Skew-Normal Approximation to the Hypergeometric Distribution

Jose Almer T. Sanqui and Dong Won Jung
Department of Mathematical Sciences, Appalachian State University, 121 Bodenheimer Drive, Boone, NC 28608

Abstract
We investigate the skew-normal approximation to the hypergeometric distribution and show that the approximation improves on the classical normal approximation.

Key Words: cumulative distribution function, Central Limit Theorem, maximum absolute error, skewness

1. Introduction

1.1 Related Literature

The skew-normal distribution, first investigated by Azzalini (1985) is a fairly recent distribution that includes the standard normal distribution as a special case. In its basic form, its probability density function (p.d.f.) is given by $f(x;\lambda) = 2\varphi(x)\Phi(\lambda x)$ for all real numbers x and λ, with the latter determining the skewness of the distribution and φ and Φ being the p.d.f. and the cumulative distribution function (CDF) of the standard normal distribution, $N(0,1)$. The distributed is denoted as the $SN(\lambda)$ distribution and it's clear that $SN(0) = N(0,1)$. A characterization for this distribution is given in Gupta et al (2004). If $X \sim SN(\lambda)$ then $Y = \frac{X-}{\sigma}$ will have a shifted and scaled Skew-Normal distribution with location parameter and scale parameter σ and its distribution is denoted by $SN(\ ,\sigma,\lambda)$. All the central moments of this distribution have closed forms. Chang et al (2008) showed that this skew-normal distribution with the location and scale parameters provides an improved approximation to the $Binomial(n,p)$ distribution compared with the classical normal approximation when the $Binomial(n,p)$ is not symmetric, which is in a way not surprising since the Skew-Normal distribution is also not symmetric when $\lambda \neq 0$. Chang et al (2010) also showed the skew-normal distribution provides an improved approximation to the negative binomial distribution compared with the classical normal approximation. This paper will show that the skew-normal distribution also provides an improved approximation to the hypergeometric distribution compared with the classical normal approximation based on the Central Limit Theorem.

1.2 The Hypergeometric Distribution

The probability mass function of the hypergeometric distribution with parameters n, K and N is given by $f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$ where N is the population size, K is the number of success states in the population, n is the number of draws and x is the number of observed successes. Like the skew-normal distribution with location and scale parameters, all the central moments of this discrete distribution have closed forms. This distribution is commonly introduced in elementary statistics and probability courses in

context of sampling without replacement. For example, in a deck of cards containing 20 cards where 6 are red and 14 are black, if 5 cards are drawn randomly without replacement the probability that exactly 4 red cards are drawn is $f(4) = \frac{\binom{6}{4}\binom{14}{1}}{\binom{20}{5}} = .013544$. As another example, in a jar that contains 1000 marbles of which 400 are red and 600 are black, if 200 marbles are randomly selected again without replacement, the probability of selecting less than 90 red marbles is $\sum_{x=0}^{89} f(x) = F(89) = .9369$ where f and F are the probability mass function and cumulative distribution function, respectively, of the hypergeometric distribution with parameters $n = 200, K = 400$ and $N = 1000$.

1.3 The Classical Normal Approximation

From the Central Limit Theorem, the CDF $F(x)$ of the hypergeometric distribution with parameters n, K and N is approximated by $F(x) \approx \Phi(y)$ where Φ is the standard normal CDF and $y = \frac{(x+0.5)-(\frac{nK}{N})}{\sqrt{\frac{nx}{N}\frac{N-K}{N}\frac{N-n}{n-1}}}$. For the marble example in the previous section, the normal approximation for the probability of selecting less than 90 marbles is .9373 which is pretty close to the exact hypergeometric probability of .9369. This approximation is expected to be good when the hypergeometric distribution is not too skewed. When the distribution is skewed, we expect the skew-normal distribution will provide an improved approximation.

2. Skew-Normal Approximation

2.1 The Matching Skew-Normal Distribution

For a given *Hypergeometric(n,K,N)* distribution, the parameters of the approximating (or, matching) *SN*(, σ,) distribution are found by equating the first three central moments of the two distributions. The solution of the resulting three nonlinear equations for the three unknowns , σ and are given by

$$= \{(\frac{\sqrt{2/}\,(\frac{4}{}-1)}{\sqrt{\frac{N}{nK}}\sqrt{\frac{N}{N-K}}\sqrt{\frac{N-1}{N-n}}(\frac{K}{N}-\frac{(N-K)}{N})\frac{N-2n}{N-2}})^{2/3} + \frac{2}{} + 1\}^{-1/2},\ \sigma = \sqrt{\frac{\frac{nK(N-K)(N-n)}{N\quad N\quad N-1}}{\frac{1+(1-\frac{2}{})^{2}}{1+\ ^{2}}}} \text{ and}$$

$$= \frac{nK}{N} - \sigma\sqrt{\frac{2}{}}\frac{}{\sqrt{1+\ ^{2}}}.$$ Once the parameters , σ, are obtained from the above equations, the hypergeometric CDF $F(x)$ is approximated using by $F(x) \approx \psi_\lambda(y)$ where ψ_λ is the CDF of SN( ) and $y = (x + .5 - \ )/\sigma$.

2.2 Skew-Normal Approximation Example

For the marble example in section 1.2, the skew-normal approximation for the probability of selecting less than 90 marbles is .9368 which is closer to the exact hypergeometric probability of .9369 compared to the classical normal approximation of .9373. It would be interesting to see if this phenomenon is typically true hence we need to use an evaluation criterion described in the next section.

2.3 Skew-Normal versus Normal Approximation of Hypergeometric Distribution

For different values of the hypergeometric distribution parameters n, K and N, we compare the accuracy of the classical normal approximation given in Section 1.3 and the accuracy of the skew-normal approximation given in Section 2.1 using the same Maximum Absolute Error (MABS) criterion used in Chang et al (2008) defined as follows: $\mathrm{MABS}(n, K, N) = \max_x |F_{n,K,N}(x) - F^*_{n,K,N}(x)|$ where $F^*_{n,K,N}(k) = \Phi\left(\frac{(x+0.5)-\frac{nK}{N}}{\sqrt{\frac{nxN-KN-n}{N\ \ N\ \ N-1}}}\right)$ for the normal case and $F^*_{n,K,N}(x) =\ \ \ \ ((x+0.5)-\ \ )/\sigma)$ for the skew-normal case. We calculated the MABS for the normal and skew-normal approximations for varying values of n, K and N and a typical result is given in Section 3.

3. Results

Figure 3.1 Hypergeometric(n=10, N=50, K=5) pmf with matching normal pdf and and skew-normal pdf

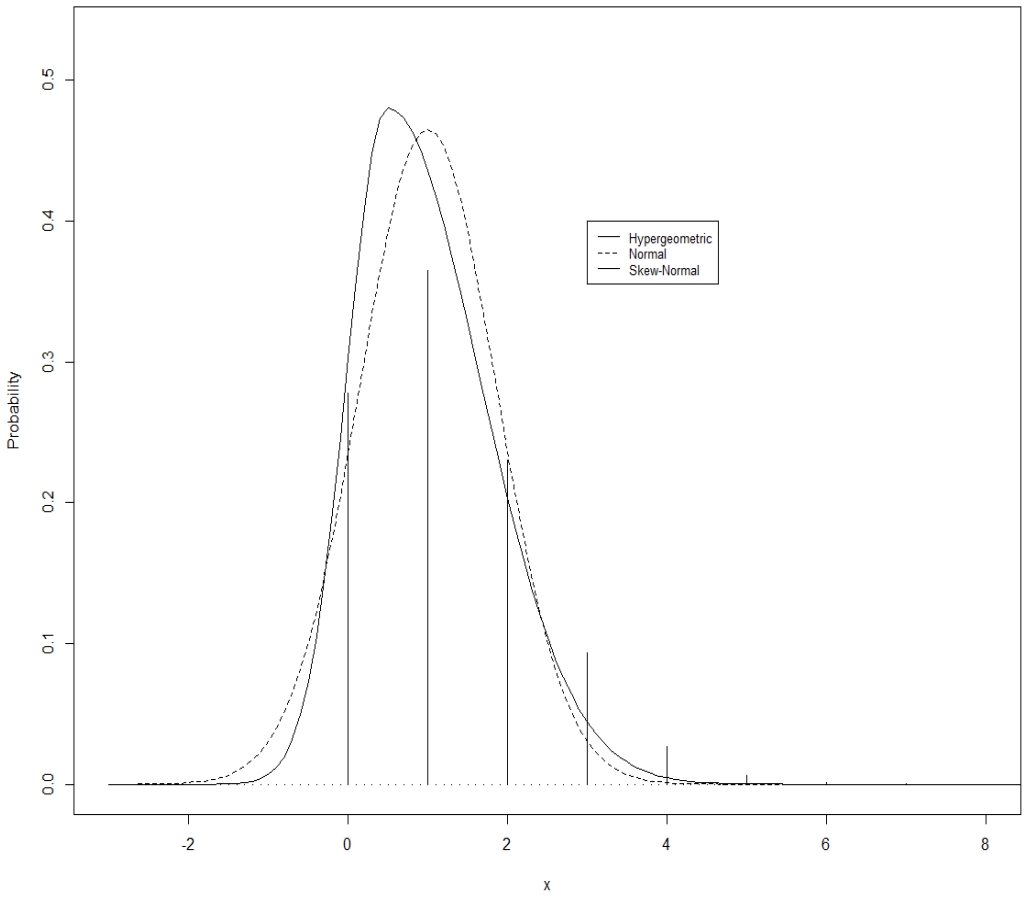
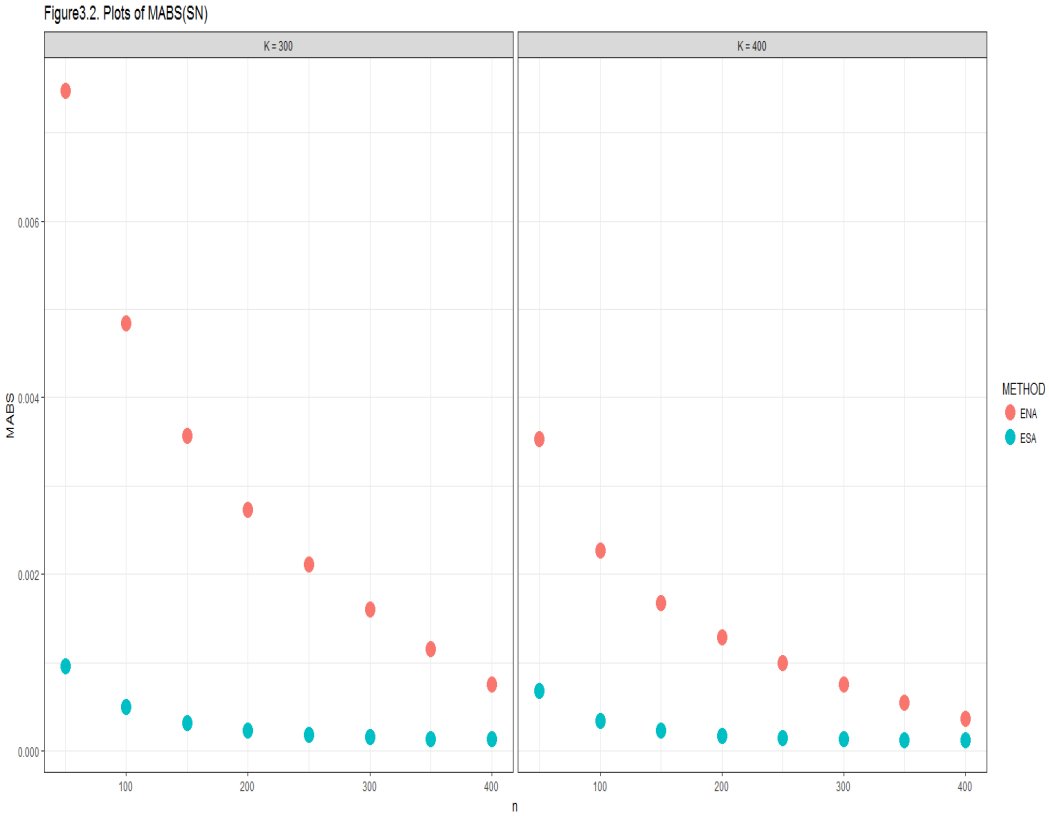


Figure 3.1 shows graph of the pmf of a hypergeometric distribution with parameters n=10, N=50 and K=5 together with the matching normal and skew-normal distribution. Since in general, it is not easy to verify which approximation is more accurate by simply looking on graphs like in this figure, we calculated the MABS for the normal and skew normal approximations for varying various of the hypergeometric parameters of n, K and N. A typical result is given in Figure 3.2.



From Figure 3.2, we can see that the MABS for the skew-normal approximation denoted by the blue dots is typically smaller than the MABS for the normal approximation denoted by the red dots. This figure shows the result of the approximations when N=1000, K=300, 400 and n = 50, 100, 150, 200, 250, 300, 350, 400. We repeated this comparisons for other values of n, K and N and in general we got the same results showing that the skew-normal approximation gives an improvement over the classical normal approximation in general.

4. Conclusion

In this paper, we obtained the formula needed to approximate the CDF of the hypergeometric distribution with parameters n, K and N and showed that the resulting skew-normal approximation is generally better than the classical normal approximation based on the Central Limit Theorem using the maximum absolute error as the criterion. It would be interesting to investigate how confidence interval estimation methods for the parameters of the hypergeometric distributions based on this skew-normal approximation perform compared with existing confidence interval estimation methods. Finally as in Chang et al (2008), we end this paper by noting that the methods used in this paper can be adopted to obtain skew-normal approximations to other discrete distributions.

References

Azzalini, A. (1985) A Class of Distribution that Includes the Normal Ones. Scandinavian Journal of Statistics, 12(2), 171-178.

Chang, C. H., Lin, J. J., Pal, N. and Chiang, M. C. (2008). A Note on Improved

Approximation of the Binomial Distribution by the Skew-Normal Distribution. The American Statistician, 62(2), 167-170.

Chang, C. H., Lin, J. J., Jou Rosemary (2010). A Note on Skew-Normal Distribution Approximation to the Negative Binomial Distribution. WSEAS Transactions On Mathematics, 9(1), 89-179.

Gupta, A.K., Nguyen, T. T. and Sanqui, J. A. T. (2004). Characterization of the Skew-Normal Distribution. Annals of the Institute of Statistical Mathematics, 56(2), 351-360.