COMMUN. STATIST.-THEORY METH. 18(1), 355-366 (1989)

ON SOME DISTRIBUTIONS ARISING IN INVERSE CLUSTER SAMPLING

Evdokia Xekalaki

Athens School of Economics and Business Studies, Greece

J. Panaretos

University of Patras, Greece

Keywords and Phrases: Cluster negative binomial distribution; generalized Poisson distribution; stuttering generalized Waring distribution; cluster logarithmic series distribution.

ABSTRACT

In this paper, distributions of items sampled inversely in clusters are derived. In particular, negative binomial type of distributions are obtained and their properties are studied. A logarithmic series type of distribution is also defined as a limiting form of the obtained generalized negative binomial distribution.

1. INTRODUCTION

In a recent paper, Panaretos and Xekalaki (1986a) introduced some generalizations of existing sampling schemes which lead to extended forms of several well known probability distributions.

One of these sampling schemes leads to an interesting negative binomial type of distribution which not only contains the ordinary negative binomial as a special case but also it pocesses properties analogous to those of the latter. It was shown there that appropriate limiting operations on its zero-truncated form lead to a generalized log series distribution through which most aspects of the existing interrelationships between the ordinary Poisson, logarithmic and negative binomial distributions are preserved under the transition to the triplet (generalized Poisson distribution, generalized logarithmic distribution, generalized negative binomial distribution).

The generalized negative binomial distribution in question is defined by the probability function given by

$$P(X=x) = \sum_{\Sigma i x_i = x} \binom{\Sigma x_i + m - 1}{x_1, \ldots, x_k, m-1} \left(\frac{\alpha}{k\beta+\alpha}\right)^m \left(\frac{\beta}{k\beta+\alpha}\right)^{\Sigma x_i} \qquad (1.1)$$

$$x=0,1,2,\ldots$$

and arises in the context of cluster sampling . In particular, from an urn containing numbered balls, α marked 0 and β marked i, $i=1,2,\ldots,k$ $(k \in I^+)$ draws are made at random , one at a time and with replacement till m balls marked 0 are observed. Then (1.1) represents the probability distribution of X, the total sum of the sampled numbers. This urn scheme could possibly be used to model biological situations where the frequency distribution of cells forming clusters is of interest.

Restricting X to be non-negative and allowing m to become very small ($m \to 0$) Panaretos and Xekalaki (1986a) obtained the generalized logarithmic distribution mentioned before defined by the probability function

$$P(X=x) = \sum_{\Sigma i x_i = x} \frac{(\Sigma x_i - 1)!}{-\ln(\alpha/(\beta k+\alpha))} \frac{(\beta/(\beta k+\alpha))^{\Sigma x_i}}{x_1! \ldots x_k!} \qquad (1.2)$$

$$x=1,2\ldots$$

In this paper a sampling scheme that gives rise to a more general form of (1.1) and subsequently of (1.2) is introduced.

The interrelationship of the thus arising distributions as well as their relationship to other existing distributions is examined. Further, a recurrence relationship is derived for the probabilities of the extended form of (1.1) which can be useful in applying the distributions to actual data.

2. THE CLUSTER NEGATIVE BINOMIAL DISTRIBUTION

Consider an urn containing balls marked with integer numbers in the range $(0,1,2,\ldots,k)$, $k>0$. Suppose that α balls are marked 0 and β_i balls are marked i, $i=1,2,\ldots,k$. A ball is drawn at random, its number is recorded and the ball is returned to the urn before the next ball is drawn. Let X be the sum of numbers sampled before the m-th zero. Then the following theorem can be shown to hold.

Theorem 2.1: The probability of the event $(X=x)$ is given by

$$P(X=x)=\sum_{\Sigma i x_i=x}\binom{\Sigma x_i+m-1}{x_1,x_2,\ldots,x_k,m-1}\left(\frac{\alpha}{\Sigma\beta_i+\alpha}\right)^m\left(\frac{\beta_1}{\Sigma\beta_i+\alpha}\right)^{x_1}\cdots\left(\frac{\beta_k}{\Sigma\beta_i+\alpha}\right)^{x_k}$$

$$x=0,1,2,\ldots \qquad (2.1)$$

where $\sum$ stands for summation over all values $1,2,\ldots,k$ of the index of the summand.

It can be shown that $\sum_{x=0}^{\infty} P(X=x)=1$. Hence the probability of the event $(X=x)$ as given by (2.1) defines a proper probability distribution. This suggests that a new probability distribution can be defined through (2.1). Though here m is an integer, the argument still holds true for any positive real number m. The form of the probability function suggests the use of this distribution for the description of the distribution of the total number of items whenever these occur in clusters jointly distributed in the negative multinomial form. Moreover, when $k=1$ (2.1) reduces to the ordinary negative binomial distribution. Hence the following definition is in order.

Definition 2.1: A non-negative, integer valued random variable X is said to have the cluster negative binomial distribution with parameters k, m, $q_1, q_2, \ldots, q_k$ denoted by $NB_k(m, q_1, \ldots, q_k)$ if and only if

$$P(X=x) = \sum_{\Sigma i x_i = x} \binom{\Sigma x_i + m - 1}{x_1, \ldots, x_k, m-1} p^m q_1^{x_1} q_2^{x_2} \ldots q_k^{x_k} \qquad (2.2)$$

$$x=0,1,2,\ldots;\ p, q_i>0,\ i=1,2,\ldots,k;\ \textstyle\sum q_i = 1-p,\ m>0.$$

Obviously when $q_i=q$, $i=1,2,\ldots,k$ (2.2) reduces to Panaretos and Xekalaki's (1986a) generalized negative binomial distribution while under the transformation $x \to x+|km|$ relationship (2.2) gives rise to a special case of Aki's (1985) extended negative binomial distribution for $p = \prod_{i=1}^{k} p_i$ and $q_i = (1-p_i)\prod_{j=1}^{i-1} p_j$, $0<p_i<1$, $i=1,\ldots,k$. It should be noted that (2.2) can also arise as a limiting form of a more general distribution introduced by Steyn (1956). Also when using straight sampling the urn scheme considered in this section gives rise to the cluster binomial distribution studied by Panaretos and Xekalaki (1986b).

It is interesting to note at this point that if m=1 i.e. if sampling is stopped once the first 0 ball is observed then (2.2) reduces to a distribution which contains the ordinary geometric distribution as a special case. In the sequel, this distribution will be referred to as the cluster geometric distribution.

Definition 2.2: A non-negative, integer-valued random variable X will be said to have the cluster geometric distribution with parameters k, $q_1, \ldots, q_k$ if and only if its probability function is given by

$$P(X=x) = p \sum_{\Sigma i x_i = x} \binom{\sum_{i=1}^{k} x_i}{x_1, \ldots, x_k} q_1^{x_1} \ldots q_k^{x_k} \qquad (2.3)$$

Theorem 2.3: Let X be a random variable having the cluster negative binomial distribution defined by (2.2). Then its probability generating function is given by

$$G_X(s) = p^m (1-\textstyle\sum q_i s^i)^{-m}$$

or equivalently by

$$G_X(s) = \left[1 + \frac{1}{p} \sum_{i=1}^{k} q_i (1-s^i) \right]^{-m} \tag{2.4}$$

Proof:

$$G_X(s) = \sum_{x=0}^{\infty} s^x \sum_{\Sigma i x_i = x} \binom{\Sigma x_i + m - 1}{x_1, \ldots, x_k, m-1} p^m q_1^{x_1} \ldots q_k^{x_k}$$

$$= \sum_{x=0}^{\infty} \sum_{\Sigma i x_i = x} \binom{\Sigma x_i + m - 1}{x_1, \ldots, x_k, m-1} p^m (q_1 s)^{x_1} \ldots (q_k s^k)^{x_k}.$$

Then letting $x_i \to x_i$ and $x \to x + \sum (i-1)x_i$ we obtain

$$G_X(s) = p^m \sum_{x=0}^{\infty} \frac{m_{(x)}}{x!} \sum_{\Sigma i x_i = x} \binom{x}{x_1, \ldots, x_k} (q_1 s)^{x_1} \ldots (q_k s^k)^{x_k}$$

$$= p^m \sum_{x=0}^{\infty} \frac{m_{(x)}}{x!} \left(\sum_{i=1}^{k} q_i s^i \right)^x .$$

Hence the result.

Corollary 2.1: Let X be defined as in theorem 2.3. Then

$$E(X) = \frac{m}{p} \sum_{i=1}^{k} i q_i \tag{2.5}$$

$$V(X) = \frac{m}{p} \left\{ \sum i^2 q_i + \frac{1}{p} \left(\sum i q_i \right)^2 \right\}$$

Theorem 2.4: Let $X_1, X_2, \ldots, X_m$ be non-negative integer valued random variables that are identically and independently distributed according to a cluster geometric distribution with probability function as given by (2.3). Then the random variable $X = X_1 + X_2 + \ldots + X_m$ has a cluster negative binomial distribution with parameters $k, m, p_1, p_2, \ldots, p_k$.

3. THE CLUSTER LOGARITHMIC SERIES DISTRIBUTION

Let X be a random variable defined as in the previous section and let $P_m(X=x)$ denote its probability function as given by (2.2). Then the limiting distribution of X conditional on the event $\{X>0\}$ as m tends to 0 leads to an extended version of the logarithmic series distribution i.e.,

$$\lim_{m\to 0} P_m(X=x|X>0) = \sum_{\Sigma i x_i = x} \frac{(\sum x_i - 1)!}{-\ln p} \prod_{i=1}^{k} \frac{q_i^{x_i}}{x_i!} \qquad (3.1)$$

$$x=1,2,\ldots$$

When $q_i=q$, $i=1,2,\ldots,k$ (3.1) reduces to Panaretos and Xekalaki's (1986a) logarithmic series distribution of order k while for k=1 it reduces to the ordinary logarithmic series distribution. Since this distribution originates from the cluster negative binomial distribution, the following definition may be given.

Definition 3.1: A positive, integer-valued random variable X will be said to have the cluster logarithmic series distribution with parameters k, $q_1,\ldots,q_k$ if and only if its probability function is given by

$$P(X=x) = \sum_{\Sigma i x_i = x} \frac{(\Sigma x_i - 1)!}{-\ln p} \frac{q_1^{x_1}}{x_1!} \cdots \frac{q_k^{x_k}}{x_k!} \qquad (3.2)$$

$$x=1,2,\ldots \quad q_i>0,\ i=1,2,\ldots,k\ ;\ p=1-\Sigma q_i>0.$$

It should be noted at this point that (3.2) can be considered as a reparameterized version of Aki's (1985) extended logratithmic series distribution for $p = \prod_{i=1}^{k} p_i$ and $q_i = (1-p_i)\prod_{j=1}^{i-1} p_j$, $0<p_i<1$, $i=1,2,\ldots,k$.

Theorem 3.2: Let $G_X(s)$ denote the probability generating function of a random variable X having the cluster logarithmic distribution with probability function given by 3.2 . Then

$$G_X(s) = \frac{\ln(1-\sum q_i s^i)}{\ln(1-\sum q_i)} \qquad (3.3)$$

Proof:

$$G_X(s) = \sum_{x=r}^{\infty} s^x \sum_{\Sigma I x_i = x} \frac{(\sum x_i - 1)!}{-\ln p} \prod_{i=1}^{k} q_i^{x_i} / x_i!$$

$$= \frac{1}{-\ln p} \sum_{x=1}^{\infty} \sum_{\Sigma I x_i = x} \frac{1}{\Sigma x_i} \binom{\Sigma x_i}{x_1, \ldots, x_k} \prod_{i=1}^{k} q_i^{x_i} .$$

Letting $x_i \to x_j$ and $x \to x + \sum(i-1)x_i$ we obtain from the last equation

$$G_X(s) = -\frac{1}{\ln p} \sum_{x=1}^{\infty} \frac{1}{x} \sum_{\Sigma x_i = x} \binom{x}{x_1, \ldots, x_k} \prod_{i=1}^{k} (q_i s^i)^{x_i}$$

$$= -\frac{1}{\ln p} \sum_{x=1}^{\infty} \frac{1}{x} \left(\sum q_i s^i \right)^x .$$

which leads to (3.3) and establishes the theorem .

It is obvious from the form of (3.3) that the cluster logarithmic series distribution admits a random sum representation since X can be represented by the sum $Y_1 + Y_2 + \ldots + Y_Z$ where Z, $Y_1, Y_2, \ldots$ are independendent random variables such that Z has the ordinary logarithmic distribution with parameter 1-p while $Y_1, Y_2, \ldots$ are identically distributed on $\{1,2,\ldots,k\}$ according to Hirano's (1986) k-point distribution with probability function $P(Y_i = r) = q_r/(1-p)$, $r = 1,2,\ldots,k$, $i = 1,2,\ldots$.

4. THE RELATIONSHIP OF THE CLUSTER NEGATIVE BINOMIAL DISTRIBUTION TO GENERALIZED POISSON OR MIXTURES OF GENERALIZED POISSON DISTRIBUTIONS.

In this section, the cluster negative binomial distribution is shown to arise as a mixture of the generalized Poisson distribution or as a Poisson sum of cluster logarithmic random variables. Furthermore, the generalized Poisson distribution is obtained as a limiting form of the cluster negative binomial

distribution. It is also shown that mixing latter distribution with respect to the parameters q_i $i=1,2,\ldots,k$ leads to the stuttering generalized Waring distribution introduced by Panaretos and Xekalaki (1986c) and studied by Panaretos (1987a,b).

Theorem 4.1: Let X be a non-negative, integer-valued random variable Assume that conditional on a positive random variable λ X follows a generalized Poisson distribution with probability generating function

$$G_{X|\lambda}(s) = \exp \lambda\left\{\sum_{i=1}^{k} \alpha_i(s^i-1)\right\}$$

$\lambda>0,\ a_i>0;\ i=1,2,\ldots,k$

Assume further that λ follows a gamma distribution with parameters m and ν and probability density function

$$f_\lambda(u) = \frac{m^\nu}{\Gamma(\nu)}\ u^{\nu-1}\ e^{-mu}\ ,m>0,\ \nu>0,\ n>0 \qquad (4.1)$$

Then the unconditional distribution of X is the cluster negative binomial distribution with parameters k, ν, $\frac{\alpha_1}{m+\Sigma\alpha_i},\ldots,\frac{\alpha_k}{m+\Sigma\alpha_i}$.

Theorem 4.2: Let $X_1,X_2,\ldots$ be a sequence of independent random variables distributed identically according to a cluster logarithmic series distribution with parameters k, $q_i>0$, $i=1,2,\ldots,k$; $1-\sum q_i=p$, $0<p<1$ and probability function given by (3.2). Let N be a non-negative random variable distributed independently of $X_1,X_2,\ldots$ according to the Poisson distribution with parameter $-\lambda\ln p$, $\lambda>0$. Then the distribution of the random variable $X=X_1+X_2+\ldots+X_N$ is the cluster negative binomial with parameters $k,\lambda,\ q_1,q_2,\ldots,q_k$.

The above two theorems demonstrate the fact that the mixing and random sum representations of the ordinary negative binomial distribution are preserved under the transition to its generalization as given by (2.2) or (2.3). For $q_i=q$, $i=1,2,\ldots,k$ theorem 4.2 leads to Panaretos and Xekalaki's (1986a) results highlighting the association of (1.1) and (1.2). It should

perhaps be noted here that in the statement of the relevant result (Panaretos and Xekalaki, (1986), theorem 4.2)' the expression for the probability function of each of the summands in $Y_1+\ldots+Y_N$ should read

$$\sum_{\Sigma x_i=y} \frac{(\Sigma x_i-1)!}{\ln(\vartheta k+1)} \frac{(\vartheta/(\vartheta k+1))^{\Sigma x_i}}{x_1!+\ldots+x_k!}$$

instead of the incorrectly printed expression.

Theorem 4.3: Let X be a non-negative, integer valued random variable having the cluster negative binomial distribution with parameters k, m, $q_1, q_2, \ldots, q_k$. Then if $m\to\infty$ and $q_i\to 0$ so that $mq_i\to\lambda_i$ for some fixed value λ_i in $(0,+\infty)$, $i=1,2,\ldots,k$, the distribution of X tends to the generalized Poisson distribution with parameters λ_i, $i=1,2,\ldots,k$.

Proof: Let $\lim_H$ stand for limit as $m\to\infty$, $q_i\to 0$ so that $mq_i\to\lambda_i$, $i=1,2,\ldots,k$ and observe that using (2.3) the probability generating function of X can be written as

$$G_X(s) = \exp\left\{ -m\ln\left[1+\frac{1}{p}\sum_{i=1}^{k} q_i(1-s^i)\right]\right\}.$$

Since

$$-m\frac{\sum_{i=1}^{k}\frac{q_i}{p}(1-s^i)}{1+\sum_{i=1}^{k}\frac{q_i}{p}(1-s^i)} \geq -m\ln\left\{1+\frac{1}{p}\sum_{i=1}^{k} q_i(1-s^i)\right\} \geq -m\sum_{i=1}^{k}\frac{q_i}{p}(1-s^i),$$

we have upon taking the limit under H that

$$\lim_H G_X(s) = \exp\left\{\sum_{i=1}^{k}\lambda_i(s^i-1)\right\}.$$

This establishes the result.

It would now be interesting to examine whether appropriate mixtures of the generalized form lead to distributions which are generalizations of the corresponding mixtures of the ordinary negative binomial distribution. The theorem that follows shows that this is the case. In particular, it is shown that if $(q_1, q_2, \ldots, q_k)$ is a Dirichlet random vector the resulting mixed

distribution is the stuttering generalized Waring distribution with probability function given by

$$G_X(s) = \frac{c_{(\Sigma b_i)}}{(m+c)_{(\Sigma b_i)}} F_D(m; b_1, b_2, \ldots, b_k; m+\Sigma b_i+c; s, s^2, \ldots, s^k) \tag{4.2}$$

where F_D represents Lauricella's hypergeometric series of type D defined by

$$F_D(\alpha; \beta_1, \ldots, \beta_k; \gamma; z_1, \ldots, z_k) =$$

$$\sum_{r_1=0}^{\infty} \cdots \sum_{r_k=0}^{\infty} \frac{\alpha_{(\Sigma r_i)} (\beta_1)_{(r_1)} \cdots (\beta_k)_{(r_k)}}{\gamma_{(\Sigma r_i)}} \frac{z_1^{r_1}}{r_1!} \cdots \frac{z_k^{r_k}}{r_k!} \tag{4.3}$$

$$\gamma-\alpha-\sum_{i=1}^{k} \beta_i>0, \quad |z_i|\leq 1, \quad i=1,2,\ldots,k$$

Theorem 4.4: Let X be a non-negative integer valued random variable having the cluster negative binomial distribution with parameters $k, m, q_1, q_2, \ldots, q_k$ and probability generating function given by (2.3). Let the parameters $q_1, q_2, \ldots, q_k$ be themselves random variables jointly distributed according to the Dirichlet (multivariate beta) distribution of the first kind with probability density function given by

$$f(q_1, \ldots, q_k) = \frac{\Gamma(c+\sum b_i)}{\Gamma(c) \prod_{i=1}^{k} \Gamma(b_i)} \prod_{i=1}^{k} q_i^{b_i-1} \left[1 - \sum_{i=1}^{k} q_i\right]^{c-1} \tag{4.4}$$

$$c, \; q_i, b_i>0, \; i=1,2,\ldots,k; \; \sum_{i=1}^{k} q_i<1$$

Then the resulting distribution of X is the stuttering generalized Waring distribution with probability generating function given by (4.2).

Proof: Obviously the cluster negative binomial distribution represents the conditional distribution of X on the random vector

$q_1, q_2, \ldots, q_k$. Therefore

$$G_X(s)=\frac{\Gamma\left(c+\sum_{i=1}^{k} b_i\right)}{\Gamma(c)\prod_{i=1}^{k}\Gamma(b_i)}\int\cdots\int_{\substack{\sum q_i<1\\ q_i>0,\\ i=1,\ldots,k}}\left[\prod_{i=1}^{k} q_i^{b_i-1}\right](1-\sum q_i)^{m+c-1}\times (1-\sum q_i s^i)^{-m}\,dq_1\ldots dq_k$$

$$=\frac{C_{(\Sigma b_i)}}{(m+c)_{(\Sigma b_i)}}\sum_{x_1=0}^{\infty}\cdots\sum_{x_k=0}^{\infty}\frac{m_{(\Sigma x_i)}\,(b_1)_{(x_1)}\cdots(b_k)_{(x_k)}}{\left(m+\sum_{i=1}^{k} b_i+c\right)_{(\Sigma x_i)}}\times\frac{s^{x_1}}{x_1!}\cdots\frac{s^{kx_k}}{x_k!}\,.$$

Using (4.3) the above relation leads to (4.2) and hence the theorem has been established.

For k=1 the result of the treorem reduces to the derivation of the generalized Waring distribution as a beta mixture of the ordinary negative binomial distribution.

BIBLIOGRAPHY

Aki, S. (1985). Discrete distribution of order k on a binary sequence. Ann. Inst. Statist. Math., A, 37, 205-224.

Hirano, K. (1986). Some properties of the distribution of order k. Fibonacci Numbers and their Applications. Philippou, A.N., Bergum, G.E. and Horadam, A.F. (eds), D. Reidel, 1986, pp. 43-53.

Panaretos, J. (1987a). On the Relationship of the Stuttering Generalized Waring Distribution to the Generalized Poisson Distribution. Proceedings of the 47th Session of the International Statistical Institute, Tokyo, Japan, 341-342.

Panaretos, J. (1987b). Some Properties of the Stuttering Generalized Waring Distibution. (Submitted).

Panaretos, J. and Xekalaki, E. (1986a). On Some Distributions Arising from Certain Generalized Sampling Schemes. Commun. Statist. Theor. Meth. 15, 873-891.

Panaretos, J. and Xekalaki, E. (1986b). On Generalized Binomial and Multinomial Distributions and Their Relation to Generalized Poisson Distributions. Ann. Inst. Statist. Math. A, 38, 223-231.

Panaretos, J. and Xekalaki, E. (1986c). The Stuttering Generalized Waring Distribution. Statist. and Probab. Letters, 4(6), 313-318.

Steyn, H. S. (1956). On the Univariable Series $F(t)=F(a;b_1,b_2,\ldots,b_k;\ c;\ t,t^2,\ldots,t^k)$ and its Application in Probability Theory. Proc. Kon. Ned. Akad. V. Wetensch., Ser. A, 59, 190-197.

Steyn, H. S. (1963). On Approximations for the Discrete Distributions obtained from Multiple Events. Proc. Kon. Ned. Akad. V. Wetensch., Ser. A, 66, 85-96.

Received by Editorial Board member January 1988; Revised August 1988.

Recommended by Kathleen Kochenlakota, University of Manitoba, CANADA.

Refereed by Ramesh C. Gupta, University of Maine, Orono, MA.