A new look at inference for the Hypergeometric Distribution.

February 23, 2009

William M. Briggs

New York Methodist Hospital, Brooklyn, NY
email: matt@wmbriggs.com

and

Russell Zaretzki

Department of Statistics, Operations, and Management Science
The University of Tennessee
331 Stokely Management Center, Knoxville, TN, 37996

email: rzaretzk@utk.edu

February 23, 2009

Summary: The problem of inference for the proportion of successes in a finite populations is given much less attention in both inference and application courses than the related infinite population problem. However, the finite population problem is both interesting and useful in its own right while also providing insight into various approaches for the infinite population case. This article explores exact Bayesian inferences for understanding a finite population proportion. The derivations of both posterior and posterior predictive distributions provide excellent insight for students regarding the updating nature of the Bayesian paradigm. The novel approach presented for deriving the posterior predictive distribution also allows students to overcome difficult combinatorial calculations. Accessible limiting arguments also can be used to justify the use of flat priors in the infinite populations. Finally, the results may be compared to conventional inferences based upon the use of finite population correction factors or exact frequentist intervals.

Key words: Binomial, Hypergeometric, Bayesian, Marginal Likelihood, Posterior Predictive Distribution, Objective prior

1 Introduction

The problem of inference for the proportion of successes in finite populations is given much less attention than the related infinite population problem. For example, there exists many studies that present practical approaches to improving the accuracy of confidence intervals for the proportion of successes in an infinite population, see Caffo and Agresti (2000); Brown et al. (2001). Although inference for proportions in finite populations is less common, it is still both important and interesting; see Wright (1992). Modern computing power also relieves the tedium and effort required for finding exact answers.

The problem of inference in a hypergeometric distribution provides an excellent opportunity to explore discrete distributions, learn about Bayesian updating of information through posterior distributions, explore and understand the derivation of a marginal distribution which is critical for normalization of a posterior, and derive results for the infinite dimensional case through straightforward limiting arguments.

The hypergeometric distribution is most frequently referenced as the null distribution for Fisher's exact test in 2×2 tables; see Simonoff (2002). It is also studied in the context of capture-recapture experiments where the goal is to infer the size of the total population, N. In Section 2, we focus instead on inference for the number of successes in a population of known size, often discussed historically in the context of studying defective items in a lot; see Chung and Delury (1950). We show how to derive the posterior distribution on the total number of successes, which principally involves computing the marginal distribution of successes given a flat prior. Once the posterior has been computed, we compute the posterior predictive distribution and discuss its utility in statistical inference. The simple derivation of the posterior predictive distribution below is much more intuitive than earlier derivations and allows us to test intuitive understanding of Bayesian reasoning. A somewhat surprising feature of the posterior predictive distribution of the number of successes in a future sample is that it is independent of the size of the population. The explanation of this phenomenon by Bose and Kedem (1996); Bose (2004) is a bit advanced, so in Section 3, we analyze this result with more elementary tools. We also investigate the limiting value of the posterior distribution as the population and number of successes go to infinity while the proportion of successes and failures is held constant. A brief discussion of standard and Bayesian intervals is given in Section 4 and the importance of predictive distributions is illustrated with an example.

2 Posterior and Predictive Distributions

Suppose we collect a sample of size n from a finite-population urn containing N balls and wish to infer M, the number of successes (balls labeled “1”), given that we saw n_1 successes in the sample. We use the notation n_0 to denote failures, so that $n_1 + n_0 = n$. We parameterize the number of successes as $M = N\theta$, with θ taking only one of the $N + 1$ discrete values of the form $i/N, \quad i \in 0, 1, \ldots N$. The probability of seeing $N\theta = j$ successes is computed with the hypergeometric probability mass function

$$Pr(n_1 = j|n, \theta, N) = \frac{\binom{N\theta}{j}\binom{N-N\theta}{n-j}}{\binom{N}{n}} = \frac{\binom{n}{j}\binom{N-n}{N\theta-j}}{\binom{N}{N\theta}} \tag{1}$$

for $\max[0, n + N\theta - N] \leq j \leq \min[n, N\theta]$. The second expression can be derived through direct manipulation and simplifies the expression slightly in terms of $N\theta$.

Assume that we have reason to believe that no number of successes is more likely than any other so that $\Pr(\theta) = 1/(N + 1)$. Instead of a prior on θ, we might choose to put one on $N\theta$. Given N, these quantities are 1-to-1 functions of each other, so we focus on θ. Incidentally, unlike continuous distributions, the parameter here is *observable*.

After n balls have been removed, the posterior parameter distribution of θ is produced in the obvious way:

$$\begin{aligned}
\Pr(\theta = \frac{j}{N}|n, n_1, N) &\propto \Pr(n_1|n, \theta = j/N, N)\Pr(\theta|n, N) \\
&= \frac{\binom{n}{n_1}\binom{N-n}{j-n_1}}{\binom{N}{n}}\frac{1}{N+1} \\
&= \binom{N-n}{j-n_1}\frac{\beta(j+1, N-j+1)}{(n+1)\beta(n_1+1, n_0+1)}
\end{aligned} \tag{2}$$

for $j = n_1, n_1 + 1, \ldots, N - n_0$, and $\beta(\cdot)$ denotes the beta function. If $n_1 > 0$ then $N\theta$ can be no smaller than n_1. Similarly, if $n_0 > 0$ then $N\theta$ can be no larger than $(N - n_0)$. In particular, if $n_1 + n_0 = N$ then $N\theta = n_1$ as desired.

In real applications, it makes sense to focus on $d = N\theta - n_1$, the number of remaining successes. Replacing $N\theta$ with $d + n_1$ in (2), we find,

$$\Pr(d|n_1, n, N) \quad \propto \quad \frac{1}{n+1} \frac{\binom{n_1+d}{d}\binom{N-n_1-d}{N-n-d}}{\binom{N+1}{n+1}} \tag{3}$$

where $d \in \{0, \ldots, N - n_0\}$. Expression (3) presents the un-normalized posterior in the form of a negative hypergeometric, which is simply a alternative name for the beta-binomial distribution (Terrell, 1999).

Before we have removed any of the balls from the urn, we can compute the prior predictive distribution of n_1 in the initial sample. This quantity is critical since it forms the normalization constant of the posterior distribution in eq. (2). To derive this we simply sum over the possible values of θ in expression (3) which is equivalent to summing a beta-binomial mass function over its range. The result is

$$\Pr(n_1 = j|n, N) = 1/(n + 1), \qquad j \in \{0, \ldots n\} \tag{4}$$

Note that under our prior, when $n = 1$, $P(n_1 = 1) = 1/2$. That is, if we reach in and grab out just one ball, the chance that it is a 1 or 0 is 1/2, no matter what N is. Furthermore, if we grab $n > 1$ balls, the result says that n_1 is equally likely to be any result in $\{0, 1, \ldots, n\}$, and is also independent of N. This is intuitive, since we began with all proportions θ being equally likely a priori and have not yet collected data that suggests otherwise.

Given the original sample A whose size is now denoted by n_a, with n_{1a} the number of successes and n_{0a} the number of failures, let a new sample of size $n_b \leq N - n_a$ be collected. We want to know the distribution of successes, n_{1b}, in this new sample. It is known that $0 \leq n_{1b} < N\theta - n_a$. Given d, the distribution of n_{1b} is clearly hypergeometric.

This allows us to compute the posterior predictive distribution on n_{1b}:

$$\Pr(n_{1b}|n_b, n_{1a}, n_a, N) = \sum_{d=0}^{N-n_a} \Pr(n_{1b}|d, n_b, n_{1a}, n_a, N)\Pr(d|n_{1a}, n_a, N)$$

$$= \sum_{d=0}^{N-n_a} \frac{\binom{n_b}{n_{1b}}\binom{N-n_a-n_b}{d-n_{1b}}}{\binom{N-n_a}{d}} \binom{N-n_a}{d} \frac{\beta(n_{1a}+d+1, n_{0a}+(N-n_a)-d+1)}{\beta(n_{0a}+1, n_{1a}+1)} \quad (5)$$

Direct calculation of this expression looks quite daunting. Bratcher et al. (1971) computed the sum analytically using combinatorial identities. However, using simple Bayesian arguments the following result is easily achieved:

Theorem 2.1. *The sum in eq. (5) is exactly*

$$\binom{n_b}{n_{1b}} \frac{\beta(n_{1a}+n_{1b}+1, n_{0a}+n_{0b}+1)}{\beta(n_{1a}+1, n_{0a}+1)} \quad (6)$$

which is a beta-binomial distribution with parameters $(n_b, n_{a1}+1, n_{a0}+1)$.

Proof. Assume that a sample $A \cup B$ of size $n_a + n_b \leq N$ is collected with $n_{1a} + n_{1b}$ successes observed. The posterior distribution for d can easily be derived from eq. 3 giving,

$$\Pr(d|n_a+n_b, n_{1a}+n_{1b}, N) = \binom{N-n_a-n_b}{d-n_{1b}} \frac{\beta(d+n_{1a}+1, N-(d+n_{1a})+1)}{\beta(n_{1a}+n_{1b}+1, n_{0a}+n_{0b}+1)} \tag{7}$$

Next, consider an alternative route to the same sample. Suppose sample A is collected first and the posterior in eq. (3) is computed. This posterior $p(d|n_{1a}, n_a, N)$ is then updated using the likelihood from the new sample, B. That is, the initial posterior becomes the prior and our new posterior distribution $p(d|n_{1a}, n_{1b}, n_a, n_b, N)$ is computed based upon the data collected in sample B. Since both posteriors integrate the same sample and prior information, they must be the same.

Because the posterior predictive distribution is simply the normalizing constant of this second sequentially defined posterior distribution, it follows from Bayes theorem that

$$\Pr(n_{1b}|n_b, n_{1a}, n_a, N) = \frac{\Pr(n_{1b}|n_b, d, n_{1a}, n_{0a}, N)\Pr(d|n_a, n_{1a}, N)}{\Pr(d|n_a+n_b, n_{1a}+n_{1b}, N)} \tag{8}$$

The result follows from taking this ratio. □

In the next section we discuss the surprising result that the predictive distribution is independent of N. This means the result also holds in the limit as $N \to \infty$. Hence, it matches the conventional beta-binomial result which is typically obtained by combining a binomial predictive distribution with a beta posterior.

3 A Curious Property of the Predictive Distribution

Bratcher et al. (1971) derives (8) and notes that it is identical to the infinite population result based on beta-binomial Bayesian inference. This highlights the fact that the predictive distribution does not depend upon the overall population size N, which implies that two persons using totally different assumptions regarding population size but the same methodology would arrive at the same solution. Bose and Kedem (1996); Bose (2004) use the concept of recursive generation of a distribution to prove that under certain classes of priors—which include the discreet uniform—the joint distribution of (n_{1a}, n_{1b}) is independent of population size N. While the proof offered is accessible, we attempt to provide a somewhat more intuitive result based upon the Bayesian argument presented in Proof 2.1.

Rewriting our earlier expression, and letting $\delta = (n_a, n_b)$, then

$$\begin{aligned}
\Pr(n_{1b}|n_{1a}, \delta, N) &= \frac{\Pr(n_{1b}|d, n_{1a}, \delta, N)\Pr(n_{1a}|d, n_a, N)\Pr(d|N)/\Pr(n_{1a}|\delta, N)}{\Pr(n_{1a}+n_{1b}|d, n_{1a}, \delta, N)\Pr(d|N)/\Pr(n_{1a}+n_{1b}|\delta, N)} \\
&= \frac{\Pr(n_{1a}, n_{1b}|d, \delta, N)\Pr(n_{1a}+n_{1b}|\delta, N)}{\Pr(n_{1a}+n_{1b}|d, \delta, N)\Pr(n_{1a}|\delta, N)} \\
&= \Pr(n_{1b}, n_{1a}|n_{1a}+n_{1b}, d, \delta, N)/\Pr(n_{1a}|n_{1a}+n_{1b}, \delta, N) \qquad (9)
\end{aligned}$$

In the case of a flat prior, the conditional $\Pr(n_{1a}|n_{1a}+n_{1b}, \delta, N)$ is just $(n_a+1)/(n_a+n_b+1)$. The first term, although nominally conditional on both d and N, is also conditional on the sum of successes; once this sum is known (along with samples sizes), then information about the overall population (N, d) adds no information, so that the joint distribution is simply a hypergeometric.

When predicting the number of future successes given the results of a previous sample, intuition suggests removing the initial sample and imagining what would happen when drawing a new sample from the remaining population. One would naturally assume that the probability of success depends upon $N - n_a$, which is the size of the remaining population. The decomposition of the joint distribution in (9) suggests a different situation. More accurately, we are selecting a subset of size $n_a + n_b$ from the population which contains $n_{1a} + n_{1b}$ (n_{1b} being fixed by the hypothetical posed when asking $P(n_{1b}|n_{1a}, \delta, N)$. The joint distribution gives the probability of seeing n_{1a} of these successes in the first sample and n_{1b} in the second sample, which follows a hypergeometric distribution. The conditional distribution follows after dividing by the conditional distribution that n_{1a} successes are observed in the first sample given the total number of successes. Key to understanding this is that when we draw the total sample of size $n_a + n_b$ we assume that the successes will be distributed homogenously in both subsamples A and B. Hence, the joint, and conditional distributions will be maximized when the proportions $n_{1a}/n_a = n_{1b}/n_b$.

A well known limiting argument relates the beta-binomial distribution

(2) with the standard binomial. The same argument can be used to find the limiting posterior distribution of θ as $m \to \infty$.

Theorem 3.1. *The function given in eq. (2) converges to the cumulative beta distribution $B(n_1+1, n_0+1)$ in the limit as $m \to \infty$.*

Proof. A very intuitive result (Ross, 1988) states that as m and $m\theta$ get very large in relation to the sample size n, the consequences of sampling without replacement are diminished and the hypergeometric distribution behaves increasingly like a binomial distribution with sample of size n and probability θ. Applying this result to derive the limiting posterior distribution for θ we can write,

$$\lim_{m\to\infty} \Pr(\theta|m, n_0, n_1) = \lim_{m\to\infty} \frac{\binom{n}{n_1}\binom{m-n}{m\theta-n_1}}{\binom{m}{m\theta}}$$

Again, following Ross (1988), we note that the righthand side can be

$$\begin{aligned}
\lim_{m\to\infty}\left[\frac{\binom{n}{n_1}\binom{m-n}{m\theta-n_1}}{\binom{m}{m\theta}}\right] &= \lim_{m\to\infty}\binom{n}{n_1}\frac{(m\theta)!}{(m\theta-n_1)}\frac{(m-m\theta)!}{(m-n-(m\theta-n_1)}\Big/\frac{m!}{(m-n)!} \\
&= \lim_{m\to\infty}\binom{n}{n_1}\frac{m\theta}{m}\frac{m\theta-1}{m-1}\cdots\frac{m\theta-n_1+1}{m-n_1+1}\times \\
&\quad \frac{m(1-\theta)}{m-n_1}\frac{m(1-\theta)-1}{m-n_1-1}\cdots\frac{m(1-\theta)-(n-n_1-1)}{m-n_1-(n-n_1-1)} \\
&= \binom{n}{n_1}\theta^{(n_1+1)-1}(1-\theta)^{(n-n_1+1)-1} \qquad (10)
\end{aligned}$$

From this we see that the posterior distribution for θ is a beta distribution, $\beta(k+1, n-k+1)$. □

We conclude that if $n_1 + n_0 << m$, for large m, the posterior on θ follows a beta distribution with parameters $n_1 + 1$, $n_0 + 1$. This result coincides with the standard infinite population problem with a “flat” prior on θ, i.e. $\alpha = \beta = 1$ in (10). The fact that in the finite sample case, some values of $\theta \in [0, 1]$ are impossible after we have taken some data (drawn some balls) contrasts with the infinite sample approximation, where the parameter always has positive probability of being smaller or larger than any given value in $(0, 1)$. There is also 0 probability for the values $\{0, 1\}$, which are still possible in the finite case.

4 Teaching Students about Inference

[Emphasize inference and practical significance are united. Discuss two welding methods A and B(cheaper newer method.) Want to verify that B is just as safe as A. How to compare these two, simulation or computer, analytical would seem to be difficult. Emphasize differences with classical inferences as well as conclusions from continuous posterior approximations.]

Bayesian and Frequentist confidence intervals for the number of successes in a discrete sample are discussed by Steck and Zimmer (1968). The standard form of the upper $100(1-\alpha)$ confidence interval is simply $M - 1$ where M is the smallest value of $M = N\theta$ which satisfies

$$\sum_{M=0}^{j} Pr(n_1 = j|n, M, N) \geq \alpha \qquad M \leq N - n \tag{11}$$

Interestingly, comparisons show that Bayesian intervals are highly dependent upon the prior chosen and that the uniform prior distribution can be significantly less conservative than the hypergeometric prior.

Burstein (1975) investigates frequentist inference and compares the use of approximate intervals based on applying the finite population correction factor to exact inferences in the case where the finite population size N is not vastly larger than the sample n. Given standard binomial intervals $[L, U]$, the finite population correction takes the form,

$$L_{Fin} = \hat{p} - (\hat{p} - L)[(N-n)/(N-1)]^{1/2} \tag{12}$$

$$U_{Fin} = \hat{p} + (U - \hat{p})[(N-n)/(N-1)]^{1/2} \tag{13}$$

$$\tag{14}$$

The methodology used in Burstein (1975) is very easy to reproduce using modern computational resources and may provide a good exercise for students. While the basic FPC is highly accurate if $N > 1000$, using a continuity type correction, replacing $\hat{p}$ with $\hat{p} - .5/n$ in the lower case and $\hat{p} + \hat{p}/n$ in the upper case, results in increased accuracy and more conservative intervals.

From an applied point of view, the posterior predictive distribution deserves more attention in contrast to intervals. The obvious advantage of (posterior) predictive distributions is that they do not depend upon any unknown and unobservable parameters. They also directly address the question that is of most interest, which is "What will actually be observed in a new sample?" For example, Wright (1992) considers a situation in which an inspector is trying to ensure that all the welds on a nuclear reactor are properly formed. Suppose that the number of welds, $N = 20$, is known. A small number of welds, $n = 5$, are scheduled for testing, which is very expensive and time consuming. After the sample is collected, and is found to have no faults, we would like to know the probability the remaining $N - n$ welds contain a fault. A standard 95% frequentist upper confidence bound for the actual number of defectives can be computed using the `phyper` command in `R-language` and gives $U = 10$ while a 95% posterior interval for the number of faults based upon eq. 2 is 9. Using the posterior predictive distribution based on flat prior (from eq. (3)) we compute the probability of one or more faults to be 71%.

Contrast this with a conclusion based upon a posterior computed by approximating the distribution for the number of faults with a binomial(n, θ) and adopting a standard beta prior (with $\alpha = \beta = 1$). Conjugate calculations give a beta posterior with hyperparameters $(1, n + 1)$. This posterior distribution is trickier because it only says something about an unobservable parameter, θ. We know that there are $N - n = 15$ welds left. If none are bad, the fraction of bad welds is, of course, 0. If one or more are bad then the fraction of bad ones is at least 1/15. So we can ask, what is the probability that the posterior on θ is greater than this. The answer is 66%, a sizeable difference.

Equally important is the two sample problem. Consider comparing the numbers of faulty welds when using a standard welding method A versus a newer, less expensive method B. We would like to verify that method B is at least as effective as method A. A typical frequentist approach would either use a two sample test of proportions, implicitly assuming that the population is infinite, or might use Fisher's exact test. In either case, a conclusion that a test statistic is "statistically significantly" does not, and could not, document how different the methods are and what any difference means from a practical point of view (Briggs, 2008). A major advantage of the predictive distribution is that there is no distinction between statistical and practical because differences are measured on the observable scale, i.e the actual difference in the number of defectives.

Continuing the example: it was desired to do a complete test of the new welding method so all the welds on two reactors were checked. After the testing, one more reactor will be ordered (the plant needs three), made by either technique A or B. Which to buy? Let $f_A = 4$ be the number of faulty welds found in the first reactor and $f_B = 1$ be the number in the second, with both having taken a sample of $n = N = 20$ welds each. Extending the continuous approximation, with the parameters θ_A and θ_B being of interest, the posterior probability that $\theta_A > \theta_B$ is 91%. However, a better question is, "Which technique, A or B, is likely to produce more faulty welds in the next reactor ordered?" Since the number of welds is discrete and small, there is a non trivial probability that the number of faulty welds are equal. That is the case here, with an 8.6% chance of this occurring. (We note that this probability cannot be calculated in the continuous case.) There is also a 79% chance that the number of faulty welds using technique A will be larger than those using technique B. Or a 79+8.6=88% chance that technique A has at least as many faulty welds as technique B.

Does the 3% difference between the continuous approximation/parameter-focused method make a difference? It might. It depends on the costs of the reactors, the cost of repair, and so on. The main point is that the parameter-focused method will *always* give an answer that is more certain than observable-focused method. Which is another way of saying you will be overconfident using the parameter-focused method when engaged in real-life decisions.

References

Bose, S. (2004). On the robustness of the predictive distribution for sampling from finite populations. *Statistics and Probability Letters*, 69:21–27.

Bose, S. and Kedem, B. (1996). Non-dependence of hte predictive distribution on the population size. *Statistics and Probability Letters*, 27:43–47.

Bratcher, T. L., Schucany, W. R., and Hunt, H. H. (1971). Bayesian prediction and population size assumptions. *Technometrics*, 13(3):678–681.

Briggs, W. M. (2008). *Breaking the Law of Averages: Real-Life Probability and Statistics*. Lulu, New York.

Brown, L. D., Cai, T. T., and Dasgupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16:101–133.

Burstein, H. (1975). Finite population correction for binomial confidence limits. *Journal of the American Statistical Association*, 70(349):67–69.

Caffo, B. S. and Agresti, A. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician*, 54:280–288.

Chung, J. H. and Delury, D. B. (1950). *Confidence Limits for the Hypergeometric Distribution*. University of Toronto Press, Toronto, CA.

Ross, S. (1988). *A First Course in Probability*. Macmillan Publishing Company, New York, third edition.

Simonoff, J. (2002). *Analyzing Categorical Data*. Springer, New York, NY.

Steck, G. P. and Zimmer, W. J. (1968). The relationship between neyman and bayes confidence intervals for the hypergeometric parameter. *Technometrics*, 10(1):199–203.

Terrell, G. R. (1999). *Mathematical Statistics*. Springer Texts in Statistics. Springer, New York, NY.

Wright, T. (1992). A note on sampling to locate rare defectives with strong prior evidence. *Biometrika*, 79(4):685–691.