INTERNATIONAL RESEARCHERS

CONSERVATIVE CONFIDENCE INTERVALS FOR THE HYPERGEOMETRIC DISTRIBUTION USING R

Margot Tollefson

Volume No.4 Issue No.3 September 2015

www.iresearcher.org

ISSN 2227-7471

THE INTERNATIONAL RESEARCH JOURNAL "INTERNATIONAL RESEACHERS"

www.iresearcher.org

INTERNATIONAL RESEARCHERS is peer-reviewed, supported by rigorous processes of criterion-referenced article ranking and qualitative commentary, ensuring that only intellectual work of the greatest substance and highest significance is published.

INTERNATIONAL RESEARCHERS is indexed in wellknown indexing diectories

INDEX COPERNICUS INTERNATIONAL JOURNAL MASTER LIST **with ICV value 5.90**

ULRICHSWEB GLOBAL SERIALS DIRECTORY

Journal Seek

Directory of Research Journals Indexing

and moniter by

THOMSON REUTERS

CONSERVATIVE CONFIDENCE INTERVALS FOR THE HYPERGEOMETRIC DISTRIBUTION USING R

Margot Tollefson, Ph.D.

Vanward Statistics, PO Box 286, Stratford, Iowa 50249-0286

(UNITED STATES OF AMERICA)

margot@vanwardstat.com

ABSTRACT

In this paper, a method is given for finding conservative confidence intervals for the number within a population with an attribute. The sampling is assumed to be such that the hypergeometric distribution holds. Before describing the method for the hypergeometric distribution, a method to find confidence intervals for the median of the lognormal distribution is given. The method for the log normal distribution is then used to justify the method for the hypergeometric distribution. In the paper, first confidence and conservative confidence intervals are defined with regard to the probability that the parameter being estimated will be in the interval. Then two sided, symmetric confidence intervals for the lognormal distribution are found. This is followed by a method to find conservative confidence intervals for the number with the attribute in the population, using the hypergeometric distribution. Then an example and justification are given for the method. Next an R function to find the intervals is given, along with an example of output from the function. The paper ends with some conclusions.

Keywords: conservative confidence interval, hypergeometric distribution, lognormal distribution

1. INTRODUCTION

The objective of this paper is to make available a simple method to find confidence intervals for the hypergeometric distribution, which does not exist at this time. A lesser objective is to demonstrate the derivation of confidence intervals for asymmetric distributions. The paper is, also, intended to demonstrate the use of probability in the construction of confidence intervals, both for continuous, asymmetric distributions and discrete distributions with a discrete parameter space.

A method is given for finding a conservative confidence interval for the number of members with a given attribute within a population. The hypergeometric distribution is used to model the problem. An R function is given which finds conservative confidence intervals for the number of members of a population with an attribute, where the sample and population sizes are known. The R function gives intervals for all possible outcomes from the sample.

First, definitions of confidence intervals and conservative confidence intervals are given. Next, finding a confidence interval for an unsymmetrical, continuous distribution is described and justified, using the lognormal distribution. Third, the hypergeometric distribution is described. Fourth, a method for finding a conservative confidence interval for the number of members of a population with an attribute, using the hypergeometric distribution, is presented. Fifth, an example is given, along with a justification of the method for finding the conservative confidence intervals. Sixth, an R function to find the conservative confidence intervals is given, along with a printout from the function. Last, there is a discussion of the results.

2. DEFINITIONS OF CONFIDENCE INTERVALS AND CONSERVATIVE CONFIDENCE INTERVALS

When estimating a parameter of a distribution, the uncertainty in the value of the estimate is of interest. A confidence interval gives an idea of how accurate an estimator is. A confidence interval consists of an interval between a lower and an upper limit. The interval always contains the estimator of the parameter. The lower and upper limits of a confidence interval are random variables, that is, the limits depend on the data and are not a function of the parameter being estimated.

Let θ be the parameter of interest, let X be the random variable to be used to find the lower and upper limits of the confidence interval, let L(X) be the lower limit, let U(X) be the upper limit and let α be the value such that (1 - α) 100% is the level of confidence. Then,

$$P(L(X) \leq \theta \leq U(X)) = (1 - \alpha) \quad (2.1)$$

for any correctly specified L(X) and U(X).

Note here that the L(X) and U(X) in equation (2.1) are unrealized random variables. The probability only holds before the experiment is run. After the experiment has been run, the interval either contains the parameter or the interval does not. After the experiment, the statistician uses the word confidence to describe how certain the statistician is that the interval contains the parameter. Under repeated sampling, correctly specified confidence intervals will contain the parameter an average of (1 - α) 100% of the time.

For continuous distributions with a continuous parameter space, one should be able to find an infinity of pairs, L(X) and U(X), that satisfy equation (2.1), since, given the distribution and either of the limits, the other limit, with the correct amount of probability, can be found. For discrete distributions with a discrete parameter space, there usually is no pair, since both the parameters and the level of probability change in discrete jumps. For discrete distributions, replacing the equal sign in equation (2.1) with a greater than or equal to sign will give a conservative confidence interval; that is, the confidence will be at least (1 - α) 100% . See Finkelstein, Tucker, and Veeh (2000).

There are three common methods of forming confidence intervals; two sided, symmetric intervals (here, the symmetry refers to probability) and two types of one sided intervals. For continuous distributions with continuous parameter spaces, two sided, symmetric intervals are formed so that

$$P(L(X) > \theta) = \frac{\alpha}{2} \quad (2.2)$$

And

$$P(U(X) < \theta) = \frac{\alpha}{2}. \quad (2.3)$$

The two types of one sided intervals are the right tailed interval and the left tailed interval. For the right tailed interval, equation (2.2) holds with α/2 replaced by α and U(X) set equal to the maximum value of θ (possibly infinity). For the left tailed interval, equation (2.3) holds with α/2 replaced by α and L(X) set equal to the minimum value of θ (possibly minus infinity). The actual intervals are found by replacing X in L(X) and U(X) with the realization of X from the data.

See Agresti (2013, Chapter 1, section 4.4 and Chapter 16, section 6), Tang, He, and Tu (2012, section 2.1.1.2) and Tufféry (2011, sections A.2.5 and A.2.6) for treatments of confidence intervals and the concept of conservative in statistics.

3. CONFIDENCE INTERVALS FOR CONTINUOUS DISTRIBUTIONS DEMONSTRATED WITH THE LOGNORMAL DISTRIBUTION

Confidence intervals for continuous distributions with a continuous parameter space are straightforward to find. The median of the lognormal distribution will be used in the following description of finding a confidence interval. Given an estimator of the median of a lognormal distribution, the distribution of the estimator can be used to form a confidence interval for the value of the median.

A lognormal distribution is a continuous probability distribution defined on the interval (0 , ∞) . Forbes, Evans, Hastings and Peacock, (2010, Chapter 29)., give the properties of the lognormal distribution described here. For a random variable distributed with the lognormal distribution, the log of the variable is distributed normally. Three parameters associated with the lognormal distribution are the median of the distribution (m), the mean of the normal distribution associated with the distribution (μ), and the variance of the normal distribution associated with the distribution (σ^2). In what follows, we assume that σ^2 is known. The distribution has the property that

$$\mu = \log(m). \quad (3.1)$$

The probability distribution for the lognormal is

$$f(x) = \frac{1}{x\sigma(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{\left(\log\left(\frac{x}{m}\right)\right)^2}{2\sigma^2}\right), \quad (3.2)$$

where zero is less than x is less than infinity. Given a sample, Forbes et al (2010), give the estimator of the median m as the geometric mean of the observations,

$$\hat{m} = [\prod_{i=1}^{n} x_i]^{\frac{1}{n}}, \quad (3.3)$$

where the x_i, i = 1 . . . n, are n independent realizations from the lognormal distribution. The estimator, $\hat{m}$, has a lognormal distribution, with the median equal to m and the mean and variance of the associated normal distribution equal to μ and σ^2/n, respectively. Let M be the random variable that estimates m. Then, $\hat{m}$ is a realization of M.

For the symmetric, two sided confidence interval, the limits of the confidence interval have the property that

$$P(L(M) > m) = \frac{\alpha}{2} \quad (3.4)$$

and

$$P(U(M) < m) = \frac{\alpha}{2}, \quad (3.5)$$

from which follows

$$P\big(L(M) \leq \theta \leq U(M)\big) = (1 - \alpha). \quad (3.6)$$

The following steps give a result from which the lower and upper limits can be found. To find L(M) and U(M), taking the log of the elements within the probability statement is useful. The log transformation is one to one and monotonic, so taking logs will not affect the associated probabilities in equation (3.6). Taking the log transformation gives

$$P(\log(L(M)) \leq \log(m) \leq \log(U(M)) = (1 - \alpha). \quad (3.7)$$

Multiplying the elements of the probability statement in equation (3.7) by minus one, which reverses the inequalities, gives

$$P(-\log(U(M)) \leq -\log(m) \leq -\log(L(M)) = (1 - \alpha). \quad (3.8)$$

Adding log (M) to the elements of the probability statement in equation (3.8) gives

$$P(\log(M) - \log\big(L(M)\big) \leq \log(M) - \log(m) \leq \log(M) - \log(U(M)) = (1 - \alpha). \quad (3.9)$$

Dividing the elements of the probability statement in equation (3.9) by $\sigma\sqrt{n}$ gives

$$P\left(\frac{\log(M) - \log(U(M))}{\sigma^2/\sqrt{n}}\right) \leq \frac{\log(M) - \log(m)}{\sigma^2/\sqrt{n}} \leq \frac{\log(M) - \log(L(M))}{\sigma^2/\sqrt{n}}\Big) = (1 - \alpha). \quad (3.10)$$

But, the second element of equation (3.10) is

$$\frac{\log(M) - \log(m)}{\sigma^2/\sqrt{n}} \quad (3.11)$$

which is a standard normal variate. Since equation (3.11) is a standard normal variate and equations (3.4) and (3.5) hold, the first element in equation (3.10) equals $z_{\alpha/2}$, that is

$$\frac{\log(M) - \log(U(M))}{\sigma^2/\sqrt{n}} = z_{\frac{\alpha}{2}}, \tag{3.12}$$

and the third element of equation (3.10) equals $z_{(1-\frac{\alpha}{2})}$, that is

$$\frac{\log(M) - log(L(M))}{\sigma^2/\sqrt{n}} = z_{\left(1-\frac{\alpha}{2}\right)}, \tag{3.13}$$

where z_β is the βth percentile of the standard normal distribution. Then,

$$\log(L(M)) = \log(M) - \frac{\sigma}{\sqrt{n}} z_{(1-\frac{\alpha}{2})} \tag{3.14}$$

and

$$\log(U(M)) = \log(M) - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}, \tag{3.15}$$

so

$$L(M) = M\exp\left(-\frac{\sigma}{\sqrt{n}} z_{(1-\frac{\alpha}{2})}\right) \tag{3.16}$$

and

$$U(M) = M\exp\left(-\frac{\sigma}{\sqrt{n}} z_{(\frac{\alpha}{2})}\right). \tag{3.17}$$

If M is replaced by $\hat{m}$ in equations (3.16) and (3.17), the lower and upper limits of the confidence interval result.

Note that, if the lower limit is assumed to be the actual median, the probability of seeing an M greater than the observed $\hat{m}$ is α / 2. To see the result, let $L(m_L)$ be the lower limit, that is,

$$L(m_L) = L(M). \tag{3.18}$$

The value for which the area to the right under the probability distribution equals α/2 for a normal distribution with mean equal to $\log(L(m_L))$ and variance equal to σ^2/n is

$$\log(L(m_L)) + \frac{\sigma}{\sqrt{n}} z_{(1-\frac{\alpha}{2})}. \tag{3.19}$$

But, from equations (3.14) and (3.18), substituting $\hat{m}$ for M in equation (3.14),

$$\log(L(m_L)) = \log(\hat{m}) - \frac{\sigma}{\sqrt{n}} z_{\left(1-\frac{\alpha}{2}\right)}. \tag{3.20}$$

The result follows, as can be seen by adding $\frac{\sigma}{\sqrt{n}} z_{\left(1-\frac{\alpha}{2}\right)}$ to both sides of equation (3.20).

Similarly, if the upper limit is assumed to be the actual median, the probability of seeing an M less than the observed $\hat{m}$ is α / 2. See Figure 1 for an example in which the observed $\hat{m}$ equals five, σ equals one, n equals nine, and α equals 0.05

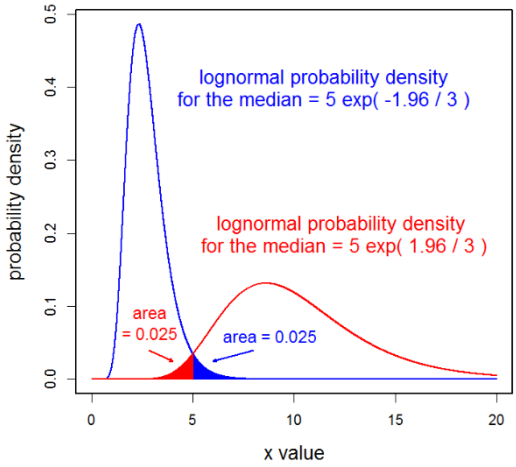

Fig. 1. *Lognormal distributions setting the median equal to the lower and upper limits of the confidence interval:* $\hat{m} = 5$, $\sigma = 1$, $n = 9$, *confidence = 95%.*

Also, note that the level of confidence for any M in the interval, given that the M is the true value of the median, does not depend on the value of M. We will see that for the hypergeometric distribution, interval size can vary with the value of the parameter in the interval.

See Lehmann (1986, pp. 89-96) for a discussion of the theoretical basis for confidence intervals and the interrelationship of the parameter space with the sample space.

4. THE HYPERGEOMETRIC DISTRIBUTION

The hypergeometric distribution is a finite, discrete distribution with a finite, discrete parameter space. The hypergeometric distribution models random samples (without replacement) from finite populations for which the members of the population either exhibit an attribute or do not exhibit an attribute. An example might be a classroom in which the students are either girls or boys. The attribute is then one or the other of the sexes. See Forbes et al. (2010, Chapter 24), for the properties given below of the hypergeometric distribution.

There are three parameters for a hypergeometric distribution, two of which are usually know. The parameters are the size of the population (N), the number of members of the population that exhibit the attribute (M), and the size of the sample (n). Assume that N and n are known in the following. Define the random variable, X, as the number of members in the sample that exhibit the attribute. The random variable X can take on integer values from the maximum of zero and (n-N+M) to the minimum of M and n, inclusive. A realization of X is denoted as x.

The probability density distribution of the hypergeometric distribution, where we assume that the sample is selected randomly without replacement from the population, is

$$f(X) = \frac{\binom{M}{X}\binom{N-M}{n-X}}{\binom{N}{n}}, \qquad (4.1)$$

where X can take on the values given above. Since both the distribution and the parameter space are discrete and finite, given N and n, the probabilities of observing each possible X for each possible M can be enumerated.

The estimator of M, using the minimum variance, unbiased method, is the maximum integer less than or equal to Nx/n {Forbes et al., Chapter 24}. The estimator is not used in the following derivation of the conservative confidence intervals.

5. CONSERVATIVE CONFIDENCE INTERVALS FOR THE HYPERGEOMETRIC DISTRIBUTION

Let the conservative confidence interval for the parameter, M, of the geometric distribution have the property that

$$P(L(X) > M) \leq \frac{\alpha}{2} \tag{5.1}$$

and

$$P(U(X) < M) \leq \frac{\alpha}{2}, \tag{5.2}$$

for all M's in the confidence interval. Then,

$$P(\, L(X) \leq M \leq U(X)\,) \geq (1-\alpha) \tag{5.3}$$

for each M in the confidence interval. Here, we are assuming that for each M, for finding the probability, the value of M is the actual value of the median. Note that the actual probabilities in equations (5.1), (5.2), and (5.3) will vary with the value of M, unlike in the continuous case.

To find the lower side of a conservative confidence interval for M, one approach would be to find the largest L(x) such that the probability of observing a value of X larger than the observed value, x, is less than or equal to α/2 given M equals L(x). Similarly, U(X) can be found by finding the smallest U(x) such that the probability of observing a smaller value of X than the observed value, x, is less than or equal to α/2 given M equals U(x). This approach is the same as the approach for the lognormal distribution shown in Figure 1.

Cochran (1977, Chapter 3, section 6) follows a similar approach. Cochran (1977) defines conservative confidence interval limits for the number in the population with the attribute M, using the hypergeometric distribution, as the L(x) and U(x) for which

$$\sum_{i=x}^{n} \frac{\binom{L(x)}{i}\binom{N-L(x)}{n-i}}{\binom{N}{n}} \leq \frac{\alpha}{2} \tag{5.4}$$

and

$$\sum_{i=0}^{x} \frac{\binom{U(x)}{i}\binom{N-U(x)}{n-i}}{\binom{N}{n}} \leq \frac{\alpha}{2}, \tag{5.5}$$

where x is the realization of X found from the sample. Here, Cochran's (1977) notation is adapted to this article. Note that Cochran (1977) uses zero and n for the limits in the two summations rather than the maximum of zero and (n-N+M) and the minimum of M and n.

For a given x, Cochran's (1977) intervals can include more M's than are necessary for each M in the confidence interval to satisfy equations (5.1) and (5.2), since x is included in finding the probability level in the tails of the probability distributions. This author has found that one way to get smaller confidence intervals is to replace L(x) by (L(x) - 1) in equation (5.4) and U(x) by (U(x) + 1) in equation (5.5). The correction will be justified below. Equations (5.4) and (5.5) then become,

$$\sum_{i=x}^{\min(M,n)} \frac{\binom{L(x)-1}{i}\binom{N-(L(x)-1)}{n-i}}{\binom{N}{n}} \leq \frac{\alpha}{2} \tag{5.6}$$

and

$$\sum_{i=\max(0,(n-N+M))}^{x} \frac{\binom{U(x)+1}{i}\binom{N-(U(x)+1)}{n-i}}{\binom{N}{n}} \leq \frac{\alpha}{2}, \tag{5.7}$$

where L(X) is the largest possible L(X) satisfying equation (5.6) and U(X) is the smallest U(X) satisfying equation (5.7).

For values of x and M close to or equal to their minimum values, finding an L(x) such that equation (5.6) holds can be impossible. In this case, this author's solution is to find a one-sided interval by setting L(x) equal to the lowest possible value of L(x) and finding the U(x) which satisfies equation (5.7). Similarly, for values of x and M close to or equal to their maximum values, finding an U(x) such that equation (5.7) holds can be impossible. In this case, this author's solution is to find a one-sided interval by setting U(x) equal to the largest possible value of U(x) and finding the L(x) which satisfies equation (5.6). Using α rather than α/2 in equations (5.6) and (5.7) for one-sided intervals is also a solution. However, this author has observed that in this case the minimum value for the confidence can be less than (1-α) 100%, with a lower bound of (1-2α) 100%.

Sometimes neither equation (5.6) or (5.7) can be satisfied, in which case all possible values of M are included in the confidence interval, as will be seen as part of the example.

If M equals the minimum value, then the probability that X equals the minimum value equals one. Similarly, if M equals the maximum value, then the probability that X equals the maximum value is one. In either case, a one sided interval will be chosen since one will always be larger than α/2. As a result, there is no conflict with (L(X)-1) being less than zero or (U(X)+1) being greater than N.

6. AN EXAMPLE

In this section, an example of conservative confidence intervals for a hypergeometric distribution is given, where the population size, N, equals sixteen and the sample size, n, equals six. The value of α is set to 0.05. Table 1 contains the complete enumeration of the probabilities associated with the possible values of the number in the sample with the attribute, X, and the number in the population with the attribute, M. The columns contain the probabilities associated with the seven possible values of X and the rows contain the probabilities associated the seventeen possible values of M. Each row is the probability distribution for X given M equals the M of the row. If a combination is impossible, say M equals three and x equals four, an 'NA' is listed in the table. Note that probabilities only have meaning across rows.

Table 1. Probability densities for the hypergeometric distribution with N=16 and n=6. Each row contains the density for the corresponding value of M. The conservative confidence intervals for each x are highlighted in grey.

M \ x	0	1	2	3	4	5	6
0	1.000	NA	NA	NA	NA	NA	NA
1	0.625	0.375	NA	NA	NA	NA	NA
2	0.375	0.500	0.125	NA	NA	NA	NA
3	0.214	0.482	0.268	0.036	NA	NA	NA
4	0.115	0.396	0.371	0.110	0.008	NA	NA
5	0.058	0.288	0.412	0.206	0.034	0.001	NA
6	0.026	0.189	0.393	0.300	0.084	0.007	0.000
7	0.010	0.110	0.330	0.367	0.157	0.024	0.001
8	0.003	0.056	0.245	0.392	0.245	0.056	0.003
9	0.001	0.024	0.157	0.367	0.330	0.110	0.010

10	0.000	0.007	0.084	0.300	0.393	0.189	0.026
11	NA	0.001	0.034	0.206	0.412	0.288	0.058
12	NA	NA	0.008	0.110	0.371	0.396	0.115
13	NA	NA	NA	0.036	0.268	0.482	0.214
14	NA	NA	NA	NA	0.125	0.500	0.375
15	NA	NA	NA	NA	NA	0.375	0.625
16	NA	NA	NA	NA	NA	NA	1.000

The elements of Table 1 highlighted in grey are the confidence intervals given that the x of the column has been observed. The confidence intervals are read down the rows. In a given column, the M's associated with the grey probabilities are those values of M for which equations (5.1) and (5.2) hold given the value of x; or there is no M for which either equation (5.1) or (5.2) holds, but the other equation holds; or there is no M for which either equation holds. In the example, if x equals zero, the 95% conservative confidence interval for M is (0,6). The value for L(0) is the lower limit, zero, since the probability that x equals zero when M equals zero is one. U(0) was found using equation (5.7).

For a given M, the level of confidence is $P(L(X) \leq M \leq U(X))$, which is the probability that the M falls in at least one of the confidence intervals. In Table 1, the confidence level for a given M is the sum over the grey probabilities within the row associated with the M. To see that the intervals obey equations (5.1), (5.2), and (5.3), for each M, look at the row associated with M. Say M equals three. Then X can take on the value zero, one, two, or three. L(0) equals zero; L(1) equals one; L(2) equals two; and L(3) equals three. For each of the possible values of X, $P(L(X) \leq 3)$ is one.

Looking at U(X), U(0) equals six; U(1) equals eight; U(2) equals eleven; and U(3) equals thirteen. For each of the possible values, $P(3 \leq U(X))$ is one. This means that the probability in equation (5.3) is one. It follows that the probabilities in equations (5.1) and (5.2) are zero. As a result, equations (5.1) and (5.2) are satisfied, since zero is less than $\alpha/2$ and equation (5.3) is satisfied since one is greater than $(1-\alpha)$.

In Table 2 are the probabilities x falls to the left of the intervals, within the intervals, and to the right of the intervals, for each of the values of M as well as the level of confidence for each of the M's. The probabilities are the values of equations (5.1), (5.3), and (5.2). By inspection, each interval satisfies equations (5.1), (5.2), and (5.3).

Table 2. Probabilities of the x being to the left, within, and to the right of the conservative confidence intervals, along with the level of confidence, for each M.

M	left tail	middle	right tail	confidence
0	0.000	1.000	0.000	100.0%
1	0.000	1.000	0.000	100.0%
2	0.000	1.000	0.000	100.0%
3	0.000	1.000	0.000	100.0%
4	0.000	0.992	0.008	99.2%
5	0.000	0.999	0.001	99.9%

6	0.000	0.992	0.008	99.2%
7	0.010	0.965	0.024	96.5%
8	0.003	0.993	0.003	99.3%
9	0.024	0.965	0.010	96.5%
10	0.008	0.992	0.000	99.2%
11	0.001	0.999	0.000	99.9%
12	0.008	0.992	0.000	99.2%
13	0.000	1.000	0.000	100.0%
14	0.000	1.000	0.000	100.0%
15	0.000	1.000	0.000	100.0%
16	0.000	1.000	0.000	100.0%

From Table 2, for some of the M's the confidence that the confidence interval will contain M is 100%. To see why, again look at M equal to three. From Table 1, X can only take on the values zero through three and the probability of observing each x is greater than α/2 for all four possible outcomes of X. Therefore, if M equals three, M will always be in one of the four possible intervals. The same logic holds for M equal to the other values for which the confidence equals 100%.

In the previous section, the justification for using equations (5.6) and (5.7) instead of equations (5.4) and (5.5) was deferred until later. Tables 3 and 4 will be used to explain the justification. Table 3 gives the probability of observing an X greater than or equal x for each possible value of x and M. Table 4 gives the probability of observing an X less than or equal to x for each possible value of x and M. Table 3 is used to find the L(X)'s and Table 4 is used to find the U(X)'s. In both Tables 3 and 4, the conservative confidence intervals are highlighted in grey.

Table 3. Probabilities of observing an X greater than equal to x for the given values of M. Conservative confidence intervals are highlighted in grey.

M \ x	0	1	2	3	4	5	6
0	1.000	0.000	0.000	0.000	0.000	0.000	0.000
1	1.000	0.375	0.000	0.000	0.000	0.000	0.000
2	1.000	0.625	0.125	0.000	0.000	0.000	0.000
3	1.000	0.786	0.304	0.036	0.000	0.000	0.000
4	1.000	0.885	0.489	0.118	0.008	0.000	0.000
5	1.000	0.942	0.654	0.242	0.036	0.001	0.000
6	1.000	0.974	0.785	0.392	0.092	0.008	0.000
7	1.000	0.990	0.879	0.549	0.182	0.024	0.001

8	1.000	0.997	0.941	0.696	0.304	0.059	0.003
9	1.000	0.999	0.976	0.818	0.451	0.121	0.010
10	1.000	1.000	0.992	0.908	0.608	0.215	0.026
11	1.000	1.000	0.999	0.964	0.758	0.346	0.058
12	1.000	1.000	1.000	0.992	0.882	0.511	0.115
13	1.000	1.000	1.000	1.000	0.964	0.696	0.214
14	1.000	1.000	1.000	1.000	1.000	0.875	0.375
15	1.000	1.000	1.000	1.000	1.000	1.000	0.625
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Equation (5.6) states that the sum, from i equal x to the minimum of n and M, of the probabilities of observing i, given M equals (L(x)-1), should be less than or equal to α/2. For the example, α/2 equals 0.025. Table 3 contains the cumulative probabilities across rows from right to left, which is from n to x.

Look at the columns of Table 3, ignoring the first column, which will always be a column of ones. In each column, the cumulative probabilities never decrease and, at some point, increase down the columns until one is reached. The rule in equation (5.6) states that, for a given x, the row of (L(x)-1) must have the cumulative probability less than or equal to α/2 in the x column. But, since the cumulative probabilities increase down the column and (L(x)-1) in the largest M for which equation (5.6) holds, the cumulative probability in the row of L(x) and the column of x is greater than α/2.

Look left to right along the diagonals of Table 3. For any diagonal whose elements are not identically equal to one or zero, the cumulative probabilities never increase and, at some point, decrease in value as x and M increase. For example, for the diagonal starting at x equal to zero and M equal to nine, the diagonal elements are 1.000, 1.000, 0.999, 0.992, 0.964, 0.875, and 0.625. Having found the maximum (L(x)-1) and using the fact that the cumulative probabilities decrease along the diagonal, the cumulative probability to the right of column x in the row of L(x) will be always be less than α/2, which satisfies equation (5.1).

Table 4 is used to find U(x) for each observed x and contains the cumulative probabilities across rows from left to right. Ignoring the last column, which is a column of ones, the elements of the columns in Table 4 never increase and, at some point, decrease from one to zero as M increases.

From equation (5.7), the sum, from i equal the maximum of zero and n minus N plus M to x, of the probabilities of observing i, given M equals (U(x)+1), should be less than or equal to 0.025. So, (U(x)+1) is the smallest M for which the cumulative probability in the row of M and the column of x is less than or equal to 0.025.

Table 4. Probabilities of observing an X less than equal to x for the given values of M. Conservative confidence intervals are highlighted in grey.

M \ X	0	1	2	3	4	5	6
0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1	0.625	1.000	1.000	1.000	1.000	1.000	1.000
2	0.375	0.875	1.000	1.000	1.000	1.000	1.000
3	0.214	0.696	0.964	1.000	1.000	1.000	1.000

4	0.115	0.511	0.882	0.992	1.000	1.000	1.000
5	0.058	0.346	0.758	0.964	0.999	1.000	1.000
6	0.026	0.215	0.608	0.908	0.992	1.000	1.000
7	0.010	0.121	0.451	0.818	0.976	0.999	1.000
8	0.003	0.059	0.304	0.696	0.941	0.997	1.000
9	0.001	0.024	0.182	0.549	0.879	0.990	1.000
10	0.000	0.008	0.092	0.392	0.785	0.974	1.000
11	0.000	0.001	0.036	0.242	0.654	0.942	1.000
12	0.000	0.000	0.008	0.118	0.489	0.885	1.000
13	0.000	0.000	0.000	0.036	0.304	0.786	1.000
14	0.000	0.000	0.000	0.000	0.125	0.625	1.000
15	0.000	0.000	0.000	0.000	0.000	0.375	1.000
16	0.000	0.000	0.000	0.000	0.000	0.000	1.000

For Table 4, the diagonal terms which are not identically equal to zero or one, increase to one as x and M increase. For example, for the diagonal starting with x equals to zero and M equal to four, the diagonal elements are 0.115, 0.346, 0.608, 0.818, 0.941, 0.990, and 1.000. Therefore having found (U(x)+1) and using the fact that diagonal elements decrease as M and x decrease, the cumulative probability to the left of column x in the row of U(x) will always be less than 0.025, which satisfies equation (5.2).

Note from Table 1, that, for x equal to three, both the upper and lower limits are for one sided intervals. As a result, the conservative confidence interval for x equal to three contains all of the possible values of M given x equals three. As a result, neither equation (5.6) nor equation (5.7) holds for x equal to three. The tendency would be to say that the $P(L(X) \leq M \leq U(X))$ equals one. However, there is no probability metric over the M's. Probabilities can only be found over the x's. Even though we know the interval contains M, for the M's within the interval, the average probability of M being in an interval is 99.0%, not 100%.

For the example given, most of the confidence levels are quite a bit higher than 95%. However, this author has observed that as N and n increase, the levels tend to approach the level of (1-α) 100% for most of the M's. Also, the conservative confidence intervals are the shortest intervals that can be found if equations (5.1) and (5.2) hold.

7. A FUNCTION IN R TO FIND CONSERVATIVE CONFIDENCE INTERVALS FOR THE HYPERGEOMETRIC DISTRIBUTION

Below is a function which finds conservative confidence intervals for the number of members of a population with an attribute, given the size of the population and the size of the sample. The variables 'N' and 'n' are the population and sample sizes, respectively. The variable, 'alpha', sets the level of α to use in the function, where (1-α) 100% is the conservative level of confidence.

In a call to the function, if 'prt' is set equal to 'True', then the tables of the cumulative probabilities used to find the lower and upper limits are printed. Otherwise, the tables are not printed. Whether the tables are printed or not, the function prints out the conservative confidence intervals, along with the average level of confidence for the intervals over the M's, for each possible value of X, and the probability that a confidence interval contains M (the confidence level divided by 100) for each possible value of M.

```
function(N=16, n=6, alpha=.05, prt=F){ cpmat= array(1,c(N+1,n+1,2))
 lab = list(paste(0:N),paste(0:n), c("P( X >= x of column | M = M of row )",  "P( X <= x of column | M = M of row )"))
 dimnames(cpmat) = lab
 for (M in 0:N) {
 cpmat[M+1,,1] = c(1, phyper(0:(n-1),M,N-M,n,lower.tail=F))
 cpmat[M+1,,2] = phyper(0:n,M,N-M,n) }
 if (prt==T) {
cat("\nM in the rows and x in the columns\n\n")
print(round(cpmat,3))}
rmat = array(0, c(N+1, n+1,2))
rmat[,,1][ cpmat[,,1] <= alpha/2 ] = 1
rmat[,,2][ cpmat[,,2] > alpha/2 ] = 1
rt= cbind(apply(rmat[,,1],2,sum), apply(rmat[,,2],2,sum)-1)
prb = numeric(N+1)
for (M in 0:N) {
ma = rep(-1,n+1)
for (x in 0:n) {
 if ( M>=rt[x+1,1] & M<=rt[x+1,2] ) {
ma[x+1]=x }}
mas = ma[ ma>-1 ]
prb[M+1] = (phyper(max(mas), M, N-M, n)  - ifelse(min(mas)==0, 0, phyper(min(mas)-1, M, N-M, n))) }
cci = numeric(n+1)
for (x in 0:n) {
 cci[x+1] = mean(prb[rt[x+1,1]:rt[x+1,2]+1])*100}
cat("\n\nconfidence intervals for possible values of x\n\n")
pr1 = cbind(paste(0:n), paste("(",rt[,1],",",rt[,2],")"), paste(round(cci,1),"%",sep=""))
dimnames(pr1) = list(rep(" ",n+1), c("x ","confidence interval  ", "mean confidence"))
print(pr1, quote=F)
cat("\n\nprobability CI contains M under random sampling\n\n")
pr = cbind( paste( "M:", 0:N), paste(" probability:" ,round(prb,3)))
dimnames(pr) = list(rep(" ",N+1),rep(" ",2))
print(pr, quote=F) }
```

Below is the output from the function for 'N' equal to ten, 'n' equal to five, 'alpha' equal to 0.05, and 'prt' equal to True.

```
M in the rows and x in the columns

, , P( X >= x of column | M = M of row )

   0     1     2     3     4     5
0  1 0.000 0.000 0.000 0.000 0.000
1  1 0.500 0.000 0.000 0.000 0.000
2  1 0.778 0.222 0.000 0.000 0.000
3  1 0.917 0.500 0.083 0.000 0.000
4  1 0.976 0.738 0.262 0.024 0.000
5  1 0.996 0.897 0.500 0.103 0.004
6  1 1.000 0.976 0.738 0.262 0.024
7  1 1.000 1.000 0.917 0.500 0.083
8  1 1.000 1.000 1.000 0.778 0.222
9  1 1.000 1.000 1.000 1.000 0.500
```

10 1 1.000 1.000 1.000 1.000 1.000

, , P(X <= x of column | M = M of row)

	0	1	2	3	4	5
0	1.000	1.000	1.000	1.000	1.000	1
1	0.500	1.000	1.000	1.000	1.000	1
2	0.222	0.778	1.000	1.000	1.000	1
3	0.083	0.500	0.917	1.000	1.000	1
4	0.024	0.262	0.738	0.976	1.000	1
5	0.004	0.103	0.500	0.897	0.996	1
6	0.000	0.024	0.262	0.738	0.976	1
7	0.000	0.000	0.083	0.500	0.917	1
8	0.000	0.000	0.000	0.222	0.778	1
9	0.000	0.000	0.000	0.000	0.500	1
10	0.000	0.000	0.000	0.000	0.000	1

confidence intervals for possible values of x

x	confidence interval	mean confidence
0	(0 , 3)	100%
1	(1 , 5)	98.9%
2	(2 , 7)	98.3%
3	(3 , 8)	98.3%
4	(5 , 9)	98.9%
5	(7 , 10)	100%

probability CI contains M under random sampling

M: 0 probability: 1
M: 1 probability: 1
M: 2 probability: 1
M: 3 probability: 1
M: 4 probability: 0.952
M: 5 probability: 0.992
M: 6 probability: 0.952
M: 7 probability: 1
M: 8 probability: 1
M: 9 probability: 1
M: 10 probability: 1

8. CONCLUSIONS

As far as the author knows, the hypergeometric result given in this paper has not been published previously. The result in Section 3 using the lognormal distribution shows that, for asymmetrical distributions, finding confidence intervals is a bit tricky, but doable. The R function of Section 7 provides conservative confidence intervals for the geometric distribution and can be easily edited for a personal application. Having a method of calculating confidence

intervals for parameters of the hypergeometric should be of use to researchers using random sampling without replacement and categorical variables. The hypergeometric result should also generalize to any finite, discrete distribution with a finite, discrete parameter space.

For future interest, the author has made a number of statements, about the properties of the hypergeometric estimators, which deserve further study. Also of interest is finding confidence intervals for N when n and M are known.

REFERENCES

Agresti, A. (2013). *Categorical Data Analysis (3rd ed.).* Hoboken, NJ: John Wiley & Sons.

Cochran, W. (1977). *Sampling Techniques (3rd ed.).* New York, NY: John Wiley & Sons.

Finkelstein, M., Tucker, H. G., & Veeh, J. A. (2000). Confidence Intervals for a Single Parameter. *Communications in Statistics, Theory and Methods, 29*, 1911-1928.

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2010). *Statistical Distributions (4th ed.).* New York, NY: John Wiley & Sons.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses (2nd ed.).* New York, NY: John Wiley & Sons.

Tang, W., He, H., & Tu, X. M. (2012). *Applied Catagorical and Count Data Analysis.* Boca Raton, FL: CRC Press.

Tufféry, S. (2011). *Data Mining and Statistics for Decision Making.* (R. Riesco, Trans.) West Sussex, United Kingdom: John Wiley & Sons, Ltd.