Estimating a Proportion Using Stratified Data From Both Convenience and Random Samples

Todd Graves, Michael Hamada, Jane Booker, Michele Decroix, Kathy Chilcoat, and Clint Bowyer

Los Alamos National Laboratory

02.01.06 0300

Abstract

Estimating the proportion of an attribute present in a population can be challenging when the population is stratified by lots produced by a common manufacturing process and the available data arise from both random and convenience samples. Moreover, all the lots may not have been sampled. This paper proposes a Bayesian methodology for making inferences about a proportion that properly accounts for the potential bias of the convenience samples, the stratification by lots and the fact that not all the lots have been sampled. The methodology is illustrated with a simulated population; however, the solution was motivated by a similar, but proprietary, production problem.

Key Words: Bayesian, biased sampling, discrete data, extended-hypergeometric and hypergeometric distributions, finite population, MCMC.

1 Introduction

Populations are surveilled to provide confidence that they are in a good state of health. For example, missile and weapon stockpiles are surveilled to assess that they will perform when necessary. Through surveillance, it may be observed or determined through analysis that some sampled systems have a component which has an attribute (either present or absent). A natural question arises as to what proportion of the population has components with the attribute present. When the population is large and the sampling is completely random, a binomial distribution provides a good approximation for the observed number of systems with the attribute present. In this situation, estimating the proportion is a simple task; the maximum likelihood estimate is the observed proportion.

In this paper, however, we consider a more complicated situation which requires a rather sophisticated analysis. Suppose that the components in question have been manufactured in small lots of varying sizes by a common manufacturing process. Not all of the manufactured components end up in manufactured systems. Some are used for monitoring quality control and others are designated for various ongoing studies. We refer to such components as arising from convenience samples because it is not known how or why they are chosen for these studies. Besides these convenience samples, the manufactured systems are randomly sampled from the population over time.

While these convenience samples have been chosen stochastically, they may not have been chosen completely randomly, or may have been chosen for certain characteristics particular to the studies which may or may not related to the attribute. Consider the situation where a quality control sample of the lot is taken and inspected. Those components which are “interesting” are kept for further study; the “uninteresting” components are released and are built up into systems. If the attribute is related to the “interesting” components (i.e., a higher proportion of “interesting” components have the attribute present than that for the “uninteresting” components), then these samples provide a biased estimate of the proportion of attribute present.

Ultimately, we are concerned with estimating the proportion p^* of the current population which have components with the attribute present. A novel aspect that we consider in this paper is that there are convenience and/or random samples available from some but not all of the lots. The challenge is to appropriately account for the potential bias in the convenience samples, i.e., components with the attribute present may appear more often in the convenience samples than by chance so that ignoring this bias would lead to overestimating p^*. In this paper, we propose a statistical approach to account for this potential bias.

This paper is organized as follows. First, we present an example motivated by a proprietary manufacturing and production problem. We present a statistical model for the population and for the data from the random and convenience samples in Section 3. In Section 4, we consider a Bayesian approach to making inferences about the proportion with an attribute present in the remaining population. We demonstrate the proposed approach with the illustrative population in Section 5. Finally, we conclude with a discussion.

2 An Example

For illustration, consider a simulated population of 5000 components, made up of 230 lots: 100 of size 10, 100 of size 25 and 30 of size 50. A convenience sample of 100 components was taken when the systems were first manufactured. Subsequently, a random sample of size 50 was taken from the remaining 4900 components. In total, 96 of the 230 lots have been sampled – 18 lots have both random and convenience samples, 21 lots have only random samples and 57 lots have only convenience samples. Table 1 presents the population and sample data which include the lot sizes (N), the unknown number of components with the attribute present (K), the convenience and random sample sizes (n_c and n_r, respectively,) and the number of components with the attribute present (y_c and y_r) in the convenience and random samples, respectively. This population of 5000 has 513 components which have the attribute present. For Case 1 in which the convenience sample is somewhat biased, there are 16 and 6 components with the attribute present in the convenience and random samples, respectively. Thus, about 10% of the remaining population of 4850 components have the attribute present. The proprietary problem, upon which this example is based, was estimated to have had a much rarer occurrence of the attribute present. We have chosen a population for this example with about a 10% occurrence rate for ease of illustration.

Taking the data as representative of the population, an estimated proportion of 22/150 or 14.7% (with a standard deviation of 2.9%) of the remaining population of components having the attribute present might be proposed. For Case 2 in which the convenience samples are even more biased, there are 24 and 3 components with the attribute present in the convenience and random samples, respectively. Again, ignoring the bias of the convenience samples, an estimated proportion of 27/150 or 18.0% (with a standard deviation of 3.1%) of the remaining population having components with the attribute present would result. Consequently, properly accounting for such biases is one of the key motivations for this paper.

3 Statistical Models for the Population and Sampled Data

Before introducing the statistical models, we begin with some notation. The finite population consists of M lots. Let the ith lot size be denoted by N_i and the unknown number of systems in the ith lot with the attribute present be denoted by K_i. If there is a convenience sample for the ith lot, its sample size is n_{ci} and y_{ci} is the number of components in the convenience sample with the attribute present. If there is a random sample for the ith lot, its sample size is n_{ri} and y_{ri} is the number of components in the convenience sample with the attribute present. Because the convenience sample is assumed to be taken first, the ith lot size for the random sample is $N_{ri} = N_i - n_{ci}$ and $K_{ri} = K_i - y_{ci}$ is the number of components with the attribute present remaining in the ith lot after the convenience sample has been taken.

Next, we consider a model for the population. We assume that the number of components with the attribute present in the ith lot K_i to be distributed as binomial, where p_i is the probability of a component having the attribute present in the ith lot, i.e.,

$$K_i \sim Binomial(N_i, p_i). \tag{1}$$

The fact that the lots were produced by a common manufacturing process means they are related but their tendencies for having a component with the attribute present are possibly different. We express this by the p_i being exchangeable as follows:

$$p_i \sim Beta(a, b). \tag{2}$$

(1) and (2) together constitute a hierarchical model.

Now we consider statistical models for the data. For the convenience sample data, we want to account for the potential bias of sampling too many or too few components with the attribute present. This sampling mechanism is an example of a non-ignorable selection procedure (Rubin (1976), Gelman et al. (1995)) which must be accounted for in making inferences. Nonresponse in survey sampling is an example of a non-ignorable selection procedure for which that the probability that an individual responds depends on the individual's value of the binary attribute of interest (Stasny (1991), Nandram and Choi (2002)). By modeling the differential response probabilities, Stasny (1991) and Nandram and Choi (2002) properly account for the potential bias in the data from the respondents.

Similarly, we handle the potential bias in the convenience data through modeling. To do this, we use the extended-hypergeometric distribution (Harkness (1965)) for y_{ci} which is denoted by

$$y_{ci} \sim Extended - hypergeometric(K_i, N_i - K_i, n_{ci}, \theta). \tag{3}$$

The extended-hypergeometric probability mass function has the following form:

$$P(y_{ci} = y) = \frac{\begin{pmatrix} n_{ci} \\ y \end{pmatrix} \begin{pmatrix} N_i - n_{ci} \\ K_i - y \end{pmatrix} \theta^y}{\sum_{j=max(0,n_{ci}-N_i+K_i)}^{min(n_{ci},K_i)} \begin{pmatrix} n_{ci} \\ j \end{pmatrix} \begin{pmatrix} N_i - n_{ci} \\ K_i - j \end{pmatrix} \theta^j} \; ,$$

for $y = max(0, n_{ci} - N_i + K_i)$, $\ldots$, $min(n_{ci}, K_i)$. When the biasing parameter θ is equal to one, the extended-hypergeometric reduces to the hypergeometric which arises from a completely random sample; i.e., there is no biasing. When θ is greater than one, the sampling favors components with the attribute present.

Table 2 demonstrates that for $\theta > 1$, the probabilities for sampling larger number of components with the attribute present are higher. This table considers the case when the lot size is 10, half the components have the attribute present and the lot sample size is 3. Note that a common θ is assumed for the convenience sampling for all lots in this table.

The extended-hypergeometric distribution can be simulated from using the following procedure. Suppose that each component in a given lot say of size N_i is determined randomly to be included in the sample. Each component with the attribute present is included with probability π_1; each component with the attribute absent is included with probability π_0. Note that the number of components included in the sample is random, as described thus far. To obtain a sample of size n_i, one has to visualize performing this procedure until the realized sample size is n_i. For such a sample, y_i, the number of components with the attribute present in the sample has the extended-hypergeometric distribution, where

$$\theta = \frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)},$$

is the odds ratio. While this may not be the exact stochastic mechanism of the convenience sampling, the probability mass function reflects a biasing of the components with the attribute present and serves as a good approximation. See Johnson and Kotz (1969) and their references for further discussion of the extended-hypergeometric distribution.

For the random sample data, the y_{ri} follow a hypergeometric distribution denoted by

$$y_{ri} \sim Hypergeometric(K_{ri}, N_{ri} - K_{ri}, n_{ri}). \tag{4}$$

Because the lot sizes are small, binomial approximations are not appropriate.

Note that analyzing only the random sampled data is problematic because some lots had convenience samples taken from them while others did not and some lots had random samples taken from while others did not. Thus, there are lots for which either convenience samples or random samples were taken but not both, those for which both samples were taken and those which were not sampled. Consequently, the random samples (possibly after convenience samples were taken) alone are biased with respect to (2), the model for the lots before any sampling (both convenience and random) was done. The modeling above, however, properly accounts for the pattern of convenience and random sampling.

4 Bayesian Approach for Inference

To provide inference about p^*, we propose using a Bayesian approach. Bayesian inference provides uncertainty about the unknowns $\boldsymbol{\eta} = (p_1, \ldots, p_M, K_1, \ldots, K_M, a, b, \theta)$ through their joint posterior distribution. For this problem, we only need to specify a prior distribution for a, b and θ because the K_i and p_i are specified by (1) and (2), respectively, and the likelihood is given by (3) and (4) for the convenience and random data, respectively.

We specify the following priors:

- $\tilde{p} = \frac{a}{a+b} \sim Beta(\tilde{a}, \tilde{b})$
- $\nu = a + b \sim Gamma(c, d)$
- $\theta \sim Lognormal(t_0, t_1)$

Choices of $\tilde{a}, \tilde{b}, c, d, t_0, t_1$ can be made so that the prior distributions are relatively flat and do not drive the results. By letting $t_0 = 0$, the median of the θ prior is 1 in which case the convenience sample data are not biased. We prefer the $(\tilde{p}, \nu)$ parameterization for the beta parameters because $E(p_i|\tilde{p}, \nu) = \tilde{p}$. To summarize, the unnormalized posterior density is

$$\begin{aligned} \pi(\tilde{p}, \nu, \theta, p_1, \ldots, p_M, K_1, \ldots, K_M | y_{c1}, \ldots, y_{cM}, y_{r1}, \ldots, y_{rM}) = \\ \tilde{p}^{a-1}(1-\tilde{p})^{b-1}\nu^{c-1}\exp(-d\nu)(\theta_1)^{-1}\phi\left(\frac{\log\theta - t_0}{t_1}\right) \\ \times \prod_{i=1}^{M}\left\{\frac{\Gamma(\nu)}{\Gamma(\nu\tilde{p})\Gamma(\nu(1-\tilde{p}))}p_i^{\nu\tilde{p}-1}(1-p_i)^{\nu(1-\tilde{p})-1}\right.\end{aligned}$$

$$\times \begin{pmatrix} N_i \\ K_i \end{pmatrix} p_i^{K_i}(1-p_i)^{N_i-K_i} \frac{\begin{pmatrix} n_{ci} \\ y_{ci} \end{pmatrix}\begin{pmatrix} N_i - n_{ci} \\ K_i - y_{ci} \end{pmatrix}\theta^{y_{ci}}}{\sum_{j=\max(0,n_{ci}-N_i+K_i)}^{\min(n_{ci},K_i)} \begin{pmatrix} n_{ci} \\ j \end{pmatrix}\begin{pmatrix} N_i - n_{ci} \\ K_i - j \end{pmatrix}\theta^{j}}$$

$$\left. \times \frac{\begin{pmatrix} n_{ri} \\ y_{ri} \end{pmatrix}\begin{pmatrix} N_i - n_{ci} - n_{ri} \\ K_i - y_{ci} - y_{ri} \end{pmatrix}}{\begin{pmatrix} N_i - n_{ci} \\ K_i - y_{ci} \end{pmatrix}} \right\}$$

To make the desired inference, we apply Bayes Theorem to obtain the joint posterior distribution of $\boldsymbol{\eta}$. Because there are $2M{+}3$ parameters (M p's, M K's, $\tilde{p}$, ν, θ), we employ an appropriate Markov Chain Monte Carlo (MCMC) method to sample from the joint posterior distribution (Gelman et al., 1995) from which inference about the unknown parameters of interest K_i can be made. For example, the Metropolis-Hastings algorithm (Chib and Greenberg (1995)) combined with Gibbs sampling (Casella and George (1992)) provide a general way to sample from the joint posterior distribution. WinBUGS (Spiegelhalter, Thomas, and Best (2000)) was not used because it cannot handle the extended-hypergeometric and hypergeometric distributions. For both cases in the example (Section 2), we implemented the MCMC method using the YADAS statistical modeling environment (Graves, 2001, 2003a,b). The MCMC algorithms used were based on the variable-at-a-time Metropolis–Hastings algorithm, which does not require the user to be able to evaluate or sample from full conditional distributions of one or more of the parameters given the others (these full conditional distributions are intractable in this problem). In variable-at-a-time Metropolis–Hastings algorithms, one loops over the unknown parameters, proposing a new value of one of the parameters, and deciding whether to accept the new value or remain in place with probability given by the unnormalized posterior distribution at the new value divided by the unnormalized posterior distribution at the old value. In this problem, the posterior draws generated using have high autocorrelation, but saving only one of every twenty iterations is enough to provide adequately independent samples. Another algorithm, based on continuous approximations to the discretely supported K_i, is a bit more efficient and discussed in Graves (2006).

Once we obtain draws from the joint posterior distribution of the K_i, we can provide inference about the remaining components in the population as follows. Let the current ith lot size be denoted by N_i^* (usually $N_i^* = N_i - n_{ci} - n_{ri}$). Let $K_i^* = K_i - y_{ci} - y_{ri}$, define $p^* = \sum K_i^* / \sum N_i^*$ and report the posterior distribution of p^*. We obtain draws from the posterior distribution on the number of components remaining in a lot with the attribute present (K_i^*) and the overall attribute proportion (p^*) of components with the attribute present remaining in the population.

Finally, because some lots have both random and convenience data, there is information about the biasing parameter θ which can be assessed through the posterior distribution for θ.

5 Example Revisited

The Bayesian analysis described in the previous section was performed on the convenience and random samples from the simulated population presented in the discussion. Note that for this population, the p_i are distributed as $Beta(1, 9)$ whose 0.025, 0.500 and 0.975 quantiles are 0.003, 0.074, 0.336, respectively. Also, for the Case 1 and Case 2 convenience samples, $\theta = 2$ and $\theta = 5$, respectively. The following priors were used in the analysis: $\frac{a}{a+b} \sim Beta(0.3, 1.7)$, $a + b \sim Gamma(2, 5)$, $\log\theta \sim Normal(0, 1)$. In the practical application in which we applied these methods, the prior parameters for a, b to attain specified quantiles of the prior distribution of p_i. These quantiles were obtained from historical data on similar features. (In particular, the priors were not chosen with the goal of being noninformative.) The simulated example, of course, has no such historical data, and we have used a relatively noninformative prior for p_i: the variance is about 2.33 times the mean(1-mean); a pure beta prior with this property would be weighted as strongly as 1.33 data points. Results with inevitably be somewhat sensitive to the prior on θ because the data has little information about this parameter: we included the parameter in our analysis not so we could accurately estimate θ but so we could allow for the possibility that the sampling mechanism was not equivalent to random sampling, and increase our uncertainty accordingly. The prior for θ implies that 95% of its mass is between about 1/7 and 7, and if the convenience samples are small, according to the discussion in Section 3, the prior assigns probability .95 to the event that π_1/π_0 is between about 1/7 and 7. This argues that the prior for θ is also fairly noninformative if one believes, as we did, that the amount of biasing is likely to be considerably less than $\pi_1/\pi_0 = 7^{\pm 1}$. A prior for $\log\theta$ symmetric about zero is equally appropriate for situations where the sampling mechanism is less likely to

*** talk about diagnostics for both cases ***

For Case 1, see the resulting posteriors obtained for $\frac{a}{a+b}$, $a+b$, θ and p^* in Figure 1. The posterior 0.05, 0.50, 0.95 quantiles for p^* are 0.073, 0.130, 0.209, respectively. Thus, an estimate for p^* using the median of the posterior is 0.130 compared with the true p^* of 0.101. Note that the point estimate is closer to the fraction of the random samples with the feature ($6/50 = 0.12$) than the fraction of the convenience samples ($16/100 = 0.16$). Posterior medians for the p_i are included in Table 3. This table collapses all lots with the same data pattern into a single row. For each observed (N, n_c, y_c, n_r, y_r), the number of lots with that data pattern is listed; also, the p_i posterior medians are presented from analyses when θ is estimated as well as when the potential biasing is ignored by fixing θ to equal one. Note that when several lots have the same data pattern, their median p_i were averaged. The posterior for θ is not wildly different from the prior: the posterior mean and standard deviation of $\log \theta$ are 0.29 and 0.54, as compared to zero and one. This means that the data weakly suggest that some biasing is present.

When $\theta = 1$ is assumed, i.e., the convenience samples are not biased, larger estimates of p^* result; the posterior 0.05, 0.50, 0.95 quantiles for p^* are 0.101, 0.146, and 0.197, respectively, which illustrate the impact of ignoring the biasing. The point estimate of 0.146, then, is very close to the naive estimate of $22/150 = 0.147$. Table 3 also shows how the posterior medians of the individual p_i's have increased.

For Case 2, see the resulting posteriors obtained for $\frac{a}{a+b}$, $a+b$, θ and p^* in Figure 2. The posterior 0.05, 0.50, 0.95 quantiles for p^* are 0.053, 0.096, and 0.168, respectively. Thus, an estimate for p^* using the median of the posterior is 0.096 compared with the true p^* of 0.101. When $\theta = 1$ is assumed, the posterior 0.05, 0.50, 0.95 quantiles for p^* are 0.138, 0.185, and 0.239, respectively, which illustrate the seriousness of ignoring the substantial biasing in this case. Table 4 presents the data patterns (N, n_c, y_c, n_r, y_r) and p_i posterior medians for this case.

6 Discussion

Estimating a proportion in a population can be challenging when the population is stratified by lots produced by a common manufacturing process and the available data arise from both random and convenience samples. Moreover, all the lots may not have been sampled. In this paper, statistical models have been proposed for the population which reflect the stratification by lots and for the small samples taken from small lots which, in the case of the convenience samples, are biased. We have shown how a Bayesian approach appropriately handles the desired inferences, especially for providing an estimate of the proportion in the remaining population.

There are a number of topics for future research. These include accounting for the amount of biasing in the convenience samples varying by lot and handling continuous rather than binary attributes of interest. Also, how to use this modeling approach to make recommendations for additional random sampling of the remaining population is of interest. The first two topics require appropriate modeling changes. This paper provides a basis for the third topic: the need to analyze the resulting data. A Bayesian approach is an obvious choice.

Acknowledgements

We thank Dee Won for her support and encouragement of this work, and the engineers in the Engineering Sciences and Applications Division. We also thank Steve Thompson for his insightful comments on an earlier version of this paper.

References

Casella, G. and George, E. (1992). "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167–174.

Chib, S. and Greenberg, E. (1995). "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 49, 327–335.

Gelman, A.B., Carlin, J.S., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*, Boca Raton: Chapman and Hall/CRC.

Graves, T.L. (2001). "YADAS: An Object-Oriented Framework for Data Analysis Using Markov Chain Monte Carlo," Los Alamos National Laboratory Technical Report LA-UR-01-4804.

Graves, T.L. (2003a). "A Framework for Expressing and Estimating Arbitrary Statistical Models Using Markov Chain Monte Carlo," Los Alamos National Laboratory Technical Report LA-UR-03-5934.

Graves, T.L. (2003b). "An Introduction to YADAS," `yadas.lanl.gov`.

Graves, T.L. (2006). "MCMC Algorithms for Correlated Discrete and Continuous Parameters," Los Alamos National Laboratory Technical Report.

Harkness, W.L. (1965). "Properties of the Extended Hypergeometric Distribution." *Annals of Mathematical Statistics* 36, 938–945.

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

Nandram, B. and Choi, J.W. (2002),"A Bayesian Analysis of a Proportion Under Non-Ignorable Non-Response," *Statistics in Medicine*, 21, 1189–1212.

Rubin, D.B. (1976). "Inference and Missing Data." *Biometrika* 63, 581–592.

Spiegelhalter, D., Thomas, A. and Best, N. (2000). *WinBUGS Version 1.3 User Manual*.

Stasny, E.A. (1991),"Hierarchical Models for the Probabilities of a Survey Classification and Nonresponse: an Example from the National Crime Survey," *Journal of the American Statistical Association*, 86, 296–303.

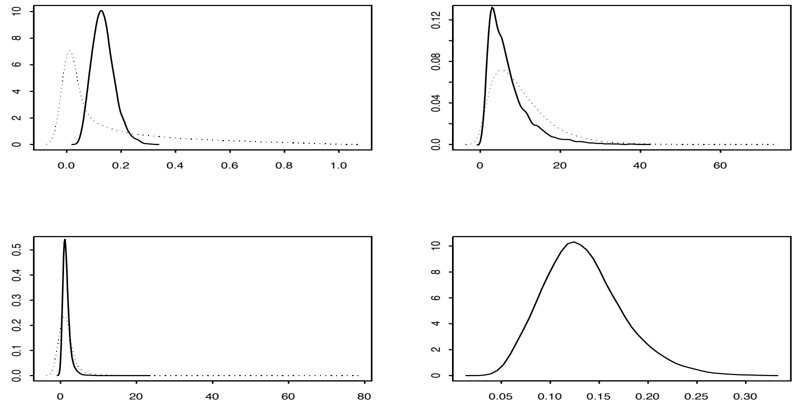

Figure 1: Case 1 posteriors for $\frac{a}{a+b}$, $a+b$, θ and p^* when θ is estimated. The priors (dotted) are overlaid in the first three plots.

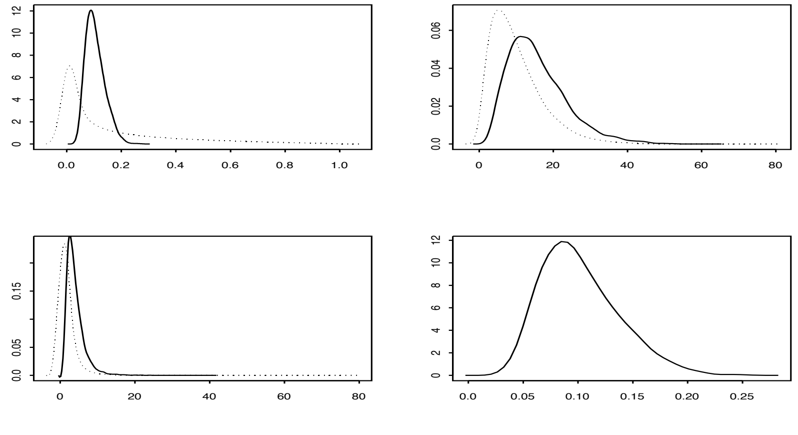

Figure 2: Case 2 posteriors for $\frac{a}{a+b}$, $a+b$, θ and p^* when θ is estimated. The priors (dotted) are overlaid in the first three plots.

Table 1: Example Population and Sample Data

					Case 1		Case 2							Case 1		Case 2	
Lot	N	K	n_c	n_r	y_c	y_r	y_c	y_r	Lot	N	K	n_c	n_r	y_c	y_r	y_c	y_r
1	10	0	0	0	0	0	0	0	61	10	2	0	0	0	0	0	0
2	10	0	0	0	0	0	0	0	62	10	1	0	0	0	0	0	0
3	10	4	0	0	0	0	0	0	63	10	0	0	0	0	0	0	0
4	10	3	0	1	0	0	0	0	64	10	1	0	0	0	0	0	0
5	10	3	1	0	1	0	1	0	65	10	0	0	0	0	0	0	0
6	10	0	0	0	0	0	0	0	66	10	2	0	1	0	0	0	0
7	10	0	0	0	0	0	0	0	67	10	5	0	0	0	0	0	0
8	10	1	0	0	0	0	0	0	68	10	0	0	0	0	0	0	0
9	10	3	0	0	0	0	0	0	69	10	1	0	0	0	0	0	0
10	10	1	0	0	0	0	0	0	70	10	0	0	0	0	0	0	0
11	10	0	0	0	0	0	0	0	71	10	0	0	0	0	0	0	0
12	10	1	2	0	0	0	1	0	72	10	0	0	0	0	0	0	0
13	10	3	0	0	0	0	0	0	73	10	2	0	0	0	0	0	0
14	10	3	0	0	0	0	0	0	74	10	2	0	0	0	0	0	0
15	10	2	0	0	0	0	0	0	75	10	3	1	0	0	0	1	0
16	10	0	0	0	0	0	0	0	76	10	0	0	2	0	0	0	0
17	10	3	0	0	0	0	0	0	77	10	0	0	0	0	0	0	0
18	10	1	0	0	0	0	0	0	78	10	1	0	0	0	0	0	0
19	10	0	1	0	0	0	0	0	79	10	0	1	0	0	0	0	0
20	10	0	0	0	0	0	0	0	80	10	1	0	0	0	0	0	0
21	10	1	0	0	0	0	0	0	81	10	3	0	0	0	0	0	0
22	10	0	1	0	0	0	0	0	82	10	3	1	1	0	1	1	0
23	10	6	0	0	0	0	0	0	83	10	0	0	0	0	0	0	0
24	10	2	0	0	0	0	0	0	84	10	0	0	0	0	0	0	0
25	10	0	0	0	0	0	0	0	85	10	1	0	0	0	0	0	0
26	10	0	0	0	0	0	0	0	86	10	2	0	0	0	0	0	0
27	10	0	0	0	0	0	0	0	87	10	1	1	0	0	0	0	0
28	10	2	0	0	0	0	0	0	88	10	0	0	0	0	0	0	0
29	10	5	0	0	0	0	0	0	89	10	3	1	1	1	1	0	0
30	10	1	0	0	0	0	0	0	90	10	1	0	0	0	0	0	0
31	10	0	0	0	0	0	0	0	91	10	0	0	0	0	0	0	0
32	10	0	0	0	0	0	0	0	92	10	0	0	0	0	0	0	0
33	10	0	0	0	0	0	0	0	93	10	2	0	0	0	0	0	0
34	10	1	0	0	0	0	0	0	94	10	1	0	0	0	0	0	0
35	10	0	0	0	0	0	0	0	95	10	0	0	1	0	0	0	0
36	10	0	0	0	0	0	0	0	96	10	1	0	0	0	0	0	0
37	10	1	0	0	0	0	0	0	97	10	0	0	0	0	0	0	0
38	10	3	0	0	0	0	0	0	98	10	1	2	0	1	0	0	0
39	10	1	0	0	0	0	0	0	99	10	3	0	0	0	0	0	0
40	10	0	0	0	0	0	0	0	100	10	1	0	0	0	0	0	0
41	10	0	0	0	0	0	0	0	101	25	3	1	0	0	0	0	0
42	10	0	0	0	0	0	0	0	102	25	1	1	0	0	0	0	0
43	10	1	0	0	0	0	0	0	103	25	1	0	1	0	0	0	0
44	10	2	0	0	0	0	0	0	104	25	0	0	0	0	0	0	0
45	10	0	1	0	0	0	0	0	105	25	4	0	0	0	0	0	0
46	10	1	0	0	0	0	0	0	106	25	7	1	0	1	0	1	0
47	10	1	0	0	0	0	0	0	107	25	0	1	0	0	0	0	0
48	10	1	0	1	0	0	0	0	108	25	14	0	0	0	0	0	0
49	10	1	1	0	1	0	0	0	109	25	0	1	0	0	0	0	0
50	10	3	0	0	0	0	0	0	110	25	2	0	0	0	0	0	0
51	10	0	1	0	0	0	0	0	111	25	1	1	0	0	0	0	0
52	10	1	0	0	0	0	0	0	112	25	1	1	1	0	0	0	0
53	10	1	1	0	0	0	0	0	113	25	5	1	0	1	0	0	0
54	10	1	0	0	0	0	0	0	114	25	1	1	0	0	0	0	0
55	10	0	0	0	0	0	0	0	115	25	3	1	0	1	0	1	0
56	10	1	1	0	0	0	0	0	116	25	1	1	0	0	0	0	0
57	10	1	0	0	0	0	0	0	117	25	5	0	0	0	0	0	0
58	10	0	0	0	0	0	0	0	118	25	6	0	0	0	0	0	0
59	10	4	0	1	0	0	0	0	119	25	4	2	0	0	0	1	0
60	10	0	0	0	0	0	0	0	120	25	1	0	0	0	0	0	0

Table 1 (Continued): Example Population and Sample Data

					Case 1		Case 2	
Lot	N	K	n_c	n_r	y_c	y_r	y_c	y_r
121	25	2	1	0	0	0	0	0
122	25	0	0	0	0	0	0	0
123	25	0	2	0	0	0	0	0
124	25	1	1	1	0	0	1	0
125	25	0	1	1	0	0	0	0
126	25	0	0	0	0	0	0	0
127	25	3	0	0	0	0	0	0
128	25	1	0	0	0	0	0	0
129	25	4	0	2	0	1	0	1
130	25	1	0	0	0	0	0	0
131	25	6	0	0	0	0	0	0
132	25	0	0	0	0	0	0	0
133	25	5	1	0	0	0	1	0
134	25	1	0	0	0	0	0	0
135	25	5	0	0	0	0	0	0
136	25	1	1	0	0	0	0	0
137	25	2	0	0	0	0	0	0
138	25	4	0	1	0	0	0	0
139	25	0	2	1	0	0	0	0
140	25	5	0	0	0	0	0	0
141	25	2	1	0	0	0	0	0
142	25	0	1	0	0	0	0	0
143	25	5	0	1	0	1	0	0
144	25	3	0	0	0	0	0	0
145	25	5	0	1	0	0	0	0
146	25	2	0	0	0	0	0	0
147	25	0	0	0	0	0	0	0
148	25	3	1	0	0	0	1	0
149	25	1	0	0	0	0	0	0
150	25	6	0	0	0	0	0	0
151	25	4	0	0	0	0	0	0
152	25	0	1	0	0	0	0	0
153	25	0	2	1	0	0	0	0
154	25	0	0	0	0	0	0	0
155	25	1	1	0	0	0	0	0
156	25	4	0	1	0	0	0	0
157	25	8	1	0	1	0	1	0
158	25	0	0	0	0	0	0	0
159	25	0	3	0	0	0	0	0
160	25	3	1	0	0	0	1	0
161	25	2	0	0	0	0	0	0
162	25	1	0	1	0	0	0	0
163	25	9	0	1	0	0	0	0
164	25	1	0	0	0	0	0	0
165	25	2	0	3	0	0	0	1
166	25	0	1	0	0	0	0	0
167	25	0	1	0	0	0	0	0
168	25	3	0	0	0	0	0	0
169	25	0	0	0	0	0	0	0
170	25	1	1	0	0	0	0	0
171	25	1	0	0	0	0	0	0
172	25	1	1	1	0	0	0	0
173	25	8	0	0	0	0	0	0
174	25	9	0	0	0	0	0	0
175	25	1	4	0	0	0	1	0
176	25	3	0	0	0	0	0	0
177	25	0	1	1	0	0	0	0
178	25	0	1	0	0	0	0	0
179	25	7	0	0	0	0	0	0
180	25	1	1	1	0	0	0	0
181	25	2	0	0	0	0	0	0
182	25	1	0	0	0	0	0	0
183	25	0	0	0	0	0	0	0
184	25	4	0	0	0	0	0	0
185	25	1	0	0	0	0	0	0
186	25	0	1	1	0	0	0	0
187	25	3	0	0	0	0	0	0
188	25	1	0	0	0	0	0	0
189	25	0	1	0	0	0	0	0
190	25	0	0	0	0	0	0	0
191	25	1	0	0	0	0	0	0
192	25	1	1	0	0	0	0	0
193	25	0	0	0	0	0	0	0
194	25	2	1	0	0	0	1	0
195	25	9	1	0	1	0	1	0
196	25	0	0	0	0	0	0	0
197	25	9	2	0	2	0	1	0
198	25	0	0	0	0	0	0	0
199	25	8	0	1	0	0	0	0
200	25	1	0	0	0	0	0	0
201	50	6	0	1	0	0	0	0
202	50	7	1	1	1	0	1	0
203	50	1	2	3	0	0	0	0
204	50	6	0	1	0	0	0	0
205	50	0	0	0	0	0	0	0
206	50	3	2	0	0	0	1	0
207	50	0	1	0	0	0	0	0
208	50	9	1	1	0	1	1	0
209	50	10	0	3	0	0	0	0
210	50	8	1	0	1	0	1	0
211	50	7	0	0	0	0	0	0
212	50	0	1	0	0	0	0	0
213	50	6	3	0	0	0	0	0
214	50	10	3	0	0	0	0	0
215	50	2	1	0	0	0	0	0
216	50	16	3	0	2	0	2	0
217	50	1	2	0	0	0	0	0
218	50	6	0	0	0	0	0	0
219	50	5	0	0	0	0	0	0
220	50	5	1	0	0	0	0	0
221	50	5	0	1	0	0	0	0
222	50	7	1	3	0	1	0	1
223	50	6	0	1	0	0	0	0
224	50	0	2	1	0	0	0	0
225	50	5	1	0	0	0	0	0
226	50	8	1	1	1	0	1	0
227	50	6	0	0	0	0	0	0
228	50	9	0	0	0	0	0	0
229	50	2	4	2	0	0	1	0
230	50	2	0	0	0	0	0	0

Table 2: Extended-hypergeometric Probabilities for a lot with $N = 10$, $K = 5$, $n = 3$ ($\theta = 1$ corresponds to hypergeometric probabilities arising from completely random sampling)

	θ					
y	0.1	0.5	1.0	1.5	2.0	5.0
0	0.645	0.205	0.083	0.043	0.026	0.004
1	0.322	0.513	0.417	0.324	0.256	0.091
2	0.032	0.256	0.417	0.487	0.513	0.453
3	0.001	0.026	0.083	0.146	0.205	0.453

Table 3: Case 1 Data Patterns and p_i Posterior Medians

Pattern						p_i Posterior Median	
N	nc	yc	nr	yr	Count	θ estimated	$\theta = 1$ fixed
50	4	0	2	0	1	0.033	0.049
50	2	0	3	0	1	0.040	0.055
25	4	0	0	0	1	0.043	0.059
50	3	0	0	0	2	0.048	0.066
25	3	0	0	0	1	0.049	0.067
25	2	0	1	0	2	0.050	0.067
50	2	0	1	0	1	0.051	0.067
50	0	0	3	0	1	0.053	0.068
25	0	0	3	0	1	0.054	0.067
25	2	0	0	0	2	0.057	0.078
10	2	0	0	0	1	0.059	0.074
25	1	0	1	0	7	0.060	0.075
50	2	0	0	0	2	0.060	0.078
10	0	0	2	0	1	0.062	0.073
25	1	0	0	0	23	0.068	0.088
10	1	0	0	0	9	0.069	0.088
50	1	0	0	0	5	0.070	0.088
50	0	0	1	0	4	0.071	0.089
10	0	0	1	0	5	0.072	0.088
25	0	0	1	0	7	0.072	0.088
10	0	0	0	0	79	0.087	0.105
25	0	0	0	0	48	0.087	0.105
50	0	0	0	0	7	0.088	0.105
50	1	0	3	1	1	0.155	0.166
10	2	1	0	0	1	0.191	0.211
10	1	0	1	1	1	0.197	0.214
50	1	0	1	1	1	0.198	0.21
50	1	1	1	0	2	0.198	0.21
25	0	0	2	1	1	0.200	0.21
10	1	1	0	0	2	0.232	0.248
25	1	1	0	0	5	0.232	0.247
50	1	1	0	0	1	0.235	0.244
25	0	0	1	1	1	0.242	0.253
50	3	2	0	0	1	0.274	0.302
25	2	2	0	0	1	0.330	0.348
10	1	1	1	1	1	0.338	0.342
5000	100	16	50	6	230		

Table 4: Case 2 Data Patterns and p_i Posterior Medians

Pattern						p_i Posterior Median	
N	nc	yc	nr	yr	Count	θ estimated	$\theta = 1$ fixed
50	2	0	3	0	1	0.050	0.115
50	3	0	0	0	2	0.050	0.13
25	3	0	0	0	1	0.052	0.13
50	2	0	1	0	1	0.055	0.129
25	2	0	1	0	2	0.056	0.13
50	2	0	0	0	1	0.058	0.139
25	2	0	0	0	1	0.060	0.14
10	2	0	0	0	1	0.061	0.141
10	1	0	1	0	1	0.064	0.139
25	1	0	1	0	6	0.064	0.14
50	0	0	3	0	1	0.065	0.13
25	1	0	0	0	20	0.067	0.153
50	1	0	0	0	5	0.067	0.153
10	1	0	0	0	9	0.068	0.152
10	0	0	2	0	1	0.071	0.14
25	0	0	1	0	8	0.075	0.153
10	0	0	1	0	5	0.076	0.152
50	0	0	1	0	4	0.076	0.153
10	0	0	0	0	79	0.081	0.166
25	0	0	0	0	48	0.081	0.166
50	0	0	0	0	7	0.081	0.167
50	4	1	2	0	1	0.083	0.165
25	4	1	0	0	1	0.095	0.187
50	1	0	3	1	1	0.111	0.187
50	2	1	0	0	1	0.113	0.22
25	2	1	0	0	2	0.116	0.22
10	2	1	0	0	1	0.120	0.218
50	1	1	1	0	3	0.125	0.219
25	1	1	1	0	1	0.127	0.219
10	1	1	1	0	1	0.128	0.218
25	0	0	3	1	1	0.129	0.207
50	1	1	0	0	1	0.131	0.239
25	1	1	0	0	8	0.133	0.238
10	1	1	0	0	2	0.136	0.237
25	0	0	2	1	1	0.138	0.218
50	3	2	0	0	1	0.153	0.273
5000	100	24	50	3	230		