COMPOUND HARRIS DISTRIBUTION

Latha C M [1] **and Sandhya E**[2]

[1] Department of Statistics, St.Thomas College, Pala-686574, India,

[2] Department of Statistics, Prajyothi Nikethan College, Pudukkad-680301, India.

Abstract The compound Harris (CH) distribution is introduced. Probabilities of the distribution are evaluated. Some of the statistical, distributional and reliability properties of the distribution are discussed. A characterization of the distribution is established using its S- function. Simulation and estimation of parameters are done using method of moments and probability generating function (pgf) based BHHJ method. Fitting of the distribution is carried out using a real life data.

Keywords: CH distribution, Panjer's formula, Fast Fourier Transform (FFT), R and S functions, Reliability, DFR and IFR, Moment estimator, pgf based BHHJ estimator, Compound Extended Geometric (CEG) distribution, Compound Negative Binomial (CNB) distribution.

.

1. Introduction

A larger class of distributions can be created by the process of compounding any two discrete distributions. The compound distribution arises as $_{i=1}$ follows. Let N be counting random variable with pgf $Q_N(t)$ and X_1, X_2, ... be independent and identically distributed (iid) counting random variables with pgf Q(t). Assuming that X_is' are independent of N, the random sum $Y = \sum_{i=1}^{N} X_i$ $(where\ N = 0\ implies\ Y = 0)$. The term 'compounding' reflects the idea that the pgf of the new distribution $Q_Y(t)$ is written as $Q_Y(t) = Q_N(Q(t))$ where $Q_N(t)$ and Q(t) are called primary and secondary distributions respectively.

In insurance context, this distribution can arise naturally. If N represents the number of accidents arising in a portfolio of risks and X_i, $i = 1, 2, ...N$ represent the number of claims from the accidents, then Y represents the total number of claims from the portfolio.

One of the standard stochastic models used in various areas of applied probability such as insurance risk theory and queuing theory is the random sum (compound) model. The literature on such models is voluminous, both from an analytic and a numerical view point. Compound distributions are specially useful for modeling outcomes exhibiting overdispersion. ie. a greater amount of variability than would be expected under a certain model.

Willmot (1989) derived a number of asymptotic formulae for some discrete compound distributions which have been found to be useful for modeling. These formulae provide insight into the distributional form and often complement recursive computational algorithm which are cubersome in the right tail of the distribution. Many distributional and reliability aspects of a geometric sum have been obtained in the literature. Shanthikumar (1988) has proved that DS- DFR property is preserved under geometric sum. Brown (1990) showed that a geometric sum $Y = \sum_{i=1}^{N} X_i$ always have the new worse than used (NWU) property whatever the distribution of X is. Other distributional properties of geometric sum can be found in Kovats et al. (1992), Willmot (2002) and Willmot et al. (2001).

Methods for analyzing dental caries and associated risk indicators have evolved considerably in recent decades. The use of zero- inflated or hurdle models is increasing. Vergenes et al. (2016) showed that zero-inflated and hurdle models can both be expressed as a compound sum. Using the same compound sum, they fitted the compound negative binomial distribution for dental caries data. Associated with the notion of bulk queue, Romeo (2015) derived the distribution of number of customers say, Y(t) that arrive in an arbitrary bulk arrival queue system during any period of time t where Y(t) can be considered as a compound random variable.

Here we introduce a new distribution, namely, Compound Harris (CH) distribution by compounding Harris distribution given by (1.1) with a standard discrete distribution and discuss some of its properties. Harris in 1948 introduced a pgf given by

$$Q_N(t) = \left(\frac{p}{1 - qt^k}\right)^{1/k}, \quad k > 0\ integer, 0 < p < 1, p + q = 1. \tag{1.1}$$

He introduced this pgf while considering a simple discrete branching process where a particle either splits into (k+1) identical particles or remains the same during a short time interval Δt. The probability distribution corresponding to the pgf $Q_N(t)$ is called Harris distribution and is denoted by $H_0(p, \mathrm{k}, \frac{1}{k})$. This distribution has support 0, k, $2k$,… where k is a positive integer. When $k = 1$, it reduces to geometric distribution.

Harris distribution plays a key role in schemes with random sums in general and in branching processes and time series models in particular. Its pgf had been discussed in the context of branching processes (Harris (1948)), N- sums and N-extremes (Satheesh et al. (2002), Satheesh et al. (2002a)) where N is Harris distributed. Satheesh et al. (2005) have developed a time series model that has an inherent N- sum structure. Satheesh et al. (2004) have shown that a Harris- sum of Harris distribution is again Harris.

Harris distribution had been used to demonstrate the notion of random infinite divisibility with respect to non- negative integer- valued random variables (Sandhya (1996)). Sherly (2007) has shown that H_0 $(p, k, \frac{1}{k})$ is infinitely divisible and self-decomposable. Also they have presented the Harris distribution as an appropriate marketing distribution for a specific manufacturing unit.

The discussion begins with the expression for probabilities of CH distribution. The difficulty in evaluation of probabilities arises from the presence of convolutions, but such evaluation is important for many applications. Panjer's (1981) recursive formula is a good choice. However, another technique known as Fast Fourier Transform (FFT) technique is employed here to evaluate the probabilities, following Embrechts et al. (2009). Section 2 deals with the numerical evaluation of CH probabilities. The graphs are also plotted. Statistical and distributional properties are discussed in sections 3 and 4 respectively. Section 5 deals with the reliability properties of the distribution. Estimation of parameters using moment method and BHHJ method are done using simulated data in section 6. Section 7 deals with fitting of the distribution using a real life data set.

2. Compound Harris Distribution

Definition 2.1

A counting random variable Y is said to follow compound Harris (CH) distribution if its pgf is given by

$$Q_Y(t) = \left(\frac{p}{1 - q[Q(t)]^k} \right)^{\frac{1}{k}}$$

where Q(t) is the pgf of X_i , i = 1, ...N, a set of random variables independent of N. Here Y is a random sum and we write $Y \sim \underset{\sim}{C}H$ (k, p, p') where p' is the parameter of distribution of X_i and Y is said to have gap k. The distributions of N, X_i and Y are called primary, secondary and compound distributions respectively.

Compound Harris Probabilities

Let $Q(t) = \sum_{i=0}^{\infty} q_i t^i$ be the pgf of secondary distribution. Then the CH probabilities g_0, g_1, g_2, ... are the coefficients of t^0, t^1, t^2,... in the expansion of Q_Y (t) and we get

$$g_0 = p^{\frac{1}{k}} \left[1 - q q_0{}^k\right]^{\left(\frac{-1}{k}\right)}$$

$$g_1 = p^{\frac{1}{k}} q_1 q q_0{}^{k-1} \left[1 - q q_0{}^k\right]^{-\left(\frac{1}{k}+1\right)}$$

$$g_2 = p^{\frac{1}{k}} q q_0{}^{k-2} [1 - q q_0{}^k]^{-\left(\frac{1}{k}+2\right)} \left[(k-1) q_1{}^2 + 2 q_0 q_2 + 2 q q_0{}^k [q_1{}^2 - q_0 q_2]\right]$$ and so on.

It is quite difficult to get compact expressions for higher probabilities. So they are evaluated numerically.

Evaluation of Probabilities

Panjer's recursive formula is a widespread standard technique for the evaluation of compound probabilities. This algorithm has received great attention in many areas such as actuarial science and operations research. Kaas et al. (2008) gave a detailed treatment for both its theory and application. Das et al. (2011) reviewed and extended the formula for evaluation of compound negative binomial distribution. Here we evaluate the CH probabilities using FFT. For example, assume that Q(t) is the pgf of geometric distribution with success probability p′= 0.5 and support {0, 1, 2,...}.

The following R commands are used to evaluate the corresponding CH probabilities

CH (*k, p, 0.5)*

1. $M \leftarrow 128$

2. $k \leftarrow 2$

3. $f \leftarrow dgeom\ (0 : (M - 1), prob = 0.5)$

4. $f\,hat \leftarrow f\,f\,t\,(f, inverse = FALS\ E)$

5. $f\,khat \leftarrow f\,hat * f\,hat$

6. $\text{u} \leftarrow \left(\frac{p}{1 - (q * fkhat)} \right)^{\frac{1}{k}}$

7. $g \leftarrow (1/M) * f\,f\,t\ (u, inverse = TRUE)$

The vector g contains the probability masses on 0, 1, 2, ... (*M*-1) where M is a truncation point. The probabilities are not displayed here as it takes much space. The probabilities in the case of any discrete secondary distribution can be evaluated in similar way. In this work we concentrate on geometric secondary distribution on 0, 1, 2,.. . The graphs of CH probabilities with geometric, binomial and uniform secondary distributions at different values of *p* are plotted below.

Fig 2.1 CH with secondary distribution geometric *(k=2, p′=0.5)*

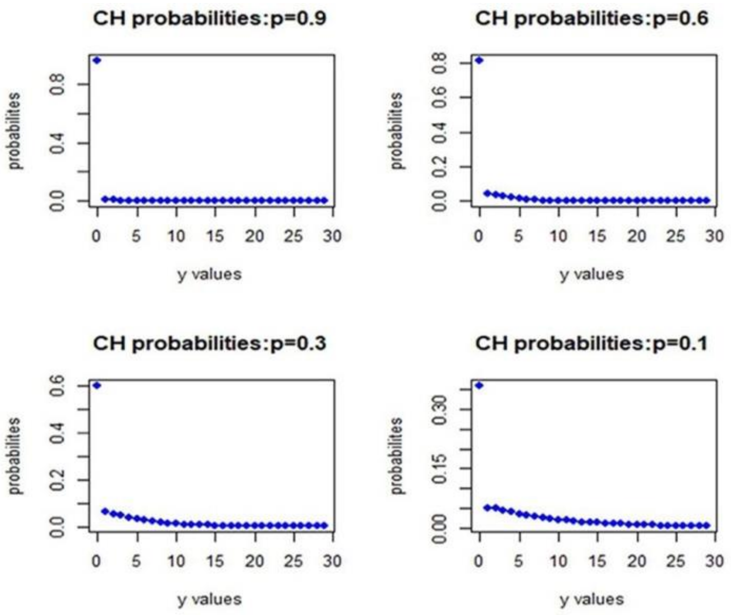


Fig 2.2 CH with secondary distribution Binomial *(k=2, p′=0.5, size=10)*

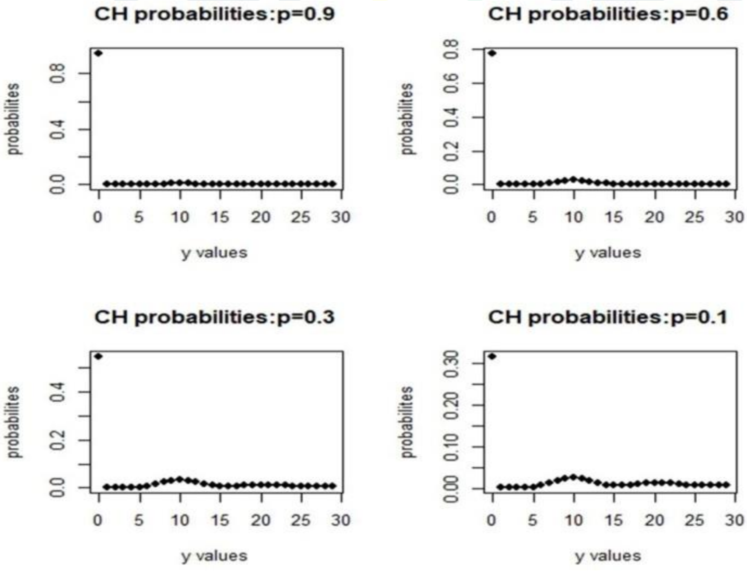
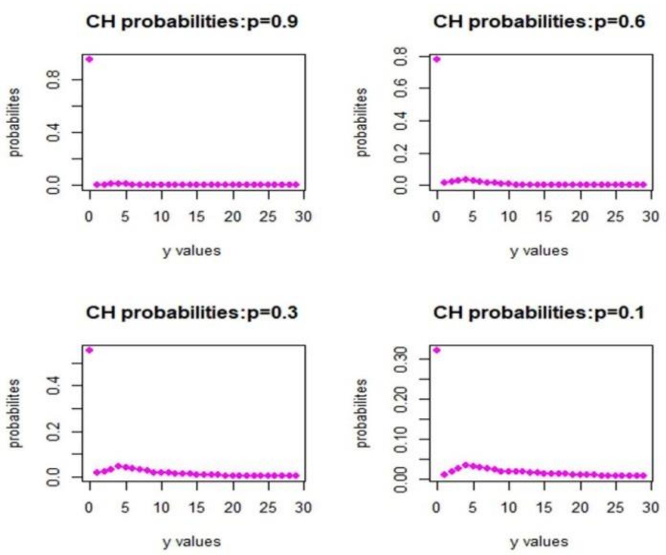


Fig 2.3 CH with secondary distribution Uniform *(k=2, p′=0.2, N=5)*



Remark 2.1 Irrespective of any secondary distribution, CH distribution has mode at $Y = 0$.

Remark 2.2 Even though Harris distribution has its probabilities at 0, *k*, *2k* ... , CH distribution has probabilities at all points 0, 1, 2 ...

3. Statistical Properties

Quantiles

The quantile function is one way of prescribing a probability distribution and it is an alternative to the pmf and the cumulative distribution function (cdf). The discrete cdf is a step function, so it does not have an inverse function. Given a probability p_0, the quantile for p_0 is defined as the smallest value of the random variable Y for which $F(y) \geq p_0$.

Closed form expression for quantiles are not easy to derive as the distribution function is not in a compact form. The quantile values at different probabilities for CH (2, p, 0.5) are tabulated below for simulated samples.

Table3.1 Quantiles for *k*=2

				p_0					
probabilities	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99
p=0.8	0	0	0	0	0	0	0	0	6
p=0.5	0	0	0	0	0	0	1	4	12
p=0.2	0	0	0	0	2	4	8	12	32
p=0.1	0	0	1	3	6	10	15	26	65

Remark 3.1 It is evident from the tables that skewness of CH distribution becomes higher as the parameter *p* increases. For *p=0.8* the quantile values up to 0.9 are zero whereas for *p=0.1* the quantile value corresponding to 0.4 is not zero.

Moments

$$\mu_1'(\mathbf{Y}) = \mu_1'(\mathbf{N})\mu_1'(\mathbf{X})$$

$$= \frac{q}{p}\, \mu_1'(\mathbf{X})$$

$$\mu_2(\mathbf{Y}) = \mu_2(\mathbf{N})[\,\mu_1'(\mathbf{X})]^2 + \,\mu_1'(\mathbf{N})\mu_2(\mathbf{X})$$

$$= \frac{kq}{p^2}\,[\,\mu_1'(\mathbf{X})]^2 + \frac{q}{p}\mu_2(\mathbf{X})$$

$$= \frac{k}{q}\,[\,\mu_1'(\mathbf{Y})]^2 + \frac{q}{p}\mu_2(\mathbf{X})$$

$$\geq \frac{k}{q}\,[\,\mu_1'(\mathbf{Y})]^2 \qquad [\because \textbf{The second term being non} - \textbf{negative}]$$

$$> [\mu_1'(\mathbf{Y})]^2 \qquad [\because k > \mathbf{q}]$$

$$\Rightarrow \frac{\mu_2(\mathbf{Y})}{[\mu_1'(\mathbf{Y})]^2} > 1$$

i.e. CV(Y) > 1.

Skewness

The third central moment μ_3 is given by

$$\mu_3(Y) = \mu_3(\text{N})\,[\mu_1'(X)]^3 + 3\,\mu_2(\text{N})\,\mu_1'(X)\mu_2(\text{X}) + \mu_1'(N)\,\mu_3(\text{X})$$

$$= \frac{k^2q(1+q)}{p^3}\,[\mu_1'(X)]^3 + \frac{3kq}{p^2}\,\mu_1'(X)\mu_2(\text{X}) + \frac{q}{p}\,\mu_3(\text{X})$$

Obviously, the nature of skewness depends upon the behaviour of secondary distribution. The various states of skewness of CH distribution are displayed in the body of the following table.

	$\mu_1'(X) > 0$	$\mu_1'(X) = 0$	$\mu_1'(X) < 0$
$\mu_3(X) > 0$	*positively skewed*	*positively skewed*	*Nothing can be in f erred*
$\mu_3(X) = 0$	*positively skewed*	*symmetric*	*negatively skewed*
$\mu_3(X) < 0$	*Nothing can be inf erred*	*negatively skewed*	*negatively skewed*

4. Distributional Properties

Let us see whether negative binomial distribution having pgf $\left(\frac{p'}{1-q't}\right)^{\frac{1}{k}}$ is compound Harris. Consider the pgf of CH distribution, given by

$$Q_Y(t) = p^{\frac{1}{k}}\,[1 - q[Q(t)]^k]^{\frac{-1}{k}}\,, \qquad \text{where Q(t)} = \textstyle\sum_{i=0}^{\infty} q_i'\ t^i$$

$$\text{Then Q(t)} = q^{\frac{-1}{k}}\left[1 - \frac{p}{[Q_Y(t)]^k}\right]^{\frac{1}{k}}$$

$$\text{Taking } Q_Y(t) = \left(\frac{p'}{1-q't}\right)^{\frac{1}{k}}, \quad \text{the pgf of negative binomial distribution.}$$

$$\text{We get Q(t)} = a^{\frac{1}{k}}\,[1 + bt]^{\frac{1}{k}} \qquad \text{where a} = \frac{p'-p}{p'q},\ \text{b} = \frac{p'q}{p'-p}$$

$$= a^{\frac{1}{k}}\left[1 + \binom{1}{k} bt + \frac{\left(\frac{1}{k}\right)\left(\frac{1}{k}+1\right)}{2!} b^2t^2 + \cdots\right]$$

For Q(t) to be a pgf, all terms on the RHS should be positive. But this happens when $p' > p$. Also $Q(1) = 1$. Hence the given negative binomial distribution is CH when the condition $p' > p$ is satisfied.

Relationship between CH and CEG distributions

Sandhya et al.(2019) defined CEG distribution as follows.

A discrete random variable is said to have CEG distribution if it admits the pgf given by $\text{p}[1 - q[Q(t)]^k]^{-1}$ where Q(t) is the pgf of secondary distribution, $k > 0$, integer, $o < p < 1$ and $p + q = 1$.

Let Y_1, Y_2, ... be iid CH random variables on {0,k,2k,…}.Then $\sum_{i=1}^{k} Y_i$ has CEG distribution on {0, k, 2k, ...}.

Proof:

$$\text{We have} \quad Q_{Y_i}(t) = p^{\frac{1}{k}}\,[1 - q[Q(t)]^k]^{\frac{-1}{k}}\,, \quad \text{i= 1, 2, 3,…}$$

$$\text{Then} \quad Q_{\sum_{i=1}^{k} Y_i}(\text{t}) = \textstyle\prod_{i=1}^{k} Q_{Y_i}(t)$$

$$= \text{p}[1 - q[Q(t)]^k]^{-1}$$

which is the pgf of CEG distribution on {0, k, 2k, ...}. Hence $\sum_{i=1}^{k} Y_i$ follows CEG distribution.

Remark 4.1 When N is Harris H_0 $(p, k, \frac{1}{k})$, Sherly et al. (2007) proved that $\frac{N}{k}$ is negative binomial with index parameter $\frac{1}{k}$. But this property is not preserved under compounding.

R and S functions

R and S functions are generating functions which may be used to characterize a probability distribution. The R function is given by

$$R_Y(t) = \frac{q}{p}\ [Q_Y(t)]^k\ [Q(t)]^{k-1}\ Q'(t)\ , \qquad 0 \leq t < 1 \qquad (4.1)$$

All terms in the RHS of (4.1) are absolutely monotone. Steutel (2004) has proved that absolute monotonicity of R- function is a necessary and sufficient condition for the infinite divisibility of a pgf. Thus we can state the following result on CH distribution.

Proposition 4.1 *CH distribution is infinitely divisible.*

The S- function, $S_Y(t)$ of CH distribution is the generating function of the sequence (s_j) where $s_j = q^{\frac{1}{k}}\ q_{j+1}$, $j = 0, 1, 2....$ Here we assume that Q(t) has support {1, 2, ...}.

Then $S_Y(t) = \sum_{j=0}^{\infty} s_j t^j$

$$= \sum_{j=0}^{\infty} q^{\frac{1}{k}}\ q_{j+1} t^j$$

$$= \frac{q^{\frac{1}{k}}}{t}\ \mathrm{Q(t)}$$

$$\Rightarrow t\, S_Y(t) = q^{\frac{1}{k}}\, \mathrm{Q(t)} \qquad (4.2)$$

The following theorem characterizes CH distribution based on its S- function.

Theorem 4.1 *A positive function Q_Y with $Q_Y(1-) = 1$ is the pgf of CH distribution iff Q_Y has the form*

$$Q_Y(t) = Q_Y(0)[1 - [tS_Y(t)]^k]^{\frac{-1}{k}}\ ,\ 0 \leq t < 1$$

with S_Y an absolute monotone function.

Proof: Let $Q_Y(t)$ be the pgf of CH distribution

i.e. $Q_Y(t) = Q_Y(0)\,[1 - q[Q(t)]^k]^{\frac{-1}{k}}$ where $Q_Y(0) = p^{\frac{1}{k}}$

$= Q_Y(0)[1 - [tS_Y(t)]^k]^{\frac{-1}{k}}$ from (4.2)

where $S_Y(t)$ is absolutely monotone by its construction.

On the otherhand, let $Q_Y(t) = Q_Y(0)\,[1 - [t\,S_Y(t)]^k]^{\frac{-1}{k}}$

Substituting for $t\,S_Y(t)$ from (4.2), we get $Q_Y(t) = Q_Y(0)\,[1 - q[Q(t)]^k]^{\frac{-1}{k}}$

which is the pgf of CH distribution.

Remark 4.2 $Q_Y(t) = Q_Y(0)\,[1 - [tS_Y(t)]^k]^{\frac{-1}{k}}$

which in turn imply that $S_Y(t) = \frac{1}{t}\left[1 - \left[\frac{Q_Y(0)}{Q_Y(t)}\right]^k\right]^{\frac{1}{k}}$

5. An AR(1) process corresponding to CH distribution

Satheesh et al. (2006) discuss an AR(1) process wherein the sequence $\{Y_{n,i}\}$ of random variable satisfies

$$\sum_{i=1}^{k} Y_{n,i} = b \sum_{i=1}^{k} Y_{n-1,i} \qquad \text{with probability p}$$

$$= \text{b} \sum_{i=1}^{k} Y_{n-1,i} + \sum_{i=1}^{k} \in_{n,i} \qquad \text{with probability 1-p}$$

with innovation sequence $\{\varepsilon_{n,i}\}$ and some $0 < b < 1$.

Assuming $Y_{0,\,i} \underline{\underline{d}} \in_{n,i}$ and marginal stationarity of $\{Y_{n\ i}\}$, we get for $n = 1$,

$$\Phi(\text{t}) = \left\{\frac{\Phi^{k}(bt)}{a-(a-1)\Phi^{k}(bt)}\right\}^{\frac{1}{k}} , \qquad \text{a}=\frac{1}{p} \tag{5.1}$$

Here $\varphi_{.}(.)$ is the characteristic function of $Y_{n,\,i}$ $\forall\, i = 1, 2, \dots n$. They also mention that for (5.1) to be satisfied, $Y_{1,i}$ is Harris (1,a,k) sum stable. Following the discussion therein, it has been shown that under the assumption $Y_{0,\,i} \underline{\underline{d}} \in_{n,i}$ a sequence $\{Y_{n,i}\}$ of random variable defines the stationary AR(1) scheme (5.1) iff $Y_{n\ i}$ is generalized semi- α- Laplace $\left(\frac{1}{p}, b, k\right)$.

A discrete analogue of (5.1) gives the distribution of $\{Y_{n,i}\}$ as discrete generalized semi- Mittag Leffler (a,b,k) law with pgf $\left\{\frac{1}{1+\psi(1-s)}\right\}^{\frac{1}{k}}$, $0 < S < 1, where\ \psi(1-s)$ satisfies $\psi(1-s) = a\psi\big(b(1-s)\big)$, $ab^2 = 1$, for some $\alpha \in (0,1]$.

Satheesh et al.(2006) also conclude that the discrete analogue of AR(1) scheme (5.1) defines the stationary AR(1) process for all b$\in (0, 1)$ iff $Y_{n,i}$ is discrete generalized ML with pgf

$$\left\{\frac{1}{1+\lambda(1-s)^{\alpha}}\right\}^{\frac{1}{k}}, \text{ k>0 integer , } \alpha \in (0,1] \text{ and } \lambda>0.$$

If $Y_{n,i}$ is to have a finite mean,, the above scheme characterizes the negative binomial law with pgf $\left\{\frac{1}{1+\lambda(1-s)}\right\}^{\frac{1}{k}}$, $\lambda>0$.

When $X_n = \sum_{i=1}^{k} Y_{n,i}$, X_n could be the quantity of water flowing through a river with k tributes or the number of patients in a hospital with k different specialities and so on.

Thus we have the following scheme.

Consider AR(1) process $\{Y_{n,i}\}$ with innovation sequence $\{\in_{n,i}\}$ given by

$$\sum_{i=1}^{k} Y_{n,i} = 0 \qquad \text{with probability p}$$

$$= \sum_{i=1}^{k} Y_{n-1,i} + \sum_{i=1}^{k} \in_{n,i} \qquad \text{with probability 1-p} \tag{5.2}$$

Here k is a positive integer, $\in$ is the innovation sequence and assume that $Y_{0\ i} \underline{\underline{d}}\, \Upsilon_{n-1,i}$, $\forall$ i.

$$\text{Then } [\text{Q}_{\text{Y}}(\text{t})]^{k} = \text{p} + [\text{Q}_{\text{Y}}(\text{t})]^{k}\ [\text{Q}_{\in}(\text{t})]^{k}\ (1-p)$$

$$\Longrightarrow \quad \text{Q}_{\text{Y}}(\text{t}) = \left[\frac{p}{1-q[\text{Q}_{\in}(\text{t})]^{k}}\right]^{\frac{1}{k}} \qquad \text{where q=1-p}$$

Thus we have the following theorem.

Theorem 5.1 *A sequence $\{Y_{n,i}\}$ given by (5.2) defines a stationary AR(1) process for some p iff it is Harris sum of innovations $\{\varepsilon_{n,i}\}$*

6. Reliability

Reliability theory has grown in the last decades into an independent discipline by drawing tools from several areas including mathematics, statistics, probability theory and actuarial science. In the recent past, special roles of discrete distribution is getting recognition from the analysis in the field of reliability theory. In this context, the well known distributions namely, geometric, negative binomial and their compounds are known discrete alternatives for distributions such as exponential, gamma and so on. Reliability classification of compound geometric distribution has been considered by various authors including Shanthikumar (1988), Brown (1990), Cai et al. (2000), Willmot (2001). Brown (1990) demonstrated that compound geometric distribution is NWU. This result was generalized by Cai et al. (2000). Willmot (2002) has derived an explicit convolution representation for the equilibrium residual life time distribution of compound zero modified geometric distribution. Second order reliability properties are proved to be preserved under geometric compounding.

Some of the basic reliability properties of CH distribution are discussed in this section.

Hazard Rate

The hazard function also known as failure rate is defined as the ratio of probability mass function (pmf) and the survival function. As no closed form for the pmf is not available, hazard rate values for given values of Y are calculated and tabulated below. The values are also plotted.

Fig 6.1 Hazard function graph

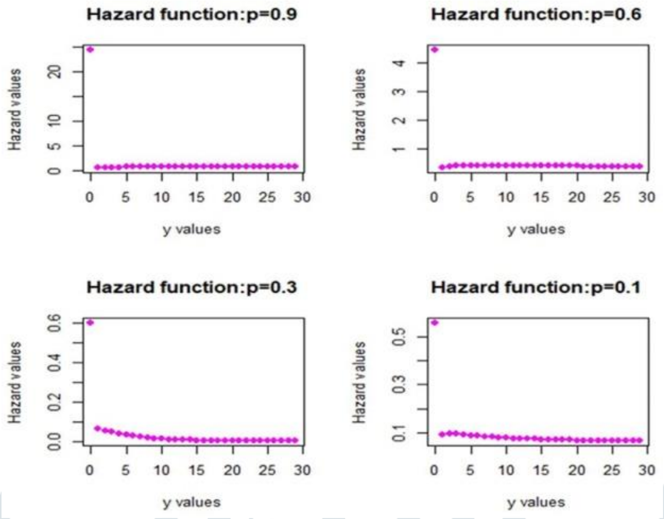


Table 6.1 Hazard function

Hazard rate values at	*p* = 0.9	*p* = 0.6	*p* = 0.3	*p* = 0.1
0	24.4899959	4.4494897	1.5190360	0.560617
1	0.4576713	0.3283632	0.1920512	0.0886152
2	0.5433406	0.3787203	0.2115220	0.0938000
3	0.5970280	0.4006159	0.2171446	0.0936868
4	0.6325582	0.4119836	0.2169806	0.0906794
5	0.6568934	0.4167057	0.2145456	0.0878098
6	0.6739232	0.4178774	0.2113831	0.0850768
7	0.6859900	0.4171841	0.2081523	0.0826301
8	0.6945860	0.4155576	0.2051147	0.0804847
9	0.7007034	0.4135134	0.2023550	0.0786119
10	0.7050240	0.4113336	0.1998825	0.0769730
11	0.7080281	0.4091683	0.1976768	0.0755310
12	0.7100612	0.4070935	0.1957083	0.0742543
13	0.71137506	0.4051436	0.1939467	0.0731165
14	0.7121556	0.4033304	0.1923641	0.0720964
15	0.7125411	0.4016533	0.1909364	0.0711765
16	0.7126352	0.4001059	0.1896427	0.0703427
17	0.7125157	0.3986788	0.1884656	0.0695834
18	0.7122413	0.3973619	0.1873903	0.0688890
19	0.7118562	0.3961450	0.1864042	0.0682513
20	0.7113939	0.3950187	0.1854968	0.0676638
21	0.7108796	0.3939743	0.1846590	0.0671207
22	0.7103321	0.3930037	0.1838831	0.0666772
23	0.7097658	0.3920999	0.1831626	0.0661492
24	0.7091912	0.3912566	0.1824915	0.0657132
25	0.7086160	0.3904682	0.1818651	0.0653062
26	0.7080459	0.3897296	0.1812789	0.0649254
27	0.7074851	0.3890365	0.1807292	0.0645685
28	0.7069364	0.3883848	0.1802126	0.0642336
29	0.7064018	0.3877710	0.1797263	0.0639188

DFR/IFR property

In this section we examine whether the DFR/FR property of secondary distribution is preserved under Harris compounding. Willmot et al. (2000) has established that compound geometric distribution is DFR if the secondary distribution consideredis DFR. But this is not the true in the case of CH distribution. We have $Q_Y(t) = \left(\frac{p}{1-q[Q(t)]^k}\right)^{\frac{1}{k}}$ is the pgf of CH random variable Y, where k is positive integer and Q(t) is the pgf of a DFR random variable. Here $[Q(t)]^k$ is the pgf of the sum $\sum_{i=1}^{k} X_i$ where $X_i\ \ s'$ are iid random varaibles. Let $Q_1(t) = [Q(t)]^k$ and $Q_1(t)$ is the pgf of DFR random variables as DFR property is preserved undersummation.

$$\text{Now } [Q_Y(t)]^k = \frac{p}{1-qQ_1(t)}$$

Here LHS is the pgf of $\sum_{i=1}^{k} Y_i$. But it is obtained as the pgf of compound geometric distribution. Hence$\sum_{i=1}^{k} Y_i$ is compound geometric which is DFR, provided the secondary distribution is DFR (Willmot et al. (2001)). This does not mean that each Y_i is DFR. Hence in general, CH distribution is not DFR even if the secondary distribution is DFR.

Result 6.1 *CH class is not a subclass of DFR class in general.*

The following example illustrate this result.

Let Q(t)=t^2 be the pgf of secondary distribution.

$$\text{Then } Q_Y(t) = p^{\frac{1}{k}}\left[1 + \left(\frac{1}{k}\right) q\, t^2 + \frac{\left(\frac{1}{k}\right)\left(\frac{1}{k}+1\right)}{2!}\, q^2 t^4 \ + \frac{\left(\frac{1}{k}\right)\left(\frac{1}{k}+1\right)\left(\frac{1}{k}+2\right)}{3!} q^3 t^6 \ + \cdots\right]$$

$$\Longrightarrow g_0 = p^{\frac{1}{k}}$$

$$g_2 = p^{\frac{1}{k}}\left(\frac{1}{k}\right) q$$

$$g_4 = p^{\frac{1}{k}} \frac{\left(\frac{1}{k}\right)\left(\frac{1}{k}+1\right)}{2!}\ q^2$$

And $g_1 = g_3 = g_5 = 0$

Then the tail probabilities are given by

$$a_0 = a_1 = 1 - p^{\frac{1}{k}}$$

$$a_2 = a_3 = 1 - p^{\frac{1}{k}} - p^{\frac{1}{k}}\left(\frac{1}{k}\right) q$$

$$a_4 = 1 - p^{\frac{1}{k}}\left[1 + \left(\frac{1}{k}\right) q \ + \frac{\left(\frac{1}{k}\right)\left(\frac{1}{k}+1\right)}{2!}\ q^2\right] \text{ and so on.}$$

Now $\frac{a_1}{a_0}$ =1, $\frac{a_2}{a_1} < 1 \ \ \Longrightarrow \frac{a_1}{a_0} > \frac{a_2}{a_1}$

$\Longrightarrow \frac{a_{n+1}}{a_n}$ is decreasing for n=0 which means that Y is not DFR.

Result 6.2 *CH distribution with IFR secondary distribution need not be IFR.*

The following example will illustrate this result.

Example :

Let the pmf values of secondary distribution be

$$q_0 = \frac{1}{4},\ q_n = \frac{3}{2^{n+2}}, \text{ n=1, 2, …}$$

$$\frac{q_1}{q_0} = \frac{3}{2},\ \frac{q_2}{q_1} = \frac{1}{2}$$

$$\frac{q_1}{q_0} > \frac{q_2}{q_1}$$

Also $\frac{q_{n+2}}{q_{n+1}} \le \frac{q_{n+1}}{q_n}$ for all n=0, 1, 2, ….
which implies that the secondary distribution is IFR.
Here Q(t) = $\frac{1+t}{2(2-t)}$

Then $Q_Y(t) = p^{\frac{1}{k}} [1 - q(1+t)^k 2^{-k} (2-t)^{-k}]^{\frac{-1}{k}}$

$$g_0 = p^{\frac{1}{k}} \left[1 - \frac{q}{2^{-2k}}\right]^{\frac{-1}{k}}$$

$$g_1 = 3p^{\frac{1}{k}}\, q\, 2^{-2k-1} \left[1 - \frac{q}{2^{2k}}\right]^{-\left(\frac{1}{k}+1\right)}$$

$$g_2 = 3p^{\frac{1}{k}}\, q\, 2^{-2k-2} \left[1 - \frac{q}{2^{2k}}\right]^{-\left(\frac{1}{k}-2\right)} [4q2^{-2k} + (3k-1)]$$

$$\frac{g_2}{g_1} = 2^{-1} [1 - q2^{-2k}]^{-1} [4q2^{-2k} + (3k-1)]$$

$$\frac{g_1}{g_0} = 3\, q\, 2^{-2k-1} \quad [1 - q2^{-2k}]^{-1}$$

$$\Longrightarrow \frac{g_2}{g_1} - \frac{g_1}{g_0} > 0 \quad \text{or} \quad \frac{g_2}{g_1} > \frac{g_1}{g_0}$$

which implies that Y is not IFR.

7. Simulation and Estimation

Parameter estimation enables inferences to be made regarding an unknown population from which data are observed. One of the popular classical estimation methods especially for well- behaved data sets, is the maximum likelihood estimation (MLE). Moment estimation is the well-known easiest method of estimation. The use of pgf in statistical inference has been proposed as a tool in estimation due to its simplicity compared to pmf in many instances. For CH distribution, the pmf and hence likelihood function has no closed form, but pgf has. Density based divergences such as BHHJ density power divergence proposed by Basu et al. (1998) is a familiar measure used in parameter estimation. It relies on a tuning parameter, say, ɑ, which may take any value greater than or equal to zero. It is preferable to have $0 \leq \alpha \leq 1$, to control the trade- off between robustness and efficiency. Considering the simplicity of pgf compared to pmf, a pgf- based BHHJ divergence proposed by Ying et al. (2016) is used here.

Hence we use moment estimation method and pgf- based BHHJ divergence method to estimate the parameters *k*, *p*, and p ′.

Moment estimation

Moment estimation of the parameters is done using simulated data in two cases.

Let m_1, m_2, m_3 denote the first, second and third raw moments of the observed data and let $X \sim Geom\,(p')$ with $q' = 1 - p'$

Case 1: When *p* is unknown.

In this case, the moment estimator of *p* is obtained by solving the equation

$$p\,p'\, m_1 - \; q\,q' = 0$$

The solution is given by $\hat{p} = \frac{q'}{p' m1 + q'}$

Table 7.1 Moment estimates using simulated sample of size 70 , no. of replications 50

$p' = 0.5$

p		*k*=2	*k*=3	*k*=4
	Estimate	0.8074814	0.0.8164207	0.8214758
0.8	Mean bias	0.0074814	0.0164207	0.0214757
	MSE	0.00066896	0.0091760	0.0113855
	Estimate	0.5130934	0.5158931	0.5187968
0.5	Mean bias	0.0130934	0.0158930	0.0187968
	MSE	0.0067074	0.0085674	0.0109913
	Estimate	0.2079216	0.2112153	0.2148126
0.2	Mean bias	0.0079216	0.0112153	0.0148126
	MSE	0.0015232	0.0024962	0.0036479

Case 2: When all the parameters are unknown.

The estimates are obtained by solving the equations using R package namely, ”nleqslv”.The equations are given by

$$p\,p'm_1 - \; q\,q' = 0$$

$$p^2\,(p')^2\,m_2 - (k\,q + q^2)(\,q')^2 - \; p\,q\,q' = 0$$

$$p^3\,(p')^3\,m_3 - (k^2\,q + (k^2 + 3k)\,q^2 + q^3)(\,q')^3 - ((3\,k\,p + p^2)\,q + 3pq^2)\,(q')^2 - q\,p^2q' = 0$$

Table 7.2.

Moment estimates using simulated sample of size 100 and no. of replications 20.

			k = 2			k = 3		
p'	$\hat{k}$		$\hat{p}$	$\widehat{p'}$	$\hat{k}$	$\hat{p}$	$\widehat{p'}$	
		Mean estimate	1.5164424	0.2080640	0.2135049	1.9042805	0.2132912	0.2227791
0.2		*Mean bias*	0.4835575	−0.0080640	−0.0135049	1.0957194	−0.0132291	−0.0227791
		MS E	0.3503904	0.0002091	0.0002985	1.2959158	0.0004110	0.0007374
		Mean estimate	2.8191141	0.3388437	0.3502273	3.6960500	0.3169342	0.3310487
0.2	0.5	*Mean bias*	−0.8191141	0.1611562	−0.1502273	−0.6960500	0.1830657	−0.1310487
		MS E	1.1385585	0.0266163	0.0234378	0.9970731	0.0340820	0.0180187
		Mean estimate	5.8810397	0.5092079	0.5359977	10.7943664	0.5397125	0.5687469
	0.8	*Mean bias*	−3.8810397	0.2907920	−0.335997	−7.6443664	0.3002874	−0.3587469
		MS E	21.4376947	0.0875494	0.1159118	78.3795997	0.0898922	0.1272151
		Mean estimate	1.438309	0.338274	0.341241	2.211987	0.338938	0.343955
	0.2	*Mean bias*	0.561691	0.138274	0.158758	0.788012	0.138937	0.156045
		MS E	0.465817	0.019449	0.025618	0.3081907	0.1091214	0.0023525
		Mean estimate	1.834514	0.505018	0.520036	3.151498	0.489018	0.500321
0.5	0.5	*Mean bias*	0.165486	0.00502	0.02004	0.151500	0.010982	0.000320
		MS E	0.437821	0.000977	0.001606	0.9441009	0.001367	0.001845
		Mean estimate	3.684506	0.659172	0.683049	4.571494	0.668147	0.619301
	0.8	*Mean bias*	1.684506	0.1408275	0.183049	1.571494	0.131852	0.119930
		MS E	0.437821	0.000977	0.001606	0.9441009	0.001367	0.001845
		Mean estimate	0.8020734	0.5056412	0.5104788	1.5026967	0.4882644	0.4928845
	0.2	*Mean bias*	1.1979266	0.3056412	0.2895211	1.4973032	0.2882644	0.3071154
		MS E	1.5819736	0.0939173	0.0845165	2.512619	0.0836085	0.09518216
		Mean estimate	1.1053586	0.6599926	0.6724095	1.5285452	0.6633971	0.6851016
0.8	0.5	*Mean bias*	0.8946413	0.1599926	0.1275904	1.4714547	0.16339712	0.1148984
		MS E	1.1415040	0.02668992	0.0178603	2.9954447	0.0281028	0.0153317
		Mean estimate	1.9752651	0.7791797	0.8073766	2.1911323	0.7813517	0.7276945
	0.8	*Mean bias*	0.0247048	0.0208202	0.0073766	0.8088676	0.0186482	0.0723054
		MS E	1.7345995	0.0010850	0.0009111	4.5742116	0.0023653	0.0632000

Remark 7.1 Moment estimation gives best results when p, and p' are almost equal to 0.5, as evident from table 7.2

BHHJ estimation

BHHJ divergence based on pgf is defined as

$$d_\alpha(g_n, g) = \int_0^1 \left[g^{1+\alpha}(t;\theta) - \left(1 + \frac{1}{\alpha}\right) g_n(t) g^\alpha(t;\theta) + \frac{1}{\alpha} {g_n}^{1+\alpha}(t) \right] dt \text{ , } \alpha>0 \qquad (7.1)$$

$where\ g(t;\theta) = E_\theta(t^x)$, ($\theta \in \Theta$, the parameter space) is the pgf and $g_n(t) = \frac{1}{n}\sum_{i=1}^{n} t^{x_i}$, 0<t<1 is the empirical probability generating function (epgf). BHHJ estimates are obtained by minimizing equation (7.1) using ”nloptr” package in R.

Case 1: When *p* is unknown.

Table 7.3 BHHJ estimates using simulated sample of size 70 , no. of replications 50

p' = 0.5

p		*k*=2	*k*=3	*k*=4
	Estimate	0.7925005	0.8042435	0.8138073
0.8	Mean bias	-0.00074994	0.0042435	0.0138073
	MSE	0.0005762	0.0082556	0.0111561
	Estimate	0.509999	0.5086318	0.5071159
0.5	Mean bias	0.0099990	0.0086317	0.0071159
	MSE	0.0060967	0.0078889	0.01037649
	Estimate	0.2065957	0.2118345	0.2200707
0.2	Mean bias	0.0065956	0.0118344	0.0.020070
	MSE	0.001983	0.0033062	0.0054444

Case 2: When all the parameters are unknown.

Table 7.4 BHHJ estimates using simulated sample of size 100 and no. of replications 20 (p' = 0.5)

		k = 2			*k* = 3		
p		$\widehat{k}$	$\hat{p}$	$\widehat{p'}$	$\hat{k}$	$\hat{p}$	$(\widehat{p'}$
	Mean estimate	3.2125076	0.0539601	0.1095357	3.0534499	0.0470384	0.0596125
0.2	*Mean bias*	–1.2125076	0.1460398	0.3904642	0.0965500	0.1629615	0.4653874
	MS E	3.2178643	0.0250673	0.1761482	0.6967293	0.0275795	0.2128576
	Mean estimate	1.3039595	0.3850635	0.4332183	1.4302299	0.3315287	0.3726456
0.5	*Mean bias*	0.6960404	0.11493646	0.0667816	–1.5697008	0.1684712	0.1273543
	MS E	0.5149379	0.0137210	0.0049671	2.5094628	0.0285802	0.0165706
	Mean estimate	3.0544282	0.5965741	0.6453194	3.9033686	0.4944442	0.5944632
0.8	*Mean bias*	–0.9544282	0.2434258	–0.1203194	–0.9033686	0.3055557	–0.0944637
	MS E	4.7671513	0.0850707	0.0272367	5.87606592	0.11205037	0.02428556

8. Real life data set

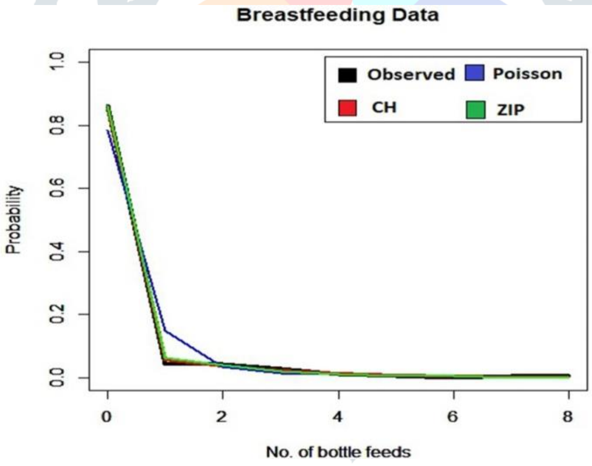
International health authorities recommend that infants be exclusively breastfed for 6 months, then introduction of complementary foods and continued breastfeeding until 12 months of age and then thereafter as long as mutually desired. In order to determine factors affecting the frequency of formula feeds by breastfeeding women, a longitudinal infant feeding study was conducted in Perth, Australia in 1992-93. The analysis was based on 209 subjects and the variable of interest was the number of bottle feeds that an infant had received in the 24th week of birth. Lee et al.(2006) used this data to fit zero-inflated Poisson (ZIP) regression model. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are not used here to measure goodness of fit as the the likelihood function cannot be put in a closed form. Hence fitting of distributions are done using Chi square test. Expected frequencies for fitting Poisson and ZIP models are displayed in columns 4 and 5 of table 7.1 as given by Lee et al. The values of *k*, *p*, and p^{J} of CH distribution are obtained as k∼2, p = 0.624041, p' = 0.633006 by method of moments. The given data is overdispersed. ZIP model is also overdispersed. The p values show that CH distribution provide a better fit.

Table 8.1 Fitting of CH distribution.

No. o f bottle feeds	*Observed frequency*	*Expected frequency*		
		CH	*Poisson*	*ZIP*
0	180	179.1466	164	180
1	9	11.6610	31	13
2	9	7.5578	7	8
3	6	4.5181	3	4
4	2	2.6174	2	2
5	1	1.4983	1	1
6	0	0.8549	1	1
7	1	0.4880	0	0
8	1	0.6579	0	0
Total	209	209	209	209
Chi square value		1.576329	20.03102	2.480769
d. f		1	2	1
p value		0.2092903	0.0000447	0.1152459

Here the estimate of $k = 2$ has a practical significance. When carrying out the study, the sampling can be done in groups of two, which will lead to less labour in sampling but more efficient results.

Fig 8.1 Breastfeeding Graph



Conclusion

Compounding of probability distributions, especially discrete distributions, gives rise to a richer class of probability distributions other than classical distributions. Keeping this in mind, we have introduced a new distribution, namely, CH distribution, by compounding Harris distribution with a standard discrete distribution.As the derivation of pmf of the distribution is not easy, we have evaluated the CH probabilities using FFT technique.The graphs of pmf values show that the mode is at $Y = 0$ and the distribution is skwed. The overdispersion property of the distribution gives an indication to the applicability of the distribution in the fields like actuarial science, medical science, biology etc. Discussion on the basic reliability properties brings out the point that DFR property of secondary distribution is not preserved under Harris compounding as geometric compounding. Moment estimation gives best results particularly when the probability p lies in the neighborhood of 0.5. A real life data set of breast feeding infants in Australia is analyzed and CH distribution is seen to be a best fit for the data. We may be able to use this model in similar situations when the data is zero-inflated.

REFERENCES

Andy, H. Lee, Jane A Scott, Kelwin K W Geoffrey JM .(2006). Multi- level zero- inflated Poisson regression modeling of correlated count data with excess zeros, Statistical methods in medical research, 15, p: 47-61. doi:10.1191/0962280206sm429oa

Basu, A., Harris, I.R., Hjort, N.L, and Jones, M.C., (1998). Robust and efficient estimation by minimizing a density power divergence, Biometrika, 85, p: 549-559.

Brown, M., (1990). Error bounds for exponential approximations of geometric convolutions, The Annals of Probability, 18, p: 1388-1402.

Cai, J. and Kalashnikov, V., (2000). NWU property of a class of random sums, Journal of Applied probability, 37, p: 283-289.

Harris, T. E., (1948). Branching Processes, Annals of Mathematical Statistics, 19, p: 474-494.

J. George Shanthikumar, (1988). DFR Property of First-Passage Times and its Preservation Under Geometric Compound- ing, The Annals of Probability, 16, p: 397-406.

Kovats, A., and Mori, T. F. (1992). Ageing properties of certain dependent geometric sums, Journal of Applied Probability, 29, p: 655666. doi:10.1017/s0021900200043473

Kumer Pial Das, Shamim Sarker, and norou Diawara, (2011). Further review of Panjer's recursion for evaluation of compound negative binomial distribution using R, 23.

Panjer, H.H., (1981). Recursive evaluation of a family of compound distributions, ASTIN Bulletin, 12, p:22-26.

Paul Embrechts and Marco Frei, (2009). Panjer recursion versus FFT for compound distributions, Mathematical Methods of Operations Research, 69, p: 497-508.

Romeo Mestrovic, (2015). On some compound random variables motivated by Bulk queues, Hindawi publishing corpo- ration Mathematical problems in engineering, http://dx.doi.org/10.1155/2015/291402

Sandhya, E.,(1996). On a generalization of geometric infinite divisibility, Proceedings of the 8th Kerala Science Congress, January-1996, p: 355357.

Sandhya, E. and Latha, C.M., (2019). Compound extended geometric distribution and some of its properties, International Journal of Statistics and Probability, 8.

Satheesh, S., Nair N. U., Sandhya E., (2002). Stability of random sums, Stochastic Modeling and Applications, 5, p: 1726.

Satheesh, S., Nair N. U., (2002a). A Note on maximum and minimum stability of certain distributions, Calcutta Statistical Association Bulletin, 53, p: 249-252.

Satheesh, S., Nair N. U., (2004). On the stability of geometric extremes, Journal of the Indian Statistical Association, 42, p: 99109.

Satheesh, S., Sandhya E., Sherly Sebastian, (2005). Time series models motivated by the stability of Harris sums and extremes, Submitted.

Satheesh, S., Sandhya E., Sherly Sebastian, (2006). A generalization of Stationary AR(1) Schemes, Statistical Methods, 8(2), p: 213-225

Sherly Sebastian (2007). Harris family of discrete distributions and processes, Department of Statistics, University of Calicut. http://hdl.handle.net/10603/61908.

Steutel F.W. and van Harn k., (2004). Infinite Divisibility of Probability Distributions on the Real Line, Pure and Applied Mathematics, p: 259.

Tay Siew Ying, Ng Choung Min, and Ong Seng Huat (2016). Parameter estimation using probability generating function based minimum power divergence. American Institute of Physics, AIP Conference Proceedings 1750, 060008, doi: 10.1063/1.4954613

Vergnes, J.N., Boucher, J.P., Lelong, N., Sixou, M., and Nabet, C., (2016). Discrete Distribution Based on Compound Sum to Model Dental Caries Count Data, Caries Research, 51(1), 6878. doi:10.1159/000450891

Willmot, G.E., (1989). Limiting tail behavior of some discrete compound distributions, Insurance: Mathematics and Economics, 8, 3, p: 175-185. doi:org/10.1016/0167-6687(89)90055-3

Willmot, G. E., and Cai, J., (2001). Ageing and other distributional properties of discrete compound geometric distribu- tions, Insurance Mathematics and Economics, 28, p: 361-379.

Willmot, G.E., Lin, X.S., 2001. Lundberg approximations for compound distributions with applications. Springer-Verlag, New York.

Willmot, G. E., (2002). On higher order properties of compound geometric distributions, Journal of Applied Probability, 39, no.2, p: 324-340. doi: 10.1239/jap/1025131429