Journal of Applied Sciences 4 (4): 554-560, 2004
ISSN 1607-8926
© 2004 Asian Network for Scientific Information

A Note on Mixed Distributions Involving Two Related Negative Hypergeometric Distributions

Titi Obilade
Department of Mathematics, Obafemi Awolowo University, Ile-Ife, Nigeria

Abstract: The study considered the mixed version of distributions of likelihood functions of two related hypergeometric distributions. This arises from a consideration of sampling without replacement from a finite population with balls of different colours and in different proportions but stopping only after some sufficient specific (possibly equal) number of balls of different colours might have been obtained. The resulting sample may be large or small relative to the specific stopping value depending mainly on the actual proportions of the different balls. Our interest is in the distribution of two likelihood functions, being normalised quotients of maximal or the minimal distribution with the distribution for draws for the rarest of the colours. With the aid of some simple recurrence relations and identities we obtain, in the case of two colours with the possibility of some inflation by some non-specific extraneous factors, the moments of the resulting distributions for the necessary number of draws. We also drive necessary equations for the estimation of the relevant parameters.

Key words: Mixed distribution, negative hypergeometric distributions, maximal negative hypergeometric, minimal negative hypergeometric

INTRODUCTION

Consider an urn that contains N + 2c balls of which the majority m+ c(2m > N) are of one colour (white) and the minority N–m+c are of another colour (black) such that $1 \le c \le m < N < 2m$. The distribution $h_{max}(y)$ for the number Y_{max} of draws beyond 2c until at least c balls of colour white and c balls of colour black are selected without replacement is the maximal negative hypergeometric defined by:

$$h_{max(y)} = \Pr\{y_{max} = y/c, m, N\} = \begin{cases} h_1(y)+h_2(y), & y=0,1\ldots,N-m \\ h_2(y), & y=N-m+1,\ldots,m \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where, $h_1(y)$ ($h_2(y)$) are, respectively the negative hypergeometric distributions corresponding to the number of draws $Y_1(Y_2)$ before obtaining c of colour white (black). In particular $h_1(y)$ is given as[1]:

$$h_1(y) = \Pr\{Y_1 = y\} = \begin{cases} \binom{y+c-1}{c-1}\binom{N+c-y}{m} \Big/ \binom{N+2c}{m+c} \\ y=0,1,\ldots,N+2c-m \\ 0, \quad \text{otherwise} \end{cases} \quad (2)$$

We mention in passing that the underlying distribution for Y_1 is essentially given by the transition probabilities

$$p(j,m+c-i,N-m+c-j) = p(i-1,j,m+c-i+1,N-m+c-j)\frac{(m+c-i+1)}{N+2c-i-j+1} + p(i,j-1,m+c-i,N-m+c-j+1)\frac{(N-m+c-j+1)}{N+2c-i-j+1}$$

where, p (i, j, k, l) is the probability of i colour white balls and j colour black balls leaving k white balls and l black balls in the urn at a point after i + j draws have been made. Of course this will be subject to some boundary conditions.

The corresponding distribution $h_{min}(y)$ for the number Y_{min} of draws beyond c until at least c of either one of the colours are selected without replacement is the minimal negative hypergeometric distribution defined by:

$h_{min}(y) = \Pr\{Y_{min} = y/c, m, N\}$:

$$h_{min}(y) = \begin{cases} h_1(y)+h_2(y), & y=0,1,2,\ldots,c-1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Given a situation where some primary screening exercise has produced a number N + 2c (say) of patients for a medical condition one may be interested in a

Corresponding Author: Dr. Titi Obilade, Department of Mathematics, Obafemi Awolowo University, Ile-Ife, Nigeria
E-mail: tobilade@yahoo.com

sub-sample for further intensive clerking and examination. It is natural to aim that the resulting sub-sample contains a reasonable number from the different strata. We note that in this instance the balls in the urn model stand for the patients while the colours depict the different strata. Two possible pairs of alternative policies that may be compared for a benefit analysis are:

(i) Sample until one gets at least a number c (say) of both strata (Y_{max} case) versus
Sample until at least the number 2c of the minority stratum (Y_2 case) are included in the selection and
(ii) Sample until one gets a number c (say) of either one of the strata (Y_{min} case) versus
Sample until at least the number c of the minority stratum (Y_2 case) are included in the selection.

We thus define the following likelihood functions for consideration:

R = Normalized relative likelihood ratio of Y_{max} in relation to the minority option Y_2.
P = Normalized relative likelihood ration of Y_{min} in relation to the minority option Y_2.

We give the properties and derive the relevant recurrence relations for an iterative determination of the moments. We also consider the problem of estimation of the relevant parameters.

Distribution of likelihood ratio R: The distribution of R is given by:

$$h_R^*(y) = \Pr\{R=y\} = \begin{cases} \dfrac{h_1(y)+h_2(y)}{\varphi h_2(y)}, & y=0,1,\ldots,N-m \\ \dfrac{1}{\varphi}, & y=N-m+1,\ldots,m \end{cases} \quad (4)$$

where, $\phi = \phi(m, N) = (m+1)(2m-N+2)/(2m-N+1)$. Equation (4) is equal to

$$h_R^*(y/m,N) = \frac{1+F(y)}{\phi}, \quad y=0,1,\ldots,N-m,\ldots,m \quad (5)$$

where, $F(y) = \binom{m-y}{2m-N} \Big/ \binom{m}{2m-N}$ and F (y) actually denotes the hypergeometric function F(y,N-2m, N-m+1, 1).

In fitting $h_R^*(y)$ to the incidence of a disease within units in a population it is plausible that only some proportion w of the population are susceptible to the disease. A mixed version $h_R(y)$ of $h_R^*(y)$ which invariably implies an inflation of the zero-class by a factor 1–w for the probability of being non-susceptible along the lines of Cohen[2] is given by $h_R(y) \equiv h_R(y, w, m, N)$ as:

$$h_R(y) = \begin{cases} (1-w)+wh_R^*(0), & y=0 \\ wh_R^*(y), & y=1,2,\ldots,m \end{cases} \quad (6)$$

or rather by $h_R(y) \equiv h_R(y, \theta, m, N)$ as:

$$h_R(y) = \begin{cases} (1-\theta), & y=0 \\ \dfrac{\theta}{\varphi-2}(1+F(y)), & y=1,2,\ldots,m \end{cases} \quad (7)$$

where, $\theta = \omega(1-h_R^*(0)) \le 1-h_R^*(0) = (\phi-2)/\phi$
and $\phi-2 = m+(N-m)/2m-N+1$

We note as follows:

(i) The distribution $h_R^*(y)$ is equivalent to a special case of $h_R(y)$ when $\theta = (\phi-2)/\phi$
(ii) The distribution $h_R^*(y)$ is inflated at y = 0 in relation to h_R^* since $h_R(0) \ge h_R^*(0)$.
(iii) The distribution $h_R(y)$ is monotonically non-increasing for y = 0, 1, 2,…,m.

This follows easily from the recurrence relations

$$h_R(y+1) = h_R(y) - \frac{\theta(2m-N)}{(\phi-2)(N-m-y)}F(y+1)$$

and

$$h_R(y+1) = h_R(y)\frac{(1+F(y+1))}{1+F(y)}$$

(iv) In effect,
$h_R(0) \ge h_R^*(0)$ but $h_R(y) \ge h_R^*(y)$ for $y > 0$.
(v) The distribution function
$H_R^*(y'/m, N) = \Pr\{Y_R \pi y'\}$ is given as:

$$H_R(y') = 1-\theta+\frac{\theta}{\phi-2}\left(y'+\frac{N-m-(m-y')F(1+y'))}{2m-N+1}\right)$$

for $y' \le m$ and with $H_R^*(0)=1-\theta$ and of course, F(y) = 0 for $y > N-m$.

The following results hold on the raw moments for the variable R.

RESULTS

The first moment E[R] for the distribution of R is given by:

$$E[R]=(m+1)\theta f \tag{8}$$

where, $1-f=\dfrac{(2m-N+1)(m(2m-N+2)+2(N-m))}{2(2m-N+2)(m(2m-N+1)+N-m)}$

Proof: We note by multiplying equation (7) by $(m-y+1)$ and summing that

$$\sum_{y=0}^{m}(m-y+1)h_R(y)=(m+1)(1+\theta)+\frac{\theta}{\phi-2}\sum_{y=1}^{m}(m-y+1)$$

$$+\frac{\theta}{\phi-2}\sum_{y=1}^{N-m}(m-y+1)F(y)$$

$$=(m+1)(1-\theta)+\frac{\theta}{\phi-2}\left(\frac{m(m+1)}{2}+(2m-N+1)\frac{\binom{m+1}{2m-N+2}}{\binom{m}{2m-N}}\right)$$

$$=(m+1)\left(1-\theta+\frac{\theta}{\phi-2}\left(\frac{m}{2}+\frac{N-m}{2m-MN+2}\right)\right)$$

Result follows on substituting ϕ - 2.

Above mentioned results can be generalized for higher moments as follows:

Let $(m-y+1)_k=(m-y+1)(m-y+2)\ldots(m-y+k)=\sum_{i=0}^{k}\phi_{ik}y^i$

The i^{th} moment $E[R^i]$ for distribution of R satisfies the recurrence relations

$$\sum_{i=0}^{k}\phi_{ik}\left(E[R^i]-\frac{\theta}{\phi-2}\sum_{y=1}^{m}y^i\right)=(m+1)_k\left(1-\theta+\frac{\theta(N-m)}{(\phi-2)(2m-N+k+1)}\right) \tag{9}$$

Proof: We note the following identity

$$\sum_{y=1}^{N-m}(m-y+1)_kF(y)=\frac{(m+1)_k(N-m)}{2m-N+k+1} \tag{10}$$

It follows by multiplying equation (7) by $(m-y+1)_k$ and summing over all y that

$$\sum_{y=0}^{m}(m-y+1)_kh_R(y)=(m+1)_k(1-\theta)+\frac{\theta}{\phi-2}\sum_{y=1}^{m}(m-y+1)_k+\frac{\theta}{\phi-2}\frac{(m+1)_k(N-m)}{2m-N+k+1}$$

The result follows:

Equation (9) completely iteratively determines moments of R in terms of Bernoulli numbers. This follows since[3]

$$\sum_{x=1}^{m}x^q=\frac{m^{q+1}}{q+1}+\frac{m^q}{2}+\sum_{p=1}^{p'}\frac{1}{2p}\binom{q}{2p-1}B_{2p}m^{q-(2p-1)}$$

(the last term contains m or m^2) where, $p'=\dfrac{q}{2}$ or $\dfrac{(q-1)}{2}$ and B_{2p} are Bernoulli numbers. In particular, putting k = 2 in equation (9), the second moment and subsequently the variance of R are given by

$$E[R^2]=\theta(m+1)(m+2)\left(-1+\frac{1}{3(\phi-2)}\left(m+\frac{3(N-m)}{2m-N+3}\right)\right)+(2m+3)E[R].$$

so that

$$Var[R]=\theta(m+1)(m+2)\left(-1+\frac{1}{3(\phi-2)}\left(m+\frac{3(N-m)}{2m-N+3}\right)\right)+E[R](2m+3-E[R]).$$

Given observations $(n_0, n_1, n_2, \ldots, n_m)$ on R such that $n = \sum_{y=0}^{m} n_y$, the maximum likelihood estimates of the parameters θ, m and N are given, if they exist, by solutions to the equations

$$\theta = 1 - \frac{n_0}{n} \tag{11a}$$

$$\sum_{y=1}^{m} \frac{n_y F(y)}{1+F(y)} (\Psi_1(y) - \Psi_1(0)) = \frac{n-n_0}{\phi - 2} \left(1 - \frac{N-m}{(2m-N+1)^2} \right) \tag{11b}$$

$$\sum_{y=1}^{m} \frac{n_y F(y)}{1+F(y)} (\Psi_2(0) - \Psi_2(y)) = \frac{n-n_0}{\phi - 2} \left(\frac{(m+1)}{(2m-N+1)^2} \right) \tag{11c}$$

where, Ψ_1 and Ψ_2 are digamma functions with $\Psi_1(y) \equiv \Psi(m-y+1)$ and $\Psi_2(y) \equiv \Psi(N-m-y+1)$.

Proof: We note that the likelihood function for R is given by:

$$L_R = (1-\theta)^{n0} \left(\frac{\theta}{\phi - 2} \right)^{n-n0} \prod_{y=1}^{N-m} (1+F(y))^{ny}.$$

Taking logarithms and differentiating w.r.t. θ, m and N, the result follows. In the next section we show that our results for R transfer naturally with some adjustments to the case for P. A major difference is that variable P depends on c while R is independent of c.

Distribution of likelihood ratio P: The distribution $P(y) \equiv P(y/c, m, N)$ is given by:

$$h_P^*(y) = \Pr\{P = y\}$$

$$= \frac{h_1(y) + h_2(y)}{D h_2(y)} \qquad y = 0, 1, \ldots, c-1$$

where, $D = D(c, m, N)$ is a normalising constant. This gives, by substitution,

$$h_P^*(y) = \frac{\binom{m}{2m-N} + \binom{m+c-y}{2m-n}}{c\binom{m}{2m-N} + \binom{m+c+1}{2m-N+1} - \binom{m+1}{2m-N+1}}, \quad y = 0, 1, \ldots c-1 \tag{12}$$

Equation (12) is equivalent to

$$h_P^*(y/c, m, N) = \frac{1 + F(y-c)}{c + ((m+c+1)F(-c) - (m+1))/(2m-N+1)} \tag{13}$$

where, $F(y) = \binom{m-y}{2m-N} \Big/ \binom{m}{2m-N}$ and F(y) denotes the hypergeometric function F (y, N-2m, N-m+1, 1). As with the case for variable R, equation (13) can be modified for a mixed version as follows:

$$h_P(y) = h_P(y/\theta, c, m, N)$$

$$= \begin{cases} 1-\theta, & y = 0 \\ \dfrac{\theta}{1 - h_P^*(0)} h_P^*(y), & y = 1, 2, \ldots, c-1 \end{cases} \tag{14}$$

or rather

$$h_P(y) = \begin{cases} 1-\theta, & y = 0 \\ \theta(1+F(y-c))/\varphi, & y = 1,2,\ldots,c-1 \end{cases}$$

where, $\varphi = \varphi(c, m, N)$

$$= c - 1 + \frac{(N-m+c)F(-c)-(m+1)}{2m-N+1}$$

and $\theta<1-h_P^*(0)$. Of course, the distribution $h_P^*(y)$ is the special case of $h_P(y)$ when θ-$1-h_P^*(0)$.
We note that the distribution function $H_P(y') = \Pr\{P \le y'\}$ is given as:

$$H_P(y') = 1 - \theta + \frac{\theta}{\varphi}\left(y' + \frac{(N-m+c)F(-c)-(m-y'+c)F(y'+1-c)}{2m-N+1} \right)$$

for $y' \le c-1$ and with $H_P(0) = 1-\theta$. Furthermore $h_P(y)$ is monotonically non-decreasing in y just as in the case for $h_R(y)$.
We also state the following results:

The first moment E[P] for the distribution of P is given as:

$$E[P] = \theta\,(m+c+1-g/\varphi) \tag{15}$$

where,

$$g = (c-1)(m+1+c/2) + (2m-N+1)\left(\frac{\binom{m+c+1}{2m-N+2} - \binom{m+2}{2m-N+2}}{\binom{m}{2m-N}} \right)$$

$$= (c-1)(m+1+c/2) + \left(\frac{(m+1)_{c+1}}{(N-m+1)_{c-1}} - (m+1)(m+2) \right) / (2m-N+2)$$

Let $(m+c-y+1)_k = (m+c-y+1)(m+c-y+2)\ldots(m+c-y+k)$

$$= \sum_{i=0}^{k} \varphi_{ik}\, y^i$$

The ith moment $E[P^i]$ for P satisfies the recurrence relation

$$\sum_{i=0}^{k} \varphi_{ik}\left(E[P^i] - \frac{\theta}{\varphi} \sum_{y=1}^{c-1} y^i \right)$$

$$= (m+c+1)_k(1-\theta) + \theta\left(\frac{(m+1)_{c+k}}{(N-m+1)_{c-1}} - (m+1)_{1+k} \right) \Big/ (\varphi(2m-N+k+1)) \tag{16}$$

Given observations $(n_0, n_1, n_2, \ldots, n_{c-1})$ on P such that $n = \sum_{y=0}^{c-1} n_y$, the maximum likelihood estimates of parameters θ, c, m and N are given, if they exist, by solutions to the equations

$$\theta = 1 - \frac{n_0}{n} \tag{17a}$$

$$\sum_{y=1}^{c-1} \frac{n_y F(y-c)}{(1+F(y-c))} \left(\Psi_1(y-c) - \Psi_2(y-c) \right)$$

$$= \sum \frac{n_y}{\varphi}\left(1+\frac{F(-c)(1+(N-m+c)(\Psi_1(-c)-\Psi_2(-c)))}{2m-N+1}\right) \quad (17b)$$

and

$$\sum_{y=1}^{c-1} \frac{n_y F(y-c)}{(1+F(y-c))}(\Psi_1(y-c)-\Psi_1(0))$$

$$= \sum_{y=1}^{c-1} \frac{n_y}{\varphi}\left(\frac{(m+1)}{(2m-N+1)^2}-\frac{1}{(2m-N+1)}\right)$$

$$+ \sum_{y=1}^{c-1} \frac{n_y (N-m+c)(-c)}{\varphi(2m-N+1)}(\Psi_1(-c)-\Psi_1(0))-\frac{1}{(2m-N+1)} \quad (17c)$$

and

$$\sum_{y=1}^{c-1} \frac{n_y F(y-c)}{(1+F(y-c))}(\Psi_2(0)-\Psi_2(y-c))$$

$$= -\sum_{y=1}^{c-1} \frac{n_y}{\varphi}\frac{(m+1)}{(2m-N+1)^2}$$

$$+ \sum_{y=1}^{c-1} \frac{n_y(N-m+c)F(-c)}{\varphi(2m-N+1)}\left(\Psi_2(0)-\Psi_2(-c)+1+\frac{1}{(2m-N+1)}\right) \quad (17d)$$

where, Ψ_1 and Ψ_2 are digamma functions defined by:

$\Psi_1(y) \equiv \Psi(m+1-y)$ and $\Psi_2(y) \equiv \Psi(N-m+1-y)$

Proof: We note that the likelihood function for P is given by:

$$L_p = (1-\theta)^{n_0}\left(\frac{\theta}{\varphi}\right)^{n-n_0}\prod_{y=1}^{c-1}(1+F(y-c))^{n_y}$$

Taking logarithms are differentiating w.r.t θ, c, m and N, the result follows. We note that equations 11(a) and 17(a) coincide with estimating θ from the proportion of zeros in the sample. An alternative feasible method of estimating parameters is the combination of the proportion of zeros and the method of moments especially if one parameter (m or N) is assumed known.

In conclusion, this short paper proposed two distributions $h_R(y)$ and $h_P(y)$ in relation to evaluating policies when sampling without replacement from an urn containing balls of two different colours. The distributions are essentially mixed versions of the negative hypergeometric distribution with one ($h_R(y)$) independent of the critical parameter c and the other ($h_P(y)$) very much dependent on c. The definitions of the distributions also involve maximal and minimal negative hypergeometric distributions in a way similar to the use in the literature[4] for the maximal and minimal negative binomial distributions. Using some identities and recurrence relations we have obtained expressions for the general raw moments of the distributions. We have also provided equations for estimation of the relevant parameters. It is our opinion that an interactive package like Maple can be used in the quick solution of the equations by numerical methods. In any case Fig. 1 indicates a scatter plot of cumulative distribution $H_R(y)$

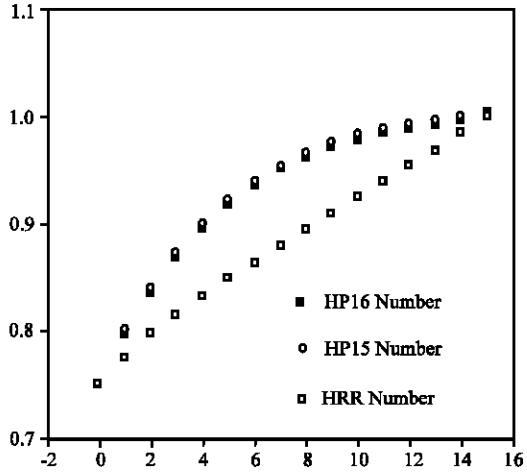

Fig. 1: Cumulative probability plots $H_r(y)$ and $H_P(y)$ for c = 15 and 16

and $H_P(y)$ for hypothetical cases N = 25, m = 15, θ = .75, c = 15, 16.

Notations and definitions

$h_1(y)$ = Neg. hypergeom. distribution for the m+c white balls in the majority

$h_2(y)$ = Neg. hypergeom. distribution for the N–m+c black balls in the minority

$h_{max}(y$ = Maximal negative hypergeometric distribution

$h_{min}(y)$ = Minimal negative hypergeometric distribution

$h_R(y)$ = Normalized likelihood ratio of $h_{max}(y)$ and $h_2(y)$

$h_P(y)$ = Normalized likelihood ratio of $h_{min}(y)$ and $h_2(y)$

$H_1(y)$ = Cumulative distribution function for $h_1(y)$

L_R(or L_P) = Likelihood function of R (or P)

$\Psi_1(y)$ = Digamma function Ψ(m+1-y)

$\Psi_2(y)$ = Digamma function Ψ(N-m+1-y)

F(y) = Hypergeometric function F (y, 2m-N, N-m+1, 1)

ϕ, φ = Functions of m, N (and c)

ϕ_{ik}, φ_{ik} = Functions of m (and c)

$(y)_k$ = y (y+1) (y+2)...(y+k-1) being polynomial of degree k in y

REFERENCES

1. Johnson, N., S. Kotz and A. Kemp, 1992. Univariate Discrete Distributions. Wiley, New York, USA.
2. Cohen, A.C., 1966. A Note on Certain Discrete Mixed Distributions. Biometrics, 22: 566-572.
3. Gradshteyn, I.S. and I.M. Ryzhik, 1980. Table of Integrals. Series and Products. Academy Press.
4. Zhang, Z., B.A. Burtness and D. Zelterman, 2000. The maximum negative binomial distribution. J. Stat. Planning and Inference, 87: 1-19.