Coverage Properties of Confidence Intervals for Proportions in Complex Sample Surveys *

Carolina Franco† Roderick J. A. Little ‡ Thomas A. Louis§ Eric V. Slud ¶

Abstract

The challenges of constructing confidence intervals for a binomial proportion and the deficiencies of the popular Wald interval in achieving its nominal coverage — particularly when the true proportion is close to 0 or 1 or when the sample size is moderate — have been well documented (Brown et al. 2001, 2002). The problem is further compounded when inference is based on complex survey data. Yet intervals resembling the Wald interval are often applied to complex surveys, an example being the confidence intervals used in the American Community Survey (ACS). In the literature, confidence intervals designed for binomial proportions with modifications based on the design effect are often used for complex surveys. Here, we adopt this approach and study the coverage and expected length properties of 7 different intervals. We focus on how phenomena such as clustering, stratification, misspecification of variances, and patterns of variation of stratum expected sampling fractions and stratum survey attribute-proportions, affect coverage. A simulation study examines the effect of such factors.

Key Words: Complex Surveys; Proportions; Confidence Intervals; Wald Interval; Design Effect; Coverage

1. Introduction

Constructing well-calibrated confidence intervals for population proportions based on survey data presents challenges unless the sample size is quite large. By well-calibrated, we mean that the interval's coverage is close to nominal across different values of the true proportion. Consider a Simple Random Sample (SRS) for a large population where the sampling fraction is negligible. Then, the problem is that of constructing a confidence interval for a binomial proportion. In this case, "exact" confidence intervals, which guarantee that the nominal coverage is met or exceeded, are available in the literature (see, for instance, Clopper and Pearson 1934, Byth and Still 1983, and Casella, 1986)[1]. However, these tend to be conservative; they often exceed the nominal coverage and have large expected lengths. Moreover, methods that guarantee nominal coverage for the binomial case do not necessarily do so when applied to complex survey data.

Exact tests aside, it has been well documented (Brown et al. 2001, 2002) that constructing confidence intervals for a binomial proportion that consistently achieve coverage that is close to nominal can be elusive unless the sample size is large, where what constitutes a "large" sample depends on how close the true proportion is to zero or one. In particular, Brown et al. highlight erratic coverage as a function of n and p of the Wald interval for a binomial proportion p based on Y successes, where $Y \sim Bin(n, p)$. They also propose alternative intervals which perform better, but still oscillate in their coverage as n and p

*This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

†US Census Bureau

‡University of Michigan

§US Census Bureau and Johns Hopkins University

¶US Census Bureau and Univerisity of Maryland, College Park

[1]For exact confidence intervals for small populations, see Wright, 1991 and Buonaccorsi,1987

vary. Namely, they propose the Jeffreys interval and the Wilson interval for small sample sizes along with the Agresti-Coull interval for large sample sizes (see Brown et al. 2001, Section 5).

The Wald interval is defined as $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$, where $\hat{p} = Y/n$. In complex surveys, intervals of the form

$$\hat{p} \pm z_{\alpha/2} SE, \tag{1}$$

where $\hat{p}$ is a survey-weighted estimator of the true proportion and SE is some estimate of its standard error, are often used. The Wald interval is a special case of (1) for SRS surveys, and its poor performance when the sample sizes are moderate raises questions about the use of intervals of the form (1) for more complex surveys, where the misspecification of the binomial assumption further confounds the problem. Examples of surveys that use intervals of the form (1) are ubiquitous, but this study was inspired by the American Community Survey (ACS). The method currently used by the ACS is of the form (1), where $\hat{p}$ is the survey-weighted estimator and the standard error is computed based on the Successive Differences Replication (SDR) Method (Fay and Train, 1995).

The ACS is the largest household sample survey in the United States, sampling approximately 3.5 million addresses yearly. It asks questions about demographic characteristics, income, education, disabilities, and health insurance, among many others. Despite its large sample size, cross-classification often leads to small sample sizes for domains of interest. The ACS is the source of billions of estimates. Due to the sheer quantity and diversity of estimates produced by the ACS, it may be inpractical to consider sophisticated methods for use across the board. Here, we consider only simple methods that would be easy to implement in ACS production. The only inputs for all methods considered are the sample size, the survey-weighted estimate of the proportion, and a measure of its standard error.

In the literature, a common approach to constructing intervals for proportions based on complex sample survey data is to modify methods designed for binomial proportions based on the design effect/effective sample size (see, for instance, Korn and Grabaurd, 1998, Kott and Liu 2009). The design effect is defined as $\textit{Deff} = \textit{Var}(\hat{p})/(p(1-p)/n)$, and the effective sample size is $n^* = n/\textit{Deff}$. The effective sample size n^* can be interpreted as the sample size one would need under an SRS design to obtain the same asymptotic CI width obtained under the complex sampling scheme. In practice, p and $\textit{Var}(\hat{p})$ are unknown and must be estimated, for instance, by the survey-weighted estimate $\hat{p}$ and a design-based estimate of its variance $\widehat{\textit{Var}}(\hat{p})$. The estimated design effect, defined as $\widehat{\textit{Deff}} = \widehat{\textit{Var}}(\hat{p})/(\hat{p}(1-\hat{p}))/n$, is used when computing the estimated effective sample size, $\hat{n}^* = n/\widehat{\textit{Deff}}$. Then n is replaced by $\hat{n}^*$ and the binomial count Y by $\hat{p} \cdot \hat{n}^*$ in formulas defining the bounds of confidence intervals for binomial proportions.

Korn and Graubard (1998) study confidence intervals for proportions in complex surveys when the expected number of successes is small. They study four intervals based on design-effect modifications, including an interval of the form (1), present simulation results, and compare the intervals using two applications. They recommend the use of the Clopper-Pearson interval with a modification to the design effect that heuristically adjusts for the variability of the variance of the estimator. Kott and Liu (2009) study one-sided intervals for proportions for SRS and stratified SRS surveys. As they note, the intervals that perform best in the one-sided case may not be the same as those that perform best in the two-sided case because the latter can have compensating one-sided errors due to the asymmetric distribution of $\hat{p}$. Here, we focus on two-sided intervals to conform to ACS practice.

In this paper, we perform a simulation that aims to control and measure the impact of the main factors that are thought to influence the performance of intervals, such as clustering, stratification, the degree of heterogeneity within and between clusters and among strata

in the population, etc. We also study the degree to which misspecification of the design effect leads to problems in coverage. We study the performance of several intervals across many scenarios through the use of a simulation with a factorial design with 864 different configurations. Some of the features of the simulation are inspired by the ACS and may be particularly relevant to household surveys, especially in cases where the sample consists of many relatively small clusters, although the results should be of general interest.

We focus on coverages that are conditional on not obtaining an estimate of 0 (or 1) for the proportion of interest. The estimated design effect is undefined in the case of zero estimates, which inevitably mandates arbitrary fix-ups in simulation studies. Instead, we regard the issue of zero counts as a separate but closely related problem, beyond the scope of this paper. In the ACS, zero counts are handled through a separate methodology (US Census Bureau, 2009), and treating their occurrence as a separate problem is consistent with that practice.

In Section 2, we discuss the details of our simulation, including a short description of 6 intervals that are candidates to improve upon the performance of the Wald interval. Our perspective is that a "good" interval generally provides close to nominal coverage without having a wastefully long expected length. In Section 3 we collect the results of the simulation, providing conclusions and suggestions for further research in Section 4.

2. A Simulation Inspired by the ACS

2.1 Simulating the Population

We first create a population of size N. The population has $J = 4$ strata, and in the jth stratum there are K_j clusters of fixed size c, and N_j units, with $N_j = c * K_j$. In alternative runs we let $c = 1$ or $c = 3$. Populations with many relatively small clusters are common for household surveys, where all members of a sampled household are often surveyed. For simplicity, we let all clusters be of equal size c and all strata contain the same number K of clusters.

The population expected proportion θ is specified within each simulation. It is apportioned into strata population proportions θ_j either equally, or growing in a linear fashion. For instance, if $\theta = 0.3$ we could have $\theta_j = 0.3$ for all j or we could have $\theta_j = 0.075, 0.225, 0.375, 0.525$. This enables the study of different patterns of variation of the expected stratum proportion and the stratum sampling fraction. More details about the specific linear relationship among the stratum proportions will be given in the next section.

A "success" in the population, with indicator variable Y_{jkl}, where j is the stratum, k is the cluster, and l is the unit within the cluster, is simulated in a hierarchical fashion as follows:

$$\begin{aligned} Y_{jkl}|p_{jk} &\sim \text{Bernoulli}(p_{jk}) \\ p_{jk}|\pi_j &\sim \text{Beta}(\tau\pi_j, \tau(1-\pi_j)) \\ \pi_j &\sim \text{Uniform}(\theta_j(1-\gamma_0),\, \theta_j(1+\gamma_0)). \end{aligned} \tag{2}$$

The parameters τ, γ_0 and $\theta_j, j = 1, \ldots, J$ are specified within each simulation, and they control features of interest within the population. For instance, notice that τ controls the degree of heterogeneity within clusters. The quantity $1/(1+\tau)$ is the intra-cluster correlation for the binary survey attribute of interest when $c > 1$. A large τ implies large within-cluster heterogeneity relative to that between clusters. The last level in the hierarchy given by expression (2) allows for further variation in the proportions than is stipulated by the linear relationship between the θ_j. This is particularly interesting in the case where the θ_j are constant over strata. The simulation parameter γ_0 can be set to 0 if no additional variation is desired.

2.2 Simulating the Sampling Design

After the population has been generated, it is sampled N_r times. The sampling design features stratification, and when a cluster is selected, all its units are selected. This setup is inspired by the ACS, with a cluster playing the role of a household.
We specify the overall sampling fraction f and the stratum sampling fractions $f_j, j = 1, \ldots, J$ as follows:

$$\theta_j = \theta \left(1 + b_0 \, (j \, - \, \frac{J+1}{2})\right) \quad , \quad f_j \propto \theta_j^{a_1} \, (2\theta - \theta_j)^{a_2} \quad , \quad 1 \leq j \leq J.$$

Note that as previously mentioned, the θ_j are either constant or increase linearly as j increases. The quantities b_0, a_1, and a_2 can be specified in each simulation run, and they are chosen to conform to four scenarios of interest regarding the relationship between the stratum sampling fractions and the expected stratum population proportions:

C:	$b_0 = a_1 = a_2 = 0$	level θ's, f's
I:	$b_0 = 3/(2J-2)$, $a_1 = 1$, $a_2 = 0$	$\nearrow$ θ's, $\nearrow$ f's
D:	$b_0 = 3/(2J-2)$, $a_1 = 0$, $a_2 = 1$	$\nearrow$ θ's, $\searrow$ f's
H:	$b_0 = 3/(2J-2)$, $a_1 = 1$, $a_2 = 1$	$\nearrow$ θ's, $\cap$ shaped f's.

The purpose of the choice $b_0 = 3/(2J-2)$ in the scenarios **I**, **D**, and **H**, when $J = 4$, is to fix the smallest of the ratios θ_j/θ as $1/4$, and the largest as $7/4$.

As an example, suppose $\theta = 0.3$ and the sampling fraction f was such that the overall sample size is $n = 80$. Then, under Scenario **C**, $n_j = 20, 20, 20, 20$, $\theta_j = 0.3, 0.3, 0.3, 0.3$, and under Scenario **D**, $n_j = 35, 25, 15, 5$, $\theta_j = 0.075, 0.225, 0.375, 0.525$. Note that with no clustering, scenario **C** would resemble most closely the case of an SRS. Under scenario **D**, on the other hand, there is large variability among stratum sample sizes, which in turn may lead to unstable estimates of the stratum variances for small overall sample sizes. This will have an impact on the estimation of the design effect, as will be discussed subsequently. The choices of parameters b_0 and γ_0 need to be constrained in order not to let attribute-proportions θ_j and π_j fall outside the range $(0, 1)$. Moreover, some rounding is necessary in the stratum sample sizes, so that the stratum and overal sampling fractions may differ slightly from those initially specified.

Once the stratum sample sizes have been determined, we take an SRS in each stratum accordingly for each of $N_r = 10,000$ simulation runs.

2.3 A Factorial Design

In the simulations presented in Section 3, some simulation parameters are fixed and some can take one of several values, creating a factorial design. We let the population size $N = 10,000$, and let the number of strata $J = 4$. Then the cluster size c is either 1 or 3, where $c = 1$ represents no clustering. The sample sizes are $n = 30, 40, 50, 80, 160, 240$, the population expected proportion $\theta = 0.05, 0.1, 0.3$ and the parameters related to the clustering and stratification are $\tau = 10, 10000$, $\gamma_0 = 0, 0.2, 0.4$. Again, there are four different scenarios that describe the relationship between the stratum sampling fractions and the stratum expected population proportions, denoted in Section 2.2 as **C**, **I**, **D**, and **H**. Note that we are studying $2 * 6 * 3 * 2 * 3 * 4 = 864$ simulation-parameter combinations.

In each simulation, we compute coverage and expected width, conditional on not obtaining an estimate of 0 or 1 for p, for the Wald interval and for the six additional methods described in Section 2.5, using both n^* and $\hat{n}^*$. The details of how n^* and $\hat{n}^*$ are computed are given in the next section.

2.4 Population and Sample Quantities of Interest

Denote the population count in stratum j and cluster k as $Y_{jk\cdot}$ and the population count in stratum j as $Y_{\cdot\cdot j}$. That is,

$$Y_{jk\cdot} = \sum_{i \in C_{jk}} y_{jkl} \quad , \qquad Y_{\cdot\cdot j} = \sum_{k=1}^{K_j} Y_{jk\cdot} = \sum_{k=1}^{K_j} \sum_{l \in C_{jk}} y_{jkl}$$

where C_{jk} denotes the set of units belonging to cluster k in stratum j.

The population total Y of the y_{ikj} counts is $Y = \sum_{j=1}^{J} Y_{\cdot\cdot j}$, and the corresponding standard sample-weighted estimator, based on SRS samples S_j of n_j^C clusters is

$$\hat{Y} = \sum_{j=1}^{J} \frac{N_j}{c\, n_j^C} \sum_{k \in S_j} \sum_{l \in C_{jk}} y_{jkl} = \sum_{j=1}^{J} \frac{N_j}{c\, n_j^C} \sum_{k \in S_j} Y_{jk\cdot}$$

The confidence intervals that we study for the population proportion $\bar{Y} = Y/N$ are based on this point estimator together with the 'model' $n^{C,*}\,(\hat{Y}/N) \sim \text{Binom}(n^{C,*}, \bar{Y})$, where $n^{C,*}$ is an appropriately defined 'effective sample size'. Note that $\bar{Y}$ is the random frame-population average, with $E(\bar{Y})$ equal to the theoretical proportion θ.

2.4.1 Effective Sample Size

Two different effective sample sizes are used in our simulations, a 'true' one based on the actual simulated (frame) population, and one which must be estimated from sampled data as must be done with real survey data.

Using the notation n^C as the number of sampled clusters, $f = c\, n^C/N$, as the overall sampling fraction, K_j as the number of clusters in stratum j, and $f_j = c\, n_j^C/N_j = n_j^C/K_j$ as the sampling fraction within the jth stratum, we have the true variance for the survey estimator $\hat{Y}$ equal to

$$V(\hat{Y}) = \sum_{j=1}^{J} \frac{K_j^2\,(1-f_j)}{n_j^C}\, s^2_{Y\cdot xj} \quad , \quad s^2_{Y\cdot xj} = \frac{1}{K_j - 1} \sum_{k=1}^{K_j} \left(Y_{jk\cdot} - \frac{Y_{j\cdot\cdot}}{K_j} \right)^2$$

The corresponding sample-estimated variance is given as

$$\hat{V}(\hat{Y}) = \sum_{j=1}^{J} \frac{K_j^2\,(1-f_j)}{n_j^C}\, \hat{s}^2_{Y\cdot xj} \quad , \quad \hat{s}^2_{Y\cdot xj} = \frac{1}{n_j^C - 1} \sum_{k \in S_j} \left(Y_{jk\cdot} - \frac{\hat{Y}_{j\cdot\cdot}}{n_j^C} \right)^2$$

In terms of these variances, whether true (known) or estimated in the simulations, the (population-level) design effect is defined as

$$\texttt{Deff} = \frac{n\, V(\hat{Y})}{Y(N-Y)\,(1-f)} \quad , \qquad \widehat{\texttt{Deff}} = \frac{n\, \hat{V}(\hat{Y})}{\hat{Y}(N-\hat{Y})\,(1-f)}.$$

The standard definition of effective sample size, which we adopt in our simulation study both in a true and estimated version is:

$$n^* = \frac{Y(N-Y)(1-f)}{\sum_{j=1}^{J}(K_j^2/n_j^C)(1-f_j)\, s^2_{Y\cdot xj}} \quad , \quad \hat{n}^* = \frac{\hat{Y}(N-\hat{Y})(1-f)}{\sum_{j=1}^{J}(K_j^2/n_j^C)(1-f_j)\, \hat{s}^2_{Y\cdot xj}}.$$

We impose the artificial fix-up in our simulations that estimated effective sample sizes are never allowed to be less than 5. In addition, we adopt the fix-up that $\hat{n}^* \leq 2\,n$. This is to avoid absurdities resulting from extreme values of the estimated design effect.

2.5 Alternatives to the Wald Interval for a Binomial Proportion

We consider 6 different alternatives to the Wald interval in our simulation. The Clopper-Pearson interval is a well-known interval which is generally regarded as conservative in the Binomial case. The Agresti-Coull interval is recommended by Brown et al. (2001) for large sample sizes and for its parsimony, and for moderately sized intervals the Wilson and Jeffreys intervals are recommended. The Jeffreys interval is actually derived from a Bayesian perspective, so that it is a credible interval rather than a confidence interval. Its name is somewhat of an abuse of nomenclature, as stated by Brown et al, since Jeffreys did not introduce this interval but rather introduced the Jeffreys prior, which is known to have good frequentist properties. The Jeffreys prior is proportional to the square root of the Fisher information, in the univariate case, and for the Binomial distribution it turns out to be a beta distribution. The beta family is the conjugate family for the binomial distribution. As an alternative credible interval to the Jeffreys Interval we use the uniform prior in the binomial-beta conjugate family. We also include an Arcsine Square Root Interval with a modification that performed well in Gilary et al. (2012). More details on these intervals, as well as their explicit formulas, are given in the appendix.

In all intervals, to adjust for the design effect due to complex sampling we replace n by the effective sample size n^* or its estimated version $\hat{n}^*$, and we replace x by the effective sample count, defined as $\hat{Y} \cdot n^*$ or its estimated version, $\hat{Y} \cdot \hat{n}^*$.

3. Analysis of Results

3.1 The Case of No Clustering

Figure 1 displays the coverages of all 7 confidence interval methods resulting from 10,000 iterations for each simulation configuration that has no clustering (c=1) under scenario $\mathbf{C}$. The vertical axis denotes the coverage based on $\hat{n}^*$ and the horizontal axis represents the coverage based on n^*. Each of the 756 points in the scatterplot shows the n^*-coverage and $\hat{n}^*$-coverage for a specific method and a fixed set of simulation parameters. Hence it is possible to compare the two types of coverage and also to make evaluations on each individually. The square in the center represents a measure of simulation error. To better visualize what the coverage is using $\hat{n}^*$, envision projecting all points to the vertical axis (and, analogously, project to the horizontal axis to visualize the n^*-coverage).

Two intervals are highlighted in red and black, the Wald Interval and the Clopper-Pearson interval, respectively. We see that overall, the Wald interval can display severe undercoverage using both n^* and $\hat{n}^*$, with some cases of extreme overcoverage. The Clopper-Pearson, on the other hand, tends to display overcoverage. This overcoverage comes at the cost of high expected widths, as can be seen in Table 1, which displays average lengths over all simulation replications for all configurations, using $\hat{n}^*$ with no clustering and $\theta = 0.1$. Note that for all sample sizes, the Clopper-Pearson interval displays the greatest average lengths. In fact because the Clopper-Pearson interval can also be expressed in terms of the beta quantiles (Brown et al., 2001), it can be shown that the Jeffreys and Uniform intervals are always contained in the Clopper-Pearson. The Wald interval has small average lengths, but this is due to its severe undercoverage. The results using n^* are very similar and only slightly narrower (not shown).

This suggests that neither of these intervals are good candidates for application to surveys such as the ACS. Similar patterns also occur in the other 3 scenarios ($\mathbf{D}, \mathbf{I}$, and $\mathbf{H}$) with no clustering (plots not shown).

Recall that scenario $\mathbf{C}$ has equal stratum proportions and stratum sample sizes. This makes this case somewhat similar to the simple binomial case, where Brown et al. also find

CI	n=30	n=40	n=50	n=80	n=160	n=240
Wald	0.203	0.178	0.162	0.132	0.093	0.076
JeffPr	0.216	0.186	0.166	0.133	0.093	0.076
UnifPr	0.223	0.191	0.170	0.135	0.094	0.076
ClPe	0.246	0.210	0.185	0.145	0.100	0.080
Wils	0.224	0.192	0.171	0.135	0.094	0.076
AgCo	0.240	0.205	0.180	0.141	0.096	0.078
Assqr	0.223	0.191	0.170	0.135	0.094	0.076

Table 1: Conditional expected lengths for $\theta = .1$, $c = 1$, averaged over all simulation configurations, computed based on $\hat{n}^*$. Configurations include $t = 10, 10000$, $\gamma_0 = 0, .2, .4$, scenario C, I, D, or H.

that the Wald interval has severe undercoverage problems, and that the Clopper-Pearson interval is wastefully conservative. Due to the equal stratum sample sizes, we would expect the variance estimator (7) to be most stable in scenario **C**, compared to scenarios where the sampling sizes differ across strata and have some strata with particularly unstable estimates of the stratum sampling variance. We see some difference between the n^*-coverages and the $\hat{n}^*$-coverages, but these are not as pronounced as what we will see under other scenarios.

Figure 2 examines how other methods do relative to each other when no clustering is present. In this picture we see some additional trends: namely, that the Agresti-Coull interval and the Arcsine Square Root Interval tend to be more conservative than the others. The Wilson, Uniform, and Jeffreys perform somewhat similarly to each other overall. Returning to Table 1 to take a closer look at the average conditional lengths, we see that the Agresti-Coull interval tends to be the widest on average, second only to the Clopper-Pearson. The average lengths for the other intervals are comparable, with the Jeffreys interval tending to be a little narrower. The results for $\theta = 0.05$ and $\theta = 0.3$ follow similar trends.

3.2 The Case of Clustering (c=3)

Figure 3 is similar to Figure 1 except it shows results with clustering, namely with $c = 3$. Again, we see that the Wald interval does very poorly overall, showing many cases of severe undercoverage regardless of whether we use the estimated or true design effect. The Clopper-Pearson again appears to have great overcoverage, particularly when using n^*. However, there are some cases of undercoverage for the Clopper-Pearson as well when it comes to the $\hat{n}^*$-coverage. Table 2 shows average lengths over all scenarios with $\theta = 0.1$ and $c = 3$ for the different CI methods over different sample sizes. Again, the Clopper-Pearson emerges as the widest interval, followed by the Agresti-Coull.

In Figure 3 a larger difference becomes apparent between the coverages based on n^* and $\hat{n}^*$ than in Figure 1. This may be attributed to two reasons: first, scenario **D** features larger differences between the stratum sampling sizes, which would result in more unstable estimates of the sampling variance for strata that have smaller sample sizes. Another explanation may be that clustering slows down the convergence of the estimated sampling variance to the true sampling variance. Figure 4 contrasts the case of clustering vs. no clustering in n^*-coverages and $\hat{n}^*$-coverages for the better behaved intervals, excluding the Wald and Clopper-Pearson. It suggests that much of the undercoverage resulting when clustering is present is due to the estimation of the design effect.

Figure 5, analogous to Figure 2, examines all scenarios with clustering ($c = 3$) for the 5 better-behaved intervals. It shows that when using $\hat{n}^*$ to compute the coverages, it is hard

CI	n=30	n=40	n=50	n=80	n=160	n=240
Wald	0.210	0.184	0.160	0.135	0.098	0.081
JeffPr	0.231	0.193	0.166	0.136	0.098	0.081
UnifPr	0.240	0.199	0.171	0.138	0.099	0.081
ClPe	0.266	0.218	0.186	0.149	0.105	0.085
Wils	0.240	0.199	0.171	0.138	0.099	0.081
AgCo	0.258	0.212	0.181	0.144	0.101	0.083
Assqr	0.241	0.199	0.170	0.138	0.099	0.081

Table 2: Conditional expected lengths for $\theta = .1$, $c = 3$, averaged over all simulation configurations, computed based on $\hat{n}^*$. Configurations include $\tau = 10, 10000$, $\gamma_0 = 0, .2, .4$, scenario C, I, D, or H.

to ascertain which interval is superior to the others, if any. Projecting all the points to the horizontal axis shows that some of the trends that we saw in the case with no clustering persist when n^* is used, such as the conservativeness of the Agresti-Coull and Arcsine Square Root intervals. Table 2 also shows a tendency for the Agresti-Coull to be a bit wider and a mild tendency for the Jeffreys to be narrower, although the Jeffreys, Uniform, Wald, and Arcsine Square Root again are comparable.

4. Conclusions

The above analysis suggests that under all scenarios, the Wald interval is unreliable and often undercovers when used in the setting of complex surveys. The Clopper-Pearson is wastefully conservative, particularly in cases where the design effect can be estimated with more accuracy, such as Scenario **C** with no clustering. Clustering leads to greater volatility of the estimated design effect, which in turns leads to lower coverages. Moreover, differential sampling fractions among strata may result in very unstable estimates of the design effect and may add to the coverage problem. Differences between coverages based on n^* and $\hat{n}^*$ suggest great benefits can potentially be reaped from improving the estimation of design effects. This paper sheds some light on some important phenomena that govern the performance of confidence intervals for proportions for complex surveys. Future research will further explore how features of the simulation, and their interactions, affect coverage. We would also like to assess the impact of heterogeneities between and within clusters and among strata on coverages and expected lengths for all the intervals.

Our preliminary comparison of the intervals leaves us with the Wilson, Jeffreys, Uniform, and Arcsine Square root as the primary contenders.

APPENDIX

A. Details of Intervals

A.1 Beta$(.5, .5)$ Conjugate Prior (Jeffreys Interval)

Let the conditional distribution of the count X given p be binomial. The Jeffreys prior distribution is assumed for p and is a Beta$(.5, .5)$.

The resulting $(1 - \alpha) * 100\%$ credible interval is:

$$L(x, n) = Beta(\alpha/2; x + 1/2, n - x + 1/2)$$

$$U(x,n) = Beta(1-\alpha/2; x+1/2, n-x+1/2)$$

where x and n represent the binomial count and number of trials, and where $Beta(\alpha; a, b)$ represents the α quantile of the Beta(a,b) distribution.

A.2 Beta$(1,1)$ Conjugate Prior (Uniform Interval)

This CI assumes a conjugate prior Beta(1,1), which is the uniform distribution. The corresponding $(1-\alpha)*100\%$ credible interval is:

$$L(x,n) = Beta(\alpha/2; x+1, n-x+1)$$

$$U(x,n) = Beta(1-\alpha/2; x+1, n-x+1)$$

A.3 Clopper-Pearson Interval

The Clopper-Pearson interval is based on exact binomial tails but can be expressed as:

$$L(x,n) = \frac{v_1 F_{v_1,v_2}(\alpha/2)}{v_2 + v_1 F_{v_1,v_2}(\alpha/2)}$$

$$U(x,n) = \frac{v_3 F_{v_3,v_4}(1-\alpha/2)}{v_4 + v_3 F_{v_3,v_4}(1-\alpha/2)}$$

where $v_1 = 2x$, $v_2 = 2(n-x+1)$, $v_3 = 2(x+1)$, $v_4 = 2(n-x)$, and $F_{d_1,d_2}(\beta)$ is the β quantile of an F distribution with d_1 and d_2 degrees of freedom (Korn and Graubard, 1998).

A.4 Wilson Interval

Like the Wald interval, the Wilson interval is derived from an asymptotic pivot. For the Wald interval the pivot is $(p-\hat{p})/(\sqrt{\hat{p}(1-\hat{p})/n})$ and for the Wilson interval the pivot is $(p-\hat{p})(\sqrt{p(1-p)/n})$. The Wilson interval bounds are

$$\frac{x+k/2}{n+k^2} \pm \frac{kn^{1/2}}{n+k^2}(\hat{p}\hat{q} + k^2/(4n))^{1/2}$$

where k represents the $(1-\alpha/2)100$th quantile of the normal distribution.

A.5 Agresti-Coull Interval

The Agresti-Coull Interval uses the same formula as the Wald interval, namely $\hat{p} \pm k\sqrt{\hat{p}(1-\hat{p})/n}$, except rather than using x/n for $\hat{p}$ it uses the center of the Wilson region. Here $\tilde{x} = x + k^2/2$ and $\tilde{n} = n + k^2$, $\tilde{p} = \tilde{x}/\tilde{n}$, and the CI is $\tilde{p} \pm k\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}$.

A.6 Arcsine Square Root Interval

The Arcsine Square Root Interval features a variance stabilizing transformation, namely $\arcsin\sqrt{p}$, with a modification. We set $\hat{p} = (x+1/2)/(n+1)$ to correct a marked anti-conservative tendency (see Gilary et al., 2012). Hence the bounds are:

$$\Big(\sin^2(\max(0, \arcsin\sqrt{\frac{x+.5}{n+1}} - \frac{z}{\sqrt{4m}})),\ \sin^2(\min(\frac{\pi}{2}, \arcsin\sqrt{\frac{x+.5}{n+1}} + \frac{z}{\sqrt{4m}})) \Big)$$

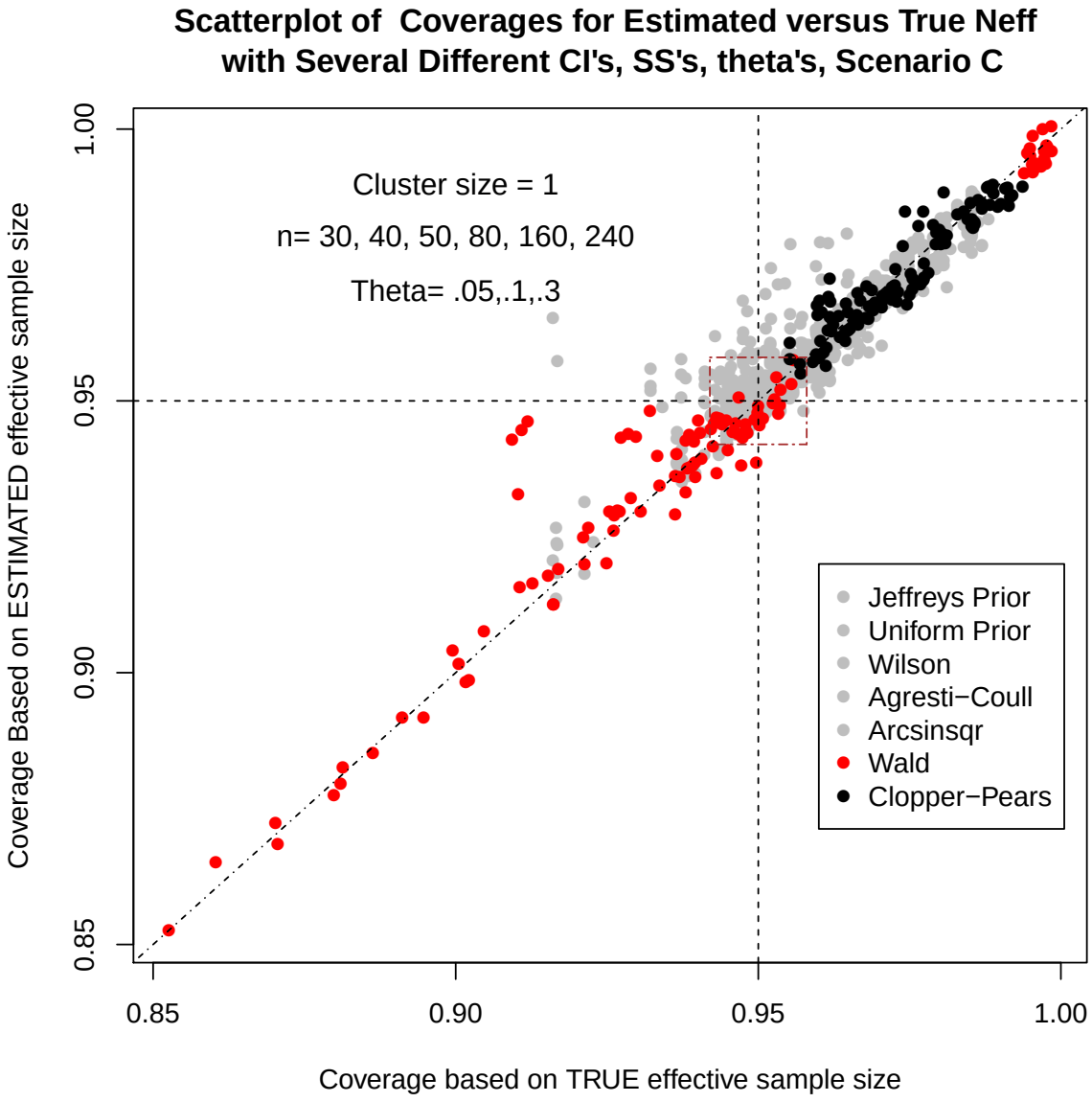

Figure 1: Problems with Wald and Clopper-Pearson Intervals in Scenario C with No Clustering

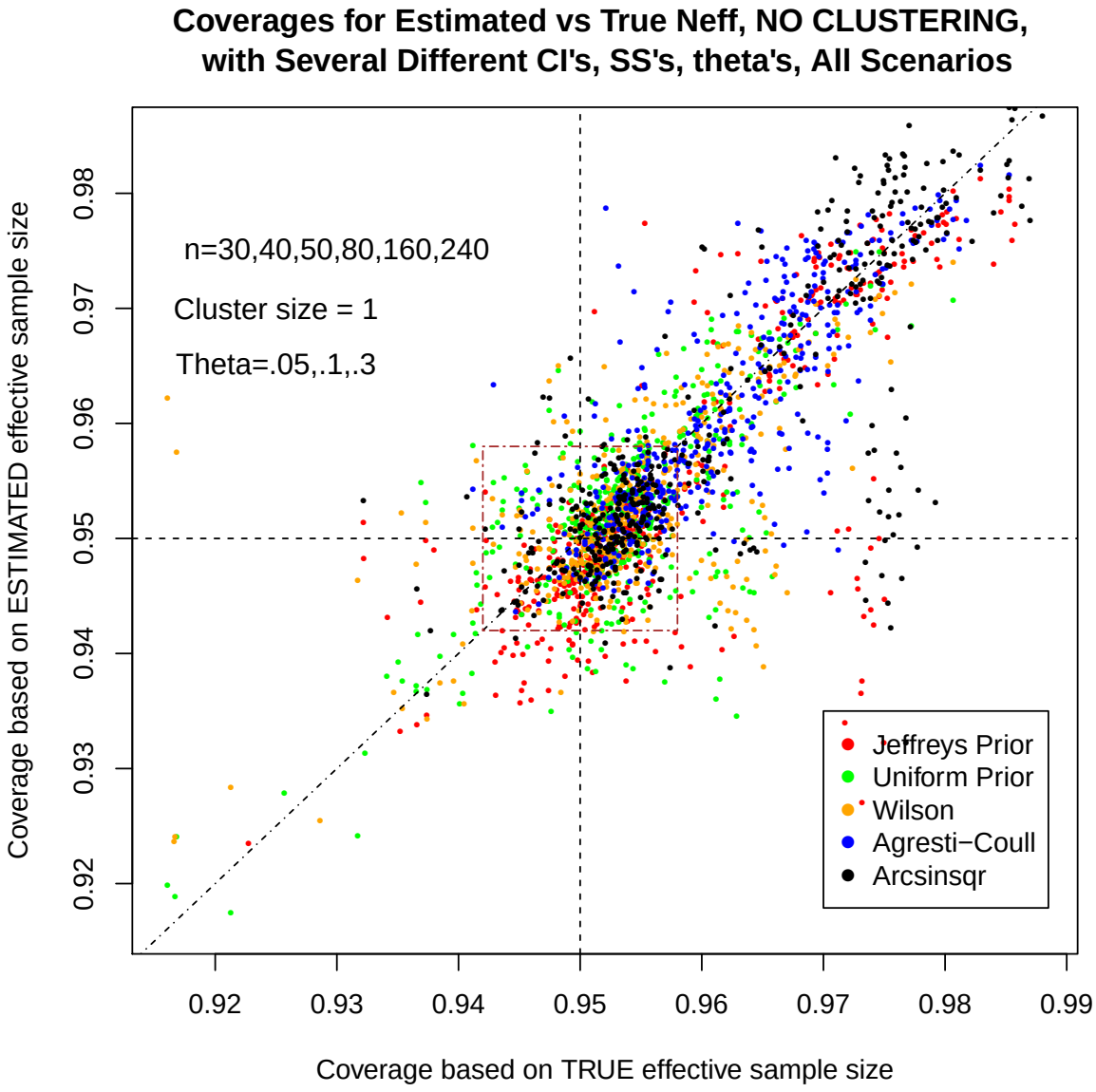

Figure 2: Scatterplots of all Scenarios with No Clustering, Excluding Wald and Clopper-Pearson Intervals

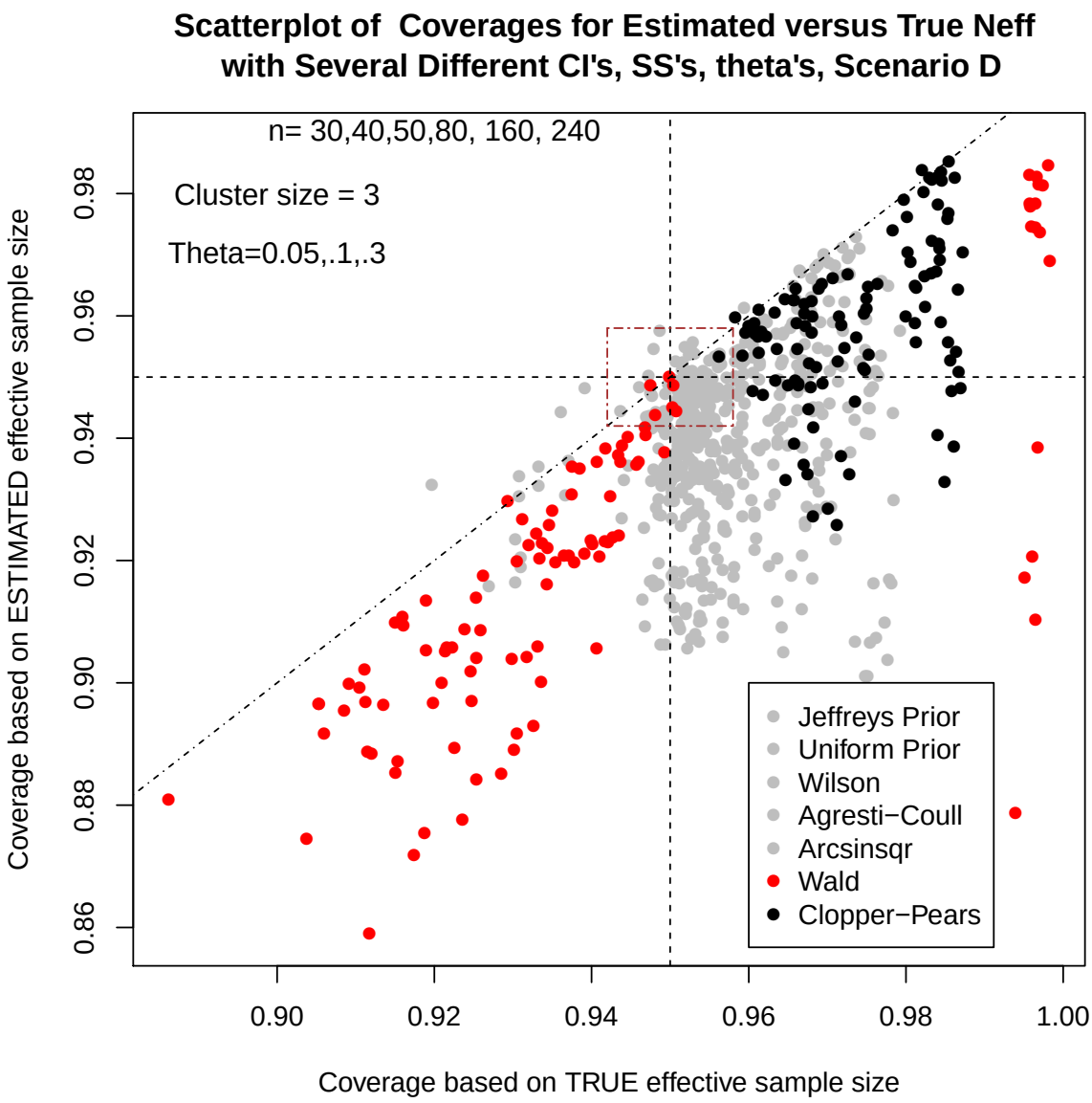

Figure 3: Scatterplot of Scenario D with Clustering

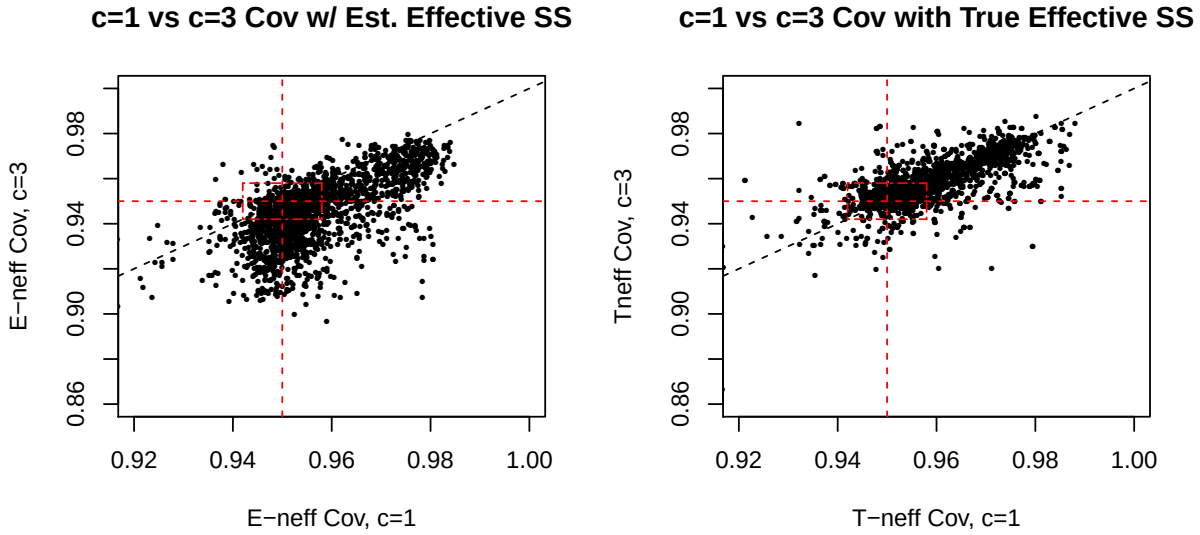

Figure 4: Scatterplots of conditional coverages using n^* and $\hat{n}^*$ contrasting the case of $c = 1$ and $c = 3$ for all intervals excluding the Wald and Clopper-Pearson

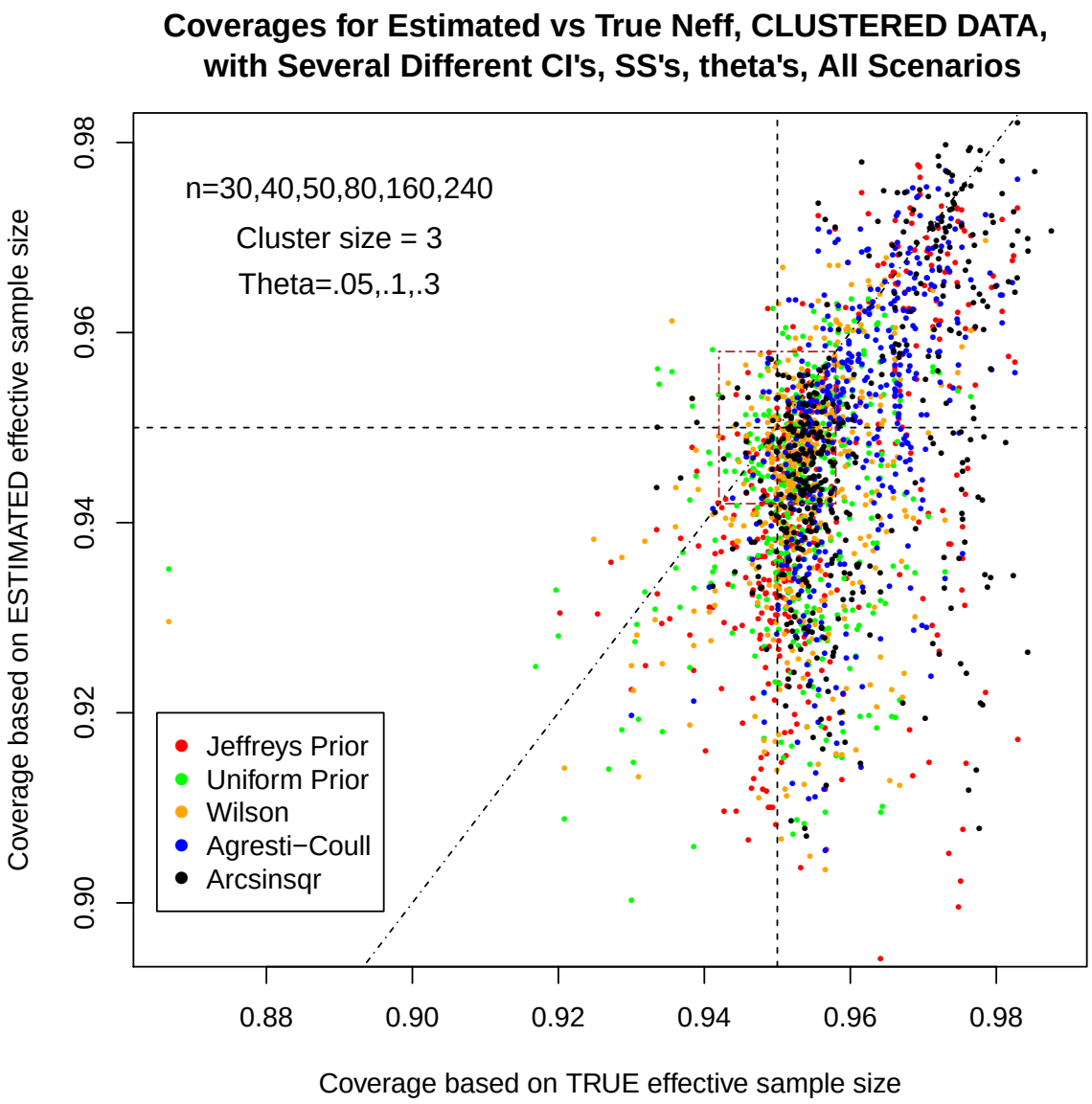

Figure 5: Scatterplots with clustering, all scenarios, excluding Wald and Clopper-Pearson

REFERENCES

U.S. Census Bureau (2009). "Design and Methodology: American Community Survey." U.S. Government Printing Office, Washington, DC.

Blyth, C. R., and Still, H. A., (1983), "Binomial Confidence Intervals," *J. Amer. Statist. Assoc.*, 78, 108-116.

Brown, L.D., Cai, T. T., and DasGupta, A. (2001), "Interval Estimation for a Binomial Proportion," *Statistical Science*, 2, 101–117.

Brown, L.D., Cai, T. T., and DasGupta, A. (2002), "Confidence Intervals for a Binomial Proportion and Asymptotic Expansions," *Ann. Statist.*, 30,1, 160-201.

Buonnacorsi, J. P. (1987), "A Note on Confidence Intervals for Proportions in Finite Population," *The American Statistician*, 41, 3, 215-218.

Casella G. (1986), "Refining Binomial Confidence Intervals," *Canad. J. Statist.*, 78, 107-116.

Clopper, C. J., and Pearson, E. S, (1934), "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, 26, 404-413.

Fay, R.E. and Train, G. (1995). "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics and States and Counties," *Proceedings of the Government Statistics Section*, Alexandria, VA: American Statistical Association, 154-159.

Gilary, A., Maples, J., and Slud E. V., "Small Area Confidence Bounds on Small Cell Proportions in Survey Populations." *JSM Proceedings*, Survey Research Section, 3541-3555.

Korn, E, and Graubard, B. I. (1998), "Confidence Interval for Proportions with Small Expected Number of Positive Counts Estimated From Survey Data," *Survey Methodology*, 24, 1030-1039.

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

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

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, 157-169

Wright, T. (1991), "Exact Confidence Bounds when Sampling from Small Finite Universes." Berlin: Springer-Verlag.