Interval Estimation for Small Area Proportions with Small True Proportions from Stratified Random Sampling Survey Data[*][†]

Carolina Franco.[‡] Partha Lahiri[§]

Abstract

Consider interval estimation of m small area proportions P_i $(i = 1, \cdots, m)$, where we assume a stratified random sampling design with equal number of observations n in each stratum, and where the domains of interest are the strata. A $100(1-\alpha)\%$ confidence interval for P_i that has appeared repeatedly in the literature and is used in application is given by $\hat{P}_i^{EB} \pm z_{\alpha/2}\sqrt{mse_i}$, where $\hat{P}_i^{EB}$ and mse_i are an empirical Bayes estimator of P_i and an associated second-order unbiased mean squared error estimator $(i = 1, \cdots, m)$. In the case where no covariates are available, the underlying model is $p_i|P_i \overset{ind.}{\sim} \mathcal{N}(P_i, \psi_i)$, $P_i \overset{ind.}{\sim} \mathcal{N}(\mu, A)$, where p_i is the sample proportion for domain i $(i = 1, \ldots, m)$; ψ_i are *known* sampling variances; and μ and A are unknown hyperparameters. We refer to models that use the normality assumption on both levels of the hierarchy as "Normal-Normal Models." The well-documented problems of the normal approximation to the binomial raise questions about the accuracy of confidence intervals based on the Normal-Normal model above when the domain sample sizes are small or when the true domain proportions are close to 0 or 1. We argue that a more reasonable model in this setting is a beta-binomial model in which the sampled stratum counts have binomial distributions and the prior distribution of the true stratum proportions follows a beta distribution. The Beta-Binomial Model has also previously appeared in the literature as a candidate for modelling small area proportions. We examine a new empirical Bayes confidence interval based on this model. We perform simulation studies under the Beta-Binomial Model that compare the peformance of this CI and an alternative CI constructed using the Normal-Normal Model.

Key Words: Complex Survey, Small Area Estimation, Proportions, Stratified Random Sampling, Empirical Bayes, Confidence Intervals, Credible Intervals, Coverage

1. Introduction

Consider interval estimation for domain proportions from data collected from a stratified random sampling survey, where the domains of interest are the strata and where the characteristic underlying the proportions is a rare event, so that all stratum proportions are small. The focus is on two-sided intervals. We assume a simple random sample is taken from each domain of interest.

Stratified random sampling designs tend to be particularly efficient in situations where the stratum means differ significantly from each other. Here we assume all stratum proportions are small, but the strata are sampled separately to ensure that data is collected from all domains of interest, so that the domains of interest are the strata. Such a situation may arise, for instance, when studying the proportion of people with a rare disease in each of several domains. For simplicity we assume equal stratum sample sizes (i.e., $n_i = n$).

[*] This report is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau

[†] This research was partially supported by NIH Grant R01 CA129101 to P. Lahiri

[‡] Bureau of the Census, 4600 Silver Hill Rd, Suitland, MD 20746

[§] Joint Program in Survey Methodology, 1218 Lefrak Hall, University of Maryland, College Park, MD 20742

Although small area models benefit greatly from the presence of relevant covariates, here we will assume that covariates are not available but that we still wish to "borrow strength" from different domains. For examples in the literature of such an approach, see Efron and Morris (1975), and Carter and Rolph (1974).

A $100(1-\alpha)\%$ confidence interval for the true proportion P_i of domain i that has appeared in the literature is $\hat{P}_i^{EB} \pm z_{\alpha/2}\sqrt{mse_i}$, where $\hat{P}_i^{EB}$ and mse_i are an empirical Bayes estimator of P_i and an associated second-order unbiased mean squared error estimator $(i = 1, \ldots, m)$. Many different choices of $\hat{P}_i^{EB}$ and mse_i are discussed in the literature. The specific estimators we will study here are described in Section 2. The underlying model for this method uses normal approximations: $p_i|P_i \overset{ind.}{\sim} N(P_i, \psi_i)$, $P_i \overset{ind.}{\sim} N(\mu, A)$, where p_i is the sample proportion for domain i based on a sample of size n_i, $i = 1, \ldots, m$, ψ_i are known smoothed sampling variances, and μ and A are unknown hyperparameters.

The Normal-Normal Model discussed above appears in Carter and Rolph (1974), where it is used to estimate the probability that an alarm signals a structural fire. A generalization of the model is the famous Fay-Herriot model (1979), where the authors exploit the availability of covariates to "borrow strength" when estimating a domain's proportion. Some of the many other papers that use Normal-Normal models for proportions are Morris (1995), Morris (1983), and Liu, Lahiri, and Kalton (2007). Moreover, a Normal-Normal model is used to estimate poverty at the state level by the Census Bureau's Small Area Income and Poverty (SAIPE) program–a linear Gaussian Fay-Herriot Model is used (National Research Council Report, 1997).

Underlying the Normal-Normal model discussed in this paper is the notion that the normal distribution is effective for approximating the binomial. As is well-known to statisticians, the normal approximation to the binomial distribution can be problematic when the probability of success is in the extremes and the number of Bernoulli trials associated with the binomial random variable is small. Brown et al. (2001, 2002) show that the actual coverage of the Wald interval, given by $Y/r \pm z_{\alpha/2}\sqrt{(Y/r)(1-Y/r)/n}$, may fall well below the nominal coverage in several examples. These include cases where p is not close to 0 or 1 and where rp and $r(1-p)$ are greater than 10, a rule of thumb sometimes given in introductory statistics books. In fact, the coverage tends to oscillate both as r increases with p being fixed and as p varies with r fixed, making the coverage for a particular problem difficult to predict. This phenomenon is due to the discreteness and skewness of the binomial distribution. The erratic coverage properties of the Wald interval raises questions about the performance of the Normal-Normal CI when the underlying true proportions are small.

It should be noted that Brown et al. (2001) discuss several other methods of constructing confidence intervals for proportions, all of which tend to perform better than the Wald interval when considering the coverage probability over the range of p. One of note is the Wilson interval, also based on a normal approximation. Like the Wald interval, the Wilson interval is derived from an asymptotic pivot. For the Wald interval the pivot is $(p-\hat{p})/\sqrt{\hat{p}(1-\hat{p})/r}$ and for the Wilson interval the pivot is $(p-\hat{p})/\sqrt{p(1-p)/r}$. The Wilson interval also shows oscillation in the true coverage as p and r vary, but it is less severe than that of the Wald interval. It is one of the two intervals recommended by Brown et al. for small sample sizes. Although both intervals are based on a normal approximations to the binomial, the Wald Interval bears a greater resemblance to the Normal-Normal CI, which is of the form $\hat{P}_i^{EB} \pm z_{\alpha/2}\sqrt{mse_i}$.

Furthermore, the support of the Normal distribution is the real line whereas a proportion must be between 0 and 1. When the true proportions are small the distribution of $p_i|P_i$ may assign significant probability to negative regions. The same is true of the distribution of the P_i under the Normal-Normal Model. Some of the choices of hyperparameters in our simulations in Section 4, which correspond to the scenarios of interest in this paper, could lead to such problems.

Negative values of $\hat{P}_i^{EB}$ can be avoided through truncating the estimators for the hyperparameter A. Further truncation may be needed to ensure the lower bound of the confidence interval is nonnegative. The need for truncation may have an impact on the coverages of the confidence intervals.

In a stratified random sampling setting with small stratum sample sizes, it is more reasonable to assume the sampled domain counts Y_i, given P_i, follow a binomial distribution, particularly when the sampling fraction is small or the sampling is with replacement, and when the proportions are in the extremes.

To apply an empirical Bayes approach, we must specify the distribution of the true proportions P_i. The beta distribution is a reasonable choice since its support is $(0, 1)$ and since its shape varies greatly with different choices of parameters, allowing some flexibility in the model. In an empirical Bayes approach to confidence interval construction, the hyperparameters, and thus the shape of the distribution, are determined by the data.

Thus, a reasonable model in this setting is:

$$Y_i|P_i \overset{ind.}{\sim} Bin(n, P_i), \tag{1}$$

and

$$P_i \overset{ind.}{\sim} Beta(a, b). \tag{2}$$

The purpose of this paper is to study the properties of a CI based on the Normal-Normal model and of a CI based on the Beta-Binomial model under the assumption that the latter model is the true model.

Many authors have considered beta-binomial models for inference on proportions in small area estimation problems. Examples are Ghosh and Lahiri (1987), Ghosh and Maiti (2004), Gilary et al. (2012), among others. Empirical and hierarchical Bayes small area beta-binomial models for domain proportions are also considered in Rao (2003). The method of estimating the hyperparameters studied here is new.

In Section 2 we give more details on the specific Normal-Normal CI studied in this paper. In Section 3 we derive an empirical Bayes confidence interval based on the assumed model given by (1) and (2), which we call the Beta-Binomial confidence interval. In Section 4 we compare through simulation studies the Normal-Normal CI to the Beta-Binomial CI, under the assumption that the true model is given by (1) and (2).

2. Normal-Normal Empirical Bayes CI

The Normal-Normal CI rests on the following assumptions:

$$\text{Level 1:} \quad p_i|P_i \overset{ind.}{\sim} N(P_i, \psi_i), \tag{3}$$

$$\text{Level 2:} \quad P_i \overset{ind.}{\sim} N(\mu, A). \tag{4}$$

The model is a special case of the famous Fay-Herriot Model (1979) and is closely related to the Efron-Morris model (1975), and to the models described in Carter and Rolph (1974). In particular, the Normal-Normal is a special case of one of four models examined in the paper by Carter and Rolph, which considers the case where the domain sample sizes are not necessarily equal. The model is used to analyze a fire alarm dataset. Three other related models are also discussed, including a two level model using an arcsine variance stabilizing transformation. Variants of the Fay-Herriot model are frequently used in surveys in small area estimation problems.

Level 1 in equation (3) is usually called the sampling model and level 2 in equation (4) is usually referred to as the linking model (Jiang and Lahiri, 2006). The sampling variabilities ψ_i are assumed to be known, although they typically need to be estimated. This is a weakness of the Fay-Herriot model since it does not incorporate the uncertainty due to estimation of ψ_i.

Returning to the empirical Bayes setup defined by (3) and (4), we note that because the Normal distribution is its own conjugate prior, the posterior distribution of $P_i|p_i$ is normal with mean

$$\gamma_i p_i + (1 - \gamma_i)\mu, \tag{5}$$

where

$$\gamma_i = \frac{A}{A + \psi_i}. \tag{6}$$

The parameter γ_i is called the shrinkage factor. Note that γ_i determines weights applied to the area-specific estimator and the prior mean μ. An estimator for P_i is given by

$$\hat{\gamma}_i p_i + (1 - \hat{\gamma}_i)\bar{p}, \tag{7}$$

where

$$\bar{p} = \frac{\sum_{i=1}^{m} p_i}{m}, \tag{8}$$

and where an estimator for γ_i will be given subsequently.

A two-sided $100(1 - \alpha)\%$ empirical Bayes CI is given by:

$$(\hat{P}_i^{EB} \pm z_{\alpha/2}\sqrt{mse_i}), \tag{9}$$

where $z_{\alpha/2}$ represents the appropriate quantile of the standard normal distribution; see Rao (2003). To estimate ψ_i we used

$$\hat{\psi}_i = (\bar{p})(1 - \bar{p})/n_i = (\bar{p})(1 - \bar{p})/n = \hat{\psi}. \tag{10}$$

The sampling variances ψ_i, are treated as known and equal to $\hat{\psi}$ in this paper. As in Carter and Rolph (1974), and Morris (1983) we estimate the sampling variance by (10). In this case because $n_i = n$ this results in equal estimators for the ψ_i. This approach is used here because this estimator is more stable than estimators based on the data from only a single domain, since $p_i(1 - p_i)/n$ should be highly variable when the domain sample sizes are small. The variability of $\hat{\psi}$ should be smaller although it should have a higher bias if the ψ_i are not approximately equal. Estimating the ψ_i by formula (10) implicitly assumes that $\psi_i = \mu(1-\mu)/n$.

Another approach would be to perform a variance stabilizing transformation. If covariates were available, better approaches for estimating the ψ_i may be possible. For instance, one may estimate the ψ_i using a generalized variance function (see Wolter 1985, Chapter 5).

The following estimators are used for A, γ_i, and the mse_i:

$$\hat{A} = \max\left\{0, (m-1)^{-1}\sum_{i=1}^{m}(p_i-\bar{p})^2 - \hat{\psi}\right\}, \tag{11}$$

$$\hat{\gamma}_i = \hat{\gamma} = \frac{\hat{A}}{\hat{\psi}+\hat{A}}, \tag{12}$$

$$mse_i^{EB} = g_1(\hat{A}) + g_2(\hat{A}) + 2g_3(\hat{A}), \tag{13}$$

where

$$g_1(\hat{A}) = \hat{\gamma}\hat{\psi}, \tag{14}$$

$$g_2(\hat{A}) = \left(\frac{\hat{\psi}}{\hat{\psi}+\hat{A}}\right)^2 \left(\sum_{j=1}^{m}\frac{1}{\hat{\psi}+\hat{A}}\right)^{-1} = \frac{\hat{\psi}^2}{m(\hat{\psi}+\hat{A})}, \tag{15}$$

and

$$g_3(\hat{A}) = \left[\frac{(1-\hat{\gamma})^2}{\hat{\psi}+\hat{A}}\right]\widehat{Var(\hat{A})} = \frac{2(1-\hat{\gamma})^2(\hat{\psi}+\hat{A})}{m}. \tag{16}$$

Discussion of formulas (11-16) can be found in Rao (2003) or Jiang and Lahiri (2006). The sum of (14) and (15) alone, as an estimator of mse_i, would be a naive estimator because it would not account for the uncertainty due to the estimation of A (Rao, 2003). If one were to use the sum of these two terms alone the bias would be of order $O(1/m)$. The estimator (13) was proposed by Prasad and Rao (1990) and has bias of order $o(1/m)$ assuming $\hat{\psi} = \psi$ is the true parameter.

The estimator for A given by (11) truncates the unbiased estimator $(m-1)^{-1}\sum_{i=1}^{m}(p_i-\bar{p})^2 - \hat{\psi}$ whenever it is negative. Such truncation guarantees that $\hat{P}_i^{EB}$ cannot be negative. However, it does not guarantee that the lower bound of the interval is nonnegative, so that the latter may need to be truncated as well.

3. Beta-Binomial Empirical Bayes CI

The Beta-Binomial confidence interval is built under the model given by (1) and (2). As previously discussed, the underlying model assumptions are more reasonable than the normal distribution assumptions, particularly in the cases of interest where the normal approximation to the binomial is inappropriate.

Since the beta distribution is a conjugate prior for the binomial, the posterior distribution $P_i|Y_i$ follows a beta distribution. If a and b were known, a credible interval for P_i would be

$$L_i = B(\alpha/2, y_i + a, n - y_i + b) \tag{17}$$

$$U_i = B(1 - \alpha/2, y_i + a, n - y_i + b) \tag{18}$$

Under our proposed method of estimation, to estimate the hyperparameters a and b we first estimate δ, given by

$$\delta = \frac{1}{a + b + 1}. \tag{19}$$

The hyperparameter δ specifies a relationship between the prior mean μ and the prior variance σ^2:

$$\sigma^2 = \mu(1 - \mu)\delta. \tag{20}$$

The hyperparameters μ and σ^2 can be expressed in terms of a and b, where

$$\mu = \frac{a}{a + b}, \quad \sigma^2 = \frac{a - 1}{a + b - 2}.$$

It is easy to see that δ has the property that $0 < \delta < 1$. Moreover, δ is directly proportional to the prior variance, so the larger the δ the less confidence it reflects on the prior distribution, i.e., the less informative the prior.

We estimate δ through the following equation:

$$\left[1 - \frac{MSW}{\bar{p}(1 - \bar{p})} - \delta\right] + \frac{C}{\delta} = 0 \tag{21}$$

where

$$MSW = \frac{n}{m(n - 1)} \sum_{i=1}^{m} p_i(1 - p_i)$$

is the mean squared error within. The above equation can be solved in closed form, and it has two solutions, one which is negative and the other one which is:

$$\hat{\delta} = \frac{K + \sqrt{K^2 + 4C}}{2}, \tag{22}$$

where $K = 1 - MSW/(\bar{p}(1 - \bar{p}))$.

When n is fixed and $m \to \infty$, as is typically assumed in small area estimation problems due to the fact that typically the domain sample size n is much smaller than the number of domains m, $MSW \xrightarrow{p} \mu(1 - \mu)(1 - \delta)$ and thus $K \xrightarrow{p} \delta$. This is easy to verify by computing $E_M(MSW)$. Note that $\hat{\delta}$ is consistent provided $C = C_m = o(1)$.

Based on extensive empirical evidence gathered through a large number of simulations, $\hat{\delta}$ appears to have the property that it is always in the desired range (i.e., $0 < \hat{\delta} < 1$) if C is appropriately chosen (C is a small positive constant). A similar technique was used by Gabler et al. (2011) in the estimation of intra-cluster correlation for the balanced one-way random effects model.

According to our simulation results, the values of C that give good coverage depend on the true parameters and on n. In our simulation studies, different values of C were examined for each table displayed in Section 4 to obtain satisfactory results. Our simulation studies suggest that in cases of interest it is possible to find a C that works well for a range of prior means and variances in the sense that the coverages nearly meet or exceed the nominal coverage. To find a C that works for a particular application, preliminary simulations are needed. This issue requires further investigation.

We could also have estimated δ more simply by:

$$1 - MSW/(\bar{p}(1-\bar{p})).$$

This corresponds to $C = 0$. The problem with this method of estimation is that it frequently yields values for δ that are outside the admissible range, particularly when μ, n, and m are small. In fact, the frequency with which δ is out of the range increases when μ approaches 0 (or 1), as n decreases and as m decreases, as is illustrated by Figure 1, which is based on simulations under the assumed model. Figure 1 shows the proportion of values of $\hat{\hat{\delta}}$ that were outside the range $0 < \hat{\hat{\delta}} < 1$. Simulations not shown here also revealed that with $C = 0$ there is significant undercoverage of the corresponding intervals.

It should be noted that by continuity of $\hat{\hat{\delta}}(C)$, there should be values of C that are small enough so that there is a positive probability of $\hat{\hat{\delta}}$ being outside the desired range. However, this was not the case with the values which were used in our simulations in Section 4.

One could arbitrarily set $\hat{\hat{\delta}}$ to be a particular constant, such as .5, whenever an inadmissible value is obtained, but this method results in undercoverage of the corresponding confidence intervals, according to our simulations.

Estimators for a and b are derived from the relations between a, b, μ and δ, as follows:

$$\hat{a} = \bar{p}\left(\frac{1}{\hat{\hat{\delta}}} - 1\right) \tag{23}$$

and

$$\hat{b} = (1-\bar{p})\left(\frac{1}{\hat{\hat{\delta}}} - 1\right). \tag{24}$$

Care must be exercised to select a C that is appropriate for the cases of interest. A poor choice of C may result in coverages that are far below the nominal.

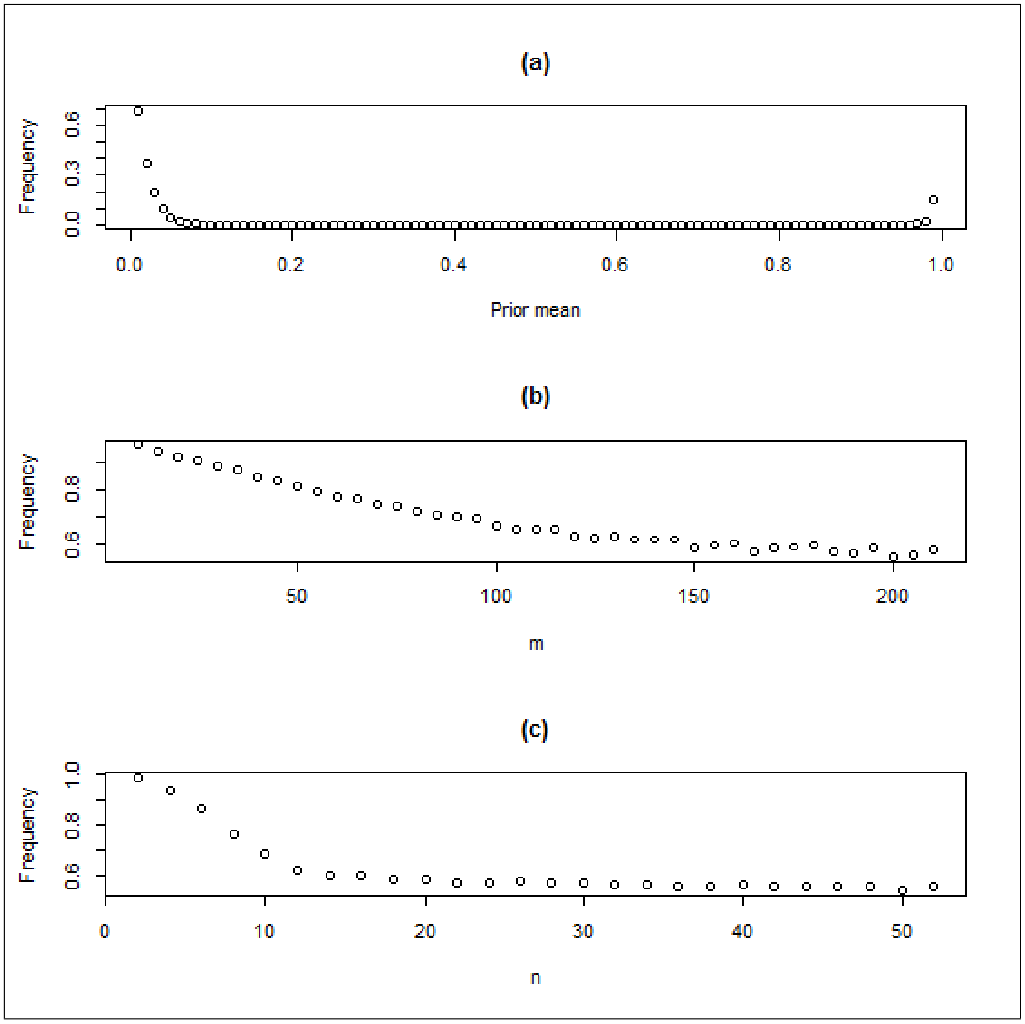

Figure 1: Relative frequency of $\hat{\delta}$ out of range for various parameters: Simulation results for $N = 10,000$ replications. (a) $m = 100$, $n = 10$, $\sigma^2 = .0099$, (b) $\mu = .01$, $\sigma^2 = .0099$, $n = 10$, (c) $\mu = .01$, $\sigma^2 = .0099$, $m = 100$. In each case C of (21) is set equal to zero.

4. Simulation Results

For each replication, we generated data using the Beta-Binomial model, with $Y_i|P_i \sim$ Bin(n, P_i) and $P_i \sim$ Beta(a, b) for a variety of choices of m, n, μ, and σ^2, focusing primarily on small-area examples with small μ. We computed coverages (computed as the proportion of replications that capture the true domain proportion) and average lengths for the Beta-Binomial CI and for the Normal-Normal CI for one domain. Each simulation was performed for N replications, where N is typically 1,000 or 10,000 depending on the desired accuracy. All our CI's have a nominal 95% coverage and are two-sided.

4.1 Robust C for Fixed n and for a Range of Small Prior Means and Variances

Table 1 displays the coverages and average lengths for the Beta-Binomial CI and for the Normal-Normal CI for domain 1.

As was previously mentioned, the optimal value of C for estimating δ depends on the prior parameters, and in fact when C is inappropriately chosen the coverages of the Beta-Binomial CI can be quite poor, according to our simulations. Table 1 suggests it is possible to choose a C that provides coverage that always nearly meets or exceeds the nominal level for a range of small prior means with small prior variances. However, we note that for very small values of μ in this table (i.e., $\mu < .01$), the Beta-Binomial CI can show significant overcoverage and higher average lengths than the Normal-Normal CI with this value of C. It should be possible to choose a C that provides lower coverages and average lengths for $\mu < .01$ for the Beta-Binomial CI, but this may result in undercoverage for $\mu > .01$.

In many cells in Table 1, the Normal-Normal CI falls below 90% coverage, although the nominal level is 95%. This table also suggests that the Normal-Normal coverage may be oscillatory as μ increases. However, we note that the oscillation and undercoverage are subtle compared to those observed for the Wald Interval in Brown et al. (2001).

Table 1: Two-sided coverages/average lengths for Beta-Binomial CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $m = 200$, $n = 20$, $c = .0001$, $\alpha = 0.05$.

σ^2 μ	0.001	.00001	.00000001
0.002	0.998 / 0.014 0.939 / 0.008	0.997 / 0.013 0.926 / 0.009	0.993 / 0.013 0.927 / 0.0089
0.005	0.987 / 0.024 0.904 / 0.017	0.985 / 0.024 0.899 / 0.019	0.986 / 0.024 0.912 / 0.019
0.01	0.975 / 0.038 0.878 / 0.032	0.976 / 0.038 0.903 / 0.034	0.971 / 0.038 0.887 / 0.034
0.02	0.969 / 0.061 0.945 / 0.062	0.951 / 0.062 0.927 / 0.063	0.955 / 0.063 0.924 / 0.063
0.03	0.957 / 0.083 0.935 / 0.088	0.95 / 0.084 0.941 / 0.089	0.945 / 0.084 0.938 / 0.089
0.04	0.946 / 0.1 0.95 / 0.11	0.946 / 0.1 0.946 / 0.11	0.946 / 0.1 0.94 / 0.11
0.05	0.943 / 0.12 0.934 / 0.13	0.952 / 0.12 0.94 / 0.13	0.954 / 0.12 0.952 / 0.13

4.2 Variation in m and n

Table 2 illustrates the impact on the coverages as m increases with n is small, and $\mu = .01$. Very large values of m are included to study the behavior of the coverage as m increases and n is fixed. In this situation, the Normal-Normal CI can exhibit undercoverage even when m is very large. The coverage of the Normal-Normal CI as m increases seems to be oscillatory when holding everything else fixed.

The Beta-Binomial CI performs well throughout Table 2 in terms of meeting or exceeding the nominal coverage, although again it shows significant overcoverage and high average lengths with $C = .001$. It may be possible to choose a C that performs better, but the purpose of this table was to study the performance of the Normal-Normal CI as m varies.

Table 2: Two-sided coverages/average lengths for Beta-Binomial CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $\mu = 0.01$, $\sigma^2 = .0001$, $c = 0.001$, $\alpha = 0.05$.

n m	3	5	10	20
100	0.976 / 0.1 0.901 / 0.055	0.979 / 0.061 0.915 / 0.048	0.993 / 0.053 0.931 / 0.043	0.979 / 0.048 0.918 / 0.036
500	0.992 / 0.059 0.882 / 0.037	0.994 / 0.059 0.857 / 0.036	0.991 / 0.055 0.85 / 0.033	0.994 / 0.047 0.923 / 0.034
1000	0.996 / 0.061 0.776 / 0.033	0.996 / 0.059 0.792 / 0.033	0.993 / 0.055 0.886 / 0.033	0.988 / 0.047 0.95 / 0.034
5000	0.998 / 0.062 0.755 / 0.029	0.997 / 0.06 0.875 / 0.033	0.995 / 0.054 0.939 / 0.035	0.996 / 0.048 0.951 / 0.034
8000	0.994 / 0.062 0.785 / 0.029	1 / 0.06 0.892 / 0.032	0.996 / 0.056 0.934 / 0.035	0.991 / 0.047 0.947 / 0.034
10000	0.994 / 0.062 0.793 / 0.029	0.998 / 0.06 0.909 / 0.033	0.997 / 0.054 0.954 / 0.035	0.986 / 0.046 0.93 / 0.035

Table 3 provides a wider range of n. Column 3, in particular, suggests that the coverage of the Normal-Normal may also be oscillatory as n increases. In these tables, however, the oscillations are much less pronounced than that of the Wald Interval for the probability of success of a single binomial random variable, most likely because the former interval "borrows strength" from other domains when estimating a domain proportion. In Table 3 we can also see that the appropriate value for C for the Beta-Binomial CI can depend on n. In this example, coverage for the Beta-Binomial CI seems to decrease as n increases, when everything else is fixed.

Table 3: Two-sided coverages/average lengths for Beta-Binomial CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $\mu = 0.01$, $\sigma^2 = .0001$, $c = 0.001$, $\alpha = 0.05$.

m n	100	300	500
3	0.983 / 0.11 0.905 / 0.053	0.992 / 0.058 0.9 / 0.038	0.994 / 0.061 0.895 / 0.037
5	0.984 / 0.059 0.913 / 0.048	0.997 / 0.058 0.914 / 0.039	0.991 / 0.058 0.875 / 0.036
10	0.991 / 0.053 0.926 / 0.041	0.992 / 0.054 0.877 / 0.036	0.996 / 0.054 0.859 / 0.034
20	0.985 / 0.046 0.919 / 0.036	0.991 / 0.046 0.905 / 0.034	0.987 / 0.047 0.924 / 0.034
30	0.982 / 0.042 0.915 / 0.033	0.98 / 0.042 0.915 / 0.033	0.972 / 0.042 0.93 / 0.033
40	0.982 / 0.038 0.924 / 0.032	0.971 / 0.039 0.928 / 0.032	0.983 / 0.038 0.948 / 0.032
50	0.963 / 0.035 0.925 / 0.031	0.972 / 0.036 0.951 / 0.031	0.956 / 0.036 0.928 / 0.031
100	0.962 / 0.028 0.941 / 0.027	0.958 / 0.028 0.949 / 0.027	0.958 / 0.029 0.948 / 0.027
500	0.947 / 0.015 0.937 / 0.016	0.939 / 0.014 0.944 / 0.016	0.954 / 0.015 0.936 / 0.016

4.3 Unlucky n?

Brown et al. (2001) show that the Wald-interval for building a confidence interval for the probability of success based on one binomial(r, p) observation can have poor coverage even when rp is fairly large. We investigate whether this phenomenon extends to our scenario. We select some of the "unlucky" pairings of (p, r) from Brown *et al.* and set our prior mean μ to equal their p and our domain sample size n to be equal to the corresponding value of r.

Table 4 shows the Normal-Normal CI can have undercoverage even when m and n are both large. The values $n = 592$ and $n = 954$ correspond to an example given Brown et al. (2001) to illustrate that the Wald interval can fail to yield the desired coverages even when rp and rq are large (a value in between these two was also included). The binomial probability of success in their example is $p = .005$, and we set our prior mean accordingly. Although the undercoverage is slight, it may be surprising due to the large n. However, this table also suggests that "borrowing strength" may improve coverage. For instance, the coverage of the Wald interval with a 95% nominal level for $n = 592$ and $p = .005$, as displayed in Brown et al. 2001, is below 80%. It should be noted that other direct intervals should perform better than the Wald interval with these parameters, according to the findings of Brown et al., such as the Wilson interval.

Another interesting observation is that in this Table the Beta-Binomial CI's average lengths seem to be slightly smaller than the Normal-Normal average lengths, despite the fact that the coverages of the latter are inferior. For this particular simulation we increased N to 10,000 to increase the accuracy, since the undercoverage is of a lesser magnitude.

Table 4: Two-sided coverages/average lengths for Beta-Binomial CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 10000$ replications, $\mu = 0.005$, $\sigma^2 = .00001$, $c = .00001$, $\alpha = 0.05$.

n m	592	700	954
50	0.953 / 0.009 0.937 / 0.0098	0.951 / 0.0083 0.936 / 0.0092	0.952 / 0.0073 0.94 / 0.0081
100	0.953 / 0.009 0.936 / 0.0098	0.952 / 0.0084 0.936 / 0.0092	0.955 / 0.0073 0.938 / 0.0081
200	0.955 / 0.0089 0.939 / 0.0098	0.954 / 0.0084 0.941 / 0.0092	0.955 / 0.0073 0.94 / 0.0081

4.4 When the True Proportion is Not Close to the Extremes

The Normal-Normal CI does not show undercoverage in all cases. Table 5 examines the situation where μ is not in the extremes, with $\mu = .4$. In this table, the Normal-Normal CI does not show undercoverage and in fact shows some overcoverage, and larger average lengths than the Beta-Binomial CI. With this choice of C and of the hyperparameters, the Beta-Binomial CI has coverages that are fairly close to the nominal.

Table 5: Two-sided coverages/average lengths for Beta-Binomial CI (top numbers in each cell) and Normal-Normal CI (bottom numbers in each cell) for one domain for $N = 1000$ replications, $\mu = 0.4$, $\sigma^2 = 0.03$, $c = 0.035$, $\alpha = 0.05$.

m n	50	100	1000
3	0.963 / 0.65 0.944 / 0.71	0.961 / 0.65 0.932 / 0.69	0.967 / 0.66 0.939 / 0.67
5	0.955 / 0.56 0.967 / 0.65	0.969 / 0.56 0.976 / 0.64	0.952 / 0.56 0.962 / 0.63
10	0.965 / 0.44 0.978 / 0.52	0.952 / 0.44 0.966 / 0.52	0.961 / 0.44 0.975 / 0.51
20	0.951 / 0.33 0.972 / 0.4	0.952 / 0.33 0.974 / 0.4	0.95 / 0.33 0.974 / 0.39
30	0.953 / 0.28 0.972 / 0.33	0.952 / 0.27 0.973 / 0.33	0.951 / 0.28 0.969 / 0.33
40	0.964 / 0.24 0.981 / 0.29	0.947 / 0.24 0.968 / 0.29	0.953 / 0.24 0.965 / 0.29
50	0.933 / 0.21 0.967 / 0.26	0.951 / 0.21 0.968 / 0.26	0.943 / 0.22 0.97 / 0.26
100	0.947 / 0.16 0.966 / 0.19	0.945 / 0.15 0.972 / 0.19	0.945 / 0.15 0.973 / 0.19
1000	0.944 / 0.05 0.968 / 0.06	0.948 / 0.05 0.972 / 0.061	0.954 / 0.05 0.977 / 0.061

5. Discussion

In this paper, we compared two models that have been used in the literature and in application to deal with small area proportions, the Normal-Normal model and the Beta-Binomial Model. Focusing on stratum proportion interval estimation for rare events from data collected from stratified random sampling surveys, we argued that the Beta-Binomial model is more reasonable in this situation, and we performed simulations based on this model to compare the performance of hierarchical Bayes' CI's based on each of the two models. The specific Normal-Normal CI uses estimators that have previously been used in the literature, while we have introduced a new CI based on the Beta-Binomial model. Of course, because the simulations were performed under the Beta-Binomial model, the conclusions hold if the statistician believes this model is appropriate for the application of interest.

Some deficiencies in using the Normal-Normal model for the setup of interest are evident. The binomial approximation to the normal distribution may not be appropriate with small domain sample sizes and/or small true proportions. Since the support of the Normal distribution is the real line, the distributions of $p_i|P_i$ and P_i may assign significant probabilities to negative regions, particularly when the true proportions are small. Due to these issues, one should not expect this model and method of confidence interval construction to perform particularly well when the true proportions are very small. Some of the undercoverage displayed in the tables in Section 4 may be explained by these issues. Despite that, the coverage problems of the Normal-Normal CI were much milder than those of the Wald interval as displayed in Brown et al. (2001).

The Binomial-Beta CI has the weakness that the statistician must have an idea of the true proportions and their variability in order to select the value of C that attains a good coverage. We have shown a certain degree of robustness of the choice of C, although to maintain the nominal coverage throughout a wide range of small proportions one must choose a C that results in overcoverage for very small values of the prior mean μ, according to our simulations. The appropriate value of C also depends on n. This value may be ascertained through simulations for a specific application, and further investigation is needed to provide guidelines on the choice of C.

Interesting areas for future research are to adapt the Binomial-Beta CI discussed here to surveys with more complex designs, and to incorporate covariates into the model.

REFERENCES

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*, 26, 2, pp. 101-133.

Brown, L. D., Cai, T. T., and DasGupta, A. (2002). Confidence Intervals for a Binomial Proportion and Asymptotic Expansions. *The Annals of Statistics*, 30, 1, pp. 160-201.

Carter, M. C., and Rolph, J. E. (1974) Empirical Bayes Methods to Estimating Fire Alarm Probabilities. *Journal of the American Statistical Association*, 69, 348, pp. 880-885.

Efron, B. and Morris, C. (1975). Data Analysis Using Stein's Estimator and its Generalizations. *Journal of the American Statistical Association*, 70, 350, pp. 311-319.

Fay, R. E. and Herriot, R. A. (1979) Estimates of Income for Small Places: An application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*. 74, 366, pp. 269-277.

Gabler, S. Ganninger, M., and Lahiri, P. (2011). A strictly Positive Estimator of Intra-Cluster Correlation for the One-Way Random Effects Model. *Preprint*.

Gilary, A., Maples, J. and Slud, E.V. (2012). Small Area Confidence Bounds on Small Cell Proportions in Survey Populations. To appear in *Proceedings of the American Statistical Association Survey Research Methods Section.*

Ghosh, M. and Lahiri, P. (1987). Robust Empirical Bayes Estimation of Means from Stratified Samples. *Journal of the American Statistical Association*, 82, 400, pp. 1153-1162.

Ghosh, M. And Maiti, T. (2004) Small-Area Estimation Based on Natural-Exponential Family Quadratic Variance Function Models and Survey Weights. *Biometrika* 91, 1, pp. 95-112.

Jiang, J., and Lahiri, P. (2006). Mixed Model Prediction and Small Area Estimation. *Sociedad de Estadistica e Investigacion Operativa*, 15, 1, pp. 1-96.

Jiang, J., and Lahiri, P. (2001). Empirical Best Prediction for Small Area Inference with Binary Data. *Ann. Inst. Statist. Math.*,53, 2, pp. 217-24

Korn, E. L., and Graubad, B. I. (1998). Confidence Intervals for Proportions with Small Expected Number of Positive Counts Estimated from Survey Data. *Survey Methodology*, 24, 2, pp. 193-201.

Liu, Y. K., and Kott, P. S. (2009). Evaluating Alternative One-Sided Coverage Intervals for a Proportion. *Journal of Official Statistics*, 25, 4, pp. 569-588.

Liu, B. Lahiri, P. and Kalton, G. (2007) *JSM Proceedings, Survey Research Section*. pp. 3181-3186.

Lohr, L., L. (2010). *Sampling: Design and Analysis, 2nd Ed.*. Boston: Brooks/Cole.

National Research Council (1997), Small Area Estimates of Children in Poverty, Interim Report 1, Evaluation of 1993 Estimates for Title I Allocations, Constance F. Citro, Michael L. Cohen, Graham Kalton, and Kirsten K. West, eds., Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics, Washington, DC: National Academy Press.

Morris, C. N. (1995). Hierarchical Models for Educational Data: An Overview. *Journal of Educational and Behavioral Statistics*, 20, 2, pp. 190-200.

Morris, C. N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*, 78, 381, pp. 47-55.

Prasad, N. G. N., and Rao, J. N. K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association* 85, pp. 163-171.

Rao, J. N. K. (2003). *Small Area Estimation*. New York: Wiley.

Slud, E. V. (2012). Assessment of Zeroes in Survey-Estimated Tables via Small-Area Confidence Bounds. *Journal of the Indian Society of Agricultural Statistics*. 66, 2. pp. 157-169.

Wolter, K. (1985). *Introduction to Variance Estimation*. New York: Springer.