A CONTINUOUS VERSION OF THE NEGATIVE BINOMIAL DISTRIBUTION

N.K. Chandra, D. Roy

1. INTRODUCTION

Continuous and discrete models in the statistical literature have their independent courses of development creating a gap between these two sets of distributions. In the recent past, attempts have been made by various authors to bridge this gap. Roy's discrete concentration approach (Roy, 1993) is aimed at generating a discrete distribution from a continuous distribution maintaining identical survival probabilities at the discrete points. Kemp's approach (Kemp, 2004) on the other hand maintains similarly with the probability density curve at discrete points and thereby generates discrete mass function under appropriate normalization. Recently, discrete distributions have been generated from the continuous set up by Roy and Ghosh (2009), maintaining similarly with the failure rate function. But all these approaches are discretization in nature in the sense these provide with discrete models from continuous models. The reverse approach of generating continuous model from discrete model has been touched upon in Roy(2002). But very little work has been done in this direction. The present work is an attempt to generate a continuous model from discrete and undertake a detailed study.

It is known that the geometric distribution is a discrete version of the exponential distribution (Roy,1993). Alternatively we may say that continuous version of the geometric distribution is the exponential distribution. Negative binomial distribution is a generalization of the geometric distribution and hence, the aim of the present paper is to develop a continuous version of the negative binomial distribution.

The negative binomial distribution has been studied by many authors from many directions. Under different parametrization, this distribution was studied by Fisher (1941), Jeffreys (1941) and Anscombe (1950). It has been shown to be the limiting form of Eggenberg and Polya's urn model by Patil et al.(1984). It has been shown to be the gamma mixture of Poisson distribution by Greenwood and Yule (1920), convolution of geometric distribution by Feller(1957) and addition of a set of correlated Poisson distribution by Martiz(1952). The negative binomial distribution also arises out of a few stochastic processes as pointed by McKendrick (1914), Irnsin (1941), Lundberg (1940) and Kendall (1949). This distribution, being more flexible than Poisson distribution, enjoys a plethora of applications. It can be used to model accident data, psychological data, economic data, consumer expenditure data, medical data, defence data and so on.

In view of its flexibility and wide applicability it may be of interest to examine a possible continuous version of the negative binomial distribution. In section 2 we propose one such continuous model along with its basic properties. In section 3 we present a special choice of this distribution from reliability point of view and ensure some interesting life distribution orderings in terms of reliability measures.

2. CONTINUOUS VERSION OF THE NEGATIVE BINOMIAL DISTRIBUTION

Let X be a nonnegative continuous random variable with survival function given by S(x). We define the continuous version of the negative binomial distribution(NBD) by the following form of the survival function:

$$S_r(x) = \begin{cases} q^x & \text{for } \quad r = 1 \\ \sum_{k=0}^{r-1} \binom{x+k-1}{k} p^k q^x & \text{for } \quad r = 2,3,..... \end{cases} \tag{1}$$

where $q = e^{-\lambda}$, $\lambda \geq 0$, $p + q = 1$ and r taking only the positive integer values.

To ensure that (1) is a valid survival function we need to verify

$$S_r(0) = 1 \tag{2}$$

$$S_r(x+0) = S_r(x) \tag{3}$$

$$S_r(x) > S_r(y) \quad \text{for } 0 \leq x < y < \infty \tag{4}$$

$$S_r(\infty) = 0 \tag{5}$$

It is easy to note (2) and (3) because $S_r(x)$ is continuous in x. To verify (4) we present the following theorem.

Theorem 2.1 $S_r(x)$ is a decreasing function in x for all $x \geq 0$.

Proof: To prove this theorem we make use of the following lemma

Lemma 2.1

$$S_r(x) = \frac{1}{B(x,r)} \int_0^q z^{x-1}(1-z)^{r-1} dz \, .$$

Proof:

$$S_r(x) = \sum_{k=0}^{r-1} \binom{x+k-1}{k} p^k q^x$$

$$= \sum_{k=0}^{r-1} \frac{(x+k-1)!}{k!(x-1)!} p^k q^x$$

$$= \frac{1}{(x-1)!(r-1)!} \sum_{k=0}^{r-1} \frac{(r-1)!(r-1-k)!(x+k-1)!}{k!(r-1-k)!} p^k q^x$$

$$= \frac{\Gamma(x+r)}{\Gamma(x)\Gamma(r)} \sum_{k=0}^{r-1} \binom{r-1}{k} \frac{\Gamma(r-k)\Gamma(x+k)}{\Gamma(x+r)} p^k q^x$$

$$= \frac{1}{B(x,r)} \sum_{k=0}^{r-1} \binom{r-1}{k} B(r-k, x+k) p^k q^x$$

$$= \frac{q^x}{B(x,r)} \sum_{k=0}^{r-1} \binom{r-1}{k} \int_0^1 t^{r-k-1} (1-t)^{x+k-1} dt \ \mathrm{p}^{\mathrm{k}}$$

$$= \frac{q^x}{B(x,r)} \int_0^1 t^{r-1} (1-t)^{x-1} \sum_{k=0}^{r-1} \binom{r-1}{k} \left[\frac{\mathrm{p}(1-\mathrm{t})}{\mathrm{t}} \right]^k dt$$

$$= \frac{q^x}{B(x,r)} \int_0^1 t^{r-1} (1-t)^{x-1} \left[1 + \frac{p(1-t)}{t} \right]^{r-1} dt$$

$$= \frac{q^x}{B(x,r)} \int_0^1 t^{r-1} (1-t)^{x-1} \left[\frac{1-q(1-t)}{t} \right]^{r-1} dt$$

$$= \frac{q^x}{B(x,r)} \int_0^1 (1-t)^{x-1} [1-q(1-t)]^{r-1} dt$$

$$= \frac{1}{B(x,r)} \int_0^q z^{x-1} (1-z)^{r-1} dz .$$

Making use of this lemma, for any $h > 0$ we get

$$\frac{1}{S_r(x+h)} = \frac{\int_0^1 z^{x+h-1}(1-z)^{r-1}\,dz}{\int_0^q z^{x+h-1}(1-z)^{r-1}\,dz}$$

$$= 1 + \frac{\int_q^1 z^h z^{x-1}(1-z)^{r-1}\,dz}{\int_0^q z^h z^{x-1}(1-z)^{r-1}\,dz}$$

$$\geq 1 + \frac{\int_q^1 z^{x-1}(1-z)^{r-1}\,dz}{\int_0^q z^{x-1}(1-z)^{r-1}\,dz}$$

$$= \frac{1}{S_r(x)}$$

Thus,

$$S_r(x) \geq S_r(x+h),$$

i.e. $S_r(x) \downarrow x$.

Lastly, we need to ensure (5). To this end we present the following theorem.

Theorem 2.2

$$S_r(\infty) = 0 .$$

Proof: We have

$$S_r(x) = \left[1 + \binom{x}{1} p + \binom{x+1}{2} \mathrm{p}^2 + \cdots + \binom{x+r-2}{r-1} \mathrm{p}^{r-1}\right] q^x .$$

But

$$1 + \binom{x}{1} p + \binom{x+1}{2} \mathrm{p}^2 + \cdots + \binom{x+r-2}{r-1} \mathrm{p}^{r-1}$$

is a polynomial function in x and q^x is an exponential function in x and hence

$$S_r(\infty) = \lim_{x\to\infty} S_r(x) = 0$$

Thus, we claim that (1) is a properly defined survival function.

To ensure that $S_r(x)$ is a continuous version of NBD we need to show that discrete concentration of $S_r(x)$ gives rise to NBD with parameters (*r, p*) which we denote by NB(*r, p*).The following theorem presents this property of $S_r(x)$.

Theorem 2.3 NB(*r*, *p*) is a discrete version of the proposed distribution with survival function as $S_r(x)$.

Proof: We need to show that discrete concentration of $S_r(x)$ is NB(*r, p*).

Here

$$S_r(x) - S_r(x+1)$$

$$= \sum_{k=0}^{r-1} \binom{x+k-1}{k} p^k q^x - \sum_{k=0}^{r-1} \binom{x+k}{k} \mathrm{p}^{\mathrm{k}} q^{x+1}$$

$$= \sum_{k=0}^{r-1} \left[\binom{x+k-1}{k} - \binom{x+k}{k} q \right] \mathrm{p}^{\mathrm{k}} q^x$$

$$= \sum_{k=0}^{r-1} \left[\binom{x+k-1}{k} - \binom{x+k}{k} + \binom{x+k}{k} p \right] \mathrm{p}^{\mathrm{k}} q^x$$

$$= \sum_{k=0}^{r-1} \left[\binom{x+k-1}{k} - \binom{x+k}{k} \right] \mathrm{p}^{\mathrm{k}} q^x + \sum_{k=0}^{r-1} \binom{x+k}{k} p^{k+1} q^x$$

$$= -\sum_{k=1}^{r-1} \binom{x+k-1}{k-1} \mathrm{p}^{\mathrm{k}} q^x + \sum_{k=0}^{r-1} \binom{x+k}{k} p^{k+1} q^x$$

$$= -\sum_{k=0}^{r-2} \binom{x+k}{k} \mathrm{p}^{\mathrm{k+1}} q^x + \sum_{k=0}^{r-1} \binom{x+k}{k} p^{k+1} q^x$$

$$= \binom{x+r-1}{r-1} \mathrm{p}^{\mathrm{r}} q^x$$

which is the pmf of a negative binomial distribution.

Thus, the discrete version of the continuous random variable *X* whose survival function is given in (1) is a NB(*r*, *p*).

3. SPECIAL CASES OF $S_r(x)$

It may be noted that for $r = 1$, $S_r(x)$ reduces to the exponential law. In this sense $S_r(x)$ is a generalization of the exponential distribution. In case $r = 2$ we can get yet another distribution which is of special interest. For $r = 2$, (1) reduces to

$$S_2(x) = (1 + px)q^x ,$$

with cdf

$$F_2(x) = 1 - (1 + px)q^x ,$$

and pdf

$$f_2(x) = -pq^x - (1 + px)q^x(-\lambda) , \because \quad \mathrm{q} = \mathrm{e}^{-\lambda}$$

$$= [(\lambda - p) + \lambda px]q^x , x \geq 0.$$

Note that

$$\lambda - p = \lambda - 1 + e^{-\lambda} = g(\lambda) \text{, say.}$$

Then

$$g(\lambda) = 0, \text{ if } \lambda = 0$$

and

$$\frac{d}{d\lambda} g(\lambda) = 1 - e^{-\lambda} \geq 0 \text{, since } \lambda \geq 0 .$$

This means $g(\lambda)$ is a non-decreasing function of λ and hence $g(\lambda) \geq 0$ for all λ i.e. $\lambda - p \geq 0$. Thus $f_2(x) \geq 0$ for all $x \geq 0$ and

$$S_2(x) - S_2(x+1) = \binom{x+1}{1} \mathrm{p}^2 q^x .$$

In view of Theorem 2.3 discretization of the continuous random variable X with survival function $S_2(x)$ is given by

$$P[X* = x^*] = \binom{x*+1}{1} \mathrm{p}^2 q^{x*} , \; x* = 0, \; 1, \; 2, \ldots$$

which is NB(2, p).

We propose to study the properties of $S_2(x)$ in details with special reference to reliability measures. Distribution of $S_2(x)$ will be referred as continuous version of NB(2, *p*) and will be abbreviated as CNB(2, *p*).

3.1. *Properties of* CNB(2, *p*)

For the subsequent presentation we can rewrite $S_2(x)$, $F_2(x)$ and $f_2(x)$ as follows:

$$S_2(x) = (1+px)q^x = (1+px)e^{-\lambda x},$$

$$F_2(x) = 1-(1+px)e^{-\lambda x},$$

$$f_2(x) = (\lambda - p + \lambda px)e^{-\lambda x}$$

where $x \geq 0$, $q = e^{-\lambda}$, $\lambda \geq 0$ and $p + q = 1$.

Remark: In the above expressions *p* has been used for notational simplifications.

To examine the shape of the density curve in respect of location, spread, skewness and kurtosis we need to derive the expressions for first four moments of CNB(*2*, *p*). By definition, moment generating function $M_X(t)$ of X following CNB(2, *p*) is given by

$$M_X(t) = \mathrm{E}e^{tx}$$

$$= (\lambda - p)\int_0^\infty e^{tx-\lambda x}dx + \lambda p\int_0^\infty xe^{tx-\lambda x}dx \text{, where } p = 1-e^{-\lambda}$$

$$= (\lambda - p)\int_0^\infty e^{-(\lambda-t)x}dx + \lambda p\int_0^\infty xe^{-(\lambda-t)x}dx$$

$$= (\lambda - p)(\lambda - t)^{-1} + \lambda p(\lambda - t)^{-2}.$$

Hence

$$M_X^{(1)}(t) = (\lambda - p)(\lambda - t)^{-2} + 2\lambda p(\lambda - t)^{-3}$$

$$M_X^{(2)}(t) = 2(\lambda - p)(\lambda - t)^{-3} + 6\lambda p(\lambda - t)^{-4}$$

$$M_X^{(3)}(t) = 6(\lambda - p)(\lambda - t)^{-4} + 24\lambda p(\lambda - t)^{-5}$$

$$M_X^{(4)}(t) = 24(\lambda - p)(\lambda - t)^{-5} + 120\lambda p(\lambda - t)^{-6}$$

where $M_X^{(r)}(t)$ is the rth order derivative of $M_X(t)$.

Then the first four raw moments are given by

$$\mu_1' = M_X^{(1)}(0) = \frac{\lambda - p}{\lambda^2} + \frac{2p}{\lambda^2} = \frac{\lambda + p}{\lambda^2}$$

$$\mu_2' = M_X^{(2)}(0) = \frac{2(\lambda - p)}{\lambda^3} + \frac{6p}{\lambda^3} = \frac{2(\lambda + 2p)}{\lambda^3}$$

$$\mu_3' = M_X^{(3)}(0) = \frac{6(\lambda - p)}{\lambda^4} + \frac{24p}{\lambda^4} = \frac{6(\lambda + 3p)}{\lambda^4}$$

$$\mu_4' = M_X^{(4)}(0) = \frac{24(\lambda - p)}{\lambda^5} + \frac{120p}{\lambda^5} = \frac{24(\lambda + 4p)}{\lambda^5}$$

Corresponding central moments can be obtained as follows:

$$\mu_2 = \mu_2' - \mu_1'^2 = \frac{2(\lambda + 2p)}{\lambda^3} - \frac{(\lambda + p)^2}{\lambda^4} = \frac{2\lambda(\lambda + 2p) - (\lambda + p)^2}{\lambda^4} = \frac{\lambda^2 + 2\lambda p - p^2}{\lambda^4}$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$= \frac{6(\lambda + 3p)}{\lambda^4} - \frac{6(\lambda + 2p)(\lambda + p)}{\lambda^5} + \frac{2(\lambda + p)^2}{\lambda^6}$$

$$= \frac{2(\lambda^3 + 3p\lambda^2 - 3p^2\lambda + p^3)}{\lambda^6}$$

$$= \frac{2[(\lambda - p)^3 + 6\lambda p(\lambda - p) + 2p^3]}{\lambda^6}$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$= \frac{3(3\lambda^4 + 12p\lambda^3 - 10p^2\lambda^2 + 4p^3\lambda - p^4)}{\lambda^8}.$$

It may be observed that μ_3 is nonnegative and hence it can be claimed that CNB(2, p) is positively skewed.

Noting that $p = 1 - e^{-\lambda}$, one can estimate λ based on the sample mean following the method of moments. Alternatively, λ can be estimated numerically from the following likelihood equation

$$\lambda - \frac{1}{\sum_{i=1}^{n} x_i} \sum_{i=1}^{n} \frac{(1 - e^{-\lambda})(1 + x_i) + \lambda e^{-\lambda} x_i}{1 - (1 - e^{-\lambda})/\lambda + (1 - e^{-\lambda}) x_i} = 0$$

based on n random sample observations $x_1, x_2, \ldots, x_n$.

The corresponding iterative equation is given by

$$\lambda(j+1)=\frac{1}{\sum_{i=1}^{n}x_i}\sum_{i=1}^{n}\frac{(1-e^{-\lambda(j)})(1+x_i)+\lambda(j)e^{-\lambda(j)}x_i}{1-(1-e^{-\lambda(j)})/\lambda(j)+(1-e^{-\lambda(j)})x_i} \tag{6}$$

where $\lambda(j)$ is the iterative solution at the *j*th stage, $j = 1, 2, \ldots$..

Regarding the choice of λ

(i) we note that $E[X] \le E(X) \le E[X] + 1$, where the Gauss symbol $[X]$ denotes the largest integer contained in X, and $[X]$ has NB(2, p) distribution with $E[X] = 2(1-p)/p$. Thus we may take

$$E(X) \simeq E[X]+\frac{1}{2}=\frac{4-3p}{2p},$$

and

$$\lambda(1)=\ln\frac{2\bar{x}+3}{2\bar{x}-1}$$

where $\bar{x}$ is the sample mean.

(ii) the iterative equation(6) has been framed out of the likelihood equation, following Lehmann and Casella (1998), it is expected that the iterative procedure has a solution that converges to the true value in probability.

3.2. *Reliability measures*

It is easy to note that the failure rate, $r(x)$, of CNB(2, p) is

$$r(x)=\frac{f_2(x)}{S_2(x)}=\lambda-\frac{p}{1+px}$$

which is an increasing function of x for all $x \ge 0$. Hence follows the following theorem.

Theorem 3.1 $F_2(x) \in$ IFR class.

Another measure of importance in reliability analysis is the mean residual life (MRL). By definition MRL after an elapsed time t is

$$E[X-t \mid X \ge t]=\frac{1}{S_2(t)}\int_t^{\infty} S_2(x)dx,$$

for CNB(2, p). This works out as

$$\mathrm{E}[X-t \mid X \geq t]=\frac{1}{(1+pt)e^{-\lambda t}}\int_t^{\infty}(1+px)e^{-\lambda x}dx$$

$$=\frac{1}{(1+pt)e^{-\lambda t}}\int_0^{\infty}(1+pt+px)e^{-\lambda t-\lambda x}dx$$

$$=\frac{1}{1+pt}\left[(1+pt)\int_0^{\infty}e^{-\lambda x}dx+p\int_0^{\infty}xe^{-\lambda x}dx\right]$$

$$=\frac{1}{1+pt}\left[\frac{1+pt}{\lambda}+\frac{p}{\lambda^2}\right]=\frac{1}{\lambda}+\frac{p}{\lambda^2(1+pt)},$$

providing an analytically compact form for MRL.

It may be noted that this MRL is a decreasing function in t for all $t \geq 0$. Hence CNB(2, p) has decreasing mean residual life (DMRL) property and is consistent with its IFR property.

Stochastic ordering of life distributions (see Shaked and Shantikumar, 1994) has assumed importance in the recent past. Very frequently examined orderings of distributions are in terms of survival function, failure rate (also known as hazard rate) and likelihood ratio giving rise to stochastic ordering, hazard rate ordering and likelihood ratio ordering. It is known that stochastic ordering, helps us to order the mean values. Similarly under hazard rate ordering with additional class property one may order the variables in terms of variability. Further, it may be noted that likelihood ratio ordering implies the above two special observations as it is a more restrictive condition.

Let us examine ordering of CNB(2, p) with respect to gamma distribution of special choice.

Theorem 3.2 Let X ~ CNB(2, p) and Y ~ gamma with survival function given by

$$\bar{G}(x)=(1+\lambda x)e^{-\lambda x}, \quad x \geq 0, \qquad \lambda \geq 0.$$

Then the following orderings hold true:

(i) $X \leq_{st} Y$

(ii) $X \leq_{hr} Y$

(iii) $X \leq_{lr} Y$

Proof: Here

$$S_2(t)=(1+pt)e^{-\lambda t}$$
$$\bar{G}(t)=(1+\lambda t)e^{-\lambda t}.$$

We have already shown that $\lambda - p > 0$ and hence

$$\frac{S_2(t)}{\bar{G}(t)} = \frac{1+pt}{1+\lambda t} \le 1, \ \forall \ \ t \ge 0$$

$$\Rightarrow S_2(t) \le \bar{G}(t), \ \forall \ \ t \ge 0$$

$$\Leftrightarrow X \le_{st} Y.$$

$$X \le_{hr} Y \Leftrightarrow \frac{S_2(t)}{\bar{G}(t)} \downarrow t .$$

Writing

$$h(t) = \frac{S_2(t)}{\bar{G}(t)} = \frac{1+pt}{1+\lambda t}.$$

We note that

$$\frac{d}{dt} h(t) = \frac{p(1+\lambda t) - \lambda(1+pt)}{(1+\lambda t)^2} = \frac{p-\lambda}{(1+\lambda t)^2} \le 0.$$

Hence $h(t) \downarrow t$. This proves (ii).
Further,

$$\frac{f_2(t)}{g(t)} = \frac{\lambda(1+pt)-p}{\lambda^2 t} = \frac{\lambda - p}{\lambda^2 t} + \frac{p}{\lambda} \downarrow t$$

Hence follows the proof of (iii).

Similar ordering of CNB(2, p) can also be studied with respect to the exponential distribution.

Department of Statistics NIMAI KUMAR CHANDRA
University of Kalyani, India

Center for Management Studies DILIP ROY
University of Burdwan, India

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the comments given by the referee, which have helped to improve the presentation of the paper.

REFERENCES

F.J. ANSCOMBE, (1950), *Sampling theory of the negative binomial and logarithmic series distributions*, "Biometrika", 37, pp. 358-382.

F. EGGENBERGER, G. PÓLYA, (1923), "Über die Statistik verketteter Vorgänge, Zeitschrift für angewandte Mathematik und Mechanik", 1, pp. 279-289.

W. FELLER, (1978), *An Introduction to Probability Theory and Its Applications*, Vol. I, Wiley Eastern Limited, New Delhi.

R.A. FISHER, (1941), "Annals of Eugenics", London, 11, pp. 182-187.

M. GREENWOOD, G.U. YULE, (1920), *An enquiry into the nature of frequency distributions of multiple happenings, with particular reference to the occurrence of multiple attacks of disease or repeated accidents*, "JRSS Series A", 83, pp. 255-279.

J.O. IRWIN, (1941), *Discussion on Chambers and Yuie's paper*, "JRSS" Supplement 7, pp. 101-107.

H. JEFFREYS, (1941), *Some applications of the method of minimum chi-squared*, "Annals of Eugenics", London 11, pp. 108-114.

A.W. KEMP, (2004), *Classes of discrete life distributions*, "Commu in Stat, Theo and Meth", 33, pp. 3069-3093.

D.G. KENDALL, (1949), *Stochastic processes and population growth*, "JRSS Series B", 11, pp. 230-282.

E.L. LEHMANN, G. CASELLA, (1998), *Theory of Point Estimation*, Second Edition, Springer-Verlag, New York.

O. LUNDBERG, (1940), "*On random processes and their application to sickness and accident statistics*", Uppsala, Almqvist and Wicksells.

J.S. MARTIZ, (1952), *Note on a certain family of discrete distributions*, "Biometrika", 39, pp. 196-198.

A.G. MCKENDRICK, (1914), *Studies on the theory of continuous probabilities with special reference to its bearing on natural phenomena of a progressive nature*, "Proceedings of the London Mathematical Society", 13(2), pp. 401-416.

G.P. PATIL, M. T. BOSWELL, S. W. JOSHI, M. V. RATNAPARKHI, (1984), "*Dictionary and Bibliography of Statistical Distributions in Scientific Work*", Vol. 1, Discrete Models, Fairland, MD: International Co-operative Publishing House.

D. ROY, (1993), *Reliability measures in discrete bivariate setup and related characterization results for a bivariate geometric distribution*, "Journal of Multivariate Analysis", 48(2), pp. 362-373.

D. ROY, (2002), *Discretization of continuous distributions with an application to stress-strength Reliability*, "Calcutta Statistical Association Bulletin", 52, pp. 297-313.

D. ROY, T. GHOSH, (2009), *A new discretization approach with application in reliability estimation*, "IEEE Trans. on Reliability", 58, pp. 456-461.

M. SHAKED, J. G. SHANTHIKUMAR, (1994), *Stochastic orders and their applications*. Academic Press, Inc.

SUMMARY

A continuous version of the negative binomial distribution

While discretization of continuous distributions have been attempted for many life distributions the reverse has hardly been attempted. The present endeavor is to establish a reverse relationship by offering a continuous counter part of a discrete distribution namely negative binomial distribution. Different properties of this distribution have been established for a special choice of the parametric value covering class properties, ordering and mean residual life.