<Research Paper>

Journal of the Korean
Statistical Society
Vol. 11, No. 2, 1982

# A Characterization of Negative Binomial Distribution Truncated at Zero

R. Shanmugam*

**ABSTRACT**

Analogous to Singh's(1978) characterization of positive-Poisson distribution and Shanmugam and Singh's(1982) characterization of logarithmic series distribution,a characterization and its statistical application of the negative binomial distribution truncated at zero are given in this paper. While it is known that under certain conditions the negative binomial distribution truncated at zero approaches the positive-Poisson and the logarithmic series distributions, we show here that the results of this paper approach in limit the results of Singh, and Shanmugam and Singh, respectively. Using the biological data from Sampford(1955), we illustrate our results. Also, expressions are explicitly given to test the hypothesis whether a random sample is indeed from a geometric distribution.

## 1. Introduction

Let $X_1, X_2, \cdots, X_n$, $n \geqslant 2$, be mutually independent positive integer random variables ($r.v.$'s). Singh(1978) gave the following necessary and sufficient condition for these variables to have the same positive-Poisson distribution (Poisson distribution truncated at zero): The joint conditional distribution of these variables for each fixed sum $\sum X_i$ is a "uniform" truncated multinomial distribution. Using this characterization, Singh obtained a test statistic $\sum X_i^2$ to decide whether a discrete positive random phenomenon is governed by a positive-Poisson law. Following Singh, an analogous characterization of logarithmic series distribution and its statistical application of testing whether $X_1$,

*University of Colorado at Denver

$X_2, \cdots, X_n$ do follow a logarithmic series probabilistic law were given in Shanmugam and Singh(1982).

Of interest to us here is the negative binomial probability distribution truncated at zero. When neither the positive-Poisson nor the logarithmic series distribution explains satisfactorily the data in which the zero class is either distorted or unobserved, the negative binomial distribution truncated at zero is usually considered as an alternative. Areas in which the truncated negative binomial distribution has been used are summarized in Johnson and Kotz(1969). In Section 2, the negative binomial distribution truncated at zero is defined and its characterization is given. A statistical application of our characterization with a numerical example illustrating our results is considered in Section 3. While it is known that under certain conditions the negative binomial distribution truncated at zero approaches the positive-Poisson and the logarithmic series distributions, we show in Section 4 that the results of this paper approach in limit the results given in Singh, and Shanmugam and Singh, respectively. In Section 5, expressions are explicitly given to test whether a random sample is indeed from a geometric distribution.

## 2. A Characterization of the Negative Binomial Distribution Truncated at Zero

A r.v. $X$ is said to have a negative binomial distribution truncated at zero if its probability function (p.f.) is

$$g(x\,;\theta,r)=((1-\theta)^{-r}-1)^{-1}\binom{r+x-1}{x}\theta^x;\ x=1,2,\cdots;\quad r=1,2,\cdots;\ 0<\theta<1\,. \tag{2.1}$$

If $X_1, X_2, \cdots, X_n$ are mutually independent and identical r.v.'s with $X_i$ having p.f. $g(x_i\,;\theta,r)$, then for a given sample size $n$, the random sample total $K=\sum_{i=1}^{n} X_i$ has the following p.f. (see Cacoullos and Charalambides(1975)):

$$P[\sum_{i=1}^{n} X_i=k]=((1-\theta)^{-r}-1)^{-n}r^n\ n!\ s(k,r,n)\theta^k/k!;\quad k=n,\ n+1,\ \cdots, \tag{2.2}$$

where the s-numbers are defined for all integers $r\neq 0,\ n>0,\ k>0$ as

$$s(k,r,n)=\Big[\sum_{i=1}^{n}(-1)^{n+i}\binom{n}{i}(ri+k-1)_k\Big]/n!r^n. \tag{2.3}$$

The joint conditional distribution of $X_1, X_2, \cdots, X_n$ given $\Sigma X_i=k$, for any fixed $k$, is obtained to be

$$P[X_i=x_i\,;\ i=1,2,\ldots,n\,|\,\Sigma X_i=k] \tag{2.4}$$
$$=k!\ \prod_{i=1}^{n}\binom{r+x_i-1}{x_i}/n!\ r^n\ s(k,r,n),$$

for $1\leq x_i\leq k-n+1$ and $k\geq n$; the conditional distribution of $X_1$ given $\Sigma X_i=k$ is then

$$P[X_1=x\,|\,\Sigma X_i=k]=\binom{k}{x}\ s(x,r,1)\ s(k-x,r,n-1)/ns(k,r,n)\ , \tag{2.5}$$

where $1\leq x\leq k-n+1$ and $s(x,r,1)=\frac{1}{r}(r+x-1)_x$. Using a recurrence relation of the s-numbers given in Charalambides(1977), expression (2.5) can be proved to be a p.f. Since the conditional distribution in (2.5) does not depend upon the parameter $\theta$, the random sample total $\Sigma X_i$ is a sufficient statistic. Furthermore, expression (2.5) is the minimum variance unbiased estimate of the p.f. in (2.1).

Using expression (2.5), the negative binomial distribution truncated at zero can be characterized as follows:

**Theorem 1.**

The mutually independent positive integer valued random variables $X_1$, $X_2$, ···, $X_n$, $n\geq 2$, have the same negative binomial probability distribution truncated at zero if and only if their conditional distribution for each fixed sum $\Sigma X_i=k$ is as given in (2.4).

Using a lemma given in Shanmugam and Singh(1982) the above characterization can be easily proved and is, therefore, omitted.

A goodness-of-fit application of the above characterization is discussed in the next section.

## 3. An Application

A goodness-of-fit application of our characterization is as follows: Testing the hypothesis that $X_1$, $X_2$, ···, $X_n$ is a random sample from a negative binomial distribution truncated at zero given in (2.1) can be replaced by testing whether the joint conditional distribution of $X_1$, $X_2$, ···, $X_n$ for each fixed sum $\Sigma X_i=k$ is a probability distribution given in (2.4).

Since $\Sigma X_i=k$, we note that $nE(X_i|k)=k$ regardless of the underlying probability distribution for the random sample. However, expressions for the $\mathrm{var}(X_i|k)$ and the $\mathrm{cov}\{(X_i, X_j)|k\}$ would depend upon the underlying probability distribution. We suggest a test statistic that would take into consideration the difference between the observed and expected values when properly weighted by the variance-covariance matrix. Ther-

efore, for any fixed $k$, we propose to use $Q=(\underline{X}-\underline{\mu})'\ \Sigma^{-}(\underline{X}-\underline{\mu})$ as a test statistic, where $\underline{X}'=(X_1, X_2, \cdots, X_n)$ is the observed vector, $\underline{\mu}'=(\mu_1, \mu_2, \cdots, \mu_n)$ with $\mu_i=E(X_i|k)$ is the vector of expected values, and $\Sigma^{-}$ is a generalized inverse of the dispersion matrix $\Sigma=\{\mathrm{cov}(X_i, X_j)|k\}$ is to be the matrix of weights. The rank of $\Sigma$ is only $(n-1)$. To compute $Q$, we need to know $\mu$ and $\Sigma$ in addition to $X_i$'s. To use $Q$ as a statistic, we need its distribution. It is easy to see that $\underline{\mu}'=\left(\frac{k}{n}, \frac{k}{n}\cdots, \frac{k}{n}\right)$. It is not so easy, however, to obtain expressions for the elements of $\Sigma$. To find the elements of $\Sigma$, we proceed as follows. Note that $EX_1(X_1-1)=\underset{k}{E}E(X_1(X_1-1)|k)$; comparing the coefficients of $\theta^k$ on both sides, we obtain that

$$E(X_1(X_1-1)|k)=\frac{k!\ r(r+1)}{n!\ r^n\ s(k,r,n)}\sum_{j\geq n-1}(n-1)!\ \frac{r^{n-1}s(j,r,n-1)}{j!}\binom{r+k-j-1}{k-j-2}$$

Using the following asymptotic property of these s-numbers (see Cacoullos and Charalambides(1975)),

$$s(k,r,n)\simeq(rn+k-1)_k/n!r^n \text{ for } k\gg n, \tag{3.1}$$

and a formula in Gould(1972), we can further simplify it to

$$E(X_1(X_1-1)|k)\simeq(r+1)k(k-1)/n(rn+1).$$

Therefore,

$$\mathrm{var}(X_1|k)=E(X_1(X_1-1)|k)+E(X_1|k)-(E(X_1|k))^2\simeq\frac{(n-1)}{n^2}\left(\frac{rn+k}{rn+1}\right)k$$

and

$$\mathrm{cov}((X_1, X_2)|k)=\underset{X_1}{\mathrm{cov}}((X_1, E(X_2|X_1))|k)\simeq-\left(\frac{1}{n-1}\right)\mathrm{var}(X_1|k).$$

The structure of the dispersion matrix $\Sigma$ is of the intraclass correlation matrix type. Using a generalized inverse of $\Sigma$, for any fixed but sufficiently large $k$, we find that,

$$Q\simeq Q^*=\left(\frac{rn+1}{rn+k}\right)\left[\frac{n}{k}\Sigma X_i^2-k\right]. \tag{3.2}$$

It is interesting to note that for computing $Q^*$, we do not need the elements of $\Sigma$, but to use it we need the sampling distribution of $Q^*$. In fact, the sampling distribution of $V=\Sigma X_i^2$ for a fixed $k$ is needed, because $V$ and $Q^*$ are linearly related.

Before discussing further the asymptotic distribution of $V$, let us find its mean and variance. Repeating the same technique which we used to find expressions of $\Sigma$, we obtain

$$E(\Sigma X_i^2|k)\simeq k+\left(\frac{r+1}{rn+1}\right)(k)_2 \tag{3.3}$$

and

$$(3.4)\quad \mathrm{var}(\sum X_i^2|k)=\sum_{i=1}^{4}G(4,i)\frac{\binom{r+i-1}{r}(k)_i}{\binom{rn+i-1}{rn}}+\frac{(n-1)rk}{(r+1)}\sum_{i=1}^{3}\frac{(r+1)^i\binom{k-1}{i}\binom{2}{i-1}}{\binom{rn+i}{i}}$$
$$-[E(\sum X_i^2|k)]^2,$$

where $G(4,i)$ are the Stirling numbers of the second kind. We now return to the distribution of $V=\sum X_i^2$. For fixed $\sum X_i=k$, the r.v.'s $X_1$, $X_2$, $\cdots$, $X_n$ are mutually dependent. However, they are asymptotically mutually independent due to the following reason: For any $i\neq j$, $i,j=1,2,\cdots,n$, note that

$$P[X_i=a,\ X_j=b|\sum X_i=k]=\phi\ P[X_i=a|k]P[X_j=b|k],$$

where

$$\phi=\binom{n}{n-1}\ \frac{(k-a)!}{s(k-a,r,n-1)}\ \frac{(k-b)!}{s(k-b,r,n-1)}\ \frac{s(k-a-b,r,n-2)}{(k-a-b)!}\ \frac{s(k,r,n)}{k!}.$$

As $k\to\infty$, $\frac{n}{\ln k}\to 0$, we note that, due to expression (3.1), $\phi$ approaches one. Because $V$ is the sum of the "asymptotically" independent r.v.'s for given $k$, it would asymptotically follow a normal distribution with mean and variance given in expressions (3.3) and (3.4), respectively.

For a given significance level $\alpha$, we could reject the null hypothesis that a random sample $X_1$, $X_2$, $\cdots$, $X_n$ is from a negative binomial distribution truncated at zero if $|\{\sum X_i^2-E(\sum X_i^2|k)\}/\sqrt{\mathrm{var}(\sum X_i^2|k)}|$ exceeds the normal ordinate $Z_{\alpha/2}$, where $E(\sum X_i^2|k)$ and $\mathrm{var}(\sum X_i^2|k)$ are as expressed in (3.3) and (3.4), respectively. The advantages of this approach over the usual chi-square goodness-of-fit approach are the following: First, computation is simpler in this method, whereas in the chi-square approach one has to estimate the parameter $\theta$, the expected frequencies and the goodness-of-fit measure. Secondly, unlike in the chi-square method, no grouping of cells having small counts is required in using our test statistic $V$.

To illustrate our results, we consider the biological data given in Sampford(1955). In an investigation of chromosome breakage, the following sample distribution of breaks per cell was obtained.

| | $x$: No. of Breaks per Cell | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 6 | 8 | 9 | 11 | 13 |
| $n_x$: No. of Cells | 11 | 6 | 4 | 5 | 1 | 2 | 1 | 1 | 1 |

Note that $n=32$, $k=\sum X_i=110$, $\sum X_i^2=686$, and an estimate (see Sampford(1955)) of $r$

is 0.633. Using expressions (3.3) and (3.4), the mean and variance of $\sum X_i^2$ are computed; they are 1031.13 and 56191.48, respectively. The test score $Z=1.45$ confirms Sampford's conclusion that the negative binomial distribution truncated at zero fits the data well.

While it is known that under certain conditions the negative binomial distribution truncated at zero approaches the positive-Poisson and logarithmic series distributions, we show in the next section that our results approach in limit the results of Singh (1978) and Shanmugam and Singh(1982), respectively.

## 4. Limiting Distributions

As $r\to\infty$ and $\theta\to 1$ such that $r(1-\theta)=\lambda$, $0<\lambda<\theta$, the p.f. given in (2.1) approaches the positive-Poisson p.f.

$$p(x\,;\,\theta)=(e^{\lambda}-1)^{-1}\lambda^x/x!,\quad x=1,\ 2,\ \cdots. \tag{4.1}$$

A characterization of the positive-Poisson distribution and its goodness-of-fit application have been given by Singh(1978). Using his characterization, Singh showed that the hypothesis of $X_1, X_2, \cdots, X_n$ being a random sample from a positive-Poisson distribution can be rejected at a significance level $\alpha$ if a test statistic $|\{\sum X_i^2-E(\sum X_i^2|k)\}/\sqrt{\mathrm{var}(\sum X_i^2|k)}|$ exceeds the normal ordinate $Z_{\alpha/2}$ where, for large $k$,

$$E(\sum X_i^2|k)\simeq k+\frac{(k)_2}{n} \tag{4.2}$$

and

$$\mathrm{var}(\sum X_i^2|k)\simeq 2(n-1)(k)_2/n^2. \tag{4.3}$$

Using the fact $s(k,r,n)\simeq r^{k-n}G(k,n)$ as $r\to\infty$ (see Charalambides(1977)) where $G(k,n)$ is the Stirling number of the second, we obtain the expression (2.3) in Singh of the positive-Poisson distribution as a limiting case of our joint conditional distribution given in (2.4). Also our expressions (3.3) and (3.4) for the conditional mean and variance of $\sum X_i^2$ converge in limit to the corresponding expressions (4.2) and (4.3) of the positive-Poisson distribution.

Let us now consider the limit of the p.f. given in (2.1) when $r\to 0$. We find that

$$\lim_{r\to 0} g(x\,;\,\theta,r)=(-\ln(1-\theta))^{-1}\theta^x/x\,;\ x=1,\ 2,\ \cdots,$$

which, in fact, is the logarithmic series distribution discussed in Shanmugam and Singh(1982). Based on their characterization of the logarithmic series distribution, Shan-

mugam and Singh showed that the hypothesis of $X_1, X_2, \cdots, X_n$ having a logarithmic series distribution can be rejected at a significance level $\alpha$ if a test score $|\{\sum X_i^2 - E(\sum X_i^2|k)\}/\sqrt{\text{var}(\sum X_i^2|k)}|$ exceeds the normal ordinate $Z_{\alpha/2}$ where, for large $k$,

$$E(\sum X_i^2|k) \simeq k(k-n+1) \simeq k^2 \tag{4.4}$$

$$\text{var}(\sum X_i^2|k) \simeq (n-1)k\binom{k-n+1}{3}. \tag{4.5}$$

Note that as $r \to 0$, the s-numbers become (see Charalambides(1977)) the signless Stirling numbers of the first kind, $F(k, n)$. Using this fact, we obtain not only the joint conditional distribution of $X_1, X_2, \cdots, X_n$ for a fixed $\sum X_i = k$ given in Shanmugam and Singh(1982) for the logarithmic series distribution as a limiting case of our expression (2.4), but also the conditional mean and variance of $\sum X_i^2$ given in (4.4) and (4.5) of the logarithmic series distribution as limits of our expressions (3.3) and (3.4), respectively.

These then conclude that the goodness-of-fit application of the characterization of positive-Poisson and logarithmic series distributions given in Singh(1978) and Shanmugam and Singh(1982), respectively, are the limiting cases of our results in this paper.

In the next section, we consider a characterization of the geometric distribution and its statistical application.

## 5. A Characterization of the Geometric Distribution: A Particular Case

When $r=1$, the p.f. given in (2.1) becomes

$$g(x\,;\,\theta) = (1-\theta)\theta^{x-1}\,;\; x=1,\ 2,\ \cdots;\ 0<\theta<1\ , \tag{5.1}$$

which is the geometric distribution. Note that $s(k, 1, n) = L(k, n)$ (see Charalambides (1977)) where $L(k, n)$ is the signless Lah numbers. While $X_1, X_2, \cdots, X_n$ are mutually independent r.v.'s having the p.f. given in (5.1), the probability distribution of its total can be shown by substituting $r=1$ in our expression (2.2). That is,

$$P[\sum_{i=1}^{n} X_i = k] = ((1-\theta)^{-1}-1)^{-n}\ n!\ L(k, n)\theta^k/k!;\ k=n,\ n+1,\ \cdots. \tag{5.2}$$

Ahuja(1974) derived the expression (5.2) in a general set up using the generalized power series principle.

Repeating the arguments given in Section 2 and our expressions (5.1) and (5.2), a characterization of the geometric distribution can be stated as a corollary to Theorem 1 as follows:

**Corollary 1.** The mutually independent positive integer valued random variables $X_1$, $X_2$, $\cdots$, $X_n$; $n\geq 2$ have the same geometric probability distribution if and only if their conditional distribution, for each fixed sum $\sum X_i=k$ is

$P[X_i=x_i,\ i=1,\ 2,\ \cdots,\ n|\sum X_i=k]=k!/n!\ L(k,n)$ for $1\leq x_i\leq k-n+1$ and $k\geq n$.

Proceeding in the same fashion as we did in Section 3, we obtain the following goodness-of-fit application of our characterization. That is: the hypothesis of $X_1, X_2, \cdots X_n$ being a random sample from a geometric distribution can be rejected if a test score $|\{\sum X_i^2-E(\sum X_i^2|k)\}/\sqrt{\mathrm{var}(\sum X_i^2|k)}|$ exceeds the critical normal ordinate $Z_{\alpha/2}$ with a significance level $\alpha$ where $E(\sum X_i^2|k)\simeq k+\frac{2(k)_2}{n+1}$ and

$$\mathrm{var}(\sum X_i^2|k)\simeq\sum_{i=1}^{4}G(4,i)\ \frac{i(k)_i}{\binom{n+i-1}{n}}+\frac{(n-1)k}{2}\sum_{i=1}^{3}2^i\frac{\binom{k-1}{i}\binom{2}{i-1}}{\binom{n+i}{i}}-[E(\sum X_i^2|k)]^2$$

for a large $k$.

## REFERENCES


(1) J. C. Ahuja and E. A. Enneking (1974) Convolution of Independent Lefttruncated Negative Binomial Variables and Limiting Distributions, *Ann. Inst. Stat. Math.*, 26, p. 265.

(2) T. Cacoullos and Ch. Charalambides (1975) On Minimum Variance Unbiased Estimation for Truncated Binomial and Negative Binomial Distributions, *Ann. Inst. Stat. Math.*, 27, p. 235.

(3) Ch. A. Charalambides (1977) A New Kind of Numbers Appearing in the n-fold Convolution of Truncated Binomial and Negative Binomial Distributions, *SIAM J.Appl. Math.*, 33, no. 2, p. 279.

(4) H. W. Gould (1972) *Combinatorial Identities*, West Virginia University Press, Morgantown.

(5) N.L. Johnson and S. Kotz (1969). *Discrete Distributions*, Houghton Mifflin, Boston.

(6) M. R. Sampford (1955) The Truncated Negative Binomial Dirtribution, *Biometrika*, 42, p. 58.

(7) R. Shanmugam and J. Singh (1982) A Characterization and Its Statistical Applications of Logarithmic Series Distribution, To appear in *Communications in Statistics.*

(8) J. Singh (1978) A Characterization of Positive Poisson Distribution and Its Statistical Application, *SIAM J. Appl. Math.*, 34, p. 545.