Science and Technology, 5 (2000) 115-123

A Generalization of the Hypergeometric Distribution

E.K. Elsheikh* and A. Benmerzouga**

**Salalah Teachers College of Education, Salalah, Sultanate Of Oman*
***Department of Mathematics and Statistics, Sultan Qaboos University, P. O. Box 36 Al-Khod, P. C. 123 Muscat, Sultanate of Oman*

تعميم للتوزيع فوق الهندسي

الفضل خليفة الشيخ , على بن مرزوقة

خلاصـة : نقدم في هذا البحث تعديلا للتوزيع فوق الهندسي يأخذ بعين الاعتبار الحالة التي تكون فيها وسيلة المعاينة منحازة إلى أحد النوعين الذين يجري منهما الاختيار وذلك في مقابلة التوزيع الهندسي الذي تكون تحته كل العينات متساوية الاحتمال. ندرس خصـائص هـذا التوزيع الذي أسميناه التوزيع فوق الهندسي المعمم بما يشمل استخلاص وتقييم نوعاً من التقريب الطبيعي لهذا التوزيع.

ABSTRACT : In this paper we introduce a modification of the hypergeometric distribution that caters for the case when the sampling scheme favours the inclusion of units of one of the two types involved, as opposed to the hypergeometric distribution under which all samples are equally likely. The properties of the resulting distribution, termed the generalized hypergeometric, are studied, including the derivation and numerical assessment of a normal approximation of the distribution.

KEYWORDS: Hypergeometric Distribution; Recurrence Relations; Normal Approximation.

The celebrated capture re- capture scheme (see , for example, Tuckwell 1995), as applied to the estimation of the number of fish in a lake, runs as follows. Catch a number of fish S, say, from the lake, mark them and then release them in the lake. Now, re-catch another sample of fish, of size N, say, from the lake. If X is the number of marked fish in this latter sample, then assuming X has a hypergeometric distribution, the total number of fish in the lake can be estimated. However, after the trauma of being caught and marked, the marked fish might, perhaps, be more likely to be caught again compared to the fish that escaped that experience. If that were the case, the hypergeometric model would not be applicable. The correct model should make samples with *few* marked fish less likely and samples with *many* marked fish more likely as compared to the hypergeometric model. One such model is the generalization of the hypergeometric distribution that we consider here.

The distribution first arose when considering a stochastic model for two species competing on a fixed number of sites. This model will be described in section 2. The assumptions of the model and the resulting equilibrium distribution resemble their counterparts in the stochastic approach to the kinetics of chemical reactions (See, for example, McQuarrie 1967, Formosinho and Miguel 1979, and Hall 1983). We make full use of the techniques used in studying those distributions to investigate the properties of the generalized hypergeometric distribution.

A Stochastic Model for Two Competing Species

Consider two species A and B with total fixed sizes S and M, respectively, competing on N fixed sites. Initially some of the sites, r_0 say, are occupied by species A and the rest by species B. We assume that all of the sites will continuously be occupied, but the number of sites occupied by either species may increase or decrease, respectively, by pushing out, or being pushed out by, members of the other

species. More specifically, denoting by $X(t)$ the number of sites occupied by species A at time t, we assume that:

(i) The probability that X(t) increases by one in $(t, t+h)$ is

$$K_1(S - X(t))(N - X(t))h + o(h);$$

(ii) The probability that $X(t)$ decreases by one in $(t, t+h)$ is

$$K_2\, X(t)(M - N + X(t))h + o(h);\ \text{and}$$

(c) The probability of more than one event in $(t, t+h)$ is $o(h)$, where K_1 and K_2 are positive constants, and $o(h)$ is such that $\lim_{h\to 0}\frac{o(h)}{h} = 0.$

It might help the argument to think, tentatively, of the two species as political parties and the N sites as seats of a parliament that is being continuously updated according to the rules specified in (i), (ii), and (iii). The parameters K_1 and K_2 are then measures of competitiveness of the two parties.

Differential Difference Equation and Equilibrium Solution

Let $p_r(t) = P(X(t) = r \mid X(0) = r_0)$. It easily follows from the assumptions that p_r satisfies the differential difference equation

$$\begin{aligned}\frac{d\,p_r(t)}{dt} = {} & K_1(S - r + 1)(N - r + 1)\,p_{r-1}(t) + K_2(r+1)(M - N + r + 1)\,p_{r+1}(t) \\ & - (K_1(S-r)(N-r) + K_2 r(M - N + r))\,p_r(t).\end{aligned} \tag{2.1}$$

The equilibrium solution is obtained by putting the derivative equal to zero and solving for $p_r = p_r(\infty)$. It is easy to see that it satisfies

$$(r+1)(M - N + r + 1)\,p_{r+1} = K(S - r)(N - r)\,p_r, \tag{2.2}$$

where

$$K = K_1/K_2. \tag{2.3}$$

Putting $r = 0,1,...,$ and making successive substitutions leads to the equilibrium solution

$$p_r = \frac{K^r S!N!(M - N)!}{r!(S - r)!(N - r)!(M - N + r)!}\,p_0. \tag{2.4}$$

The constant p_0 can be obtained by noting that the summation of p_r over all values of r is equal to 1. This distribution is a generalization of the hypergeometric distribution and can justifiably be called a *generalized hypergeometric* (gh) distribution. However, the term generalized distribution may mean a different thing in the literature. See for example, Johnson and Kotz (1969), pp. 158-60. The hypergeometric distribution corresponds to the special case $K = 1$. In fact, stressing the dependence on K, p_r can be written as

$$p_r(K) = K^r \binom{S}{r}\binom{M}{N-r} p_0(K) \Big/ \binom{M}{N} \tag{2.5}$$

Putting $K = 1$ and summing over all r gives

$$p_0(1) = \binom{M}{N} \Big/ \binom{S+M}{N}. \tag{2.6}$$

That is , the distribution reduces to the hypergeometric distribution. Thus, if K =1, the number of sites occupied by species A is effectively determined by taking a *simple* random sample of size N from the $S+M$ units in the two species and counting the units that belong to species A. In this case all samples of size N are equally likely. The case $K > 1$ is more favorable to inclusion of units of type A than the case $K = 1$. One should expect that situations in which a *few* sites are occupied by A are less likely, and those in which *many* sites are occupied by A are more likely, compared to the hypergeometric case. The reverse should be expected when $K<1$. It is reasonable to interpret the number of sites r occupied by species A as few or many according as r is less than or greater than the average number of sites, μ, occupied by species A. One obvious result to expect is that $\mu(K)$ should be an increasing function of K. We now see how these expectations are met by the gh distribution.

Let $G(x, K)$ and $H(x)$ denote, respectively, the probability generating functions (pgf) of the gh distribution and the hypergeometric distribution. Thus $H(x) = G(x,1)$. Now, from (2.5), we have

$$p_r(K) = K^r\, p_r(1)\, p_0(K) \big/ p_0(1)\,. \tag{2.7}$$

Summing over all r gives

$$H(K) = p_0(1) \big/ p_0(K) \tag{2.8}$$

Likewise, multiplying both sides of (2.7) by r, summing over all r and using the above result gives

$$\mu(K) = K\, H'(K) \big/ H(K)\,. \tag{2.9}$$

Note also that

$$p_r(K) = K^r\, p_r(1) \big/ H(K)\,. \tag{2.10}$$

Multiplying both sides by x^r and summing over all r expresses the pgf G in terms of H as

$$G(x,K) = H(Kx) \big/ H(K)\,. \tag{2.11}$$

Using (2.11), or multiplying both sides of (2.2) by x^r and summing over all r, leads to the following differential equation of $G(x, K)$

$$KSNG - (K(S + N - 1)x + M - N + 1)\frac{dG}{dx} + (Kx - 1)x\frac{d^2G}{dx^2} = 0\,. \tag{2.12}$$

Let $\psi_r(K) = p_r(1) \big/ p_r(K)$ for r a non-negative integer in the support of p_r. It follows from (2.10) that

$$\psi_r(K) = H(K) / K^r \tag{2.13}$$

We can now prove the following theorem.

***Theorem* 1**: Keeping the other parameters fixed:

(a) $\psi_r(K)$ is an increasing or decreasing function of K according as $r < \mu(K)$ or $r > \mu(K)$;

(b) $\mu(K)$ is an increasing functions of K.

Proof: The first assertion follows by differentiating (2.13) with respect to K, noting where the derivative is positive or negative, and expressing the condition in terms of $\mu(K)$, using (2.9). The assertion concerning $\mu(K)$ follows if we can establish that the derivative with respect to K is positive. It follows from (2.9) that

$$\frac{d\mu(K)}{dK} = \{H(K)[H'(K) + KH''(K)] - K[H'(K)]^2\} \big/ \left[H(K)\right]^2 \tag{2.14}$$

Differentiating both sides of (2.11) twice with respect to x and putting $x = 1$ gives

$$K^2H''(K) \big/ H(K) = E(X(X - 1)) = \sigma^2 + \mu^2 - \mu\,. \tag{2.15}$$

Substituting for $\mu(K)$ its expression in (2.9) leads to

$$\frac{d\mu(K)}{dK} = \frac{\sigma^2(K)}{K} > 0 . \tag{2.16}$$

The essence of the comparison between the gh and the hypergeometric distributions is captured in the following corollary that easily follows from *Theorem* 1 .

Corollary:

(a) $p_r(1) > p_r(K)$ if and only if $r < \mu(K)$;
(b) $\mu(K) > \mu(1)$ if and only if $K > 1$.

Moments of the gh Distribution

It is difficult to express the moments of the gh distribution, including the mean and variance, in simple algebraic functions of the parameters. In this section we develop relations that provide convenient means of evaluating the mean and variance without the need to compute the probabilities in (2.4). Note that putting $x = 1$ in (2.12) leads to the relationship

$$\mu(M - N + \mu) + \sigma^2 = K(S - \mu)(N - \mu) + K\sigma^2 . \tag{3.1}$$

We need to eliminate σ^2 in order to get a relation involving μ alone. This can be achieved by investigating the variation of the mean with each of the four parameters of the distribution keeping the others fixed. In the case of K this leads to a first order differential equation, while for M, S, and N it leads to recurrence relations. As for other moments, it can be shown, by differentiating (2.12) r times and putting $x = 1$, that the factorial moments satisfy the relations

$$K(S - r)(N - r)m_r - (K(S + N - 2r - 1) + M - N + r + 1)m_{r+1} + (K - 1)m_{r+2} = 0 , \quad r = 0, 1, 2,... \tag{3.2}$$

Variation of the Mean with *K*

Substituting (2.16) in (3.1) and noting that when $K = 1$ the mean reduces to that of the corresponding hypergeometric distribution, we get the following theorem.

***Theorem* 2**: The mean of the gh distribution varies with K according to the differential equation

$$K(K - 1)\frac{d\mu}{dK} = \mu(M - N + \mu) - K(S - \mu)(N - \mu) , \tag{3.3}$$

with initial value

$$\mu(1) = NS/(S + M) . \tag{3.4}$$

It is difficult to solve this differential equation analytically but a numerical solution should be possible. Once the mean is computed, $\sigma^2(K)$ can be evaluated from (3.1). Note also, from (2.16), that $\sigma^2(K)$ can be obtained from the slope of the curve of $\mu(K)$ at K.

Recurrence Relations for the Mean

First note that, considering the dependence of p_r on S, for fixed K, M, and N, we have from (2.4)

$$S\, p_r(S - 1) = (S - r)\, p_r(S)\, p_0(S - 1)/p_0(S) , \tag{3.5}$$

true for all r . Summing over all r gives

$$S = [S - \mu(S)]\, p_0(S-1) \big/ p_0(S) . \tag{3.6}$$

Multiplying (3.5) by r and summing over all r leads to

$$S\mu(S-1) = [S\mu(S) - \mu^2(S) - \sigma^2(S)]\, p_0(S-1) \big/ p_0(S) . \tag{3.7}$$

Substituting (3.6) in (3.7) we get

$$\sigma^2(S) = [S - \mu(S)].[\mu(S) - \mu(S-1)] . \tag{3.8}$$

Following similar steps, it can be shown that

$$\sigma^2(M,N) = [N - \mu(M,N)].[\mu(M,N) - \mu(M-1, N-1)] . \tag{3.9}$$

Finally, by interchanging M and S, and μ and μ_B, where μ_B is the average number of sites occupied by species B, it follows from (3.8) that

$$\sigma^2(M) = [M - \mu_B(M)].[\mu_B(M) - \mu_B(M-1)] . \tag{3.10}$$

But $\mu_B(M) = N - \mu(M)$. Thus

$$\sigma^2(M) = [M - N + \mu(M)].[\mu(M-1) - \mu(M)] . \tag{3.11}$$

Substituting (3.8), (3.9), and (3.11) in (3.1) and introducing suitable starting values for computational purposes leads to the following theorem.

***Theorem* 3**: The mean varies with each of the parameter(s) shown as argument, with the other parameters fixed, according to the following recurrence relations

(i) $\mu(S) = [KSN + S\mu(S-1) - KS\mu(S-1)]/[M - N + S + KN + (1-K)\mu(S-1)]$, with $\mu(0) = 0$. (3.12)

(ii) $\mu(M) = [KSN - (M-N)(1-K)\mu(M-1)]/[K(S+M) + (1-K)\mu(M-1)]$, with $\mu(0) = N$. (3.13)

(iii) $\mu(M,N) = [KSN + N(1-K)\mu(M-1, N-1)]/[M + KS + (1-K)\mu(M-1, N-1)]$, (3.14)
with $\mu(0, M) = 0$ and $\mu(N, 0) = N$.

It is desirable to have a relation that describes the variation of the mean with N alone. This easily follows by suitably substituting (3.14) in either of (3.12) or (3.13). The resulting relation is given in the following corollary.

Corollary: The mean varies with N according to the following relation

$$\mu(N) = KN[S - \mu(N-1)]/[KS + M - N + 1 + (1-K)\mu(N-1)], \text{ with initial value } \mu(0) = 0. \tag{3.15}$$

Elsheikh (1997) has derived similar recurrence relations for the equilibrium means of some distributions arising in chemical reactions. These relations are best illustrated by an example.

Example

Consider a gh distribution with parameters $M = 6$, $S = 5$, $N = 4$, and $K = 2$. The distribution can be worked out using (2.2). It turned out that it has mean and variance given by $\mu = 632/275$, and $\sigma^2 = 51576/2752$. Using relation (3.14), the most feasible in this case, with the given starting value, one can directly verify that $\mu(1) = 5/8$, $\mu(2) = 28/23$, $\mu(3) = 87/49$, and finally, $\mu(4) = 632/275$, in agreement with the correct value. The variance now can be obtained from (3.1). The other relations given by *Theorem* 3 give exactly the same values. Note that for the latter relations the variance can also be obtained from (3.8), or (3.9), or (3.11) according to the relation that was used to obtain the mean.

Approximating the Mean and Variance

Note that when the minimum of *N, S, M* is fairly large then $\mu(S)$ should not differ much from $\mu(S\text{-}1)$. Using this in (3.12), the mean can be expected to be well approximated by the positive root of the equation

$$\mu(M - N + \mu) = K(S - \mu)(N - \mu). \tag{3.16}$$

Differentiating (3.16) with respect to K and using (2.16) gives

$$[(M - N + \mu + K(S - \mu) + K(N - \mu) + \mu]\sigma^2 = K(S - \mu)(N - \mu). \tag{3.17}$$

Using (3.1) again we get

$$\frac{1}{\sigma^2} = \frac{1}{\mu} + \frac{1}{N - \mu} + \frac{1}{S - \mu} + \frac{1}{M - N + \mu}. \tag{3.18}$$

Numerical computation using the recurrence equations derived in section 3 indicates that the error in these approximations does not exceed 1, irrespective of the values of the parameters. Noting that the random variable is integer-valued, the approximations are thus very good.

Maximum Likelihood Estimation

First assume that all parameters of the distribution are known except K. This situation may arise when we suspect that the sampling scheme is favorable to one of the two types concerned and we want to quantify the extent of that. The likelihood function is conveniently expressed in (2.10). Differentiating with respect to K and equating the derivative to zero gives the maximum likelihood estimate (MLE) of K as the solution of the equation

$$rH(K) = K' H(K). \tag{4.1}$$

Using (2.9), we have the MLE of K as the solution of

$$\mu(K) = r \tag{4.2}$$

Using (4.2) with (3.16), we get an approximate expression of the MLE of K as

$$\hat{K} = r(M - N + r)/(S - r)(N - r). \tag{4.3}$$

The second situation to consider is when all parameters are known except M. We think of M here as the number of unmarked fish, and we assume that K is known from previous experiences. It is straightforward from (2.4) that

$$M(M - N)P_0(M)P_r(M - 1) = (M - N + r)P_0(M - 1)P_r(M). \tag{4.4}$$

Summing over all r gives

$$M(M - N)P_0(M) = (M - N + \mu)P_0(M - 1). \tag{4.5}$$

Using (4.5) in (4.4) we can express the likelihood ratio (LR) as

$$LR(M) = \frac{P_r(M)}{P_r(M - 1)} = \frac{M - N + \mu}{M - N + r}. \tag{4.6}$$

The MLE of M is given as

$$\hat{M} = \max\{M : LR(M) \geq 1\}. \tag{4.7}$$

This simplifies, in light of (4.6), to

$$\hat{M} = \max\{M: \mu(M) \geq r\}. \tag{4.8}$$

Since μ obviously increases with M (see (3.10)), it follows that $\hat{M}$ is given by the integer part of the solution for M of the equation

$$\mu(M) = r \tag{4.9}$$

Using (4.9) with (3.16) gives an approximate expression of $\hat{M}$ as

$$\hat{M} = \left[\frac{(N - r)(K(S - r) + r)}{r}\right]. \tag{4.10}$$

where [x] is the integer part of x. It is interesting to note that when $K = 1$ the approximate MLE of M reduces to that given by the hypergeometric distribution.

It is to be noticed from the preceding derivation that the same likelihood equation resulted in the estimation of K and M. If both of them are unknown, it is necessary to take more than one observation from the distribution.

Binomial and Normal Approximations of the Distribution

Assume K and N are finite, $S \to \infty$, $M \to \infty$, such that $S/M = \lambda$. It follows from (2.2) that

$$(r + 1)\, p_{r+1} = K\lambda(N - r)\, p_r. \tag{5.1}$$

This corresponds to a binomial distribution with parameters N and $p = K\lambda /(1+K\lambda)$. If N is large the normal approximation of the binomial distribution should be in effect.

A normal approximation of the distribution, not necessarily through the binomial, is also possible. The derivation that follows is essentially due to Dunstan and Reynolds (1981). We have , from (2.2),

$$\frac{p_r}{p_{r-1}} = \frac{K(S - r + 1)(N - r + 1)}{r(M - N + r)}. \tag{5.2}$$

If the minimum of N, S, and M is large, the mode m defined by

$$m = \max\{r : p_r / p_{r-1} > 1\}, \tag{5.3}$$

approximately satisfies the equation

$$m(M - N + m) = K(S - m + 1)(N - m + 1). \tag{5.4}$$

It follows from (5.2) that, for $j > 0$

$$\frac{p_{m+j}}{p_m} = \prod_{i=1}^{j} \frac{p_{m+i}}{p_{m+i-1}} = \prod_{i=1}^{j} T_i \tag{5.5}$$

where

$$T_i \cong (1 - \frac{i}{S - m + 1})(1 - \frac{i}{N - m + 1})(1 + \frac{i}{m})^{-1}(1 + \frac{i}{M - N + m})^{-1}. \tag{5.6}$$

Thus,

$$T_i \cong \exp[-i(\frac{1}{m}+\frac{1}{N-m}+\frac{1}{S-m}+\frac{1}{M-N+m})]. \quad (5.7)$$

It follows that

$$\frac{p_{m+j}}{p_m} \cong \exp(-\frac{j^2}{2\sigma^2}), \quad (5.8)$$

where

$$\frac{1}{\sigma^2}=\frac{1}{m}+\frac{1}{N-m}+\frac{1}{S-m}+\frac{1}{M-N+m}. \quad (5.9)$$

Hence the distribution can be approximated by a normal distribution with mean μ and variance σ^2 as given by (5.4) and (5.9). As the mean of the distribution is well approximated by the solution of (3.16), the approximate mean can be used in place of the mode.

As noticed by Hall (1983), the derivation above is rather vague about the relative sizes of the parameters for which the approximation holds, apart from the requirement that the values of *M, S*, and *N* should be large, which is almost always met as far as chemical reactions are concerned. Hall has provided conditions on the parameters under which the distribution converges to normality. The approach we adopt here is to resort to numerical computations.

Numerical Assessment of the Normal Approximation

The analysis is restricted to $N \le S, M$. We considered values of N from as small as 10. The cumulative distribution function corresponding to (2.4) was computed at each possible integral value and compared to the value given by the normal distribution function with mean and variance given by (3.16) and (3.18), respectively. A continuity correction was used. For each set of parameters the maximum difference was recorded. The approximation was taken as satisfactory so long as the maximum absolute difference does not exceed 0.001.

Initially, extensive exploratory computations were carried out by varying the parameters *N*, *M*, *S*, and *K*. It was noticed that the quality of the approximation was closely related to the values of *N* and μ/N, where μ is the approximate mean given by the positive solution of (3.16). A program was then written to facilitate the investigation of the range of values of μ/N, for each *N,* under which the approximation is satisfactory. The calculations were run separately for $K < 1$ and $K > 1$. It was found that the approximation can be good for values of *N* as small as 30.
The results can be summarized for the two cases of *K* as follows.

Case of K > 1:

(i) For $N \ge 30$ the approximation is found to be satisfactory for $\mu/N \in [0.23, 0.59]$ and it gets better as *K* approaches 1.
(ii) For $N \ge 50$ also the approximation is found to be satisfactory provided $\mu/N \in [0.15, 0.69]$ and it gets better as *K* approaches 1.
(iii) For $N \ge 100$ also the approximation is found to be satisfactory provided $\mu/N \in [0.086, 1.0]$ and for all values of *K*.

Case of K < 1:

(i) For $N \ge 30$ the approximation is found to be satisfactory for $\mu/N \in [0.399, 1.0)$ and it gets better as *K* approaches 1.
(ii) For $N \ge 50$ also the approximation is found to be satisfactory provided $\mu/N \in [0.282, 1.0)$ and it gets better as *K* approaches 1.
(iii) For $N \ge 100$ also the approximation is found to be satisfactory provided $\mu/N \in [0.092, 1.0)$ and for all values of *K*.

Acknowledgements

Thanks are due to anonymous referees for their helpful comments and suggestions.

References

DUNSTAN, F. D. J. and REYNOLDS, J. F. 1981. Normal Approximations for Distributions Arising in the Stochastic Approach to Chemical Reactions Kinetics. J. *Appl. Prob*. **18:** 263-267.

ELSHEIKH, E. K. 1997. Recurrence Relations for the Equilibrium Means of Distributions Arising in Chemical Reactions. *Sultan Qaboos University Journal for Scientific Research – Science and Technology*. **2:** 77-85.

FORMOSINHO, S. J. and MIGUEL, M. DA G. M. 1979. Markov Chains for Plotting the Course of Chemical Reactions. J. *Chem. Education* **56:** 582-585.

HALL, A. 1983. On the Roles of the Bessel and Poisson Distributions in Chemical Reactions Kinetics. J. *Appl. Prob*. **20:** 585-595.

JOHNSON, N. L. and KOTZ, S. 1969. *Discrete distributions*. John Wiley & Sons. New York.

McQUARRIE, D. A. 1967. Stochastic Approach To Chemical Kinetics. J. *Appl. Prob*. **4:** 413-478.

TUCKWELL, H. C. 1995. *Elementary Applications of Probability Theory.* Chapman & Hall. New York.

Received 20 January 1999
Accepted 25 June 2000