Approximate Test for Comparing Parameters of Several Inverse Hypergeometric Distributions

Lei Zhang[1], Hongmei Han[2], Dachuan Zhang[3], and William D. Johnson[2]

1. Mississippi State Department of Health, Office of Health Data and Research, 570 East Woodrow Wilson, Jackson, Mississippi, 39215, USA
2. Pennington Biomedical Research Center, 6400 Perkins Rd, Baton Rouge, LA 70808, USA
3. Louisiana State University, 3357 Highland Road, Baton Rouge, LA 70802, USA

ABSTRACT

Consider the case of two-stage sampling where (first stage) m bins are selected from a large population of M bins each containing N items, (second stage) inverse subsampling is performed without replacement from each of the m bins, and a binary observation is made on each subsample item (inverse hypergeometric distribution). Denote the two types of items *Red* and *Blue*. Assuming N is known but the number of each type is not, we consider the hypothesis that the number of *Red* items is the same in all M bins. We employ an unbiased parameter estimator and use the Delta Method to approximate the estimator's variance. We then propose a large sample statistic for testing the hypothesis. We selected various parameter values for the inverse hypergeometric distribution to empirically investigate performance of the test in terms of exact calculations of the ability of the test to maintain significance at the nominal 0.05 level under the null hypothesis and inform power of the test for specified parameter values when the null hypothesis is false. The empirical findings provide pragmatic validation of the merits of the test.

Key Words: Empirical power, Inverse sampling, Large sample theory, Negative hypergeometric distribution.

1. INTRODUCTION

In a previous paper, we developed the unbiased estimator of the parameter of Inverse Hypergeometric Distribution (IHG) and then used the Delta method to develop approximations for the variance of this unbiased estimator. Here we present a large sample statistic for testing the hypothesis that the parameter has a specific value. We extent the testing to the case of two-stage sampling and evaluate its performance in terms of exact calculations of expected values for probability of Type 1 errors under the null hypothesis and power estimated under selected values of the parameter under the alternative hypothesis. We begin in Section 2 with an overview of the salient characteristics of the distribution.

2. THE INVERSE HYPERGEOMETRIC DISTRIBUTION

Consider a bin that contains a total of N balls where R balls are red, B balls are blue and $R + B = N$. Suppose we wish to select a random sample from the bin and observe the number of balls of each color in the selected sample. Our goal might be, for example, to estimate the number of red balls in the bin where N is known and R (hence, B) is not.

Suppose the balls are well mixed in the bin and a given trial of an "experiment" is as follows: we randomly select N balls, sampling one at a time without replacement, until we obtain a fixed number of red balls (successes), denoted as r, where $r \in \{1, 2, \ldots, R\}$. Let $X \in \{0, 1, \ldots, B\}$ denote the number of blue balls that must be drawn to get r red balls. Note that we stop selecting balls when the r^{th} red ball is chosen so that some permutation of $r - 1$ red balls and x blue balls will be chosen in the first $r + x - 1$ selections and the last ball drawn will always be red. Let A_1 be the event that $r - 1$ red balls are drawn in the first $r + x - 1$ trials and let A_2 be the event that the r^{th} red ball is drawn at the $(r + x)^{th}$ trial given that event A_1 has occurred. Now, the probability $X = x$ is $P(X = x) = P(A_1) \times P(A_2 \mid A_1)$ which can be expressed as

$$P(X = x) = \left[\frac{\binom{R}{r-1}\binom{N-R}{x}}{\binom{N}{r+x-1}} \right] \frac{R-r+1}{N-r-x+1} \quad , \quad \mathrm{x} \in \{0, 1, \ \ldots \ , \ N-R\}.$$

This expression represents the probability distribution function (pdf) for the random variable X. For given N, R and r, we refer to the non-zero probabilities determined by the pdf for all values in the domain of the random variable, together with the corresponding values of the random variable that occur with these non-zero probabilities, as the IHG distribution. IHG distributions are skewed to the left when $R < B$ and to right when $R > B$, but when N is large and R and B are approximately equal, the probability distributions are close to being bell-shaped and resemble normal distributions. The mean and variance of X are, respectively,

$$\mu_x = E(X) = \frac{rB}{R+1}$$

and,

$$\sigma_x^2 = V(X) = \frac{rB(R-r+1)(N+1)}{(R+2)(R+1)^2}$$

3. TESTING THE HYPOTHESIS H$_0$: $\boldsymbol{R = R_0}$

Let

$$\hat{R}_X = \left(\frac{r-1}{r+X-1}\right)$$

and note that

$$E(\hat{R}_X) = \sum_{x=0}^{x=B} P(X = x) \left(\frac{r-1}{r+X-1}\right) N = R$$

indicating that $\hat{R}_X$ is an unbiased estimator for R. Moreover,

$$V(\hat{R}_X) = \sum_{x=0}^{x=B} P(X = x) \left(\hat{R}_x - R\right)^2$$

Using the Delta method, an estimator for the variance of the unbiased estimator is given by

$$V\left(\hat{R}_u\right) \approx \frac{(r-1)^2 N^2 \; (R+1)^2 \; r \; (N-R) \; (N+1) \; (R-r+1)}{(R+2) \; (rN-R+r-1)^4}$$

and, since *R* is unknown we use the approximation

$$V\left(\hat{R}_u\right) \approx \frac{(r-1)^2 \; N^2 (\hat{R}_u+1)^2 r\left(N-\hat{R}_u\right)(N+1)(\hat{R}_u-r+1)}{(\hat{R}_u+2)(rN-\hat{R}_u+r-1)^4}$$

For large samples, the statistic

$$Z = \frac{\hat{R}_X - R_0}{\sqrt{V(\hat{R}_X)}}$$

has a distribution that is approximately Gaussian with mean = 0 and standard deviation = 1 and therefore can be used to test the null hypothesis

$$H_0\text{: } R = R_0$$

4. TESTING THE HYPOTHESIS H_0: $\mu_R = R_0$

Consider the case of two-stage sampling where (first stage) *m* bins are selected from a large population of *M* bins each containing *N* items, (second stage) inverse subsampling is performed without replacement from each of the *m* bins, and a binary observation is made on each subsample item (inverse hypergeometric distribution). Denote the two types of items *Red* and *Blue*. Assuming *N* is known but the number of each type is not, we consider the hypothesis that the number of Red items is the same in all *M* bins. Let X_h denote the blue balls selected before the r^{th} red ball is chosen from the h^{th} bin, where *h* = 1, 2, … *m*, ; further, let $\hat{R}_h$ and $\hat{V}_h$ denote sample estimates and variances based on the unbiased estimator of *R* and the approximate variance estimator, respectively, for the h^{th} bin. Let

$$\bar{R} = \frac{\sum_{h=1}^{m} \hat{R}_h}{m} \text{ and } V_{\bar{R}} = \left(1 - \frac{m}{M}\right) V\left(\frac{\sum_{h=1}^{m} \hat{R}_h}{m}\right)$$

represent the average estimate and variance, respectively, of the number of red balls in the *m* bins. If *M* is large, and $\frac{m}{M} \approx 0$, the variance of $\hat{R}$ is approximately

$$V_{\bar{R}} \approx V\left(\frac{\sum_{h=1}^{m} \hat{R}_h}{m}\right) = \frac{\sum_{h=1}^{m} V(\hat{R}_h)}{m^2} = \frac{\sum_{h=1}^{m} \sum_{x=0}^{B} P(X_h = x)\left(\hat{R}_h - R\right)^2}{m^2}.$$

When $m > 1,\ Z_m = \frac{\bar{R} - R_0}{\sqrt{V(\bar{R})}}$ would be superior to Z for testing the stated hypothesis.

5. EXACT CALCULATIONS USED TO EVALUATE TEST PERFORMANCE

We employ an unbiased estimator of R and use the Delta Method to approximate the estimator's variance. We then propose a large sample statistic for testing the hypothesis. Specifically, Formulas for variance $V(\hat{R}_X) = \sum_{x=0}^{x=B} P(X = x)\left(\hat{R}_x - R\right)^2$ and test statistics of the form $Z = \frac{\hat{R}_X - R_0}{\sqrt{V(\hat{R}_X)}}$ were used for exact calculations with SAS 9.3 and R 3.0 software. We selected various parameter values for the IHG to empirically investigate performance of the test in terms of exact calculations of the ability of the test to maintain significance at the nominal 0.05 level under the null hypothesis and inform power of the test for specified parameter values when the null hypothesis is false. In this paper, we let R = 10, 20, 50, 60, and 90; r = 5, 15, 15, 15, and 15, respectively. For a given random variable of X, when the absolute value of test statistics Z is greater than or equal to 1.96, the power of rejecting H_0 will be assigned to 1.0. Otherwise a value of 0 will be assigned. The overall power of rejecting H_0 will be the weighted average of production between $P(X = x)$ and each rejecting power (1.0 or 0) when x = 0, 1, 2, 3, …The exact calculations will be performed for M = 1, 2, and 3, where M represents the number of bins.

6. RESULTS FROM THE EXACT CALCULATIONS

The results from exact calculations were presented in Figures 1-5.

1. The power of rejecting *H*o is at or close to nominal 0.05 level when the null hypothesis is true (Figures 1-4).

2. The power of rejection is quite strong (approaches to 1.0) as departures from the null increase (Figures 1-5).

3. When R_0 is close to N (e.g., 100) and *H*o is true, the probability of rejecting this null hypothesis is not maintained desirably close 0.05 (Figure 5).

4. Overall the larger the number of bins, the bigger the power of rejection. However, the effect of the number of bins is minimized when the null hypothesis is true or R_0 is close to N.

Figure 1. Power of Rejecting Ho: *R = R_0* (*R = 10, r = 5*)

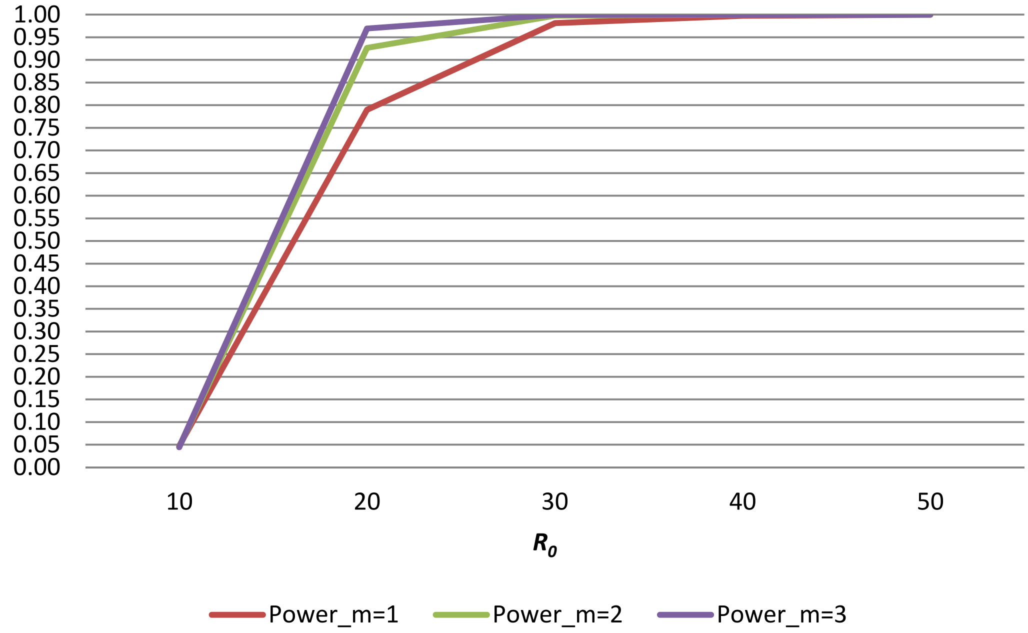


Figure 2. Power of Rejecting Ho: *R = R_0 (R = 20, r = 10*)

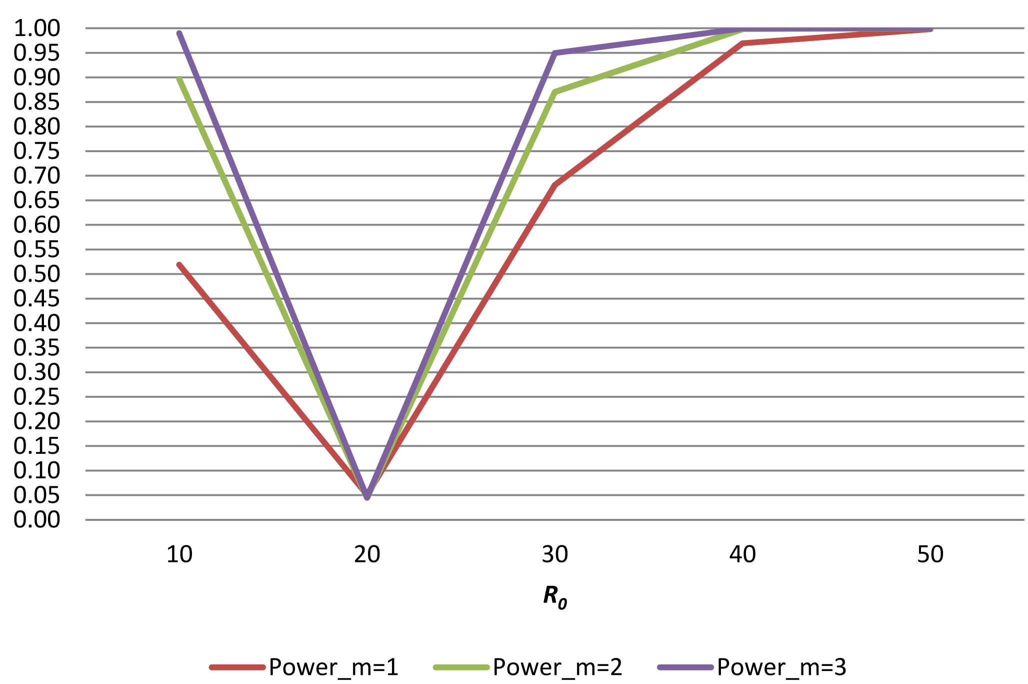


Figure 3. Power of Rejecting Ho: $R = R_0$ ($R = 50$, $r = 15$)

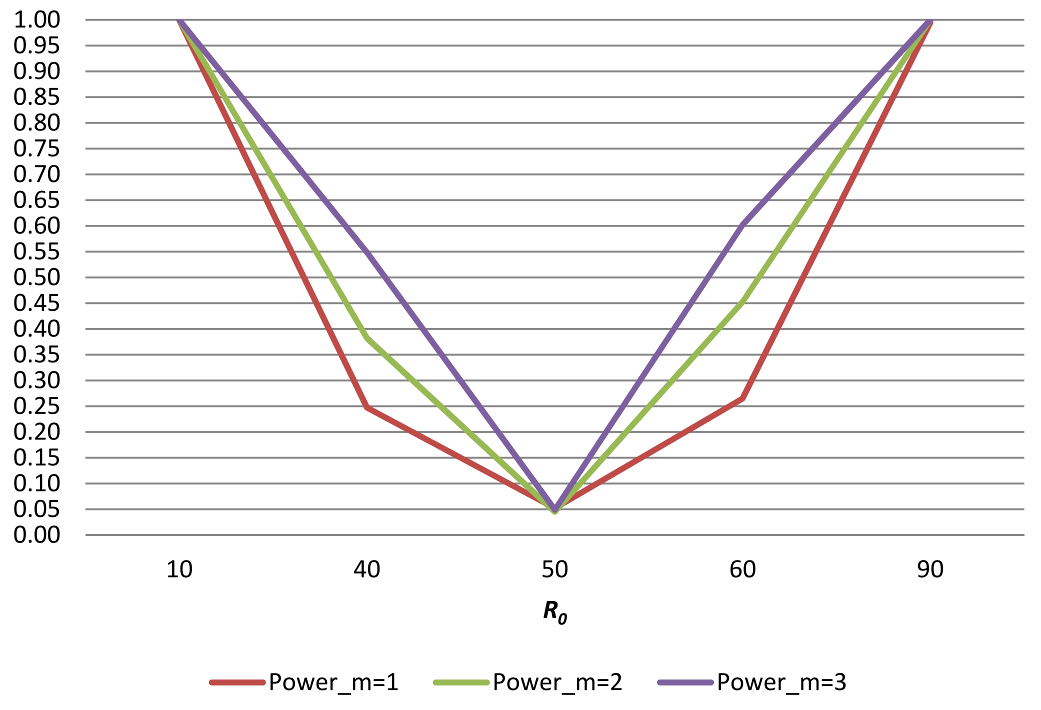
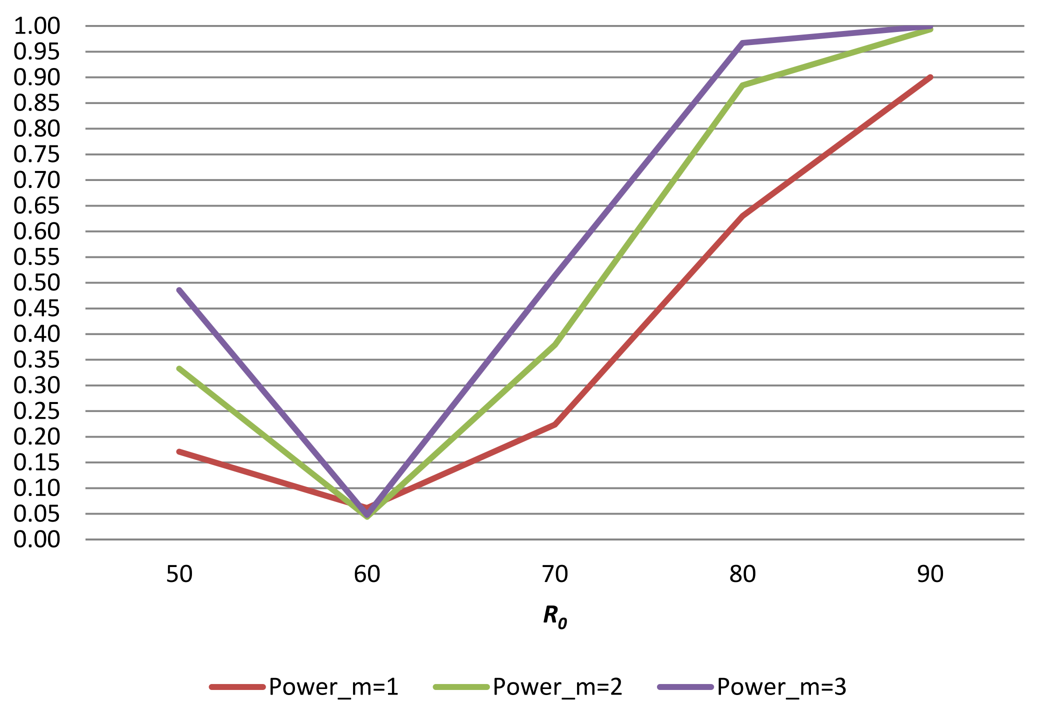
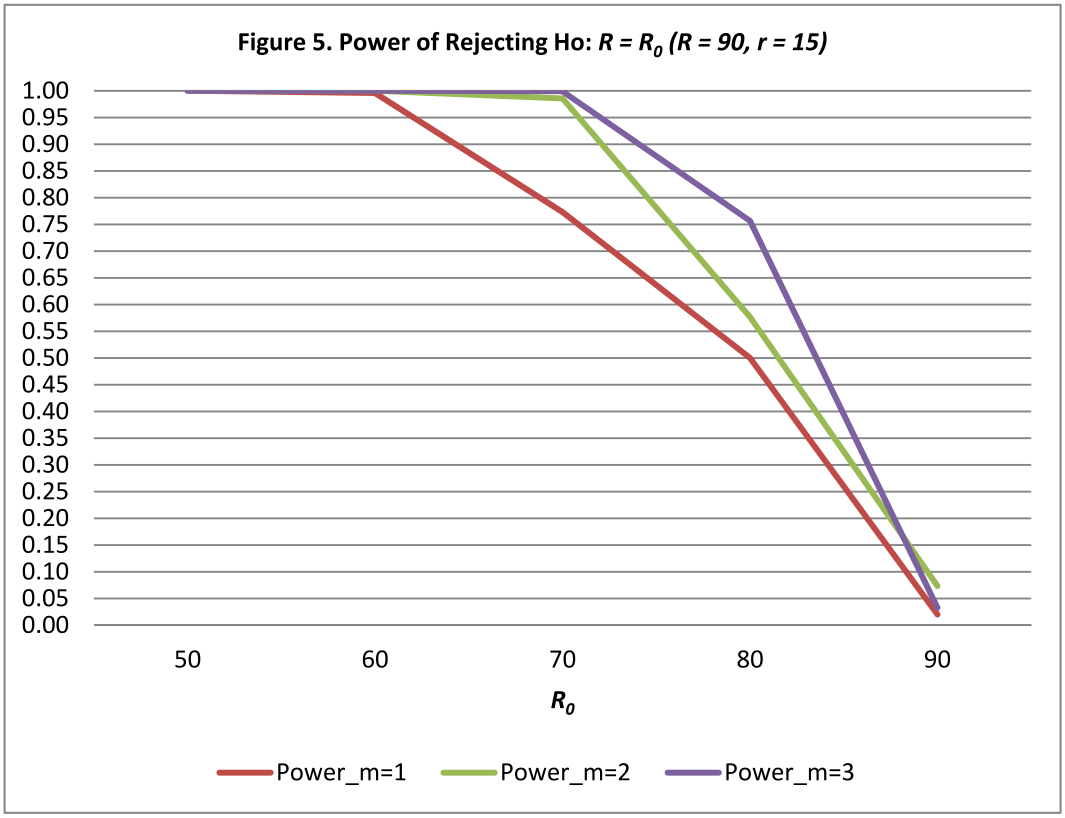


Figure 4. Power of Rejecting Ho: $R = R_0$ ($R = 60$, $r = 15$)





7. TESTING H_0: $\bar{R}_X = \bar{R}_Y$

Let M_X and M_Y denote the number of bins in two populations of bins. Suppose we select without replacement a random sample of m_X and m_Y, respectively, from each of the two populations for the purpose of testing the hypothesis that the number of red balls in the bins is the same in the two populations; i.e., H_0: $\bar{R}_X = \bar{R}_Y$ where

$$\bar{R}_X = \frac{\sum_{h=1}^{m_X} \hat{R}_h}{m_X} \text{ and } \bar{R}_Y = \frac{\sum_{h=1}^{m_Y} \hat{R}_h}{m_Y}$$

The proposed large sample test statistic is

$$Z = \frac{\bar{R}_X - \bar{R}_Y}{\sqrt{V(\bar{R}_X - \bar{R}_Y)}}$$

8. TESTING H_0: $\bar{R}_1 = \bar{R}_2 = \cdots = \bar{R}_K$

Let M_1, M_2, … , M_K denote the number of bins in K populations of bins. Suppose we select without replacement a random sample of $m_1, m_2, \dots, m_K$, respectively, from each of the K populations for the purpose of testing the hypothesis that the number of red balls in the bins is the same in the K populations.

Using formulas analogous to those given above, we can calculate $\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_K$ and $V(\bar{R}_1), V(\bar{R}_2), \ldots, V(\bar{R}_K)$, then construct the vector of means

$$\bar{\boldsymbol{R}} = \begin{bmatrix} \bar{R}_1 \\ \bar{R}_2 \\ \vdots \\ \bar{R}_K \end{bmatrix}$$

and the covariance matrix

$$\boldsymbol{V} = \begin{pmatrix} V(\bar{R}_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V(\bar{R}_K) \end{pmatrix}.$$

Consider hypotheses of the form H_0: $\boldsymbol{L}\bar{\boldsymbol{R}} = \boldsymbol{0}$, where is a $c \times K$ matrix, $c \leq K$, is a matrix of coefficients designed to form linear contrasts among the means. Let L^T and $\bar{R}^T$represent the transpose of the matrix $\boldsymbol{L}$ and vector $\bar{\boldsymbol{R}}$, respecvtively. Note that the variance of $\boldsymbol{L}\bar{\boldsymbol{R}}$ is

$V(\boldsymbol{L}\bar{\boldsymbol{R}}) = \boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^{\boldsymbol{T}}$. The statistic

$$X^2 = \bar{\boldsymbol{R}}^{\boldsymbol{T}}\boldsymbol{L}^{\boldsymbol{T}}\left(\boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^{\boldsymbol{T}}\right)^{-\boldsymbol{1}}\boldsymbol{L}\bar{\boldsymbol{R}}$$

is asymptotically distributed as Chi-square with c degrees of freedom where c is the rank of $\boldsymbol{L}$.

9. CONCLUDING REMARKS

In summary, we demonstrated the validity of a large sample test statistic for the simple case of testing the hypothesis that the parameter for a set of IHD is a fixed constant. The proposed test statistics performed well by maintaining significance at the nominal 0.05 level. The power of rejection quickly reached to 1.0 as departures from the null hypothesis. In addition, the number of bins will impact the power of rejection when departures from the null hypothesis. In Sections 7 and 8, we developed a large sample test for comparing parameters across several IHG distributions.

REFERENCES

Zhang, L., Xie, W., and Johnson, W.D. (2013). Performance of Interval Estimators for the Inverse Hypergeometric Distribution. *Communications in Statistics – Simulation and Computation* (Accepted).

D'Elia, A. (1999). A proposal for ranks statistical modeling. In: H. Friedl, A. Berghold and G. Kauermann (eds.). "Statistical Medelling. Proceedings of the 14th ISWM", University of Graz, Austria, 468-471.

D'Elia, A. (2001). A comparison between two asymptotic tests for analyzing preferences, *Quaderni di Statistica*, 127-143.

D'Elia, A. (2003). Modelling ranks using the Inverse Hypergeometric distribution, *Statistical Modelling: an International Journal*, 3, 65-78.

D'Elia, A. and Piccolo D. (2005). The moment estimator for the IHG distribution. In: C.Provasi (eds.), "Modelli complessi e metodi computazionali intensivi per la stima e la previsione", S.Co.2005, CLEUP, Padua, 245-250.

Grizzle, J.E, Starmer, C.F. and Koch, G,G. (1969). Analysis of categorical data by linear models. *Biometrics* 25:489-504.

Guenther, W.C. (1975). “The Inverse Hypergeometric – A Useful Model.” *Statistica Neerlandica*, 29: 129-144.

Johnson, N. L. and Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions.* Houghton Mifflin.

Johnson, N. L., Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions (3rd Edition).* John Wiley & Sons, Hoboken, New Jersey.

Miller, G.K. and Fridell, S.L. (2007). A forgotten discrete distribution? Reviving the negative hypergeometric model, *The American Statistician*, 61:347-350.

Moran, P.A.P. (1968). *An Introduction to Probability Theory*. Oxford, Great Britain.

Piccolo, D. (2002). Some approximations for the asymptotic variance of the maximum likelihood estimator of the parameter in the Inverse Hypergeometric random variable, *Quaderni di Statistica*, 4: 199-213.

Wilks, S. (1963). *Mathematical Statistics*, John Wiley & Sons, New York.
Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association,* 22:209-212.

Zelterman, D. (2004). *Discrete Distributions*. John Wiley & Sons, New York.

Submitting author

Lei Zhang, PhD, Office of Health Data and Research, Mississippi State Department of Health, 570 East Woodrow Wilson, Jackson, MS 39215, USA. Phone (601) 576-8165, E-mail: lei.zhang@msdh.ms.gov

Acknowledgements

Supported by 1 U54 GM104940 from the National Institute of General Medical Sciences of the NIH, which funds the Louisiana Clinical and Translational Science Center