International Journal of Scientific and Research Publications, Volume 9, Issue 12, December 2019
ISSN 2250-3153

CG and CEG Distributions with Uniform Secondary Distribution

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

[1] Department of Statistics, St.Thomas College, Pala-686574, India, cmlatha.krishnakumar@gmail.com

[2] Department of Statistics, Prajyothi Nikethan College, Pudukkad-680301, India, esandhya@hotmail.com

Abstract Compound geometric (CG) and Compound extended geometric (CEG) distributions with uniform distribution on $\{0,1,....n-1\}$ as secondary distribution are discussed. Probabilities are evaluated numerically using Fast Fourier Transform (FFT) technique. A characterization of CG distribution is established for $n=2$. Some distributional, statistical and reliability properties of CG and CEG distributions are discussed. Parameters are estimated using Moment and BHHJ methods. The distributions are fitted using real life data sets.

Keywords: Akaike/ Bayesian information criterion(AIC,BIC), CEG distribution, CG distribution, IFR/DFR distributions, Moment/BHHJ method of estimation, New worse than used (NWU), R function, S function.

1. Introduction

Geometric compounds have been discussed extensively in the literature and have many useful applications in area of applied probability including reliability, queueing theory and risk theory. Asymptotic formulae have been derived for compound distributions when the primary distribution is of a particular type. Sundt (1982) considered geometric distribution as the primary distribution. These results were generalized by Milidiu (1985) to the compound negative binomial distribution and by Embrechts et al.(1985) to a general family of compound distributions. Willmot (1989) examined the limiting tail behavior of some discrete compound distributions including CG distribution which have been found to be useful in insurance claim modeling.It is well- known that a CG distribution converges to an exponential distribution as the geometric parameter $p \to 0$. Explicit bounds in exponential approximation for CG distribution have been given by Brown (1990,2015), Bon (2006) and Pekoz and Ollin (2011). Brown's work takes advantage of reliability properties of such CG distribution. The bounds given by Pekoz and Ollin (2011) apply more generally than to CG distributions, relaxing the assumptions that the primary distribution is geometric and the secondary random variables are independent. Pekoz et al. (2013) give bounds in the geometric approximation of CG distributions. Some of the above mentioned bounds apply in the case where the geometric distribution is on $\{0,1,2,...\}$ and some in the case where it is on $\{1,2,3,...\}$. Steutel and Van Harn (2004), discuss compound geometric distribution as part of discussing infinitely divisible (id) distributions.They prove that CG distribution on $\mathbb{Z}_+$ is infinitely divisible and hence compound Poisson. Sandhya and Latha (2019) introduced CEG distribution with discrete secondary distribution. It is shown that the distribution is id. A characterization of the distribution is done using its S- function, a generating function.

Reliability studies of probability distributions are mainly concerned with the concepts of hazard and reverse hazard functions. A useful tool for reducing the range of possible candidates for fitting a distribution to observed data is provided by the shape and monotonicity of the failure rate (hazard rate) function. Let X be discrete random variable with support $\{0,1,2,...\}$ or a subset thereof. Suppose $p(x)=P(X=x)$, $F(x)=P(X\leq x)=\sum_{j=0}^{x} p(j)$ and $R(x)=P(X\geq x)=\sum_{j=x}^{\infty} p(j)$ denote, respectively, the probability mass function (pmf), distribution function and reliability (survival) function of X. Then the hazard rate denoted by h(x) is

$$h(x)=\frac{p(x)}{R(x)} \quad , \qquad x = 1,\ 2,\ ... \tag{1.1}$$

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

Reverse hazard rate denoted by H(x) is the probability that a component will fail at time $X = x$, given that it is known to have failed before x, where X represents the lifetime of the component. i.e,

$$H(x) = \frac{p(x)}{F(x)} \quad , \qquad x = 1, 2, \ldots \tag{1.2}$$

Willmot and Cai (2001) have proved that CG distribution with DFR secondary distribution is DFR. Sandhya and Latha (2019) proved the same result in the case of CEG distribution.

Auto regressive(AR) models are based on the idea that the current value of the series can be explained as a function of previous values of the series. A sequence $\{Y_n\}$ of r.v.s describes an AR(1) scheme if there exists an innovation sequence $\{\varepsilon_n\}$ of iid r.v.s satisfying

$$Y_n = bY_{n-1} + \varepsilon_n \quad \forall n > 0,$$

integer and some $0 < b < 1$. $\{Y_n\}$ is stationary if $Y_n \stackrel{d}{=} Y_{n-1} \ \ \forall n > 0$. Satheesh et.al (2006) modify this model by considering k independent AR(1) sequences $\{Y_{n,\,i}\}\ , i = 1, 2, \ldots k$ where

$$Y_{n,i} = b\, Y_{n-1,i} + \varepsilon_{n,i} \quad \forall \quad n > 0, \ \ \textit{integer and } 0 < b < 1.$$

where $\{Y_{n,\,i}\}\ , i = 1, 2, \ldots k$ are identically distributed and $\{\varepsilon_{n,i}\}$ are iid. Here we consider AR(1) processes with geometric and extended geometric sum of innovations. Sample path of a stochastic process is a particular realization of the process. It is a particular set of values y_n for all values of n, generated according to the (stochastic) rules of the process. Sample paths of AR(1) models corresponding to geometric and EG sums may be drawn and their behaviour may be analyzed.

Simulation and estimation of parameters are done using both moment and BHHJ methods. BHHJ divergence based on probability generating function (pgf) is defined as

$$d_\alpha(g_r, g) = \int_0^1 [g^{1+\alpha}(t;\theta) - (1+\tfrac{1}{\alpha}) g_r(t) g^\alpha(t;\theta) + \tfrac{1}{\alpha} g_r^{1+\alpha}(t)]dt, \quad \alpha > 0 \tag{1.3}$$

where $g(t,\theta) = E_\theta(t^x), (\theta \varepsilon \Theta$, the parameter space) is the pgf and $g_r(t) = \frac{1}{r}\sum_{i=1}^r t^{x_i}, 0 < t < 1$

is the empirical probability generating function (epgf). Parameters are estimated by minimizing equation (1.3).This optimization is done through "nloptr" package in R. Performance of estimators are measured in terms of mean squared error (MSE) to capture the precision and accuracy of estimators.

In this work, we consider discrete uniform distribution as the secondary distribution. Section 2 deals with such a geometric compound, evaluation of its probabilities and a characterization. Some of its distributional, statistical and reliability properties are discussed. In section 3, EG compound with secondary distribution as uniform is introduced and properties like distributional, statistical and reliability are discussed. AR(1) process and their sample paths with geometric and EG innovations are also discussed. Estimation of parameters using moment method and BHHJ method are done using simulated data for both the distributions. Real life data sets are analyzed and the distributions are found to fit the data sets.

2. CG Distribution with Uniform Secondary Distribution

The discrete uniform distribution is a non- parametric, symmetric distribution whereby a finite number of values are equally likely to be observed; every one of n values has equal probability $\frac{1}{n}$. Here we consider the distribution with support $\{0, 1, 2, \ldots n-1\}$. It has $\textit{mean} = \frac{n-1}{2}$, $\textit{variance} = \frac{n^2-1}{12}$ and $\mu_3 = 0$ (symmetric). This distribution is log-concave and IFR but not infinitely divisible.

2.1 Distributional Properties

We consider geometric distribution on {0, 1, 2...} having pgf given by

$$Q_N(t) = \frac{p}{1 - q\,t} \ , \ 0 < p < 1, \ p + q = 1$$

and denoted by Geo(p). Let $X_1, X_2 \ldots$ be the independent and identically distributed (iid) uniform random variables with

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

pgf Q(t), $Y = \sum_{i=1}^{N} X_i$, the compound random variable, then

$$\begin{aligned} Q_Y(t) &= \frac{p}{1-q\,Q(t)} \\ &= \frac{p}{1-\frac{q(1-t^n)}{n(1-t)}} \end{aligned} \tag{2.1.1}$$

is the pgf of CG uniform distribution and we write $Y \sim CGU\ (p,n)$.

Geometric distribution is a member of Panjer's (1981) family of distributions and hence recursive formulae are available for finding the CGU probabilities. FFT technique is another method of evaluating compound probabilities. Compared to the Panjer recursion, FFT has the main advantage that it works with arbitrary frequency distribution and it is much more efficient. Though expressions for compound probabilities are available in the literature (Johnson et.al (2005)) in the context of stopped sum distributions, numerical evaluation of probabilities is not seen anywhere. Here we derive a general expression for the CGU probabilities using the pgf $Q_Y(t)$ and evaluate the probabilities numerically using FFT technique.

If

$$g_i = P[Y=i]\ ,\ \ i=0,1,2,...$$

denote the CGU probabilities, they are obtained from the pgf $Q_Y(t)$ using the formula

$$g_r = \frac{1}{r!}\left[\frac{d^r}{dt^r}\right]_{t=0} Q_Y(t),\ \ r=1,2,..\ with\ g_0 = Q_Y(0)$$

Now

$$\begin{aligned} Q_Y(t) &= \frac{p}{1-\frac{q(1-t^n)}{n(1-t)}} \\ &= p\left[1-\frac{q}{n}(1+t+t^2+...+t^{n-1})\right]^{-1} \end{aligned}$$

Thus the CGU probabilities are given by

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

$$g_r = p\left(\frac{q}{n}\right)^{\left[\frac{r-1}{2}\right]+1} \sum_{i=0}^{\left[\frac{r}{2}\right]} \binom{r-i}{i} \left(\frac{q}{n}\right)^{\left[\frac{r}{2}\right]-i} \left(1-\frac{q}{n}\right)^{-(r-i+1)}\ \ for\ r=1,\ 2,\ 3,... \tag{2.1.2}$$

Here $\left[\frac{r}{2}\right]$ denotes the integer part of $\frac{r}{2}$

$$\begin{aligned} Then\ g_1 &= p\left(\frac{q}{n}\right)\left(1-\frac{q}{n}\right)^{-2} \\ g_2 &= p\left(\frac{q}{n}\right)\left[\left(\frac{q}{n}\right)\left(1-\frac{q}{n}\right)^{-3} + \left(1-\frac{q}{n}\right)^{-2}\right] \\ g_3 &= p\left(\frac{q}{n}\right)^2\left[\left(\frac{q}{n}\right)\left(1-\frac{q}{n}\right)^{-4} + 2\ \left(1-\frac{q}{n}\right)^{-3}\right] \\ g_4 &= p\left(\frac{q}{n}\right)^2\left[\left(\frac{q}{n}\right)^2\left(1-\frac{q}{n}\right)^{-5} + 3\ \left(\frac{q}{n}\right)\left(1-\frac{q}{n}\right)^{-4} + \left(1-\frac{q}{n}\right)^{-3}\right]\ and\ so\ on. \end{aligned}$$

The distribution function is given by

$$G_0 = g_0$$

$$and\ G_r = \sum_{j=0}^{r} p\left(\frac{q}{n}\right)^{\left[\frac{j-1}{2}\right]+1} \sum_{i=0}^{\left[\frac{j}{2}\right]} \binom{j-i}{i} \left(\frac{q}{n}\right)^{\left[\frac{j}{2}\right]-i} \left(1-\frac{q}{n}\right)^{-(j-i+1)}\ \ for\ r=1,\ 2,\ 3,... \tag{2.1.3}$$

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

Using FFT technique we evaluate the CGU probabilities, following Embrechts and Frei (2009), using the following R commands.

CGU (*p, 5)*

1. $M \leftarrow 128$
2. $f \leftarrow \ vector(length = M)$
3. $n \leftarrow 5$
4. for $(j \ in \ 1:M)$ {
   if $(j < \ n+1) \ f[j] = \frac{1}{n} \ else \ f[j] = 0$
   }
5. $fhat \ \leftarrow \ fft(f, inverse = FALSE)$
6. $u \ \leftarrow \ \frac{p}{(1-(q*fhat))}$
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. They are numerically evaluated for different values of *p*, but are not included in the paper as it takes much space. The graphs are plotted below.

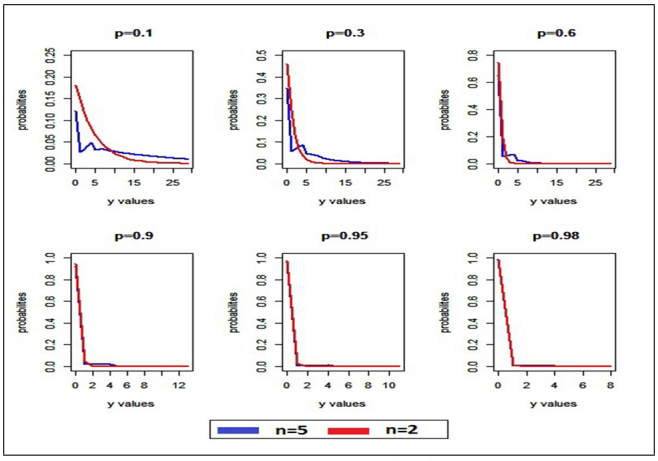

Fig 2.1.1 CGU *(p,5)*

Remark 2.1.1 As $p \rightarrow 1$, the probability graph gets the shape of L shaped curve, irrespective of any value of n.

Remark 2.1.2 Irrespective of any value of p and n, mode of the distribution is at $Y = 0$

Moments of CGU distribution in terms of geometric and uniform parameters are given below.

$$\mu_1'(Y) = \frac{q}{p}\left(\frac{n-1}{2}\right)$$
$$\mu_2(Y) = \frac{q}{12p^2}(n-1)[3(n-1)+p(n+1)]$$
$$\mu_3(Y) = \frac{q}{4p^3}(n-1)^2[n+p-1]$$
$$> 0 \ \ always,$$

implying that CGU distribution is positively skewed.

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

In order to evaluate median for different values of p and n, we make use of the idea of 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 $F(y) \geq p_0$. The quantile corresponding to $p_0 = 0.5$ gives the median.

Closed form expression for quantiles are not easy to derive as the distribution function is not in a compact form. We have simulated sample of size 100 from CGU distribution (p = 0.9, 0.6, 0.3, 0.1) and the quantile values at different probabilities are tabulated below.

Table 2.1.1

	$p_0 \rightarrow$	0.1	0.3	0.5	0.7	0.9	0.99
$n=5$	0.1	0	5	12	30	43	86
	0.3	0	0	3	6	12	26
	0.6	0	0	0	1	4	10
	0.9	0	0	0	0	0	4
$n=2$	0.1	0	1	3	5	11	22
	0.3	0	0	1	1	3	7
	0.6	0	0	0	0	1	3
	0.9	0	0	0	0	0	1

Mean ,median and variance for different values of *p and n* are tabulated below.

Table 2.1.2

	$n=2$			$n=5$		
p	*Mean*	*Median*	*Variance*	*Mean*	*Median*	*Variance*
0.1	$9/2$	3	$99/4$	18	12	378
0.3	$7/6$	1	$91/36$	$14/3$	3	$322/9$
0.6	$1/3$	0	$8/18$	$4/3$	0	$104/18$
0.9	$1/8$	0	$57/972$	$2/9$	0	$58/81$

Remark 2.1.3 As p increases, mean and variance get decreased, median is non increasing but mode remains the same.

A characterization of CGU distribution, for $n = 2$ is stated below.

Theorem 2.1.1 *A CGU (p,n) is Geo* $(p\prime)$ *where* $p\prime = \frac{2p}{2-q}$ *iff the secondary distribution is uniform with* $n = 2$. *i.e. on* $\{0,1\}$

Proof: Let $Y \sim CGU(p,2)$.

$$\begin{aligned} Then\ Q_Y(t) &= \frac{p}{1-\frac{q(1+t)}{2}} \\ &= \frac{2p}{2-q-qt} \\ &= \frac{p'}{1-q't} \ , \ where \ \ p' = \frac{2p}{2-q}, \ \ q' = \frac{q}{2-q} \ , \end{aligned}$$

which is the pgf of Geo (p').

On the other hand,

$$\begin{aligned} Let\ Q_Y(t) &= \frac{p'}{1-q't} \ \ ,where \ \ p' = \frac{2p}{2-q}, \ \ q' = \frac{q}{2-q} \\ &= \frac{p}{1-q\left(\frac{1+t}{2}\right)} \qquad , which\ is\ the\ pgf\ of\ CGU\ (p,\ 2). \end{aligned}$$

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

Result 2.1.1 CGU distribution approaches to degenerate distribution at 0 as $p \rightarrow 1$, irrespective of the values of n.

Steutel and Van Harn (2004) established id property of CG distribution using the absolute monotonicity of its R- function. For CGU distribution, the R- function is given by

$$R_Y(t) = \frac{Q_y^{'}(t)}{Q_Y(t)},\ 0 \leq t < 1$$
$$= q\,[1 - qQ(t)]^{-1} Q^{'}(t),\ \textit{which is absolutely monotone}.$$

They give a characterization of the CG distribution using generating function S of the sequence $(s_j)_{j\varepsilon\mathbb{Z}_+}$, of non negative numbers with $s_j = q\,q_{j+1},\ j\varepsilon\mathbb{Z}_+ \quad where \quad Q(t) = \sum_{i=1}^{\infty} q_i t^i$.

The characterization property is stated below.

Theorem 2.1.2 *Let $Q_Y(t)$ be a positive function on $0 \leq t < 1$ with $Q_Y(1-) = 1$ then $Q_Y(t)$ is a CG pgf iff its S function is absolutely monotone.*

The function S in this theorem is given by

$$S_Y(t) = \frac{1}{t}\left[1 - \frac{Q_Y(0)}{Q_{Y(t)}}\right] \qquad (0 \leq t < 1)$$

which is absolutely monotone, by its construction. Thus the class of CG distributions turns out to have many properties very similar to the class of all id distributions.

2.2 Reliability Properties

Reliability classification of compound geometric distribution has been considered by various authors such as Shanthikumar (1988), Brown (1990), Cai and Kalashnikov (2000), Willmot (2002) and so on. Szekli (1986) proved that complete monotonicity is preserved under geometric compounding. The reliability properties like DFR and NWU are not preserved under finite sums. But Shantikumar (1988) has shown that DFR property is preserved under geometric sum. i.e. if the secondary distribution is DFR then the geometric sum is also DFR. Cai and Kalashnikov (2000) prove that geometric sum is NWU. They have introduced a new class of discrete distributions as follows.

Definition 2.2.1 The discrete distribution $\{p_n, n \geq 0\}$ of a non- negative integer valued random variable M is said to be discrete new worse strongly than used (DS-NWU) if $a_{m+n+1} \geq a_m a_n$, $m, n = 0, 1, \ldots.$ where $a_n = \sum_{i=n+1}^{\infty} p_i$, written $M\ \varepsilon\ DS - NWU$.

They have shown that a class of random sums is NWU, if the primary distribution is DS-NWU , whatever the secondary distribution is. It is stated in the following theorem.

Theorem 2.2.1 *If $M\ \varepsilon\ DS - NWU$, then the random sum $\sum_{i=1}^{M} X_i$ is NWU.*

For geometric distribution on {0,1,2....},

$$a_n = q^{n+1}$$
$$a_{m+n+1} = q^{m+n+2} = a_m a_n$$

which imply that geometric distribution is DS-NWU. Hence by theorem (2.2.1), CG distribution is NWU and hence New worse than used in expectation (NWUE).

Willmot and Cai (2004) proved that residual life time of a compound geometric convolution is again a compound geometric convolution. Willmot and Cai (2001) established that if Y is CG, then Y is DS-NWU. Also if Y is DS-DFR, Y is DS-NWU.

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635 www.ijsrp.org

The following diagram summarizes the relationship between some discrete reliability classes.

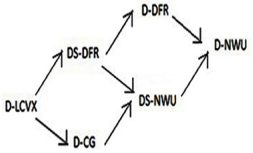

Fig 2.2.1

From (1.1) and (1.2), the hazard and reverse hazard functions are

$$h(r) = \frac{g_r}{R_r},\ for\ r = 1,2,\ldots$$

$$and \qquad H(r) = \frac{g_r}{G_r},\ for\ r = 1,2,\ldots$$

where g_r *and* G_r are given by (2.1.2) and (2.1.3) with $R_r = \sum_{j=r}^{\infty} g_j$.

The expressions for reverse hazard rate for CGU distribution for certain values of r are given below.

$$H_1 = H_2 = \frac{q}{n}$$

$$H_3 = \frac{\left(\frac{q}{n}\right)^2\left[\frac{q}{n} + 2\left(1-\frac{q}{n}\right)\right]}{\left(1-\frac{q}{n}\right) + \left(\frac{q}{n}\right)^2\left[2-\frac{q}{n}\right]}$$

$$H_4 = \frac{\left(\frac{q}{n}\right)^2\left[\left(\frac{q}{n}\right)^2 + \left(1-\frac{q}{n}\right)\left(1+\frac{2q}{n}\right)\right]}{\left(1-\frac{q}{n}\right)^2 + \left(\frac{q}{n}\right)^2\left[3-\frac{2q}{n}\right]}\ and\ so\ on.$$

Compact expression for hazard rate and reverse hazard rate are not available so their values are evaluated and plotted

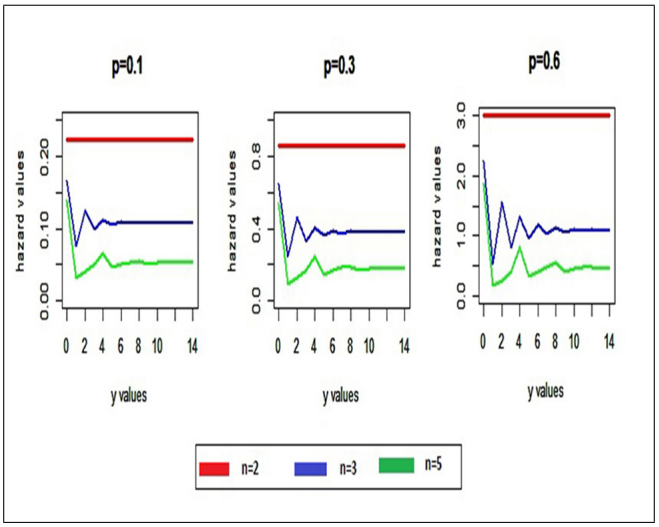

Fig 2.2.2 Hazard function graph (CGU)

Remark 2.2.1 For $n = 2$, the hazard function graph is a straight line, indicating constancy of hazard rate (this has been theoretically established by theorem 2.1.1). Also, the graph gets stabilized more rapidly for small values of p than for large values, irrespective of the value of n.

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

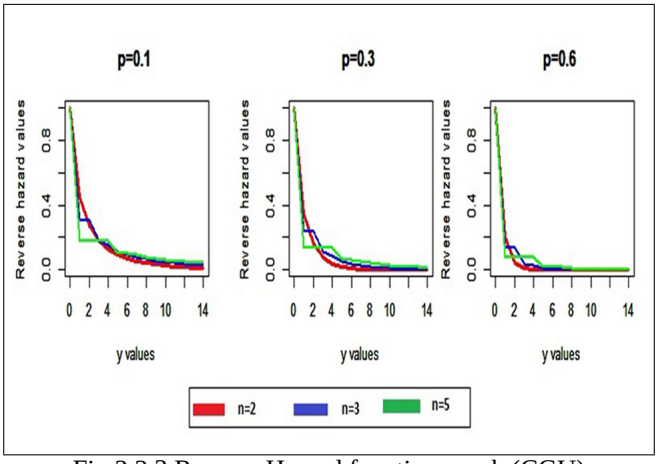

Fig 2.2.3 Reverse Hazard function graph (CGU)

Remark 2.2.2 The tail of the function becomes parallel to the X axis as p becomes higher and higher, irrespective of the value of n.

2.3 An AR(1) process corresponding to CGU distribution

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

$$\begin{aligned} Y_{n,i} &= 0 \; with \; the \; probability \; p \\ &= Y_{n-1,i} + \varepsilon_{n,i} \; with \; probability \; (1-p) \end{aligned} \qquad (2.3.1)$$

Then $Q_Y(t) = p + Q_Y(t)\, Q_\varepsilon(t)\, (1-p)$

$\Rightarrow Q_Y(t) = \frac{p}{1 - q\, Q_\varepsilon(t)}$, assuming that $Y_{n,i} \overset{d}{=} Y_{n-1,i} \;\; \forall i$

The following theorem characterizes CGU distribution.

Theorem 2.3.1 *A sequence* $\{Y_{n,i}\}$ *given by (2.3.1) defines a stationary AR(1) process for some p iff it is geometric sum of innovations* $\{\varepsilon_{n,i}\}$.

Sample path of the AR(1) process defined by (2.3.1)is displayed below for simulated data, for different values of *p* and n, by assuming CGU for $\{\varepsilon_{n,i}\}$.

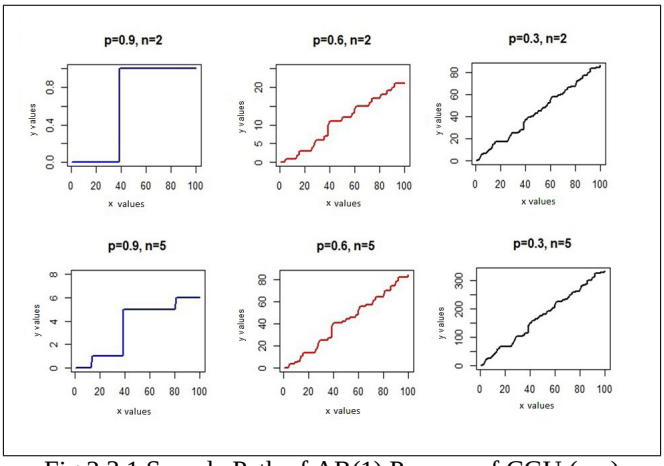

Fig 2.3.1 Sample Path of AR(1) Process of CGU (p,n)

Remark 2.3.1 As $p \to 1$, the sample path gets the shape of a step function. The no. of steps is getting increased for increased values of n.

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

2.4 Simulation and Estimation

As uniform distribution is nonparametric, CGU distribution has only one parameter. Maximum likelihood estimation (mle) method cannot be used as the pmf has no compact form. So they are estimated using (1) Moment estimation method and (2) BHHJ method.

Let m_1 denote the sample mean. Moment estimator of p is obtained by solving the following equation.

$$2pm_1 - q(n-1) = 0 \qquad (2.4.1)$$

For known values of n, the parameter p is estimated using (2.4.1). BHHJ estimate is obtained by minimizing equation (1.3) using "nloptr" package in R.

Table 2.4.1
Estimates using simulated sample of size 100 , no. of replications 50 .

$n = 5$

p		Moment	BHHJ
0.2	Estimate	0.203735	0.202942
	Mean bias	0.003735	0.002942
	MSE	0.000356	0.000684
0.5	Estimate	0.507468	0.501737
	Mean bias	0.007468	0.001737
	MSE	0.001587	0.001785
0.8	Estimate	0.808023	0.807485
	Mean bias	0.008023	0.007485
	MSE	0.001635	0.001540

2.5 Fitting of CGU using Real Life Data Set

Thalassemia is a genetic blood disorder in which the body makes an abnormal form of hemoglobin, the protein in red blood cells that carries oxygen. A study was conducted by Zafakali (2013) involving a sample of 930 for children age between 1-12 years. To build this data set, the numbers of diagnoses among children aged 1-12 years who suffer from Thalassemia were counted. The data were collected at the Medical Record Unit in Hospital Universiti Sains Malaysia (HUSM), Kubang Kerian, Kelantan in north-east Malaysia 2005 to 2010. In the data set, there were 930 patients and among these, 635 patients had no diagnosis; thus, there are 68 zero counts in the data. Frequency (percent) of patients who received a different type of diagnosis is a total of 125 (13.4 percentage), while patients who received two different types of diagnosis was 95 (10.2 percentage). For patients who received three different types of diagnosis the frequency (percent) was 58 (6.2 percentage) and patients who received four different types of diagnosis was 17 (1.8 percentage). The fitting is done using chi- square test of goodness of fit. Here both AIC and BIC methods cannot be applied here because of the complexity of pmf.

Table 2.5.1

Type of diagonosis	*Obs. freq.*	*Exp. freq.* (*pooled*)
0	68.3	72.772279
1	13.4	9.907116
2	10.2	11.255857
3	6.2	6.064743
4	1.8	
Total	100	100
d.f		2
Chi value		2.288363
p value		0.3184845

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

The parameter p is estimated using moment method for $n = 3$ and we get $p^{\Lambda} = 0.5621801$ The p value shows that CGU distribution provides a good fit for the data.

The probability graphs corresponding to observed and theoretical frequencies are displayed below.

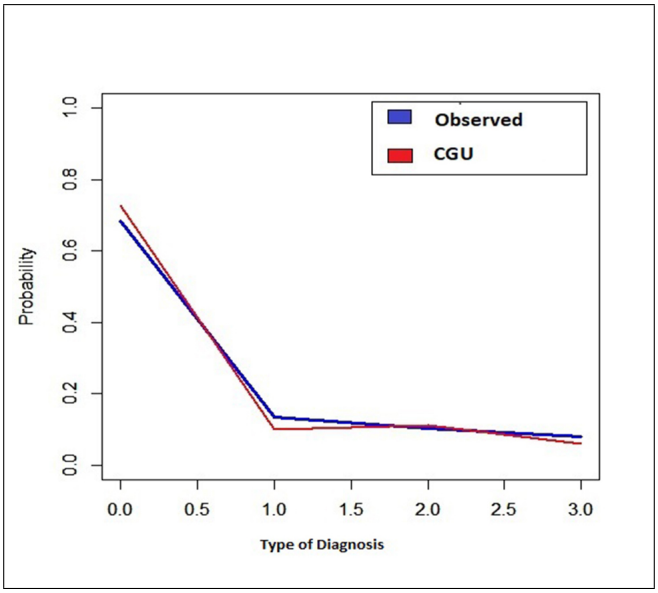

Fig 2.5.1

3. CEG Distribution with Uniform Secondary Distribution

3.1 Distributional Properties

The probability distribution having pgf given by

$$Q_N(t) = \frac{p}{1-q\,t^k} \quad , \quad k > 0 \; integer, \; 0 < p < 1, \; p+q = 1$$

is called extended geometric (EG) distribution on {0,k,2k,...}. Putting $k = 1$, it reduces to geometric distribution.

Let $X_1, X_2 \ldots$ be the independent and identically distributed (iid) uniform random variable with pgf Q(t), $Y = \sum_{i=1}^{N} X_i$ the compound random variable, then

$$Q_Y(t) = \frac{p}{1-q\,[Q(t)]^k} \qquad (3.1.1)$$

where $Q(t) = \frac{1-t^n}{n(1-t)}$, $0 < t < 1$, is the pgf of CEG uniform distribution and we write $Y \sim CEGU(k,p,n)$. Putting $k = 1$ this distribution reduces to CGU distribution.

From $Q_Y(t)$, CEGU probabilities are given by

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

$$g_1 = k\,p\,q\,n^{-k}\,\left[1-q\,n^{-k}\right]^{-2}$$

$$g_2 = \frac{k\,p\,q\,n^{-k}}{2}\left[kqn^{-k}\left(1-qn^{-k}\right)^{-3} + \left(\frac{k+1}{2}\right)\left(1-qn^{-k}\right)^{-2}\right] \; and \; so \; on.$$

Its not easy to have a general formula for the pmf of CEGU distribution. The probabilities can be evaluated numerically, using FFT, with the following R- commands. They are also plotted for different values of p.

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

CEGU (*k, p, 5)*

1. $M \leftarrow 128$
2. $k \leftarrow 2$
3. $f \leftarrow \ vector(length = M)$
4. $n \leftarrow 5$
5. for $(j \ in \ 1:M)$ {
   if $(j < \ n+1) \ f[j] = \frac{1}{n} \ else \ f[j] = 0$
   }
6. $fhat \ \leftarrow \ fft(f, inverse = FALSE)$
7. $fkhat \ \leftarrow \ fhat * fhat$
8. $u \ \leftarrow \ \frac{p}{(1-(q*fkhat))}$
9. $g \ \leftarrow \ (1/M) * fft(u, inverse = TRUE)$

The vector g contains the probability masses on $0, 1, 2, ...(M-1)$ where M is a truncation point. The probabilities are evaluated using FFT technique (following Embrechts and Frei(2009)) for specific values of *p, k* and *n*.They are not displayed here as it takes much space. Graphs are given below.

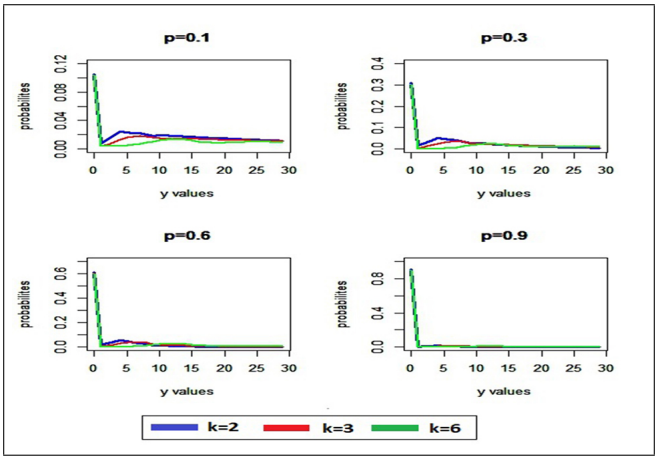

Fig 3.1.1 CEGU *(k, p, 5)*

It is observed from the graph that, change in the value of k does not affect the shape (L shape) of the probability graph for values of p approaching 1.

Moments of CEGU are given by

$$\mu_1'(Y) = \frac{kq}{p}\left(\frac{n-1}{2}\right)$$

$$\mu_2(Y) = \frac{kq}{12p^2}(n-1)[3k(n-1)+p(n+1)]$$

$$\mu_3(Y) = \frac{k^2q(n-1)^2}{8p^3}[k(n-1)(1+q)+p(n+1)]$$

$$> 0 \ \ always,$$

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

implying that CEGU distribution is positively skewed.

Irrespective of any value of k,p and n mode of the distribution is at $Y = 0$.

In order to evaluate median for different values of p and n, we make use of the idea of quantiles. Closed form expression for quantiles are not easy to derive as the distribution function is not in a compact form. We have simulated a simulated sample of size 100 from CEGU distribution (p = 0.9, 0.6, 0.3, 0.1) and the quantile values at different probabilities are tabulated below.

Table 3.1.1 ($k = 2$)

	p_0	0.1	0.3	0.5	0.7	0.9	0.99
	0.1	0	11	23	41	76	118
	0.3	0	0	6	12	24	51
$n = 5$	0.6	0	0	0	3	8	20
	0.9	0	0	0	0	0	7
	0.1	0	3	6	11	22	44
	0.3	0	0	1	3	6	13
$n = 2$	0.6	0	0	0	1	2	5
	0.9	0	0	0	0	0	2

Mean ,median and variance for $k = 2, 3$ and for different values of *p and n* are tabulated below.

Table 3.1.2

p		$p = 0.1$		$p = 0.3$		$p = 0.6$		$p = 0.9$	
		$k = 2$	$k = 3$	$k = 2$	$k = 3$	$k = 2$	$k = 3$	$k = 2$	$k = 3$
	Mean	9	$27/2$	$7/3$	$7/2$	$2/3$	1	$1/9$	$1/6$
$n = 2$	*Median*	6	9	1	2	0	0	0	0
	Variance	$27/2$	$837/4$	$7/2$	$77/4$	1	12	$1/6$	$13/36$
	Mean	36	54	$28/3$	14	$8/3$	4	$4/9$	$2/3$
$n = 5$	*Median*	23	31	6	8	0	0	0	0
	Variance	180	3294	$140/3$	294	$40/3$	44	$20/9$	$46/9$

Remark 3.1.1 Irrespective of the value of k, as p increases, mean and variance get decreased, median is non increasing but mode remains the same.

Result 3.1.1 CEGU distribution approaches to degenerate distribution at 0 as $p \rightarrow 1$, irrespective of the values of n.

Following Steutel and Van Harn (2004), Sandhya and Latha (2019) establish that CEG distribution with any secondary distribution is id using the absolute monotonicity of its R-function. Hence CEGU is also id. Another generating function $S_Y(t)$ introduced by Steutel and Van Harn (2004) also leads to a characterization for CEG distribution. In this connection Sandhya and Latha (2019) state the following theorem.

Theorem 3.1.1 *A distribution is CEG iff its S function given by*

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

is absolutely monotone.

3.2 Reliability Properties

We define DS-NWU property for distributions having support with gap , say, k as follows.

Definition 3.2.1 The discrete distribution $\{p_{nk}, (n \geq 0 \text{ and } k > 0, \text{ integer})\}$ of a non- negative integer valued random variable M is said to be DS-NWU if $a_{(m+n+1)k} \geq a_{mk}a_{nk}$, $m, n = 0, 1, ...$ where $a_{nk} = \sum_{i=(n+1)}^{\infty} p_{ik}$, written $M \; \varepsilon \; DS-NWU$.

For CEG distribution, $p_{ik} = pq^i, i = 0, 1, 2...$, $a_{mk} = \sum_{i=(m+1)}^{\infty} p_{ik} = q^{m+1}$ and $a_{(m+n+1)} = q^{(m+n+2)} = a_{mk}a_{nk}$ which implies that the distribution is DS-NWU. By theorem (2.1.1), CEG is NWU and hence NWUE.

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

Sandhya and Latha (2019) have proved that CEG distribution with DFR secondary distribution is DFR. But the IFR property of secondary distribution is not preserved, in general, under compounding. Sandhya and Latha (2019) have illustrated this in the case of CEG distribution with an IFR secondary distribution. Here we illustrate the same result using uniform secondary distribution which is an IFR distribution.

For CEGU (k, p, n),

$$\begin{aligned} \frac{g_2}{g_1} - \frac{g_1}{g_0} &= \frac{k+1}{2} \\ &\geq 0 \ always. \\ &\Rightarrow \frac{g_2}{g_1} > \frac{g_1}{g_0} \\ &\Rightarrow \frac{g_1}{g_0} \not> \frac{g_2}{g_1} \end{aligned}$$

implying that CEGU is not IFR.

Since the hazard and reverse hazard functions of CEGU have no closed form, they are evaluated using (1.1) and (1.2). Graphs are plotted below for different values of p and n, and k.

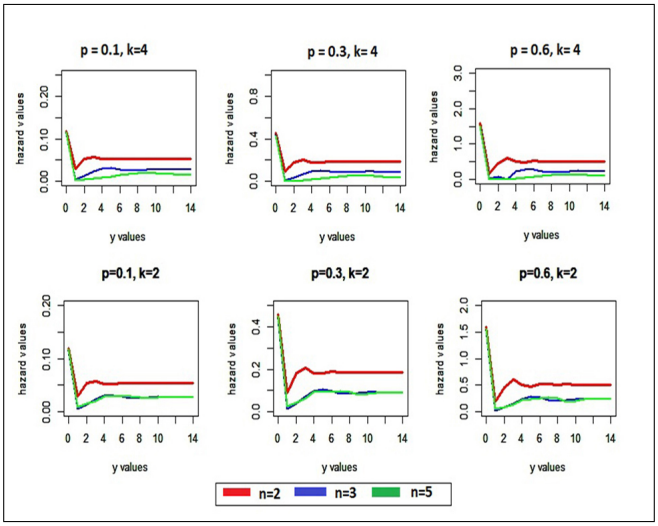

Fig 3.2.1 Hazard function graph (CEGU)

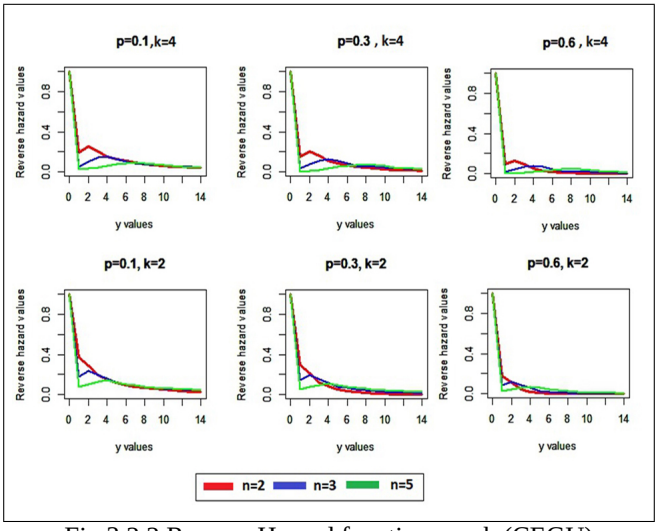

Fig 3.2.2 Reverse Hazard function graph (CEGU)

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

Remark 3.2.1 Variation in n becomes noticeable in the graph as k increases, for both hazard and reverse hazard rates. Constancy of hazard rate can be observed for $n = 2$ on the right tail of the distribution.

3.3 An AR(1) process corresponding to CEGU distribution

Sandhya and Latha (2019) defined a stationary AR(1) process with EG sum of innovations as below.

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

$$\begin{aligned} Y_{n,i} &= 0 \text{ with the probability } p \\ &= Y_{n-1,i} + \sum_{i=1}^{k} \varepsilon_{n,i} \text{ with probability } (1-p) \end{aligned} \tag{3.3.1}$$

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 $Y_{n,i} \overset{d}{=} Y_{n-1,i}\ \forall i$

Theorem 3.3.1 *A sequence $\{Y_{n,i}\}$ given by (3.3.1) defines a stationary AR(1) process for some p iff it is extended geometric sum of innovations $\{\varepsilon_{n,i}\}$.*

Sample path of the AR(1) process defined by (3.3.1)is displayed below for simulated data, for different values of p and n by assuming CEGU for $\{\varepsilon_{n,i}\}$

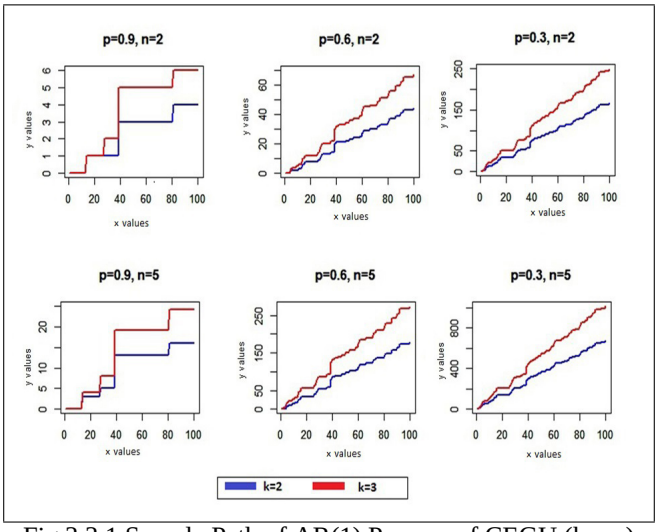

Fig 3.3.1 Sample Path of AR(1) Process of CEGU (k,p,n)

Remark 3.3.1 As in the case of CGU distribution,the sample path of CEGU AR(1) process is a step function for values of p approaching 1. Variation in k does not affect the shape of the sample path but it affects the height of the jumps.

3.4 Simulation and Estimation

As uniform distribution is nonparametric, CEGU distribution has only two parameters. Maximum likelihood estimation (mle) method cannot be used as the pmf has no compact form. So they are estimated using (1) Moment estimation method and (2) BHHJ method.

Moment estimation

Let m_1, m_2 denote the sample raw moments. Moment estimators of k and p are obtained by solving the following equations using 'nleqslv' package in R. The equations and estimates for different values of p are given below.

$$2pm_1 - kq(n-1) = 0 \tag{3.4.1}$$

$$12p^2m_2 - 3k^2q(n-1)^2(1+q) - kpq(n^2-1) = 0 \tag{3.4.2}$$

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

Case 1: When *p* is unknown.

For known values of n, the parameter *p* is estimated using (3.4.1)

Table 3.4.1
Moment estimates using simulated sample of size 100 , no. of replications 50 .

$n = 5$

p		*k*=2	*k*=3
0.2	Estimate	0.204632	0.210882
	Mean bias	0.004632	0.010882
	MSE	0.000334	0.0.000411
0.5	Estimate	0.506512	0.506507
	Mean bias	0.006512	0.006507
	MSE	0.001388	0.001320
0.8	Estimate	0.806847	0.803091
	Mean bias	0.006847	0.006563
	MSE	0.001430	0.001325

Case 2: When two parameters are unknown.

For known values of n, CEG distribution has two unknown parameters and are obtained by solving equations (3.4.1) and (3.4.2) using nleqslv package in R.

Table 3.4.2
Moment estimates using simulated sample of size 100 and no. of replications 10 $(n = 5)$.

		$k = 2$		$k = 3$	
p		$p^{\wedge}$	$k^{\wedge}$	$p^{\wedge}$	$k^{\wedge}$
0.2	*Mean estimate*	0.241371	2.607611	0.147864	2.404951
	Mean bias	-0.041371	-0.607612	0.052135	0.059504
	MSE	0.007387	1.571391	0.011350	1.014295
0.5	*Mean estimate*	0.590976	2.755884	0.556360	3.363823
	Mean bias	-0.090976	-0.755884	-0.056360	-0.363823
	MSE	0.012602	1.034691	0.007078	0.723615
0.8	*Mean estimate*	0.854548	2.829627	0.848387	3.507484
	Mean bias	-0.054548	-0.829627	-0.048358	-0.507484
	MSE	0.004188	0.52796	0.003344	1.101041

BHHJ estimation

BHHJ estimates are obtained by minimizing equation (1.3) using ”nloptr” package in R.

Case 1: When *p* is unknown.

Table 3.4.3
BHHJ estimates using simulated sample of size 100 , no. of replications 50 .

	p		0.2	0.5	0.8
$n = 5$	k=2	Estimate	0.203416	0.502162	0.804255
		Mean bias	0.003416	0.002162	0.004255
		MSE	0.000690	0.001485	0.001444
	k=3	Estimate	0.204370	0.503696	0.803091
		Mean bias	0.004370	0.003696	0.003912
		MSE	0.000778	0.001498	0.001351

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

Case 2: When two parameters are unknown.

Table 3.4.4
BHHJ estimates using simulated sample of size 100 and no. of replications 20 ($n = 5$).

		$k = 2$		$k = 3$	
p		$k^{\wedge}$	$p^{\wedge}$	$k^{\wedge}$	$p^{\wedge}$
	Mean estimate	2.1944230	0.2205859	3.1317359	0.2128203
0.2	*Mean bias*	-0.1944230	-0.0205859	-0.1317359	-0.0128203
	MSE	0.1705347	0.0018693	0.3234513	0.0020037
	Mean estimate	2.1259131	0.5207505	2.9856644	0.4975384
0.5	*Mean bias*	-0.1259131	-0.0207505	0.0143355	0.0024615
	MSE	0.0158541	0.0026717	0.2218073	0.0017699
	Mean estimate	1.9962663	0.7845363	2.9037136	0.7913346
0.8	*Mean bias*	0.0037336	0.0154363	0.0962863	0.0086653
	MSE	0.1004599	0.0013723	0.1273030	0.0015122

3.5 Fitting of CEGU using Real Life Data Set

Ridout et.al (2001) give the no. of roots produced by micro- propagated shoots of the columnar apple cultivar Trajan. The roots had been produced under an 8- h or 16-h photoperiod in culture systems that utilized one of four different concentrations of the cytokinin in BAP culture medium. Let Group I (Gr I) consist of the data produced under 8 hour period and Group II (Gr II) consist of the data produced under 16 hour photoperiod.

Table 3.5.1

Number of roots	*Obs. fr.* (GrI)	*Obs. fr.* $(GrII)$
0	2	62
1	3	7
2	6	7
3	7	8
4	13	8
5	12	6
6	14	10
7	17	4
8	21	2
9	14	7
10	13	4
11	10	2
12	2	3
13	2	0
14	3	0
15	0	0
16	0	0
17	1	0
Total	140	130

Here we analyze data with observed frequencies of Group II. The parameters k and p are estimated using moment method and are given by $p^{\Lambda} = 0.43$ and $k^{\Lambda} \simeq 2$ for $n = 3$. The model is fitted using chi square test and the details are given below.

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

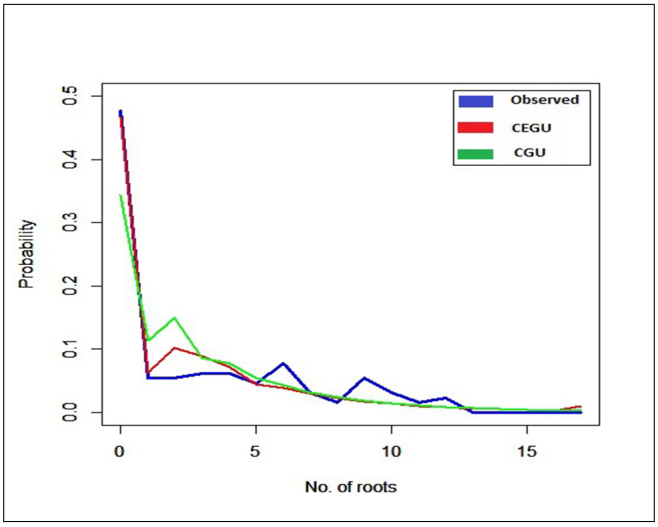

Fig 3.5.1

Table 3.5.2

Obs. fr. $(GrII)$	*Expected frequency.* $(CEGU)$	*Expected frequency.* (CGU)
62	60.708724	44.708988
7	8.089586	14.666441
7	13.212336	19.477655
8	11.467098	11.2007101
8	9.291633	10.063803
6	5.829716	6.975661
10	5.042330	5.589666
4	3.839281	4.121959
2	2.915338	3.185824
7	2.216181	2.397262
4	1.725570	1.831489
2	1.317176	1.387209
3	1.009971	1.055869
0	0.775449	0.801433
0	0.595687	0.609274
0	0.456712	0.462771
0	0.350544	0.351676
0	1.15664	1.112273
Total	130	130
$d.f$	6	7
Chi square value	10.54589	28.97532
p value	0.103466	0.0001461

The p values show that CGU is not fit for the data whereas CEGU provides a good fit.

Conclusion

Compounding changes so many aspects of the basic distribution. From a distribution, which is symmetric, we get a positively skewed distribution, a distribution which has no mode, to a distribution with a fixed mode at 0, from a distribution with finite support to a distribution with infinite support. At the same time, the calculation of probabilities become tedious due to compounding. The CGU and CEGU probability functions behave alike, approaching to degenerate distribution at 0, as $p \to 1$. The IFR property of uniform distribution is not preserved under both geometric and EG compounding.

http://dx.doi.org/10.29322/IJSRP.9.12.2019.p9635

Observing the hazard rate values of the CGU and CEGU distributions,it can be inferred that both distributions are not IFR/DFR in general. For $n = 2$, CGU distribution shows constant hazard rate. Thus by compounding a constant failure rate distribution with an IFR distribution, we get a constant failure rate distribution. For higher values of p, for different values of k, the sample path is a step function. The significance of the gap k is evident from the real life data set given by Ridout et.al (2001). CGU distribution is not fit for the data, but CEGU provides a good fit.

REFERENCES

[1] Bon, J.-L. (2006). *Error bounds for exponential approximation of large-system reliability*, Journal of mathematical Science (N.Y.) 138, p: 53665376.

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

[3] Brown, M. (2015). *Sharp bounds for exponential approximations under a hazard rate upper bound*, Journal of Applied Probability to appear.

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

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

[6] Kalashnikov,V., (1997). *Geometric sums: Bounds for rare events with applications*, Kluwer, Dordrecht.

[7] Milidui. R. ,(1985). *The computation of compound distributions with Tuoridal severities.* Ph.D dissertation, Dept. of Industrial Engineering and Operations Research, University of California at Berkeley.

[8] Johnson N.L., A.W. Kemp and Samuel Kotz, (2005). *Univariate discrete distributions*, Third editions John wily and sons.

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

[10] Paul Embrechts, Maejima, M., and Teugels, J. (1985). *Asymptotic behaviour of compound distributions*, Astin Bulletin, 15, p: 45-48.

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

[12] Pekoz, E. A. and R ollin, A.,(2011). *New rates for exponential approximation and the theorems of Renyi and Yaglom*. Annals of Probability, 39, p:587608.

[13] Pekoz, E. A., R ollin, A. and Ross, N. (2013). *Total variation error bounds for geometric approximation*. Bernoulli 19, p: 610632.

[14] Ridout, M., Hinde, J., Demetrio, C. G. B. (2001). *A score test for testing zero-inflated Poisson regression model against zero-inflated negative binomial alternatives*, Biometrics 57, p:219-223.

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

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

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

[18] Sundt, B., (1982). *Asymptotic behaviour of compound distributions and stop- loss premiums*, ASTIN Bulletin, 13, p: 89-98.

[19] Szekli,R. (1986). *On the concavity of the waiting time distribution in some GI/G/1 queues*, Journal of Applied Probability, 23, p:555-561.

[20] Willmot, G. E., (1989). *Limiting tail behaviour of some discrete compound distributions*, Insurance: Mathematics and Economics, 8, P: 175-185. https://doi.org/10.1016/0167-6687(89)90055-3

[21] Willmot, G., E., and Cai, J., (2004). *On applications of residual lifetimes of compound geometric convolutions*, Journal of Applied Probability, 41, p:802-815, DOI: 10.1239/jap/1091543427

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

[23] Willmot, G. E., (2002). *Compound geometric residual lifetime distributions and the deficit at ruin*, Insurance Mathematics and Economics, 30, p: 421-438.

[24] Zafakali, Nur Syabiha binti and Ahmad, Wan Muhamad Amir bin W ,(2013). *Modeling and Handling Overdispersion Health Science Data with Zero-Inflated Poisson Model*, Journal of Modern Applied Statistical Methods: Vol. 12 : Iss. 1 , Article 28. DOI: 10.22237/jmasm/1367382420.

First Author- Latha C M, M.Sc, M.Phil, Department of Statistics, St.Thomas College, Pala-686574, India, cmlatha.krishnakumar@gmail.com

Second Author- Sandhya E, M.Sc, Ph.D, Department of Statistics, Prajyothi Nikethan College, Pudukkad-680301, India, esandhya@hotmail.com