"Science Stays True Here"
Journal of Mathematics and Statistical Science (ISSN 2411-2518, USA), Vol.6, Issue 9, 233-240 | Science Signpost Publishing

Theoretical Results and Applications of the Negative Hypergeometric Distribution

W.H. Moolman

Walter Sisulu University, Mthatha, South Africa.

E-mail: moolman.henri@gmail.com

Abstract

The negative binomial distribution can be seen as the waiting time distribution associated with the binomial distribution. In a similar fashion the negative hypergeometric distribution can be seen as the waiting time distribution of the hypergeometric distribution. A definition and summary of the main theoretical results of this distribution will be given. Results and outputs (using R) of applications of the negative hypergeometric distribution will be shown and discussed.

Keywords: negative hypergeometric, negative binomial, approximations, moments, estimation, marketing, capture-recapture, games applications, simulation, matching.

1. Definition, Sum and Approximation of the Negative Hypergeometric Distribution

An urn has N balls consisting of M red and $N-M$ white balls. Let X denote the number of balls drawn before the k th red ball appears. The random variable X follows a **negative hypergeometric** distribution with probability mass function

$$P(X=x)=f_X(x)=P(k-1\text{ red balls in the first }x-1\text{ draws})\times P(\text{ red ball on the }xth\text{ draw})$$

$$P(X=x)=f_X(x)=\frac{{}^{N-M}C_{x-k}\,{}^{M}C_{k-1}}{{}^{N}C_{x-1}}\times\frac{M-k+1}{N-x+1}$$

$$=\frac{{}^{x-1}C_{k-1}\,{}^{N-x}C_{M-k}}{{}^{N}C_{M}}\text{ , where }x=k,k+1,\cdots,k+N-M\text{ .}$$

This distribution is also referred to as the **hypergeometric waiting time** distribution or **inverse hypergeometric** distribution. Its probability mass and distribution functions will be denoted by dnhyper (x,M,N,k) and pnhyper (x,M,N,k) respectively. To show that the abovementioned probabilities add up to 1, the following formula from Graham, Knuth and Patashnik (1994) can be used. For any natural numbers

$$a>b>0\text{ and }c>0,\ \sum_{y=0}^{a-b}{}^{a-y}C_b\,{}^{c+y}C_c={}^{a+c+1}C_{b+c+1}\,.$$

When $k=c+1$, $N=a+k$, $M=b+k$ and $y=x-k$ the above formula becomes

$\sum_{x=k}^{k+N-M} {}^{x-1}C_{k-1}\, {}^{N-x}C_{M-k} = {}^{N}C_{M}$ i.e. $\sum_{x=k}^{k+N-M} f_X(x) = 1$.

If $\frac{M}{N} \to p$ as $N \to \infty$ and $M \to \infty$, $\frac{M}{N-i} \to p$, $\frac{M-i}{N-x+k-i} \to p$ and $\frac{M-k+1}{N-x+1} \to p$,

the distribution of X becomes $f_X(x) = {}^{x-1}C_{k-1} p^{k-1}(1-p)^{x-k}\, p \; = {}^{x-1}C_{k-1} p^{k}(1-p)^{x-k}$,

which is the negative binomial distribution. A more detailed proof of the above mentioned formula can be found in stack exchange. Improved approximations to the negative hypergeometric distribution were suggested by Hu, Cui and Yin (2013), Hu and Yin (2014) and Teerapabolarn (2014). Mansuri (2012) gave a poisson approximation by using the w-function and Stein identity. Under certain conditions the normal and gamma distributions can also be used as approximations.

2. Moments of the Negative Hypergeometric Distribution

Khan (1994) showed that for the negative hypergeometric distribution

$E(X_m) = \frac{(k+m-1)!}{(k-1)!} \times \frac{(N+1)_m}{(M+1)_m}$, where $X_m = X(X+1)\cdots(X+m-1)$. From this expression all the moments can be calculated e.g. $E(X) = \frac{k(N+1)}{M+1}$ and

$$\mathrm{var}(X) = E(X_2) - E(X) - E^2(X) = E[X(X+1)] - E(X) - E^2(X) = E(X_2) - E(X) - E^2(X)$$

$$= \frac{(N+1)(N-M)}{(M+1)^2(M+2)} k(M+1-k).$$

3. Applications of the Negative Hypergeometric Distribution

3.1 Marketing – Estimation of M (N known)

A new product is being marketed. A small sample of the product was sent to each of N = 500 potential buyers. In a follow up they found that they had to contact x = 30 of these potential buyers before finding k = 5 potential buyers that stated that they will use the product. This information can be used to estimate M the number of buyers who will use the product and the proportion $p = \frac{M}{N}$. The plot below shows

1000 × dnhyper (30, M, 500, 5) versus M and suggests that M = 83 has maximum probability.

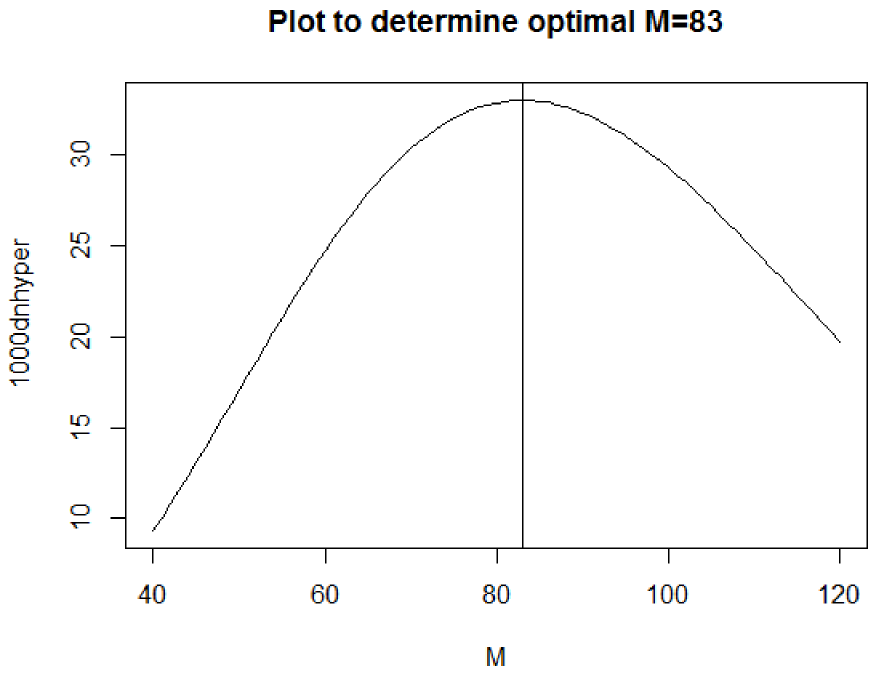

Figure 1 – Plot of 1000 × dnhyper (30, *M*, 500, 5) versus *M*

3.2 Estimation of a rare characteristic proportion

Let *M* denote the of elements, in a population of size *N*, that have a rare characteristic of interest.

Since
$$E(\frac{k-1}{x-1}) = \sum_{x=k}^{k+N-M} \frac{k-1}{x-1} \times \frac{{}^{N-M}C_{x-k}\, {}^{M}C_{k-1}}{{}^{N}C_{x-1}} \times \frac{M-k+1}{N-x+1}$$

$$= \frac{M}{N} \sum_{x=k}^{k+N-M} \frac{{}^{N-M}C_{x-k}\, {}^{M-1}C_{k-2}}{{}^{N-1}C_{x-2}} \times \frac{M-k+1}{N-x+1} = \frac{M}{N},$$

the estimator $\hat{p}_{neghyp} = \frac{k-1}{x-1}$ is an unbiased estimator of $p = \frac{M}{N}$. It can be shown that when $N \to \infty$, $\hat{\text{var}}(\hat{p}_{neghyp}) = \frac{(k-1)(x-k)}{(x-1)^2(x-2)} = \frac{\hat{p}_{neghyp}(1-\hat{p}_{neghyp})}{x-2}$. For the traditional estimator $\hat{p} = \frac{k}{x}$ of p, $\hat{\text{var}}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{x} = \frac{k(x-k)}{x^3}$. It can be shown [using calculations for various combinations of (k, x) values] that $\hat{\text{var}}(\hat{p}_{neghyp}) < \hat{\text{var}}(\hat{p}), x = 4k-1, 4k, 4k+1, \cdots$. When estimating a rare characteristic proportion, x will be large in relation to k. Therefore, in such an estimation, $\hat{p}_{neghyp}$ will be a more accurate estimator.

3.3 Capture-recapture – Estimation of N (M known)

Mark $M = t \times N_u$, where N_u is an upper bound of N based on past information and $t \le 0.10$, members of the population of interest and release them back into the population. Select a sample of size n at random without replacement from the population and observe the positions of the marked members in the sample $X_k, k = 1,2,\cdots,\ell$. Dalpatadu and Singh (2016) performed experiments with $N_u = 20\,000, n = 200$, $t = 0.03$ and $t = 0.06$. Each X_k is substituted for $E(X)$ in the formula for the negative hypergeometric mean i.e. $X_k = \dfrac{k(N_k + 1)}{M+1}$ and the equation solved for N. This gives $\hat{N}_k = \dfrac{X_k(M+1)}{k} - 1$ and $\hat{N} = \dfrac{1}{\ell}\sum_{k=1}^{\ell}[\dfrac{X_k(M+1)}{k} - 1]$. The graphs below show (i) the histogram of the simulated distribution of 20 000 values of $\hat{N}$ using $N_u = 20\,000, n = 200$, $t = 0.10$ and (ii) a Pearson type VI curve which was suggested and fitted to the frequencies by using the PearsonDS package in R.

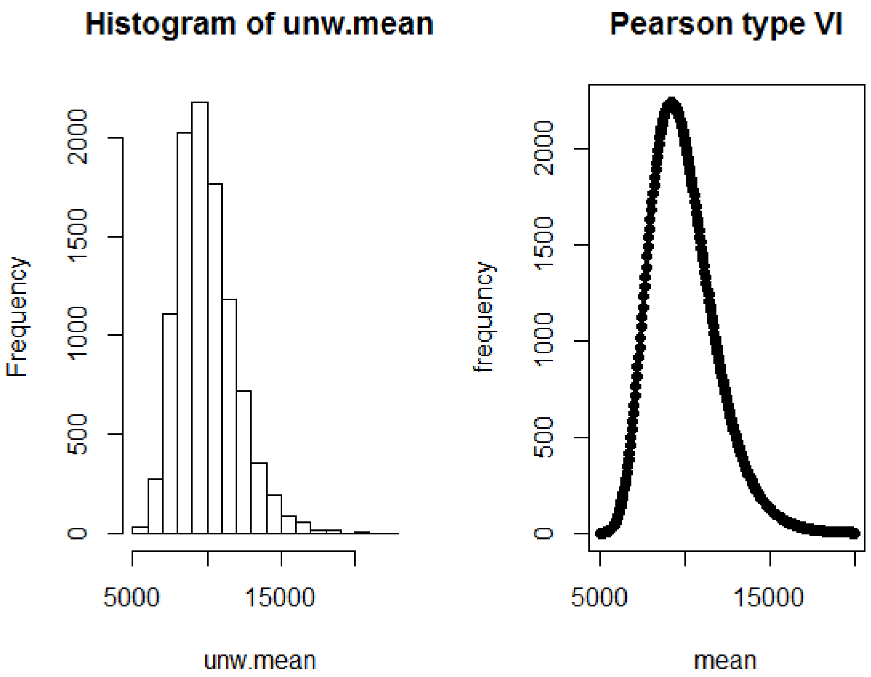

Figure 2 – Observed and fitted distributions of $\hat{N}$

3.4 Estimation of the proportion $p = \dfrac{M}{N}$ (N known)

The proportion p of marked members of the population (referred to in 3.3 above) is unknown. A random sample of size n is selected at random without replacement from the population and the positions $X_k, k = 1,2,\cdots,\ell$ of the marked members in the sample observed. Each X_k is substituted for $E(X)$ in the formula for the negative hypergeometric mean i.e. $X_k = \frac{k(N+1)}{M+1} = \frac{k(N+1)}{Np_k+1}$ and the equation solved for p_k. This gives $\hat{p}_k = \frac{1}{N}[\frac{k(N+1)}{X_k} - 1]$ and $\hat{p} = \frac{1}{N\ell}\sum_{k=1}^{\ell}[\frac{k(N+1)}{X_k} - 1]$. The graphs below show the histogram of the simulated distribution of 20 000 values of $\hat{p}$ for $N = 10\ 000$, $n = 200$, $p = 0.10$ and a Pearson type VI curve fitted to the distribution using the PearsonDS package in R

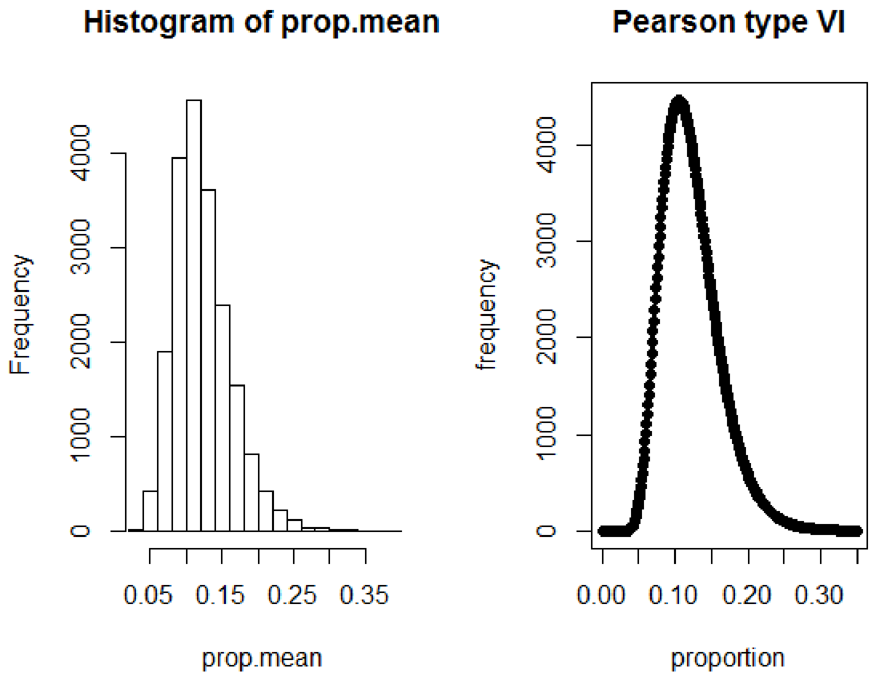

Figure 3 – Observed and fitted distributions of $\hat{p}$

3.5 Matching card game simulation

The following game is described by Dalpatadu and Singh (2016). A card is selected at random (and removed) from a standard deck of cards and the card rank (2, 3, 4, 5, 6, 7, 8, 9 ,10, jack, queen, king, ace) and suit (clubs, diamonds, hearts, spades) recorded. After the initial selection, cards are selected at random without replacement from the remaining cards in the deck until a match (in terms of rank) with the original card occurs. The random variable of interest is X the number of cards needed to achieve such a match. A contestant is rewarded according to the value of X (see table below). Under the assumption of a random selection of cards the possible rewards are set up to have an average pay off of 25 (see table below).

This experiment was simulated 100 000 times and the number of draws in each case recorded. The table below shows the theoretical probabilities (given by Dalpatadu and Singh), probabilities obtained from the simulation experiments for the X class intervals (1-10, 11-20, 21-30, 31-40, 41-49) as well as their respective awards. It can be seen that the simulated probabilities are very close to the theoretical ones.

Table 2 – Matching card games simulation results

x	**1-10**	**11-20**	**21-30**	**31-40**	**41-49**
prob (theory)	0.488115	0.296039	0.151980	0.055943	0.007923
prob (simulation)	0.48725	0.29530	0.15308	0.05664	0.00773
reward	10	15	30	100	700
reward × prob(simulation)	4.8725	4.4295	4.5924	5.664	5.411

The simulated expected award is 24.9694 , which is very close to the theoretical award of 25.

3.6 Occurrence of vowels in Psalms in the English Bible

The book of Psalms in the English Bible was downloaded and the values of X **,** the number of characters until a vowel is encountered, counted for the entire book. The frequency distribution of X is shown below.

Table 2 – Frequency distribution of number of characters until a vowel is encountered

x	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
frequency	7534	13707	15572	12445	8876	3670	1576	1037	477	285
x	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
frequency	497	617	581	309	143	60	38	12	6	5

According to theory (assuming random selections of vowels) the distribution of X should be negative hypergeometric with $k = 1$. However, this distribution will not be a good fit to the data as can be seen from the left of the plots below (The distribution is bimodal). The reason is that the vowels do not occur in a random fashion e.g. more than 86% of the vowels are encountered after counting 5 or less characters. A mixed distribution with a Pearson type VI distribution fitted to the frequencies for $x = 1,2,\cdots,9$ and a Pearson type I distribution to the frequencies for $x = 10,11,\cdots,20$ is an excellent fit to the frequencies (see right of the plots below). When fitting the mixed distribution to the Pearson type VI and I distributions respectively, the weights $w_1 = 0.962148$ and $w_2 = 1 - 0.962148 = 0.037852$ were used.

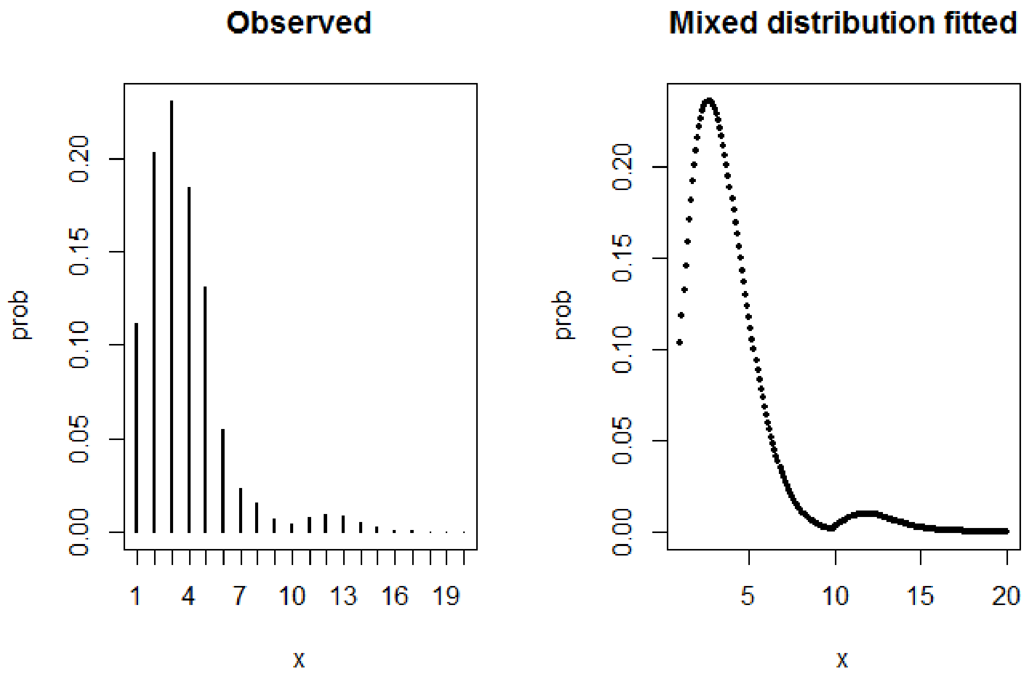

Figure 4 – Observed and fitted distributions of X

3.7 Games

3.7.1 Slot machine bonus games

An $m \times n$ rectangular grid contains s squares with jokers and various "reward" amounts on the remaining $mn - s$ squares. The game is played until (say) $j \le s$ jokers had occurred. The total payoff is the sum of all the rewards. The length of the game is a dnhyper(X, s, mn, j) random variable.

3.7.2 Battleship game

Each of 2 players secretly places pieces that make up a given set of ships on m locations on separate $n \times n$ grids. In turn players randomly select numbers between 1 and n^2 without replacement from the other player's grid. If a number containing part of a ship's location is selected, a "hit" is recorded. The game ends when a player first records $k = m$ hits on the other player's grid. For each player the number of selections to achieve m hits is a dnhyper(X, m, n^2, m) random variable. Denote the number of selections to achieve m hits by players 1 and 2 as X_1 and X_2 respectively. Then the length of the game will be expressed as the random variable $Y = \min(X_1, X_2)$. Since X_1 and X_2 are independent random variables, $F_Y(y) = P(Y \le y) = 1 - [1 -$ pnhyper $(y, m, n^2, m)]^2$, where pnhyper(X, m, n^2, m) is the distribution function of X .

References

Autin, MA and Gerstenschlager, NE (2018). Battleship and the Negative Hypergeometric Distribution. Teaching Statistics Trust, 41(1), 3-7.

Becker, M. (2017). Package 'PearsonDS'. Available at
https://cran.r-project.org/web/packages/PearsonDS/PearsonDS.pdf

Dalpatadu, RJ and Singh, AK (2016). Applications of the Negative Hypergeometric Distribution. ICAS2016, Phuket, Thailand, O95-O99.

Graham, RL, Knuth, DE and Patashnik, O (1994). Concrete Mathematics: A foundation for Computer Science (2nd ed.). Addison-Wesley, Reading, MA.
https://math.stackexchange.com/questions/2333583/negative-binomial-as-limit-of-the-negative-hypergeometric

Hu, D, Cui, Y and Yin, A (2013). An Improved Negative Binomial Approximation for Negative Hypergeometric Distribution, Applied Mechanics and Materials Online: 2013-09-27 ISSN: 1662-7482, Vols. 427-429, pp 2549-255. doi: 10.4028/www.scientific.net/AMM.427-429.2549

Hu, D and Yin, A (2014). Approximating the negative hypergeometric distribution. Int. J. Wireless and Mobile Computing, 7(6), 591-598.

Jones, SN (2013). A Gaming Application of the Negative Hypergeometric Distribution. Masters' Thesis, University of Nevada, Las Vegas.

Khan, RA (1994). A note on the generating function of a negative hypergeometric distribution. *Sankhyā* : The Indian Journal of Statistics B, 56(3), 309-313.

Mansuri, SB (2012). A Note on the Negative Hypergeometric Distribution and Its Approximation. World Academy of Science, Engineering and Technology, International Journal of Mathematical and Computational Sciences, 6(11), 1571-1573.

Miller, GK and Fridell, SL (2007). A forgotten discrete distribution? Reviving the Negative Hypergeometric Model. The American Statistician, 61(4), 347-350.

Teerapabolarn, K (2011). On the Poisson approximation to the negative hypergeometric distribution, Bulletin of the Malaysian Mathematical Sciences Society, 34, 331-336.

Teerapabolarn, K (2014). An Improved Negative Binomial distribution to approximate the negative hypergeometric distribution. International Journal of Pure and Applied Mathematics, 91(2), 231-235.