (5)). For this reason, an alternative way to use this model in a regression context is through the median (see equation (4)). Let $Q_{0.5}$ be the median of the model. The pmf of the model

with reparametrization based on $Q_{0.5}$ and β is given by

$$p_k = 2\sum_{j=0}^{1}(-1)^j \Phi\left(\tau\left(\frac{k+j}{1+Q_{0.5}}\right)^{\beta}\right), \quad k = 0,1,2,\ldots \tag{9}$$

where $\tau = 0.674489$.

A common specification for $Q_{0.5}$ is exponential, ensuring the non–negativity of this parameter. That is,

$$\log Q_{0.5i} = \sum_{s=1}^{\kappa} x_{is}\gamma_s, \quad i = 1,\ldots,t,$$

where $x_{i1}, x_{i2}, \ldots, x_{i\kappa}$ are covariates and $\gamma_1, \gamma_2, \ldots, \gamma_\kappa$ are unknown regression coefficients. The log-likelihood for the vector (γ, β) is

$$\ell(\gamma,\beta) = n\log 2 + \sum_{i=1}^{n}\log\left\{\Phi\left(\tau\left(\frac{k}{1+Q_{0.5i}}\right)^{\beta}\right) - \Phi\left(\tau\left(\frac{k+1}{1+Q_{0.5i}}\right)^{\beta}\right)\right\}. \tag{10}$$

Again, the mle of (γ, β) can be obtained maximizing (10) in relation to them.

4 Applications

This section presents applications to three real data sets.

4.1 An application in ecology

This data set (Kulasekera and Tonkyn, 1992 and Table **2** here) consists of the number of weevil eggs laid per bean and contains 193 observations.

Table 2: *Number of weevil eggs laid per bean*

Number / bean	0	1	2	3	Total
Obs. Freq.	5	68	88	32	193

To analyse the data we considered the model proposed in this paper, comparing it to the models in Kemp (2008), Gómez-Déniz et al. (2014) and in Kulasekera and Tonkyn (1992) (denoted as Kula in the tables). ML estimators of the parameters for each model are shown in Table 3. This table also shows the value of the maximized log-likelihood, L, and the Akaike information criterion, Akaike (1974), defined as $\mathrm{AIC} = 2r - 2\log L$, where r is the number of parameters. As is well-known, the model with lower AIC is preferred. Therefore, according to this criterion, the proposed model provides a better fit than the other laws. To illustrate the performance of the DGHN model for this data,

we estimate the probability of the events $X = 0$, $X = 1$, $X = 2$, $X = 3$ and $X \geq 4$ for all the models with their respective 95% confidence intervals based on the delta method (we exclude the estimations provided by Kulasekera and Tonkyn (1992) because their intervals are very wide). Results are presented in Table 4. Note that the DGHN model is the only one for which the confidence intervals always include the observed frequencies. Therefore, the proposed distribution may be an attractive alternative to models for data taking values in $\mathbb{N}_0$.

Table 3: *Model ML estimates and standard errors (in parentheses).*

	Kemp	Gómez-Déniz et al. (2014)	DGHN	Kula
	$\widehat{\theta} = 12.9970\ (15.2697)$	$\widehat{\alpha} = 54.1196\ (0.2091)$	$\widehat{\beta} = 2.8873\ (3.0927)$	$\widehat{\alpha} = 11.0943\ (13.8496)$
	$\widehat{q} = 0.1393\ (0.0490)$	$\widehat{\sigma} = 1.0860\ (0.0562)$	$\widehat{\sigma} = 2.6519\ (0.0251)$	$\widehat{q} = 0.0125\ (0.0057)$
L	–223.956	–222.9054	–218.7891	–221.9045
AIC	451.9119	449.8108	441.5782	447.8090

Table 4: *Estimated probabilities for $P(X = k), k = 0,1,2,3$, and $P(X \geq 4)$ and their 95% confidence intervals (CI).*

	$X=0$		$X=1$		$X=2$		$X=3$		$X\geq 4$	
model	point	95% CI	point	95% CI	point	95% CI	point	95% CI	point	95% CI
Kemp	0.022	(0.011,0.034)	0.291	(0.249,0.333)	0.527	(0.476,0.577)	0.133	(0.093,0.173)	0.005	(0.000,0.009)
Gómez-Déniz et al. (2014)	0.061	(0.039,0.084)	0.280	(0.235,0.324)	0.522	(0.463,0.581)	0.131	(0.090,0.172)	0.006	(0.001,0.011)
DGHN	0.048	(0.025,0.069)	0.294	(0.253,0.356)	0.505	(0.449,0.561)	0.152	(0.106,0.199)	0.001	(0.000,0.003)
observed	0.026		0.352		0.456		0.166		0.000	

4.2 A real application in the health framework

Since the seminal work of Koenker and Bassett (1978) quantile regression has attracted much research, particularly in recent years, probably due to the help of computers. This technique allows a natural generalization of the generalized linear models for certain well-known robust estimators of location. The methodology we propose in this Section is simple and, enables us to explain the median by the effects of covariate factors, as discussed in Section 3.2.

Many authors in the literature have focused on the factors that affect the mean of the dependent variable under study. The proposal presented here is based on studying the factors that can affect the median of the dependent variable. As far as we know, there are few studies in the theoretical or applied statistical literature of regression of quantiles for a discrete variable (parametric model).

A common specification for the median parameter, $Q_{0.5}$, is exponential, ensuring the non–negativity of the parameter. That is,

$$\log Q_{0.5} = \sum_{s=1}^{\kappa} x_{is}\gamma_s, \quad i = 1,\ldots,t,$$

obtaining the conventional log-linear model such that $Q_{0.5} = \exp\left\{\gamma^\top x\right\}$, where x is the vector of covariates and γ is an unknown vector of regression coefficients.

The marginal effect, which reflects the variation of the conditional median due to a one-unit change in the jth covariate $(j = 1,\ldots,\kappa)$, has a similar consideration to that in generalized linear models. For indicator variables such as $x_{\kappa i}$ which takes only the values 0 or 1, the marginal effect is $\delta_j = Q_{0.5}(k_i|x_j = 1, x_1,\ldots,x_\kappa)/Q_{0.5}(k_i|x_j = 0, x_1,\ldots,x_\kappa) \approx \exp(\beta_j)$, $i = 1,\ldots,n$; $j = 1,\ldots,\kappa$. Therefore, the conditional median is $\exp(\beta_j)$ times larger if the indicator variable is one rather than zero.

For the present purpose we used data obtained from the 1977-78 Australian Health Survey, a well-known data set previously studied by Cameron and Trivedi (1998); see also Cameron and Trivedi (1986). This data set can be downloaded from the web page

`http://cameron.econ.ucdavis.edu/racd/racddata.html`

Details of this data source can also be consulted in the "Ecdat" R (data(DoctorAUS)) package. The data set consists of 5190 elements with fifteen variables. The variable ILLNESS, the number of illnesses in past 2 weeks is taken as the dependent variable. The minimum value of this variable is 0, the maximum value 5 and the median is 1. A different count variable could be taken as the dependent variable if another study were required. Fundamentally, the convenience of this approach is based on the fact that by testing all the count variables appearing in the data, the variable ILLNESS presents a median different from zero and a larger index of dispersion.

In our study, CHCOND (chronic condition) is not considered, and INSURANCE (medlevy : medibanl levy, levyplus: private health insurance, freepoor: government insurance due to low income, freerepa : government insurance due to old age disability or veteran status) is converted into three dichotomous variables, FREEPOR, FREEREPA AND LEVYPLUS. Therefore, MEDLEVY is the reference variable.

Descriptive statistics on the variables in this dataset are given in Cameron and Trivedi (1986, p.68) (see Table 3.2 in this work). In our study the following distributions were also considered for comparison purposes: a Poisson (P) distribution with parameter $\beta > 0$; a negative binomial (NB) distribution with parameters $\beta > 0$ and mean $q > 0$; a generalised Poisson (GP) distribution with parameters $\beta > 0$ and mean $q > 0$ and of course the proposed distribution studied here. Among the various parameterisations of the generalized Poisson distribution, we used the one described in Consul and Famoye (1992).

Tables 5 and 6 show the estimation in the case of non including and including covariates, respectively. Again, in view of the maximum value of the logarithm of the likelihood function, the proposed distribution studied here is superior to the remainders. We estimated the two parameters, β and $q = Q_{0.5}$ by maximizing directly the log-likelihood function given by $L = \sum_{i=1}^{n} \log p_{k_i}$. We also show the value obtained for the Akaike Information Criterion (AIC). (Note that AIC $= 2(k - L)$, where k is the number of model

parameters and L is the maximum value of the log–likelihood function). The goodness of fit is also corroborated by looking at the graph shown in Figure 5, in which it can be observed that the model seems to be a reasonable choice for the given data.

Table 5: *Coefficient estimates and p-values for the different models considered without covariates.*

	P		NB		GP		DGHN	
Parameter	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
$\widehat{\beta}$	1.431	0.000	3.801	0.000	0.120	0.000	0.082	0.000
$\widehat{q}$			1.431	0.000	1.432	0.000	0.015	0.000
L	-8390.942		-8264.408		-8266.708		-8255.156	
AIC	16783.90		16532.80		16537.40		16514.30	

Table 6: *Coefficient estimates and p-values for the different models considered with covariates. The cases of P, NB and GP correspond to maximizing the mean link and the GHN to maximizing the median*

	P		NB		GP		DGHN	
Variable	Estimate	$\Pr > \|t\|$	Variable	$\Pr > \|t\|$	Variable	$\Pr > \|t\|$	Variable	$\Pr > \|t\|$
SEX	0.022	0.259	0.021	0.419	0.021	0.407	0.013	0.750
AGE	0.151	0.026	0.143	0.081	0.142	0.080	0.367	0.003
INCOME	-0.125	0.000	-0.125	0.001	-0.125	0.001	-0.186	0.002
HSCORE	0.082	0.000	0.084	0.000	0.084	0.000	0.126	0.000
DOCTORCO	0.043	0.000	0.045	0.000	0.045	0.000	0.060	0.002
NONDOCCO	0.009	0.253	0.008	0.415	0.008	0.384	0.000	0.962
HOSPADMI	-0.014	0.433	-0.012	0.614	-0.012	0.611	-0.011	0.716
HOSPDAYS	0.000	0.475	0.000	0.655	0.000	0.638	0.001	0.463
MEDECINE	0.071	0.000	0.072	0.050	0.072	0.000	0.095	0.000
PRESCRIB	0.077	0.000	0.078	0.037	0.078	0.000	0.097	0.000
NONPRESC	0.103	0.000	0.105	0.007	0.105	0.000	0.154	0.000
FREEPOR	0.008	0.610	0.009	0.936	0.009	0.720	-0.040	0.209
FREEREPA	0.103	0.003	0.107	0.015	0.107	0.011	0.136	0.044
LEVYPLUS	0.008	0.610	0.009	0.936	0.009	0.720	0.049	0.128
CONSTANT	-0.064	0.084	-0.068	0.122	-0.069	0.114	-0.968	0.000
$\widehat{\beta}$			38.373	0.053	0.013	0.053	1.213	0.000
L	-7590.674		-7588.737		-7588.696		-7759.528	

As can be seen, most of the covariates considered are statistically significant except SEX, NONDOCCO, HOSPADMI, HOSPDAYS, FREEPOR and LEVYPLUS in all the models used. Observe that the sign of the regressors coincides for all the models.

It can be seen that the maximum value of the log-likelihood function is lower in the case of the quantile regression although the estimates are similar in terms of sign and significance. This is not surprising since the link used affects the mean in classical models and the median in the distribution studied here. Thus from our point of view, the model is viable for cases in which classical distributions provide a poor fit of the variable to be studied, as will be seen in the last example provided in the next subsection.

The different models considered were analysed using the BFGS algorithm (Broyden, Fletcher, Goldfarb and Shanno), with RATS and Mathematica (Wolfram) software, for

both the inflated and the non-inflated models. In all of the models considered, the convergence of the algorithm is extremely fast. In general, the algorithm converged in fewer than 30 iterations.

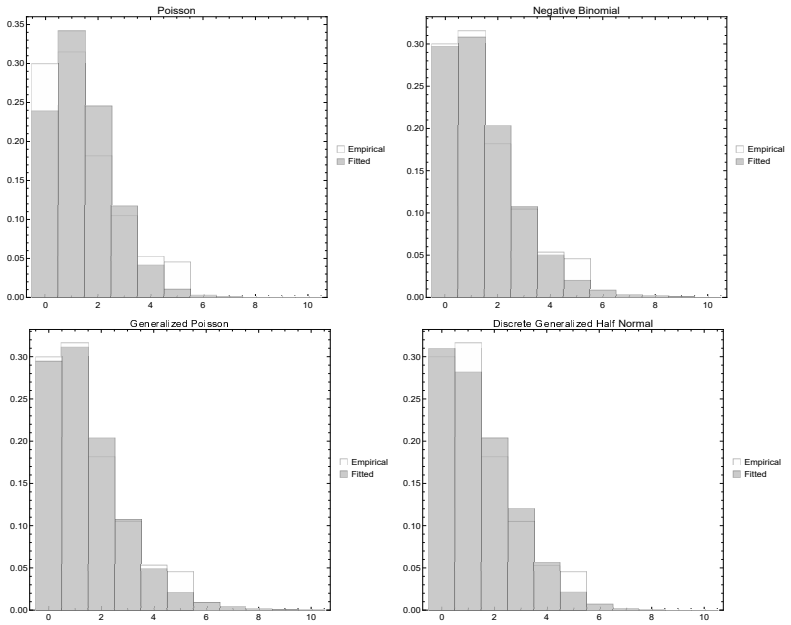

Figure 5: *Empirical and fitted data for the number of illness in the past two weeks.*

4.3 An actuarial application

Usually in automobile insurance rate-making the target is to estimate the probability of a claim in order to compute a premium according to a premium calculation principle. In this example we consider a dataset of Swedish third party automobile insurance claims which is well-known in the actuarial literature. Some of the most important factors of claim frequency will be taken into account. The variable kilometres (Km) is the kilometers travelled by a vehicle, here grouped into seven categories (category 1, less than 1000 km per year, category 2, 1000-15000 km per year, etc.); Zone gives the graphic zone, also grouped into seven categories; Bonus is a variable representing the driver claim record grouped into seven categories; Insured starts in the class 1 and is moved up one class, to a maximum of 7, for each year in which there is no claim; finally, Make represents the type of vehicle (nine specified makes of car). The dependent variable is the Number of claims. More details can be seen in Frees (2010).

For comparison purposes we have considered the Poisson and the negative binomial distributions, which are very widely used in the actuarial context, to fit the number of

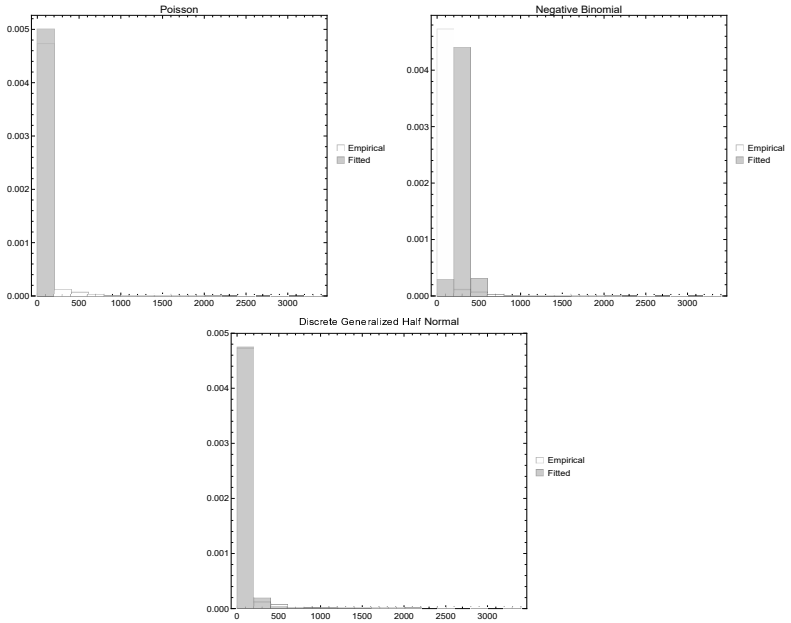

Figure 6: *Empirical and fitted data for Swedish automobile claims.*

claims. The values of the maximum of the log-likelihood function for these models are -221571.00 and -93806.541, respectively, compared to -8920.57 for the distribution proposed here. Therefore, the proposed distribution is very much superior to the others. The estimated values of the parameters are $\widehat{\beta} = 0.324\,(0.005)$ and $\widehat{q} = 1.571\,(0.197)$ (standard errors in parentheses).

Figure 6 shows the empirical and fitted distributions obtained using Poisson, negative binomial and our proposed distribution. This graphic confirms the superiority of the proposed distribution over the others.

Using a similar idea to that proposed in Heras, Moreno and Vilar-Zanón (2018), we have used the covariates explained above in order to explain the median of the dependent variable given by the number of claims. The results are shown in Table 7. As we can see the value of the maximum of the log-likelihood function has been much reduced.

It can be seen that all the variables are highly significant, and the signs (see Frees , 2010) are similar to those of the classical regression model when the Number of claims is considered as the dependent variable, except for the covariate Km; when this last is studied in detail, the interpretation is observed to be similar; the covariable Km takes values from 1 to 5, increasing with the number of kilometers traveled by the insured. The negative value of the regressor indicates that the greater the number of kilometers traveled, the smaller will be the value of the median. The insured will have better insurance terms than justified by his claim record (because he has travelled more kilometers).

Finally, the premium for this automobile insurance portfolio, which is not computed here, can be obtained by using the quantile principle used by Heras et al. (2018).

Table 7: *Parameter estimates from the new count distribution using quantile (median) regression*

Parameter	Estimate	S.E.	t–Wald	$\Pr > \|t\|$
km	–0.536	0.035	15.096	0.000
zone	–0.394	0.026	14.802	0.000
bonus	0.253	0.020	12.382	0.000
make	0.289	0.015	19.251	0.000
$\widehat{\beta}$	0.410	0.006	58.834	0.000
constant	2.485	0.183	13.581	0.000
$L = -8516.35$				
AIC = 17044.70				

Conclusions

This work introduces the discrete version of the continuous GHN distribution. We have presented its most important probabilistic properties. Parameter estimation was approached by maximum likelihood. Using three applications to real data sets, we have shown that the discrete generalized half-normal distribution proposed in this work provided a better fit than other extensions of the discrete half-normal model, illustrating that the model is competitive with other discrete models depending on two parameters.

One of the disadvantages of the discretization of a continuous variable is that the average does not appear expressed in a closed form allowing simple reparameterization of the distribution in order to incorporate covariables. However, as noted, this drawback can be avoided by carrying out quantile regression (the median in our case). This is possible due to the fact that the discretization is carried out from the distribution function, which has a simple, closed expression. This particularity has been incorporated into this work with an application in the health scenario, which take into account the fact that on many occasions the median is a more intuitive, manageable and practical characteristic than the mean.

Acknowledgments

The authors are grateful to two anonymous Referees as well as the Associate Editor for their contributions, which have improved the presentation of this work.

The research of D. I. Gallardo was supported by FONDECYT 11160670 (Chile). The research of H. W. Gómez was supported by MINEDUC-UA project, code ANT 1755 (Chile). EGD's work was partially funded by grant ECO2017–85577–P (Ministerio de Economía, Industria y Competitividad. Agencia Estatal de Investigación)). EGD also acknowledges the Departamento de Matemáticas, Facultad de Ciencias Básicas, Universidad de Antofagasta, Antofagasta (Chile) for their special support, as part of this work was done while EGD was visiting this University in 2018.

References

Akaike, H. (1974). A new look at statistical model identification, *IEEE Transactions on Automatic Control*, 19, 716-723.

Cameron, A.C. and Trivedi, P.K. (1986). Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. *Journal of Applied Econometrics*, 1, 29-54.

Cameron, C. and Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press.

Chakraborty, S. (2015). Generating discrete analogues of continuous probability distributions-A survey of methods and constructions. *Journal of Statistical Distributions and Applications*, 2, 6.

Consul, P.C. and Famoye, F. (1992). Generalized Poisson regression model. *Communications in Statistics-Theory and Methods*, 21, 89-109.

Cooray, K. and Ananda, M. (2008). A Generalization of the Half-Normal Distribution with Applications to Lifetime Data. *Communications in Statistics - Theory and Methods*, 37, 1323-1337.

Dharmadhikari, S. and Joag-Dev, K. (1988). *Unimodality, convexity and applications. Probability and mathematical statistics*. Academic Press Inc, Boston

Frees, E.W. (2010). *Regression Models with Actuarial and Financial Applications*. Cambridge University Press

Gómez-Déniz, E. and Calderín, E. (2011). The discrete Lindley distribution: properties and applications. *Journal of Statistical Computation and Simulation*, 81, 1405-1416.

Gómez-Déniz, E., Vázquez-Polo, F.J. and García-García, V. (2014). A discrete version of the half-normal distribution and its generalization with applications. *Statistical Papers*, 55, 497-511.

Heras, A., Moreno, I. and Vilar-Zanón, J.L. (2018). An application of two-stage quantile regression to insurance ratemaking. *Scandinavian Actuarial Journal*, 9, 753-769.

Keilson, J. and Gerber, H. (1971). Some results for discrete unimodality. *Journal of the American Statistical Association*, 66, 386-389.

Kemp, A.W. (2008). The discrete half-normal distribution. In: Birkhä (Ed) *Advances in mathematical and statistical modeling*, pp. 353-365.

Koender, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33-50.

Krishna, H. and Singh, P. (2009). Discrete Burr and discrete Pareto distributions. *Statistical Methodology*, 6, 177-188.

Kulasekera, K.B. and Tonkyn, D.W. (1992). A new discrete distribution with applications to survival, dispersal and dispersion. *Communications in Statistics - Simulation and Computation*, 21, 499-518.

Lariviere, M.A. and Porteus, E.L. (2001). Selling to the newsvendor: An analysis of price-only contracts. *M&SOM*, 3, 293-305.

Lariviere, M.A. (2006). A note on probability distributions with increasing generalized failure rates. *Operations Research*, 54 602-604.

Marshall, A.W. and Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. *Biometrika*, 84, 641-652.

Nakagawa, T. and Osaki, S. (1975). The discrete Weibull distribution. *IEEE Transactions on Reliability*, 24, 300-301.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/