Compound Extended Geometric Distribution and Some of Its Properties

Sandhya E[1] & Latha C M[2]

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

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

Correspondence: Latha C M, Department of Statistics, St.Thomas College, Pala-686574, India

Received: March 22, 2019 Accepted: April 25, 2019 Online Published: April 28, 2019

doi:10.5539/ijsp.v8n3p96 URL: https://doi.org/10.5539/ijsp.v8n3p96

Abstract

The compound extended geometric distribution (CEG distribution) is introduced. Probabilities of the distribution are evaluated . Certain statistical, distributional and reliability properties are discussed. The distribution is characterized using S- function. An AR(1) process corresponding to CEG distribution is derived. Simulation and estimation of parameters are done using the method of moments and BHHJ estimation.

Keywords: compound extended geometric distribution, Panjer's recursion, Fast Fourier Transform (FFT), R and S functions, reliability, DFR and IFR, moment estimator, probability generating function (pgf) based BHHJ estimator

1. Introduction

From the last few decades, researchers are busy to obtain new probability distributions by using different techniques. Compounding of probability distributions has received attention since it is an innovative and sound technique to obtain new probability distributions. In several research papers, it has been found that compound distributions are very flexible and can be used efficiently to model different types of data sets. As a result, many compound distribution has been constructed. Compound distributions arise when either

(1) all or some parameters of a distribution vary according to some probability distribution (mixture) or

(2) the number of random variables N in a sum of independent and identically distributed (i.i.d.) random variables $\sum_{i=1}^{N} X_i$, is again a random variable. The distribution of N is referred to as primary distribution and that of X_i is referred to as secondary distribution (Willmot et al.(2012)). In this paper , by compounding we mean the second one.

The compounding of probability distributions enables us to obtain both discrete as well as continuous distributions. Here we consider the compounding with discrete primary and secondary distributions. The compound Poisson distribution is often a popular choice for modeling aggregate claims in insurance. The compound negative binomial model arises naturally in several fields such as insurance mathematics and actuarial science and has been studied by several authors. Furthermore, compound geometric distribution as a special case of compound negative binomial distribution plays a vital role in analysis of ruin probabilities and related problems in risk theory. It has many reliability, queueing and insurance applications also. Some of these applications are discussed by Schafer (1996), Willmot et al.(1997), Tang (2005) and others. Even though compound Poisson distribution has many attractive properties, it is not a good model when the variance of N is greater than the mean of N. In such cases, compound negative binomial distribution is a better fit. The CEG distribution that we introduce here may also be a better fit in such cases.

Pekoz and Ross (2004) derived an identity concerning the expectation of an arbitrary function of a compound random variable and use this identity to obtain recursive formulae for the probability mass function (pmf) of compound random variable when the compounding distribution is Poisson, binomial, negative binomial etc. Numerical evaluation procedures are often necessary for most compound distributions. Recursive evaluation procedures may be used if the primary distribution belongs to the Sundt-Jewell class (Willmot et al.(1988)). Such a recursive procedure was introduced by Panjer (1981). The use of this algorithm has become a widespread standard technique for the life and general insurance problem.Another approach is to use FFT to evaluate the probabilities by inverting the characteristic function.

Infinitely divisible distributions form an important class of distribution on $\mathbb{R}$ that include the compound Poisson distribution, as well as several of the most important special parametric families of distribution. The concept of infinite divisibility is useful in characterizing compound distributions. Feller's (1968) characterization of compound Poisson distribution states that a non-negative integer valued random variable is infinitely divisible if and only if its distribution is discrete compound. More probabilistic properties of compound distributions are discussed by many authors. Reliability classifications are useful tools in analysis of compound distributions and they are considered in a number of papers. Brown (1990) proved that compound geometric distribution is new worse than used (NWU). This result was generalized by Cai et al.(2000). Distributional and ageing properties of compound geometric distribution have been discussed also by Willmot et al.(2001).

Maximum likelihood is often used to estimate the unknown parameters in models because it provides asymptotically unbiased estimators which have the lowest possible asymptotic variance. For distributions having no closed form for the pmf, estimating unknown parameters using this method is not practical. Minimum distance estimators are of interest because they have the the desirable properties of being both robust and efficient. Basu et al.(1998) proposed a generalized Hellinger divergence involving densities, for count data. Along similar lines but employing pgf, a pgf-based Minimum Hellinger Divergence (MHD) type estimation is proposed by Sim et al.(2010). Ying et al.(2016) proposed a pgf -based minimum power divergence for parameter estimation and they called this method as BHHJ-pgf divergence method.

This work is on compound extended geometric distribution with discrete secondary (severity) distribution. In section 2, we introduce compound extended geometric (CEG) distribution and evaluate probabilities using FFT. Statistical, distributional and reliability properties are discussed in sections 3, 4, 5 and 6 respectively. Section 7 discusses simulation and estimation of parameters using moment method and BHHJ-pgf divergence method. Fitting of the distribution using a real life data is done in section 8.

2. Compound Extended Geometric Distribution and Numerical Evaluation of Its Probabilities

Definition 2.1

A discrete random variable N is said to have extended geometric (EG) distribution if its pmf is given by

$$Pr[N = x] = pq^x, x = 0, k, 2k, ...$$

where k is a positive integer. Sandhya et al.(2006) has established the connection between a negative binomial random variable and geometric random variable by considering the concept of fractional successes. There are situations where a success can occur only when k (> 1 , *an integer*) fractional successes happen. For example, suppose a sales representative receives an incentive only when he sells k identical items. Here selling of each item is a fractional success and obtaining an incentive is a success and success occurs only when the k fractional successes happen i.e. when k items are sold. In similar cases extended geometric distribution may be used to model the number of successes.

A random variable Y has a compound distribution if $Y = \sum_{i=1}^{N} X_i$ where the number of terms N is a discrete random variable whose support is the set of all non negative integers. It is assumed that X_i are i.i.d. random variables and each X_i is independent of N. When N has EG distribution, Y is said to have CEG distribution.

Definition 2.2

A counting random variable Y is said to have CEG distribution if its probability generating function (pgf) admits the presentation

$$Q_Y(t) = \frac{p}{1 - q[Q(t)]^k}, 0 < p < 1, p + q = 1, \quad k, positive\ integer \tag{2.1}$$

where $Q(t) = \sum_{i=0}^{\infty} q_i t^i$ is a pgf. Here Y has support {0,1,2,...} and $Y = \sum_{i=1}^{N} X_i$ is a random sum, $Q(t)$ is the pgf of X_i , N is EG random variable and each X_i is independent of N. Thus we write Y∼ CEG *(k, p, p′)* where p' is the parameter corresponding to X_i .

Remark 2.1 When $Q(t)$ corresponds to a distribution which is having additive property (eg. Binomial), (2.1) becomes nothing but compound geometric. When X_i does not have the additive property (eg. Geometric) and $[Q(t)]^k$ conforms to the pgf of a standard distribution, then also (2.1) can be interpreted as compound geometric. But when X_i does not have additive property (eg. Uniform) and $[Q(t)]^k$ does not conform to a standard distribution, the compound extended geometric distribution becomes more relevant.

Obviously, the CEG probabilities are the coefficients of $t^0, t^1, t^2 ...$ in the expansion of $Q_Y(t)$. They are given by

$$
\begin{aligned}
g_0 &= P_r\left[Y = 0\right] = p(1 - qq_0^k)^{-1} \\
g_y &= P_r\left[Y = y\right], y \geq 1 \ \ is \ \ the \ \ sum \ \ of \ \ y \ \ terms. \\
&= p \sum_{i=1}^{\infty} (qq_0^k)^i [\binom{ik}{1} \frac{q_y}{q_0} \\
&+ \sum_{j=2}^{y} \binom{ik}{j} \sum_{a_1=1}^{y-(j-1)} \sum_{a_2=1}^{y-(j-2)-a_1} \cdots \sum_{a_{j-1}=1}^{y-1-(a_1+...a_{j-2})} \left(\frac{q_{a_1} q_{a_2} ... q_{a_{j-1}} q_{y-\sum_{i=1}^{j-1} a_i}}{q_0^j} \right)], y \geq 1
\end{aligned} \tag{2.2}
$$

Thus

$$
\begin{aligned}
g_0 &= p(1 - qq_0^k)^{-1} \\
g_1 &= pqq_1 q_0^{k-1} \ k(1 - qq_0^k)^{-2} \\
g_2 &= p \ q \ q_2 \ q_0^{k-1} \ k \ (1 - qq_0^k)^{-2} + p \ q \ q_1^2 \ q_0^{k-2} \left[\frac{k^2(1 - qq_0^k)^{-3}(1 + qq_0^k) - k(1 - qq_0^k)^{-2}}{2} \right]
\end{aligned}
$$

and so on.

Note: The compound geometric probabilities are obtained by putting k=1 in (2.2).

3. Evaluation of Probabilities

Numerical evaluation is often necessary for most compound distributions, since the pmf has no closed form especially for discrete cases. Even in the case of compound geometric distribution, the derivation of general form of pmf is not seen in any literature,to the best of our knowledge. Since the original 1981 paper of Panjer, considerable attention has been paid to Panjer's recursive formula. The Panjer's recursive formula can be successfully used to evaluate compound geometric probabilities. In practice, both recursive methods as well as transform based techniques are widely used. Willmot et al.(1988), Sundt (1992) gave modifications to this formula. The Fast Fourier Transform (FFT) technique is a viable alternative. Here we evaluate the CEG probabilities using FFT , from the pgf of Y using the discrete Fourier Transform, following Embrechts et al.(2009). For example, assume that Q(t) is the pgf of a binomial distribution with success probability $p\prime = 0.5$.

The following R commands are used to evaluate the corresponding CEG probabilities. **CEG (*k, p, 0.5)***

1. $M = 2^6$
2. $k = 2$
3. $f \leftarrow \ dbinomial\,(0 : (M - 1), size = 10, prob = 0.5)$
4. $fhat \ \leftarrow \ fft(f, inverse = FALSE)$
5. $fkhat \ \leftarrow \ fhat * fhat$
6. $u \ \leftarrow \ \frac{p}{1-(q*fkhat)}$
7. $g \ \leftarrow \ (1/M) * fft(u, inverse = TRUE)$

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 CEG probabilities with(1) Binomial, (2) Geometric and (3) Uniform secondary distributions are plotted here.

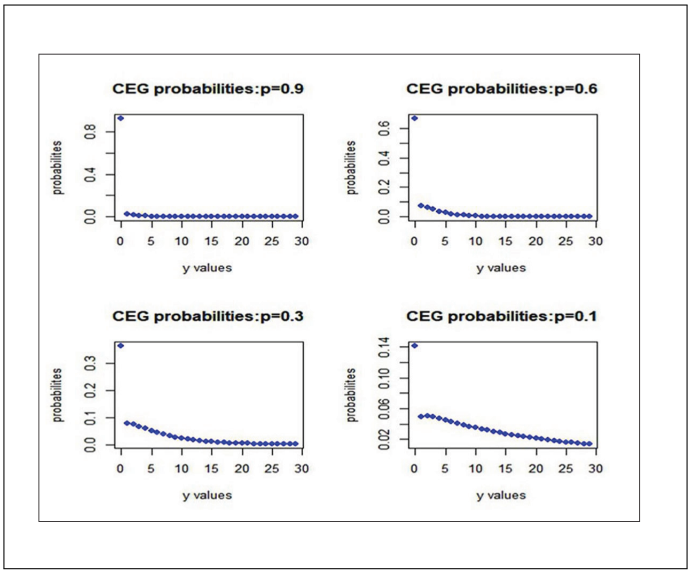
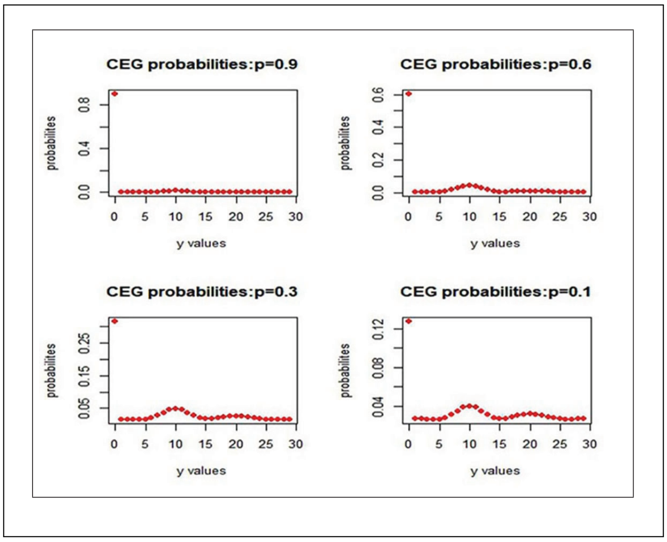


Figure 3.1. CEG with the secondary distribution geometric *(k=2, p'=0.5)*



1

Figure 3.2. CEG with the secondary distribution Binomial *(k=2, p'=0.5, size=10)*

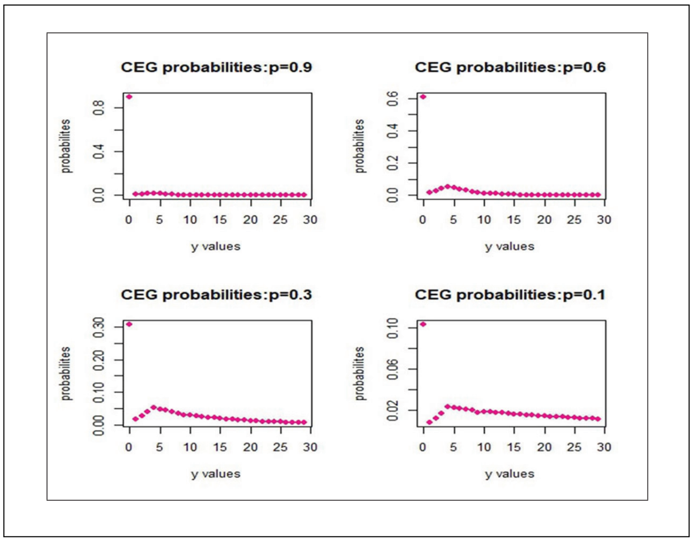


Figure 3.3. CEG with the secondary distribution Uniform *(k=2, p′=0.2, N=5)*

Remark 2.2 Irrespective of any secondary distribution, CEG distribution has mode at $Y = 0$.

Remark 2.3 Probabilities are spread out more to the right tail of the distribution in the case of binomial secondary distribution than geometric and uniform secondary distributions.

Remark 2.4 Even though extended geometric distribution has its probabilities at 0, *k, 2k ...* , CEG distribution has probabilities at all points 0, 1, 2 ...

4. Statistical Properties

Quantiles

Quantiles are useful measures because they are less susceptible than means to long-tailed distributions. 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. We have simulated 50 samples each of size 70 from CEG distribution (p = 0.9, 0.6, 0.3, 0.1) with geometric distribution (probability = 0.5) as secondary distribution and the quantile values at different probabilities are tabulated below.

Table 4.1. 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=0.9	0	0	0	0	0	0	0	0	5
p=0.6	0	0	0	0	0	1	2	5	12
p=0.3	0	0	1	2	4	6	8	13	28
p=0.1	2	5	8	12	16	22	30	44	87

Table 4.2. Quantiles for *k*=4

				p_0					
probabilities	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99
p=0.9	0	0	0	0	0	0	0	0	9
p=0.6	0	0	0	0	0	3	5	9	21
p=0.3	0	0	3	5	8	12	16	25	53
p=0.1	5	10	16	23	31	41	55	76	118

Remark 3.1 It is evident from the tables that skewness of CEG distribution becomes higher as the parameter p increases.

For *p=0.9* the quantile values up to 0.9 are zero whereas for *p=0.1* the quantile value corresponding to 0.2 itself is not zero irrespective of the value of *k*.

Cumulants

Cumulant generating function of CEG distribution is given by

$\Phi_Y(t) \;=\; \ln(p) \;-\; \ln\;(1 - q[M_X(t)]^k)$ where $M_X(t)$ is the moment generating function of secondary distribution.

Cumulants are derived using the general expression for deriving cumulants of random sum, proposed by Naoto Niki et al.(1990).

Let k_r, α_r, ν_r denote the r^{th} cumulants of Y, X and N respectively then,

$$k_1 = \nu_1\,\alpha_1$$
$$k_2 = \nu_2\,\alpha_1^2 \;+\; \nu_1\,\alpha_2$$
$$k_3 = \nu_3\alpha_1^3 + 3\nu_2\alpha_1\alpha_2 + \nu_1\alpha_3$$

Considering geometric distribution as secondary distribution,

$$\alpha_1 \;=\; \frac{q'}{p'}\;,\; \alpha_2 \;=\; \frac{q'}{(p')^2},\; \alpha_3 \;=\; \frac{q'(1+q')}{(p')^3}$$

$$\nu_1 = \frac{kq}{p}$$
$$\nu_2 = \frac{k^2q}{p^2}$$
$$\nu_3 = \frac{k^3q(1+q)}{p^3}$$

As *mean* = k_1, *variance* = k_2 and $\mu_3 = k_3$ we have:

EG distribution	
mean	$\frac{kq}{p}$
variance	$\frac{k^2q}{p^2}$
CEG distribution	
mean	$\frac{kq'}{pp'}$
variance	$\frac{k^2q(q')^2+kpqq'}{(pp')^2}$

$$k_3 = \frac{k^3q(1+q)(q')^3 + 3k^2pq(q')^2 + kp^2qq'(1+q')}{(pp')^3} \tag{3.1}$$

Putting *k*=1, we get the moments of compound geometric distribution.

Also coefficient of variation is given by

$$(CV_Y)^2 = \frac{p}{kqq'}\left[1 + \frac{kq'}{p}\right]$$

4. A Characterization of CEG Distribution

From (2.1), it follows that if $Q_Y(t)$ is a pgf of CEG (k, p, p'), then there exists another pgf Q(t) such that

$Q(t) = q^{\frac{-1}{k}}\left[1 - \frac{p}{Q_Y(t)}\right]^{\frac{1}{k}}$ for given values of the parameters.

Now, let us check whether a geometric distribution on $\{0,\ 1,\ 2, ...\}$ is CEG. Then

$$\begin{aligned} Q(t) &= q^{\frac{-1}{k}}\left[1-\frac{p(1-q\prime\,t)}{p\prime}\right]^{\frac{1}{k}} \ \textit{should be a pgf.} \\ &= a[1+bt]^{\frac{1}{k}} \ \textit{ should be a pgf, where } a = \frac{p\prime\,-p}{p\prime\,q},\ b = \frac{p\,q\prime}{p\prime - p} \\ \textit{That is,} \quad Q(t) &= a\left[1+\frac{1}{k}b\,t+\frac{(\frac{1}{k})(\frac{1}{k}+1)}{2!}b^2t^2+...\right] \end{aligned} \tag{4.1}$$

should be a pgf. For Q(t) to be a pgf, all the terms in (4.1) should be positive. But a and b need not be greater than zero always. Hence geometric distribution on{0,1,2,...} is not CEG.

Also we know that a compound Poisson is not compound geometric and it can not be CEG. Now consider $Q(t) = \left(\frac{p\prime\,t}{1-q\prime\,t}\right)^{\frac{1}{k}}$, which is the pgf of negative binomial distribution with index parameter $\frac{1}{k}$. Let us find the pgf of CEG corresponding to this Q(t).

$$\begin{aligned} \textit{Then } \ Q_Y(t) &= \left(\frac{p}{1-\frac{qp\prime\,t}{1-q\prime\,t}}\right) \\ &= p(1-q\prime\,t)[1-at]^{-1} \ \textit{where } a \ = \ q\prime + qp\prime \\ &= p\ +q\ \left(\frac{pp\prime\,t}{1-(1-pp\prime)t}\right) \\ &= p\ +\ q\left(\frac{p^*\,t}{1-q^*\,t}\right), \textit{where } p^* = p\,p\prime \textit{ and } q^* = 1\ -p\,p\prime \end{aligned}$$

which is a mixture of two generating functions.

R and S Functions of CEG Distribution

R and S functions are generating functions which enable us to analyze properties like infinite divisibility of a probability distribution.

R function

$$\begin{aligned} R_Y(t) &= \frac{Q_Y^{'}(t)}{Q_Y(t)} \\ &= k\,\frac{q}{p}\,Q_Y(t)\ \ [Q(t)]^{k-1}\,Q^{'}(t) \end{aligned} \tag{4.2}$$

where $Q_Y(t)$ and $Q(t)$ are pgfs' of Y and X_i respectively. All terms in the RHS of (4.2) are absolutely monotone from which it follows that $R_Y(t)$ is absolutely monotone. As the absolute monotonicity of R function is a necessary and sufficient condition (Steutel et al.(2004)) for the infinite divisibility of a pgf, we have:

Proposition 4.1 *CEG distribution is infinitely divisible.*

Remark 4.1 $Q_Y(t)$ can be expressed in terms of the R function $R_Y(t)$ as $Q_Y(t) = g_0\ \ exp(\int_0^t R_Y(x)dx)$ where $g_0 = P_r[Y = 0]$.

S function

Steutel et al.(2004) defined S function of compound geometric distributon on $\{0, 1, 2, ...\}$ as $S_Y(t) = \frac{1}{t}\left(1-\frac{Q_Y(0)}{Q_Y(t)}\right)\ \ 0\ \le\ t\ <\ 1$. Let us consider the S function of the CEG distribution $S_Y(t)$ as 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 $\{1, 2, ...\}$.

$$\begin{aligned} \textit{Then} \qquad S_Y(t) &= \sum_{j=0}^{\infty} s_jt^j \\ &= \frac{q^{\frac{1}{k}}}{t}\ Q(t) \end{aligned} \tag{4.3}$$

$$\textit{Now} \qquad Q_Y(t) = \frac{Q_Y(0)}{1-q[Q(t)]^k}$$

Substituting for $q[Q(t)]^k$ from (4.3) we get

$$S_Y(t) = \frac{1}{t}\left[1 - \frac{Q_Y(0)}{Q_Y(t)}\right]^{\frac{1}{k}} \tag{4.4}$$

and in turn $Q_Y(t) \ = \frac{Q_Y(0)}{1\ -\ (t\ S_Y(t))^k}$.

The following theorem characterizes CEG distribution based on its S function.

Theorem 4.1 *A distribution is CEG (k, p, p′) if and only if its S function given by (4.4) is absolutely monotone.*

Proof: Suppose that Y is CEG. Then $Q_Y(t) \ = \frac{Q_Y(0)}{1\ -\ q[\ Q(t)]^k}$ and we have $S_Y(t)$ given by (4.4) is absolutely monotone by its construction. Now assume that $S_Y(t)$ given by (4.4) is absolutely monotone. We have

$$\begin{aligned} Q_Y(t) &= \frac{Q_Y(0)}{1-\ t^k[S_Y(t)]^k} \\ &= \frac{p}{1-\ q\ [Q(t)]^k} \\ &\Rightarrow\ Y\ is\ CEG. \end{aligned}$$

Remark 4.2 If Q(t) takes values on $\{0, 1, 2, ...\}$, we can define $s_j = q^{\frac{1}{k}} q_j,\ j \ = \ 0,\ 1,\ 2, ...$
Then $S_Y(t) = [1 - \frac{Q_Y(0)}{Q_Y(t)}]^{\frac{1}{k}}$

Illustration 4.1 Let $Q(t) \ = \ \frac{p'\ t}{1-q'\ t}$, the pgf of geometric distribution on $\{1, 2, 3, ...\}$.Taking $s_j = q^{\frac{1}{k}}\ q_{j+1}$, $j = 0, 1, 2, ...$. we have

$$\begin{aligned} S_Y(t) &= \frac{q^{\frac{1}{k}}}{t}\ \frac{p'\ t}{1-q'\ t} \\ &= q^{\frac{1}{k}}\ \frac{p'}{1-q'\ t}. \end{aligned}$$

and $S_Y(t)$ is absolutely monotone.

5. An AR(1) Process Corresponding to CEG Distribution

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

$$\begin{aligned} Y_{n,i} &= 0\ with\ the\ probability\ p \\ &= Y_{n-1,i}\ + \sum_{i=1}^{k} \varepsilon_{n,i}\ with\ probability\ (1-p) \end{aligned} \tag{5.1}$$

Then $Q_Y(t) \ = \ p \ + \ Q_Y(t)\ [Q_\varepsilon(t)]^k\ (1-p)$
$\Rightarrow Q_Y(t) \ = \ \frac{p}{1\ -\ q\ [Q_\varepsilon(t)]^k}$, assuming that $Y_{n,i} \ \overset{d}{=} \ Y_{n-1,i}\ \ \forall i.$

To cite an example for such AR(1) process, consider a collection centre which renders service to farmers in storing their products which are perishable. Let $\{Y_{n,i}\}$ denote the inventory on the nth day if a power failure occurs on a day , the centre deny to store products and then $Y_{n,i} \ = \ 0$ otherwise $Y_{n,i} \ = \ Y_{n-1,i} \ + \sum_{i=1}^{k} \ \varepsilon_{n,i}$ where $\sum_{i=1}^{k} \ \varepsilon_{n,i}$ denote the total quantity of products that are received from k outlets on the day.

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

6. Reliability Properties

The well-known NWU property of geometric sums is extended to the class of random sums by Cai et al. (2000). Willmot et al. (2001) have discussed aging and other distributional properties of compound geometric distribution. Higher order reliability property of a distribution are discussed by Willmot (2002). Cai et al. (2005) showed that the aging properties of many compound distributions can be characterized uniquely by the aging properties of the the primary distribution whatever the secondary distribution is. Here we discuss some of these properties in the case of CEG distribution.

Hazard Rates

Knowledge of the functional form of hazard rate function is equivalent to that of the distribution itself. If the function has a closed form it will help to determine the distribution of the random variable uniquely. As there is no closed form here, we evaluate the hazard rate values for given values of p, k, and p' and are tabulated below. Graphs are also plotted.

Table 6.1. Hazard function

Hazard rate values at	p=0.9	p=0.6	p=0.3	p=0.1
0	12.0000000	2.0000000	0.5714428	0.1505453
1	0.4444444	0.2857143	0.1379426	0.04721529
2	0.5260870	0.3263158	0.1530256	0.05145147
3	0.5764499	0.3464567	0.1591356	0.05304683
4	0.6092096	0.3566766	0.1616304	0.05370729
5	0.6312597	0.3619219	0.1626533	0.05403935
6	0.6464447	0.3646298	0.1630748	0.05425961
7	0.6570675	0.3660320	0.1632502	0.05444578
8	0.6645806	0.3667592	0.1633253	0.05462599
9	0.6699356	0.3671366	0.1633597	0.05481069
10	0.6737735	0.3673326	0.1633780	0.05500406
11	0.6765349	0.3674344	0.1633904	0.05520806
12	0.6785275	0.3674873	0.1634012	0.05542388
13	0.6799682	0.3675148	0.1634122	0.05565251
14	0.6810114	0.3675290	0.1634245	0.05589488
15	0.6817677	0.3675365	0.1634384	0.05615196
16	0.6823163	0.3675403	0.1634546	0.05642480
17	0.6827145	0.3675423	0.1634734	0.05671450
18	0.6830037	0.3675433	0.1634952	0.05702229
19	0.6832137	0.3675439	0.1635206	0.05734951
20	0.6833663	0.3675442	0.1635501	0.05769759
21	0.6834772	0.3675443	0.1635845	0.05806811
22	0.6835578	0.3675444	0.1636245	0.05846282
23	0.6836164	0.3675444	0.1636711	0.05888361
24	0.6836589	0.3675444	0.1637253	0.05933257
25	0.6836899	0.3675445	0.1637884	0.05981202
26	0.6837124	0.3675445	0.1638619	0.06032449
27	0.6837287	0.3675445	0.1639475	0.06087282
28	0.6837406	0.3675445	0.1640472	0.06146014
29	0.6837493	0.3675445	0.1641632	0.06208995

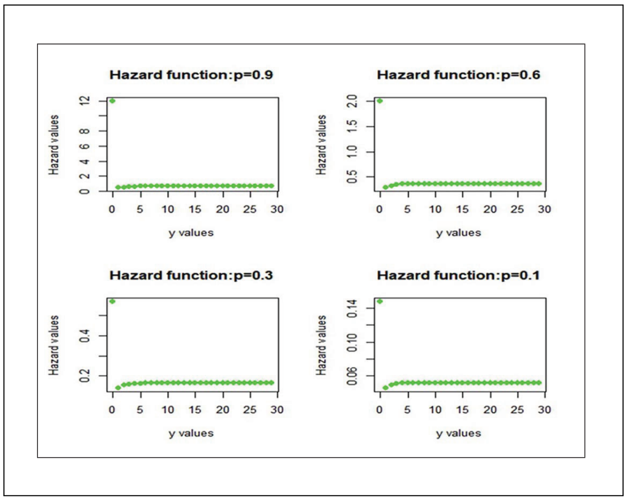


Figure 6.1. Hazard function graph

DFR/IFR property

Willmot et al.(2001) has established a remarkable theorem on the DFR property of compound geometric distribution which is stated below:

Theorem 6.1 *Compound geometric distribution with DFR secondary distribution is DFR.*

Using this theorem, the following result may be stated in the case of CEG distribution.

Result 6.1 *CEG distribution with DFR secondary distribution is DFR.*

Proof: $Q_Y(t) = \frac{p}{1-q[Q(t)]^k}$, where k is a positive integer, is the pgf of a CEG random variable Y. Here $Q_1(t) = [Q(t)]^k$ is the pgf of the sum $\sum_{i=1}^{k} X_i$ of i.i.d variables, each having DFR property. As sum of DFR random variables is again DFR, Q_1 is DFR. Thus $Q_Y(t) = \frac{p}{1-q\ Q_1(t)}$ which is the pgf of compound geometric distribution and by theorem 6.1, Y is DFR.
The following example illustrates result 6.1

Consider the negative binomial pgf

$$Q(t) = \frac{(p^*)^m}{(1-q^*t)^m}, \quad 0 \ < \ m \ < \ 1 \tag{6.1}$$

The probabilities are given by

$$q_r = \frac{m\,(m\ +\ 1)...(m+r-1)}{r!}(p^*)^m\ \ (q^*)^r, \quad r\ =\ 0,1,2,...$$

$$\begin{aligned} Then\ \ q_0 &= \ (p^*)^m \\ q_1 &= \ m\,(p^*)^m\ q^* \\ q_2 &= \ \frac{m(m+1)}{2}(p^*)^m\ (q^*)^2\ \ \ and\ \ so\ \ on. \end{aligned}$$

These probabilities are monotonically non increasing and the distribution is log convex and DFR. Then the CEG variable with this secondary distribution has pgf $Q_Y(t) = \frac{p}{1-q[Q(t)]^k}$ where $Q(t)$ is given by (6.1).

And the CEG probabilities are

$$\begin{aligned}
g_0 &= p[1 - q(p^*)^{mk}]^{-1} \\
g_1 &= k\ p\ q\ m\ q^*\ (p^*)^{mk}\ [1 - q(p^*)^{mk}]^{-2} \\
g_2 &= k\ p\ q\ \frac{m(m+1)}{2}\ (q^*)^2\ (p^*)^{mk}\ (1 - q(p^*)^{mk})^{-2} \\
&+ \frac{k^2\ p\ q\ m^2\ (q^*)^2(p^*)^{mk}}{2}(1 - q\ (p^*)^{mk})^{-3}\ (1 + q(p^*)^{mk})^2 \\
&- \frac{k\ p\ q\ m^2\ (q^*)^2(p^*)^{mk}}{2}(1 - q(p^*)^{mk})^{-2} \\
g_2 g_0 - g_1^2 &= \frac{k\ p^2\ q\ (q^*)^2(p^*)^{mk}}{2}\ (1 + m^2)\ (1 - q(p^*)^{mk})^{-3} \\
&\geq 0
\end{aligned}$$

i.e., $g_2 g_0 \ \geq\ g_1^2\ \ or\ \ \frac{g_2}{g_1}\ \geq\ \frac{g_1}{g_0}$

Thus Y is log convex. Also log convexity is a sufficient condition for a random variable to be DFR. Hence Y is DFR.

Kemp (2004) established various relationships between classes of life time distributions given by

$DFR/IFR\ \Rightarrow\ DFRA/IFRA \Rightarrow NWU/NBU \Rightarrow NWUE/NBUE \Rightarrow IMRL/DMRL.$

Using this, we state the following result.

Result 6.2 *CEG distribution with DFR secondary distribution is DFRA, NWU, NWUE and IMRL. Also it will have monotonically non increasing probabilities.*

Result 6.3 *CEG class is not a subclass of DFR class in general.*

The following example illustrates this result.

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

Then the CEG probabilities are

$g_{4i} = pq^i, i = 0, 1, 2, ...$
$\quad = 0, elsewhere$

Denoting the tail probabilities of Y by $a_n, n = 0, 1, 2, ...$

$a_n = Pr[Y > n]$
$\quad = \sum_{j=n+1}^{\infty} g_j.$

Then

$a_0\ = a_1\ = a_2\ = a_3\ =\ q$

$a_4 = a_5 = a_6 = a_7\ = q^2$
and so on

Thus we have

$\frac{a_{n+1}}{a_n} < 1$ for $n = 3 + 4j, j = 0, 1, 2, ...$
$\quad = 1$ for all other values of n.

This implies that $\frac{a_{n+1}}{a_n}$ is not non decreasing in n for $n = 0, 1, 2, ...$ which again imply that Y is not DFR.

Result 6.4 *CEG distribution with IFR secondary distribution need not be IFR.*

The following example illustrates this result.

Let the Poisson distribution with parameter λ be the secondary distribution. Then

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

Gupta et al.(1997) showed that Poisson distribution is IFR for any $\lambda > 0$.

$$\begin{aligned}
g_0 &= p\,(1 - q\,e^{-k\lambda})^{-1}\\
g_1 &= k\,p\,q\,\lambda\,e^{-\lambda}\,e^{-(k-1)\lambda}(1 - q\,e^{-k\lambda})^{-2}\\
g_2 &= k\,p\,q\,\frac{\lambda^2 e^{-\lambda}}{2}\,e^{-(k-1)\lambda}\,(1 - q\,e^{-k\lambda})^{-2}\\
&\quad + \,p\,q\,\lambda^2\,e^{-2\lambda}\frac{e^{-(k-2)\lambda}}{2}[k^2(1 - q\,e^{-k\lambda})^{-3}\,(1 + q\,e^{-k\lambda})\; - \;k\,(1 - q\,e^{-k\lambda})^{-2}]\\
g_2 g_0 \; - \; g_1^2 &= \frac{k^2\,p^2\,\lambda^2\,e^{-k\lambda}}{2}\,q\,(1 - q\,e^{-k\lambda})^{-4}\\
&\geq \;0
\end{aligned}$$

i.e. $g_2 g_0 \;\geq\; g_1^2 \;\; or \;\; \frac{g_2}{g_1} \;\geq\; \frac{g_1}{g_0}$

But for IFR distribution $\frac{g_{n+2}}{g_{n+1}} < \frac{g_{n+1}}{g_n}$ for all n.

$\Rightarrow$ Y is not IFR.

7. Simulation and Estimation

Estimation of parameters of CEG distribution is discussed in this section. Its clear from (2.2) that the expression for g_y is much complicated and so the MLE method is impractical. We use other two methods namely moment estimation method and a pgf based method. The performance of these estimators are compared using simulation.

Moment estimation

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

Case 1: When p is unknown.

Moment estimator of p for known values of p' and k are obtained by solving the non linear equation given by

$$m_1\,p\,p' - k\,(1 - p)\,(1 - p') = 0$$

where m_1 denote the sample mean. Solution is done using root-finding function in R, namely, uniroot, in the R-package, rootSolve.

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
0.9	Estimate	0.880503	0.857142	0.894568
	Mean bias	-0.000389	-0.000857	-0.000108
	MSE	0.000007	0.000036	0.0000005
0.6	Estimate	0.593220	0.711864	0.533333
	Mean bias	-0.000135	0.002237	-0.001333
	MSE	0.0000009	0.000250	0.000088
0.3	Estimate	0.3063457	0.254545	0.318543
	Mean bias	0.000126	-0.000909	0.000370
	MSE	0.0000008	0.000041	0.000006
0.1	Estimate	0.138067	0.092429	0.097323
	Mean bias	0.000761	-0.000151	-0.000053
	MSE	0.000028	0.000001	0.0000001

Table 7.2. Moment estimates using simulated sample of size 100 , no. of replications 100

$p' = 0.5$

p		k=2	k=3	k=4
	Estimate	0.895761	0.902605	0.901646
0.9	Mean bias	-0.004238	0.002605	0.001646
	MSE	0.001429	0.001336	0.001003
	Estimate	0.607815	0.604842	0.600160
0.6	Mean bias	0.007814	0.004841	0.000160
	MSE	0.002361	0.002463	0.001545
	Estimate	0.304055	0.297088	0.307787
0.3	Mean bias	0.004055	-0.002911	0.007787
	MSE	0.001100	0.000692	0.000772
	Estimate	0.100574	0.101289	0.101031
0.1	Mean bias	0.000574	0.001289	0.001031
	MSE	0.000095	0.000116	0.000113

Case 2: When the parameters *p* and *k* are unknown.

Table 7.3. Moment estimates using simulated sample of size 100

$p' = 0.5$

		k=2		k=3	
p		$p^\wedge$	$k^\wedge$	$p^\wedge$	$k^\wedge$
	Estimate	0.897189	1.483529	0.914690	2.787692
0.9	Bias	0.002810	0.516471	-0.014690	0.212308
	Estimate	0.550560	1.543492	0.547488	2.310890
0.6	Bias	0.049439	0.456508	0.0525112	0.68911
	Estimate	0.235045	1.385765	0.256971	2.327518
0.3	Bias	0.064955	0.614235	0.043029	0.672482

Table 7.4. Moment estimates using simulated sample of size 200

$p' = 0.5$

		k=2		k=3	
p		$p^\wedge$	$k^\wedge$	$p^\wedge$	$k^\wedge$
	Estimate	0.851696	1.062432	0.879149	2.109655
0.9	Bias	0.048304	0.937568	0.020850	0.890345
	Estimate	0.580577	1.702601	0.578516	2.580425
0.6	Bias	0.049422	0.297399	0.021484	0.419575
	Estimate	0.230081	1.346270	0.245584	2.190817
0.3	Bias	0.069918	0.653730	0.054415	0.809183

Table 7.5. Moment estimates using simulated sample of size 200 and no. of replications 10

$p' = 0.5$

		k=2		k=3	
p		$p^\wedge$	$k^\wedge$	$p^\wedge$	$k^\wedge$
0.6	Mean estimate	0.621392	2.198139	0.615971	3.217357
	Mean bias	-0.021392	-0.19814	-0.015971	-0.217357
	MSE	0.005840	0.499688	0.004465	0.878118
0.3	Mean estimate	0.308553	1.851169	0.293683	3.019106
	Mean bias	-0.008553	0.148830	0.006316	-0.019106
	MSE	0.008773	1.011429	0.006264	1.308463

Remark 7.1

Moment estimates of p show more efficiency but that of k show less efficiency for simulated data.

BHHJ estimation

The use of pgf in statistical inference has been proposed as a tool in estimation due to its simplicity compared to probability mass function in many instances. Application of pgf in parameter estimation through Hellinger distance is investigated in the context of discrete distributions by Sim et al.(2010). Basu et al. (1998) proposed a density power divergence method for parameter estimation, abbreviated as BHHJ estimation method. Here we use BHHJ-pgf divergence method, proposed by Ying et al. (2016). This method relies on a tuning parameter say, α. Although α may take any value greater than or equal to zero, it is preferable to have $0 \leq \alpha \leq 1$. A divergence is a measure between two probability functions which is equal to zero if and only if both the functions are exactly identical.

BHHJ divergence based on pgf is defined as

$$d_\alpha(g_n, g) = \int_0^1 [g^{1+\alpha}(t;\theta) - (1 + \frac{1}{\alpha})g_n(t)g^\alpha(t;\theta) + \frac{1}{\alpha}g_n^{1+\alpha}(t)]dt, \quad a > 0 \tag{7.1}$$

where $g(t, \theta) = E_\theta(t^x), \theta\epsilon\Theta$, 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 numerical integration and optimization techniques in R.

Case 1 : When p is unknown

Table 7.6. BHHJ estimates using simulated sample of size 70 , no. of replications 50

$p' = 0.5$

p		k=2	k=3	k=4
0.9	Estimate	0.9108968	0.9053703	0.9041653
	Mean bias	0.01089678	0.005370265	0.004165289
	MSE	0.001490264	0.00129572	0.00123835
0.6	Estimate	0.6526661	0.640099	0.6330657
	Mean bias	0.05266611	0.04009899	0.03306566
	MSE	0.005718887	0.00432516	0.003916185
0.3	Estimate	0.3479656	0.3380421	0.3337955
	Mean bias	0.04796564	0.03804213	0.03379551
	MSE	0.004348459	0.00370362	0.003410143
0.1	Estimate	0.1232959	0.1209915	0.1195978
	Mean bias	0.02329586	0.02099153	0.01959776
	MSE	0.00110247	0.001059967	0.001036039

Table 7.7. BHHJ estimates using simulated sample of size 100 , no. of replications 50

$p' = 0.5$

p		k=2	k=3	k=4
	Estimate	0.8772054	0.880645	0.8815072
0.9	Mean bias	-0.02279409	-0.01953547	-0.01849279
	MSE	0.001717514	0.001380089	0.001267511
	Estimate	0.5951251	0.5953881	0.5957902
0.6	Mean bias	-0.004874921	-0.004611889	-0.00420979
	MSE	0.001979797	0.002025841	0.002080004
	Estimate	0.2944718	0.2954374	0.296052
0.3	Mean bias	-0.005528216	-0.004562633	-0.003948003
	MSE	0.001578441	0.001710894	0.00177828
	Estimate	0.09929298	0.1005937	0.1034169
0.1	Mean bias	-0.0007070249	0.0005937333	0.003416925
	MSE	0.000301405	0.0003323098	0.003795251

Case 2 : When two parameters p and $p\prime$ are unknown

Table 7.8. BHHJ estimates using simulated sample of size 100 and no. of replications 20

p		$k = 2$		$k = 3$	
		$p^{\wedge}$	$(p\prime)^{\wedge}$	$p^{\wedge}$	$(p\prime)^{\wedge}$
	$Estimate$	0.8931849	0.4328041	0.91177768	0.4866129
0.9	$Meanbias$	−0.00681504	−0.4671958	0.01177768	−0.41338708
	MSE	0.00162173	0.2270093	0.00115606	0.17728369
	$Estimate$	0.63138006	0.50130096	0.59195481	0.5172811
0.6	$Meanbias$	0.03138006	−0.13421505	−0.00804518	−0.0827189
	MSE	0.00368531	0.0368625	0.003772016	0.00873732
	$Estimate$	0.29306584	0.5212073	0.27628933	0.51374002
0.3	$Meanbias$	−0.00693415	0.2212073	−0.02371066	0.213740025
	MSE	0.00485659	0.05474073	0.006566003	0.0492060
	$Estimate$	0.0904562	0.64514572	0.10105595	0.51724618
0.1	$Meanbias$	0.0054562	0.5451457	0.001055958	0.41724618
	MSE	0.00262791	0.30885135	0.00116145	0.18415606

8. Real Life Data Set

The number of depression symptoms can be considered as count data in order to get complete and accurate analysis findings in studies of depression. Tao et al. (2017), in his study, aims to compare the goodness of fit of four count outcomes models by a large survey sample to identify the optimum model for a risk factor study of the number of depression symptoms.

Table 8.1. Proportions and predictive probabilities of each counts (in percentage)

Count	*Observed*	*Poisson*	*NB*	*ZIP*	*ZINB*
0	39.28	28.10	36.89	39.22	39.04
1	23.74	33.19	28.02	20.63	22.67
2	15.23	21.61	16.40	18.67	17.64
3	10.38	10.42	8.83	11.75	10.58
4	6.33	4.24	4.62	5.84	5.46
5	3.21	1.58	2.41	2.47	2.58
6	1.40	0.56	1.27	0.94	1.15
7	0.43	0.20	0.68	0.33	0.50
Total	100	100	100	100	100
Chi square value		12.3153467	1.3954038	1.5389177	0.5528416
d.f		3	2	3	2
p value		0.0063774	0.497727	0.6733177	0.7584937

NB, negative binomial; ZINB, zero-inflated negaitve binomial; ZIP,zero-inflated poisson.

The second column of above table presents the observed distribution of the number of depression symptoms. Among the total of 15462 respondents, 39.28 percentage reported no depression symptoms and so on. 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 CEG distribution is done using Chi square test. Parameter p is estimated using BHHJ method.

Table 8.2. Fitting of CEG distribution

		Expected (k = 2, $p\prime$ = 0.98)	
Count	*Observed*	*unpooled*	*pooled*
0	39.28	39.6659	39.6659
1	23.74	23.2165	23.2165
2	15.23	14.2852	14.2852
3	10.38	8.7874	8.7874
4	6.33	5.4054	14.045
5	3.21	3.3251	
6	1.4	2.0454	(*for counts* $\geq$ 4)
7	0.43	3.2691	
Total	100	100	100
Chi square value			0.876162
d.f			3
p value			0.831176

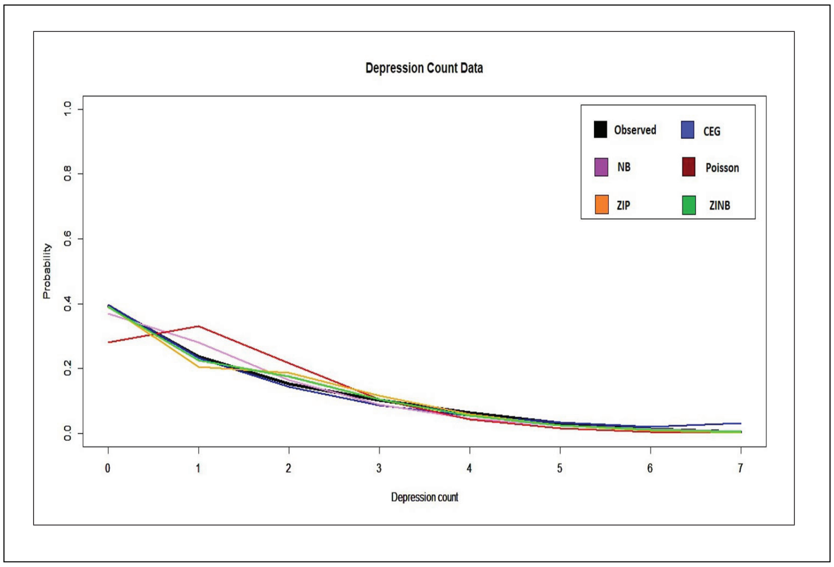


Figure 8.1. Depression count graph

9. Conclusion

This paper offers CEG distribution, which is obtained by compounding EG distribution with discrete distribution. We have derived some properties of the distribution, including reliability properties. Apart from standard distributions, pmf of many discrete compound distributions have no closed form. Hence even evaluation of probabilities is a difficult task. We have given R commands to evaluate CEG probabilities in the case of binomial secondary distribution.This can be used for any secondary distribution. Characterization of the distribution is done using S function. An AR(1) process corresponding to CEG distribution is obtained. We have tried to explore the role of CEG distribution as a reliability class of distributions and demonstrated how the distribution falls into place within the hierarchy of various reliability classes. The parameter estimation via the maximum likelihood method is not practical due to the complexity in the form of pmf. Hence it is done using method of moments and pgf based BHHJ estimation method. We conducted simulation study to compare the methods.Based on our simulations, we conclude that BHHJ method performs better than method of moments. Finally, we have fitted CEG distribution to a real data and compared it with fitting Poisson, negative binomial, zero-inflated Poisson and zero-inflated negative binomial distributions. The p values show that CEG distribution provides a better fit compared to others.

References

Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. C. (1998). Robust and efficient estimation by minimizing a density power divergence, *Biometrika, 85,* 549-559. https://doi.org/10.1093/biomet/85.3.549

Brown, M. (1990). Error bounds for exponential approximations of geometric convolutions, *The Annals of Probability, 18,* 1388-1402. https://doi.org/10.1214/aop/1176990750

Cai, J., & Kalashnikov, V. (2000). NWU property of a class of random sums, *Journal of Applied probability, 37,* 283-289. https://doi.org/10.1239/jap/1014842286

Cai, J., & Willmot, G. E. (2005). Monotonicity and aging properties of random sums, *Statistics and Probability Letters, 73,* 381-392. https://doi.org/10.1016/j.spl.2005.04.013

Erol, P., & Shehldon, M. R. (2004). Probability in the engineering and informational sciences, *18,* 473-484.

Gupta, P. L., Gupta, R. C., & Tripathi, R. C. (1997). On the monotonic properties of discrete failure rates, J Statist. *Plann. Inference, 65,* 255-268. https://doi.org/10.1016/S0378-3758(97)00064-5

Kemp, A. W. (2004). Classes of discrete lifetime distributions, *communications in Statistics, Theory and Methods, 33,* 3069-3093. https://doi.org/10.1081/STA-200039051

Naoto, N., Shigekazu, N., & Hideyuki, I. (1990). Cumulants of random sum distributions, *Communications in Statistics-Theory and Methods, 19*, 1857-1861.

Panjer, H. H. (1981). Recursive evaluation of a family of compound distributions, *ASTIN Bulletin, 12,* 22-26. https://doi.org/10.1017/S0515036100006796

Paul, E., & Marco, F. (2009). Panjer recursion versus FFT for compound distributions, *Mathematical Methods of Operations Research, 69,* 497-508.

Qihe, T. (2005). The finite- time ruin probability of the compound poisson model with constant interest force, *J. Appl. Prob., 42,* 608-619.

Sandhya, E., Sherly, S., Jos, M. K., & Raju, N. (2006). Characterizations of the extended geometric, Harris, negative binomial and gamma distributions, *STARS Int.Journal (sciences), 1,* 5-17

Schafer, A. J. (1996). Selected problems involving the probability of ruin for an insurance company, Presidential Scholars Thesis (1990-2006), 12.

Sim, S.,Z., & Ong, S. H. (2010). Simulation and Computation, *Communications in Statistics, 39,* 305-314. https://doi.org/10.1080/03610910903443980

Steutel F. W. (1990). The set of geometrically infinitely divisible distributions, Memorandom COSOR, vol. 9042.

Steutel, F. W., & van Harn, K. (2004). Infinite Divisibility of Probability Distributions on the Real Line, *Pure and Applied Mathematics*, p: 259. https://doi.org/10.1201/9780203014127

Sundt, B. (1992). On some extensions of Panjer's class of counting distributions, *ASTIN Bulletin, 22,* 61-80. https://doi.org/10.2143/AST.22.1.2005127

Tao, X., Guangjin, Z., & Shaomei, H. (2017). Study of depression influencing factors with zero-inflated regression models in a large-scale population survey. BMJ Open;7:e016471. https://doi.org/10.1136/bmjopen-2017-016471

Tay, S. Y., Ng Choung, M., & Ong, S. H. (2016). Parameter estimation using probability generating function based minimum power divergence. American Institute of Physics, AIP Conference Proceedings 1750, 060008 (2016), doi: 10.1063/1.4954613

William, F. (1968). An introduction to probability theory and its applications, vol 1, third edition, John Wiley and Sons, new York.

Willmot, G. E., & Lin, X. O. (1997). Upper bounds for the tail of the compound negative binomial distribution, Scandinavian Actuarial Journal 1997, Issue 2. https://doi.org/10.1080/03461238.1997.10413983

Willmot, G. E., Sundt,& Jewells. (1988). Family of discrete distributions, *ASTIN Bulletin, 18*, 17-29.
https://doi.org/10.2143/AST.18.1.2014957

Willmot, G. E., & Cai, J. (2001). Ageing and other distributional properties of discrete compound geometric distributions, *Insurance Mathematics and Economics, 28*, 361-379. https://doi.org/10.1016/S0167-6687(01)00062-2

Willmot, G. E. (2002). Compound geometric residual lifetime distributions and the deficit at ruin, *Insurance Mathematics and Economics, 30,* 421-438. https://doi.org/10.1016/S0167-6687(02)00122-1

Willmot, G. E. (2002). Higher order reliability properties of compound geometric distribution, *Journal of Applied Probability, 39*(2), 324-340. https://doi.org/10.1239/jap/1025131429

Willmot, G. E., Panjer, H. H., & Klugman, H. (2012). Loss Models: From Data to Decisions. Wiley, New York.

Copyrights