Comparative Study of Confidence Intervals for Proportions in Complex Sample Surveys *

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

August 2, 2018

Abstract

The most widespread method of computing confidence intervals (CIs) in complex surveys is to add and subtract the margin of error (MOE) from the point estimate, where the MOE is the estimated standard error multiplied by the suitable Gaussian quantile. This Wald-type interval is used by the American Community Survey (ACS), the largest US household sample survey. For inferences on small proportions with moderate sample sizes, this method often results in marked under-coverage and lower CI endpoint less than 0. We assess via simulation the coverage and width, in complex sample surveys, of seven alternatives to the Wald interval for a binomial proportion with sample size replaced by the 'effective sample size,' that is, the sample size divided by the design effect. Building on work of Franco et al. (2014), our simulations address the impact of clustering, stratification, different stratum sampling fractions, and stratum-specific proportions. We show that all intervals undercover when there is clustering and design effects are computed from a simple design-based estimator of sampling variance. Coverage can be better calibrated for the alternatives to Wald by improving estimation of the effective sample size through superpopulation modeling. This approach is more effective in our simulations than modifications of effective sample size proposed by Korn and Graubard (1998) and Dean and Pagano (2015). We recommend intervals of the Wilson or Bayes uniform-prior form, with the Jeffreys-prior interval not far behind.

*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

[§]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

[¶]US Census Bureau and University of Maryland, College Park

KEYWORDS: Complex Surveys; Confidence Interval for Proportion; Design Effect; Effective Sample Size; Bayesian Formalism

Word count estimate: Approximately 5700.

1 Introduction

Constructing well-calibrated confidence intervals (CIs) for population proportions based on survey data is challenging, unless the sample size is very large. With a simple random sample (SRS) for a large population where the sampling fraction is negligible, the data are approximately binomial, and CIs that guarantee at least nominal coverage are available (Clopper and Pearson 1934, Blyth and Still 1983, and Casella 1986). These CIs tend to be conservative (exceed the nominal coverage) and wide. However, methods that guarantee nominal coverage for the binomial case do not necessarily do so when applied to complex survey data. Intervals based on randomized tests– that is, tests based on exact discrete distributions with auxiliary randomization to produce rejection regions of exact size α under the null hypothesis– exist in the SRS case (see, for instance, Wright 1997). Such tests would be very difficult to construct in more complex settings, and we focus only on non-randomized intervals.

Many surveys use intervals of the form $\hat{p} \pm z_{\alpha/2} \cdot \widehat{SE}$, or equivalently $\hat{p} \pm \text{MOE}$, where $\hat{p}$ is an estimate of the proportion, $\widehat{SE}$ is an estimate of its standard error, $z_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of the normal distribution, sometimes replaced by a corresponding t-quantile, and MOE is the Margin of Error. The Wald interval is the most basic, with $\widehat{SE} = \sqrt{\hat{p}(1-\hat{p})/n}$, but it can perform poorly even in the SRS case. An improvement is provided by replacing the sample size with the effective sample size (e.g., Korn and Graubard 1998, Liu and Kott 2009, Dean and Pagano 2015), but the Wald interval still performs poorly.

We conduct an extensive simulation study to evaluate and compare coverage and width of the Wald and seven other candidate intervals, as well as the performance of three different approaches to estimating the effective sample size. The three approaches include a simple design-based approach, a modification to this design-based estimate recommended by Dean and Pagano (2015), and the use of an improved method to estimate the sampling variance based on super-population model assumptions. The intervals considered are the Wald, Agresti-Coull (1998), Clopper-Pearson (1934), Wilson (1927), Arcsine Square Root (Sokal and

Rohlf 1995, as modified by Gilary et al. 2012) and two Bayesian intervals using the Jeffreys and the uniform priors. We also study the CI for the logit-transformed proportion, which was previously considered in Liu and Kott (2009) and Dean and Pagano (2015). Results for the logit-transformed method, or Logit interval, are limited to the Online Supplement, because its performance was less promising than competitors. Our results are consistent with those of Brown et. al (2001), who find that it can produce very wide intervals in the binomial case. The Agresti-Coull, Wilson, and Jeffreys- and uniform-prior intervals previously performed well in the SRS setting (Brown et al. 2001, Carlin and Louis 2009). The Clopper-Pearson always meets or exceeds the nominal coverage in the SRS setting, and was recommended by Korn and Graubard (1998) for settings with small expected numbers of successes. Performance of the Arcsine Square Root interval was compared by simulation with a cell-based version in Gilary et al. (2012) in a small-area (Fay-Herriot) model setting. A modified Arcsine Square Root interval was used to produce the Census Bureau's May 2012 publicly-released confidence bounds for estimates of 2010 Census erroneous enumeration rates. For all these intervals, in complex surveys we replace the sample size by an estimate of the effective sample size. The Bayesian intervals are based on the Beta prior and posterior distribution from the SRS case, but also replace sample size and observed proportion by effective sample size and design-weighted estimated proportion.

This work builds on the simulation studies in Franco et al. (2014) and Dean and Pagano (2015). We evaluate the joint distribution of coverage and width, in the context of clustering and stratification, of several degrees of heterogeneity within and between clusters and among strata, and of uncertainty in estimating sampling variances. We set aside samples for which the estimated sampling variance is 0, and so our results are conditional on a positive estimated variance. Our primary objective is to find intervals that have well-calibrated coverage and controlled width. We treat a wide range of scenarios, and aim to find intervals that work well across all scenarios rather than prescribing criteria for the use of particular intervals, because the determining factors will generally not be known to the analyst.

For intervals that perform well in the SRS context, our results suggest that the principal cause of undercoverage in complex surveys is uncertainty in estimating the effective sample size. Hence, we have improved estimation of the sampling variance and consequently of the design effect and effective sample size. These improvements come by making basic assumptions about the superpopulation. We then take the expectation both with respect to the sampling design and the superpopulation model when computing the

variance of the survey weighted estimator (i.e., the "anticipated variance" of Isaki and Fuller, 1982). Chen and Rust (2017) also use superpopulation models to improve variance estimates, based on Kish (1987)'s well-known design effect formula.

Our study is motivated by the American Community Survey (ACS), the largest household sample survey in the United States, sampling approximately 3.5 million addresses annually, and producing billions of estimates (US Census Bureau, 2014). ACS publishes CIs of the form $\hat{p} \pm z_{\alpha/2} \cdot \widehat{SE}$, with $\widehat{SE}$ based on the Successive Difference Replication (SDR) Method (Fay and Train, 1995). See U.S. Census Bureau (2014). Despite its large overall sample size, the extensive cross-classification of its many demographic, personal and economic questions can generate domains with small sample sizes. Due to the sheer quantity and diversity of estimates, we consider only basic CI methods that are easy to implement and depend only on sample size, the survey-weighted estimate of the proportion, and a sampling variance estimate used to estimate the design effect and effective sample size. Three different approaches to estimating the sampling variance and effective sample size are considered.

Other authors have conducted related simulations for complex surveys. Liu and Kott (2009) and Kott and Liu (2009) compared one-sided intervals for proportions in stratified SRS surveys; we focus on two-sided intervals. Korn and Graubard (1998) studied CIs for small proportions in surveys including clusters (of sizes 10 or 100) and unequal weights by simulation and data analysis, comparing intervals based on design-effect modifications including replacement of Gaussian by t quantiles in Wald-type intervals. Dean and Pagano (2015) considered essentially the same intervals we do (excluding the Arcsine Square Root interval), varying overall prevalence p and Intracluster Correlations (ICCs) within a design of 30 primary and 7 secondary sampling units. They also have limited results related to stratification, including a case with two strata in their sampling design. Their modification of effective sample size resembles Korn and Graubard's (1998), except that in their adjustment factor (our formula (20)), they replace the t-quantile by a z-quantile in the numerator. Kott et al. (2001) also discuss confidence intervals for complex surveys, but their simulations are carried out under SRS.

The sampling design in our simulations is that of a single-stage stratified SRS sample of all-or-none clusters of identical size. The strata sampling fractions and the cluster sizes, as well as the relationship between the sample sizes and the true stratum proportions, vary among runs. Although this is more basic than the

complex designs common in practice, our setup is more complex than those of previously published simulation studies. Moreover, single stage SRS samples of appropriately chosen ultimate clusters can be used to approximate more complex epsem or stratified epsem-within-stratum designs (see for example Kalton, 1979). In an epsem design each unit in the population has an equal probability of selection.

Our research builds on previous work by: (1) a more elaborate multi-factorial simulation design that allows estimation of the main effects of scenario components and interactions, (2) assessing the impact of uncertainty in estimating the effective sample size by comparing results to those using the true effective sample size, and (3) applying superpopulation models to improve performance by better estimation of the sampling variance, and hence of the effective sample size.

Section 2 defines the eight intervals we study, and Section 3 develops estimation of the effective sample size. Section 4 provides simulation specifications. Section 5 presents results, and Section 6 draws conclusions and formulates recommendations and promising avenues for future research. The appendix includes mathematical proofs, additional details of the simulations, and a brief description of the `R` code (R Core Team, 2017) and workspace for computing design effect estimates and CIs that are included in the Supplementary Materials.

2 Candidate Intervals

We consider seven alternatives to the basic Wald interval for a binomial proportion: Jeffreys and Uniform prior Bayesian intervals; the Clopper-Pearson, Wilson, Agresti-Coull, Arcsine Square Root, and Logit intervals. Each of these interval methods is in turn treated in three ways: using a simple design-based estimate of the effective sample size, adjusting this estimate as recommended by Dean and Pagano (2015), and estimating the design effect using superpopulation model assumptions.

Here, we describe the interval construction methods first for Bernoulli sampling, with n trials and X successes; the intervals for complex surveys are obtained by replacing n by an estimate of the effective sample size n_{eff} and X by an estimate of $n_{\text{eff}} \cdot p$, where p is estimated by the survey-weighted proportion.

The different methods of estimating the effective sample size are discussed in Section 3.

2.1 Wald Interval

The Wald interval is

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n}, \tag{1}$$

with $\hat{p} = X/n$ and $z_{\alpha/2}$ the quantile of the standard normal distribution. This is a special case of what we refer to as "Wald-type intervals" for general complex surveys

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

where $\hat{p}$ is possibly survey weighted and $\widehat{SE}$ is an estimate of its standard error.

The normal quantiles are sometimes replaced by t-quantiles, with the degrees of freedom depending on the amount of clustering and stratification (Korn and Graubard, 1998). This adjustment is based on empirical evidence (Frankel 1971, ch. 7), with some formal justification under strong assumptions (Korn and Graubard, 1990).

2.2 Bayesian Intervals: Jeffreys and Uniform

With the prior $[p \mid a, b] = \text{Beta}(a, b); a, b > 0$ and data distributed as $[X \mid p, n] = \text{Binomial}(n, p)$, the posterior distribution is $[p \mid X, n] = \text{Beta}(X + a, n - X + b)$. With $qbeta(r; \cdot, \cdot)$ denoting the r quantile of a Beta distribution, the $(1-\alpha) * 100\%$ equal-tail credible interval is,

$$\begin{aligned} L(X, n) &= qbeta(\alpha/2; X + a, n - X + b) \\ U(X, n) &= qbeta(1 - \alpha/2; X + a, n - X + b) \end{aligned} \tag{3}$$

The Jeffreys interval ("JeffPr"), uses $a = b = 0.5$, and the Uniform interval ("UnifPr") uses $a = b = 1$. Carlin and Louis (2009) show that these have excellent frequentist properties for SRS sampling, making them attractive candidates in the survey context.

2.3 Clopper-Pearson Interval

The Clopper-Pearson interval ("ClPe" or "CP") is based on exact binomial tails, and can be expressed as,

$$\begin{aligned} L(X,n) &= \frac{v_1 F_{v_1,v_2}(\alpha/2)}{v_2 + v_1 F_{v_1,v_2}(\alpha/2)} &= qbeta(\alpha/2; X, n-X+1) \\ U(X,n) &= \frac{v_3 F_{v_3,v_4}(1-\alpha/2)}{v_4 + v_3 F_{v_3,v_4}(1-\alpha/2)} &= qbeta(1-\alpha/2; X+1, n-X) \end{aligned} \tag{4}$$

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). Interval endpoints in (4) are very similar to those of Jeffreys and Uniform, shown in (3), but demonstrably wider (see Appendix A for a proof).

2.4 Wilson Interval

Like the Wald interval, the Wilson interval ("Wils") can be derived from an asymptotic pivot. In place of the Wald pivot $(p-\hat{p})/\sqrt{\hat{p}(1-\hat{p})/n}$, the Wilson interval uses $(p-\hat{p})/\sqrt{p(1-p)/n}$, producing CI limits,

$$\frac{X + z^2/2}{n+z^2} \pm \frac{zn^{1/2}}{n+z^2}\left\{\hat{p}\hat{q} + z^2/(4n)\right\}^{1/2}, \tag{5}$$

where $z = z_{\alpha/2}$ from now on.

2.5 Agresti-Coull Interval

The Agresti-Coull Interval ("AgCo" or "AC") uses the same form as the Wald interval (2), replacing $\hat{p}$ with the center of the Wilson interval $\tilde{p} = (X + z^2/2)/(n+z^2)$, and n with the denominator of $\tilde{p}$, i.e. $\tilde{n} = n + z^2$. The interval is then

$$\tilde{p} \;\pm\; z\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}. \tag{6}$$

Agresti and Coull (1998) deal with the case of a 95% CI, pointing out that, at this confidence level, this is approximately the same as adding two successes and two failures and then applying the Wald interval. They also show that the center of the Wilson interval is a weighted average between $\hat{p}$ and 0.5. They note that the interval is simpler in form than the Wilson interval, is not as conservative as the Clopper-Pearson interval, and performs better than the Wald interval in the SRS case.

2.6 Arcsine Square Root Interval

The Arcsine Square Root Interval ("Assqr") uses $\text{arcsine}\sqrt{p}$ as variance stabilizing transformation, along with $\hat{p} = (X + 1/2)/(n + 1)$ (as in Jeffreys) to correct the marked anti-conservatism of the Wald interval (Gilary et al. 2012). The Wald formula (2) produces endpoints in the transformed scale which are back-transformed to produce CI limits,

$$\begin{aligned} L(X,n) &= \sin^2\left\{\max\left(0,\, \arcsin\sqrt{\frac{X+.5}{n+1}} - \frac{z}{\sqrt{4n}}\right)\right\} \\ U(X,n) &= \sin^2\left\{\min\left(\frac{\pi}{2},\, \arcsin\sqrt{\frac{X+.5}{n+1}} + \frac{z}{\sqrt{4n}}\right)\right\} \end{aligned} \tag{7}$$

2.7 Logit Interval

The Logit interval applies a logit transformation, then produces a Wald-type interval, and then back-transforms to the original scale, yielding:

$$\begin{aligned} L(X,n) &= \frac{e^{\lambda_l}}{1+e^{\lambda_l}} \\ U(X,n) &= \frac{e^{\lambda_u}}{1+e^{\lambda_u}} \end{aligned} \tag{8}$$

where

$\lambda_l = \hat{\lambda} - z\sqrt{\hat{V}}$, and $\lambda_u = \hat{\lambda} + z\sqrt{\hat{V}}$ with $\hat{\lambda} = \log(\hat{p}/(1-\hat{p}))$, and $\hat{V} = n/(X(n-X))$. Note that this interval is undefined when $\hat{p} = 0$ or $\hat{p} = 1$. We define $\hat{\lambda}$ to be $-\infty$ when $\hat{p} = 0$, ∞ when $\hat{p} = 1$, and $= \log\left(\hat{p}/(1-\hat{p})\right)$ otherwise. Such a definition does not affect our results since we condition on positive

estimated variance.

2.8 Discussion of candidate intervals

Brown et al. (2001, 2002) proposed alternative methods that ameliorate the erratic coverage of the standard Wald interval, recommending Jeffreys and Wilson for small sample sizes and Agresti-Coull for large sample sizes (Brown et al. 2001, Section 5). These intervals are appropriate for survey data with SRS designs where the sampling fraction is small or sampling is with replacement, but they are not designed to accommodate the clustering, stratification, or unequal weights of more complex sample surveys.

A common approach to constructing confidence intervals for proportions from complex sample survey data is to modify the inputs to binomial intervals, such as the Wald interval (1), to account for survey weighting and the design effect. The survey-weighted estimated proportion, $\hat{p}$, is used along with a consistent design-based estimate, $\widehat{\text{Var}}(\hat{p})$, of its variance. These combine to estimate the design effect (Kish 1965) and effective sample size,

$$\begin{aligned} \widehat{\textit{Deff}} &= \frac{\widehat{\text{Var}}(\hat{p})}{\hat{p}(1-\hat{p})/n} \\ \widehat{n}_{\text{eff}} &= \frac{n}{\widehat{\textit{Deff}}} = \frac{\hat{p}(1-\hat{p})}{\widehat{\text{Var}}(\hat{p})}. \end{aligned} \tag{9}$$

For simplicity we ignore the finite population correction in the SRS variance expression in the denominator of $\widehat{\textit{Deff}}$. In CI expressions, n is replaced by $\widehat{n}_{\text{eff}}$ and X by $\hat{p} \cdot \hat{n}_{\text{eff}}$ without rounding (e.g., Korn and Graubard 1998, Liu and Kott 2009, Dean and Pagano 2015). The effective sample size n_{eff} can be interpreted as the sample size needed under an SRS design to obtain the same large-sample CI width obtained under the complex sampling scheme. Applying the design-effect modifications to the Wald interval produces $\hat{p} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{p})}$.

3 Estimating the Effective Sample Size

Let Y_{hki} be the (binary) response for individual i in cluster k in stratum h. Denote the population count in stratum h and cluster k by Y_{hk+} and the population count in stratum h by Y_{h++}. That is,

$$Y_{hk+} = \sum_{i \in C_{hk}} Y_{hki} \quad , \qquad Y_{h++} = \sum_{k=1}^{K_h} Y_{hk+} = \sum_{k=1}^{K_h} \sum_{i \in C_{hk}} Y_{hki} \, ,$$

where C_{hk} denotes the set of units or individuals i belonging to cluster k in stratum h, and K_h is the number of clusters in stratum h. The population total is denoted by Y, and the corresponding sample weighted estimator is

$$\hat{Y} = \sum_{h=1}^{H} \hat{Y}_{h++} \quad , \qquad \hat{Y}_{h++} = \frac{K_h}{n_h^C} \sum_{k \in S_h} Y_{hk+}, \tag{10}$$

where S_h and n_h^C are the set and number of sampled clusters in stratum h, and H is the number of strata. For future reference, also define N_h to be the population size in stratum h, and denote by c the size of each cluster in the population.

The population proportion $\bar{Y} = Y/N$ has expectation $E(\bar{Y}) = \theta$, and confidence intervals for it are based on (10) together with the 'working model' $n_{\text{eff}} \cdot (\hat{Y}/N) \sim \text{Binom}(n_{\text{eff}}, \bar{Y})$, where n_{eff} is a suitable effective sample size. It is permissible for values of $n_{\text{eff}} \cdot (\hat{Y}/N)$ and n_{eff} to be non-integer within likelihood-based methods such as those implemented in `R`.

We evaluate the performance of CIs for the overall proportion $\bar{Y}$ of successes within a survey assumed to have the sampling design of a SRS of clusters, with clusters sampled all-or-none. Generalizations of design- and model-based estimators to the case of cluster sampling with unequal cluster sizes and weights within strata are given in Appendix B.

We compare coverages and widths of the intervals using the 'true' effective sample size based on the actual simulated (frame) population, and using 'estimated' effective sample sizes computed from sampled data. For the former, we compare two approaches; one with no superpopulation model assumptions, and one that makes some basic assumptions. We incorporate finite population corrections, although the sampling fractions we consider are small.

3.1 Design-based Estimate of the Design Effect

Let $f = c\,n^C/N$ denote the overall sampling fraction, with n^C the number of clusters sampled, and $f_h = c\,n_h^C/N_h = n_h^C/K_h$ the sampling fraction within the h'th stratum. The design variance $\hat{Y}$ of the survey estimator (for stratumwise SRS cluster samples) is

$$\text{Var}(\hat{Y}) \;=\; \sum_{h=1}^{H} \frac{K_h^2\,(1-f_h)}{n_h^{\mathcal{C}}}\; s^2_{Y_{h\cdot+}} \quad , \qquad s^2_{Y_{h\cdot+}} \;=\; \frac{1}{K_h-1}\sum_{k=1}^{K_h}\left(Y_{hk+} - \frac{Y_{h++}}{K_h}\right)^2. \tag{11}$$

So the true design effect and true effective sample size are,

$$\texttt{Deff} \;=\; \frac{n\;\text{Var}(\hat{Y})}{Y(N-Y)\,(1-f)} \quad , \qquad n_{\text{eff}} \;=\; n/\texttt{Deff}. \tag{12}$$

Superpopulation model-free estimates of the design effect and effective sample size, denoted $\widehat{\texttt{Deff}}$ and $\hat{n}_{\text{eff}}$, are:

$$\widehat{\text{Var}}(\hat{Y}) \;=\; \sum_{h=1}^{H} \frac{K_h^2\,(1-f_h)}{n_h^{\mathcal{C}}}\; \hat{s}^2_{Y_{h\cdot+}} \quad , \qquad \hat{s}^2_{Y_{h\cdot+}} \;=\; \frac{1}{n_h^C-1}\sum_{k\in S_h}\left(Y_{hk+} - \frac{1}{n_h^C}\sum_{l\in S_h} Y_{hl+}\right)^2, \tag{13}$$

$$\widehat{\texttt{Deff}} \;=\; \frac{n\;\widehat{\text{Var}}(\hat{Y})}{\hat{Y}(N-\hat{Y})\,(1-f)} \quad , \qquad \hat{n}_{\text{eff}} \;=\; n/\widehat{\texttt{Deff}}. \tag{14}$$

3.2 Model-based Estimate of the Design Effect

The method that Kish (1987) used to derive his famous approximate formula for design effects in terms of intra-cluster correlations (ICCs) and unit-level attribute variances can be viewed as an attempt to combine design-based variance formulas with simple modeling assumptions about the superpopulation. Like Gabler et al. (1999) and Chen and Rust (2017), we extend this method (in Appendix B) to obtain a model-based estimator of `Deff` under somewhat more general assumptions. The assumptions that we consider here are:

(A.i) $E(Y_{hki}) = \tau_h$ for all clusters k and individuals i in stratum h,

(A.ii) $\text{Var}(Y_{hki}) = \sigma_h^2$ for all k in stratum h and i in cluster C_{kh},

(A.iii) $\text{Corr}(Y_{hki}, Y_{hk'j}) \;=\; \rho$ when $i,\, j \in C_{kh}$ and $k = k'$, and Corr$= 0$ otherwise.

Assumption (A.iii) is restrictive in assuming constancy of ICCs across strata, and (A.i)-(A.ii) might also oversimplify in assuming distributional parameters of all attributes within stratum to be the same. Although a superpopulation model based on these assumptions is too simple to be realistic, we will find that the reduction in variability of the estimated design effect more than compensates for potential bias.

Remark 1 *In our setting of binary Y_{hki}, the assumptions (A.i) and (A.ii) are redundant, since $\sigma_h^2 = \tau_h(1-\tau_h)$ for all h. For this reason, the parameter estimates $\hat{\sigma}_h^2$ are defined to be $(n_h/(n_h-1))\,\hat{\tau}_h\,(1-\hat{\tau}_h)$ as in (16) below, or (as actually implemented in our simulations) by the formula $(n_h^C/(n_h^C-1))\,\hat{\tau}_h\,(1-\hat{\tau}_h)$ which inflates variances in a helpful way.* □

As justified in Appendix B, the model-based estimation formula derived from (A.i)-(A.iii) for $\mathrm{Var}(\hat{Y})$ that we implement in our simulations (specifically for the stratumwise SRS cluster sampling of equal-sized clusters of binary attributes) is closely related to Kish's formula. The variance formula is

$$\widehat{\mathrm{Var}}^{\,*}(\hat{Y}) \;=\; \sum_{h=1}^{H} \hat{\sigma}_h^2 \,\frac{K_h - n_h^C}{n_h^C}\, N_h \left(1 \,+\, (c-1)\,\hat{\rho}\right) \tag{15}$$

with σ_h^2 and ρ parameters estimated according to the formulas $\hat{\tau}_h \;=\; \sum_{k\in S_h} Y_{hk+}/n_h$ and

$$\hat{\sigma}_h^2 \;\;=\;\; \frac{1}{cn_h^C - 1} \sum_{k\in S_h} \sum_{i\in C_{kh}} \left(Y_{hki} \,-\, \hat{\tau}_h\right)^2 \;=\; \frac{n_h}{n_h-1}\,\hat{\tau}_h(1-\hat{\tau}_h) \tag{16}$$

$$1-\hat{\rho} \;\;=\;\; \frac{n-H}{2n(c-1)} \sum_{h=1}^{H} \sum_{k\in S_h} \sum_{i,j\in C_{kh}} (Y_{hki}-Y_{hkj})^2 \,\Big/\, \sum_{h=1}^{H} (cn_h^C-1)\hat{\sigma}_h^2 \tag{17}$$

(but $\hat{\rho}$ is defined as 0 when $c=1$, and in the simulations, $\hat{\rho}$ was set to 0 whenever it was negative in (17)). For the more realistic case of unequal-sized clusters and unknown cluster sizes for unsampled clusters, see (26) in Appendix B.

The corresponding estimated design effect and effective sample size are:

$$\widehat{\texttt{Deff}}^{\,*} \;=\; \frac{n\;\widehat{\mathrm{Var}}^{\,*}(\hat{Y})}{\hat{Y}(N-\hat{Y})\,(1-f)} \quad , \qquad \hat{n}_{\text{eff}}^{*} \;=\; n/\widehat{\texttt{Deff}}^{\,*} \quad . \tag{18}$$

These formulas have analogs, justified and developed more generally in Appendix B, for more complex designs. Our broader point is that generalized model-based formulas such as (18) yield CIs with better coverage properties than CIs from purely design-based estimates of effective sample size.

3.3 Adjustments to Estimated Effective Sample Size

Korn and Graubard (1998) suggested multiplying the effective sample size by a factor,

$$\hat{n}_{\text{eff}}^{df} = \hat{n}_{\text{eff}} \cdot \left\{ \frac{t_{n-1}(1-\alpha/2)}{t_d(1-\alpha/2)} \right\}^2, \tag{19}$$

where the design degrees of freedom are $d =$ #{sampled clusters} - #{strata} for a multi-stage design with stratified selection of clusters at the first stage, and $\hat{n}_{\text{eff}}$ is an estimate of the effective sample size. If $n - 1 < d$, as when there is significant clustering, the bracketed ratio will be less than 1. The effective sample size will be reduced, resulting in wider intervals, counteracting to a degree the undercoverage typically associated with clustering.

Dean and Pagano (2015) similarly define adjusted estimated effective sample size as,

$$\hat{n}_{\text{eff}}^{dfDP} = \hat{n}_{\text{eff}} \cdot \left\{ \frac{z(1-\alpha/2)}{t_d(1-\alpha/2)} \right\}^2, \tag{20}$$

which is (19) with a normal quantile in the numerator in place of the t quantile. This replacement yields a smaller ratio, smaller effective sample size, wider confidence intervals and higher coverage.

4 Simulation Study

We simulate one population for each parameter configuration, then implement sampling designs, analyze the data and summarize results.

4.1 Simulating the Population

First, we create a population of size $N = 10,000$ with $H = 4$ strata. In the h^{th} stratum there are $K_h = K$ clusters, each of size c, and N_h units with $N_h = c \cdot K_h$. We allow different sampling fractions in different strata. In separate runs, $c = 1, 3, 5$ or 7. The expected population proportion, $E(\bar{Y}) = \theta$, is specified for each simulation, where $\bar{Y}$ is the population mean of the binary attribute. Scenarios jointly specify the dependence on the stratum-specific samples n_h and population proportions of the form $\theta_h = \theta + b \cdot (h - 2.5)$ ensuring that $\bar{\theta}_h = \theta$ for various choices of b including $b = 0$ (see Section 4.2).

A "success" or "failure," $Y_{hki} \in \{0, 1\}$, for unit i in cluster k in stratum h, is generated from the model,

$$\begin{aligned} [p_{hk} \,|\, \theta_h] &\sim \text{Beta}\Big(\frac{1-\rho}{\rho}\,\theta_h,\ \frac{1-\rho}{\rho}\,(1-\theta_h)\Big) \\ [Y_{hki} \,|\, p_{hk}] &\sim \text{Bernoulli}(p_{hk}). \end{aligned} \tag{21}$$

As described in Section 4.3, parameter configurations $(\rho,\ \{\theta_h\}_{h=1}^{H},\ c)$, "scenario" and sample size n are specified once for each simulated frame population. Here ρ is the ICC for the binary attribute, which measures within-cluster heterogeneity when $c > 1$.

4.2 Simulating the Sampling Design

After generating the population, it is sampled $R = 10,000$ times for each simulation configuration. As discussed in the introduction, the sampling design is a single-stage stratified SRS sample of all-or-none clusters of identical size, where the objective is inference about the proportion Y/N.

An alternative to generating each population once and sampling repeatedly is to generate $10,000$ populations and sample each once. Our approach is consistent with the design-based philosophy prevalent among survey practitioners, in which the finite population is viewed as fixed and all randomness is ascribed to the sampling process. In our simulations, the large number of frame populations generated for different factorial combinations prevent anomalous characteristics in any single frame population from distorting the results.

With $f = n/N$ the overall sampling fraction, and $f_h = n_h/N_h$ the stratum-specific fractions, for $h = 1, \ldots, 4$, we study four scenarios:

Scenario C: $\theta_h \equiv \theta, \quad f_h \equiv f.$

Scenario I: f_h increases as θ_h increases.

Scenario D: f_h decreases as θ_h increases.

Scenario H: the relation between f_h and θ_h is quadratic and concave.

For example, if $\theta = 0.3,\ c = 1,$ and the sampling fraction f is such that $n = 80$, then under Scenario **C** the vector of stratum sample sizes is $\mathbf{n} = (20, 20, 20, 20),$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_4) = (0.3, 0.3, 0.3, 0.3)$; under Scenario **D** it is $\mathbf{n} = (28, 22, 18, 12)$, $\boldsymbol{\theta} = (0.1875, 0.2625, 0.3375, 0.4125)$. With no clustering, scenario **C** closely resembles a SRS design. By contrast, scenario **D** yields large variability among stratum sample sizes. For some simulation configurations, rounding of stratum sample sizes may cause actual total sample sizes to differ slightly from the nominal n. For more details on rounding and other aspects of the simulation design, see Appendix C.

4.3 Factorial Design

Each simulation parameter can take on several values, creating a factorial design shown in the following table. The combinations number 648, after excluding configurations with $n \leq 50$ and $c \geq 5$ because of problems such as undefined stratum sample variances due to strata with one or zero clusters.

Factor	Symbol	Levels
Cluster size	c	1, 3, 5, 7
Sample size*	n	30, 40, 50, 84, 196, 280
Scenario	–	C, I, D, H
Expected Proportion	$\theta = E(Y/N)$	0.05, 0.10, 0.30
Intra-Cluster Correlation	ρ	0.001, 0.10, 0.25

*Sample sizes $n \leq 50$ excluded when $c \geq 5$. Case $c = 1$ represents no clustering.

For each element in the factorial design, $R = 10,000$ replicated samples are drawn. In each simulated sample for which $\widehat{V}(\hat{p}) \neq 0$, the coverage indicator and interval width are computed for the Wald and seven

other 95% CI methods described in Section 2, using both n_{eff} and $\hat{n}_{\text{eff}}$, where the latter is computed from three different variance estimates: the purely design-based estimator (13), the modification by the reciprocal of the effective sample size factor of Dean and Pagano (2015) in (20), or the Kish-type formula (15) derived from superpopulation model assumptions. Empirical coverage in each simulation configuration is the percentage of replicate samples with $\hat{V}(\hat{p}) \neq 0$ for which the interval contains the true proportion Y/N. Non-coverage is (100 - coverage)%. Width is computed as the average of widths of CIs intersected with $[0, 1]$ (needed for Wald, Wilson, and Agresti-Coull).

5 Simulation Results

We present results on CI performance in four steps. First, in Section 5.1 we compare coverage of the Wald CI with coverage of the other CI methods. The clearly inferior CI coverage of Wald eliminates this method from further consideration. Second, we summarize the performance of methods other than Wald across all simulation configurations, first with design-based estimated effective sample size and then with true effective size n_{eff}. Third, in Section 5.2, we compare the Dean and Pagano estimated effective sample size (20) with the Kish-type estimates $\hat{n}^*_{\text{eff}}$ in (18). Finally, in Section 5.3, we compare the relative merits of the seven non-Wald intervals based on $\hat{n}^*_{\text{eff}}$, with results for the Logit method discussed separately in Section 5.3.1, and illustrated in the Online Supplement.

5.1 Coverage with design-based sampling variance estimate

We first examine in Figure 1 seven intervals (Wald, Uniform, Clopper Pearson, Wilson, Agresti-Coull, and Arcsine Square Root) computed from estimated (in the left panel) or true (in the right panel) effective sample size, given by (14) and (12), respectively. Figure 1 plots coverage, based on this design-based estimate of design effect, against "effective expected number of successes" $(n_{\text{eff}} \cdot \theta)$, a feature which increases with n/c and θ. There are 7 plotted points for each element of our factorial design, plotted red for Wald intervals and gray for the others, and each point summarizes $10,000$ samples.

All intervals with estimated effective sample size under-cover, especially for small $n_{\text{eff}} \cdot \theta$, but the lesser

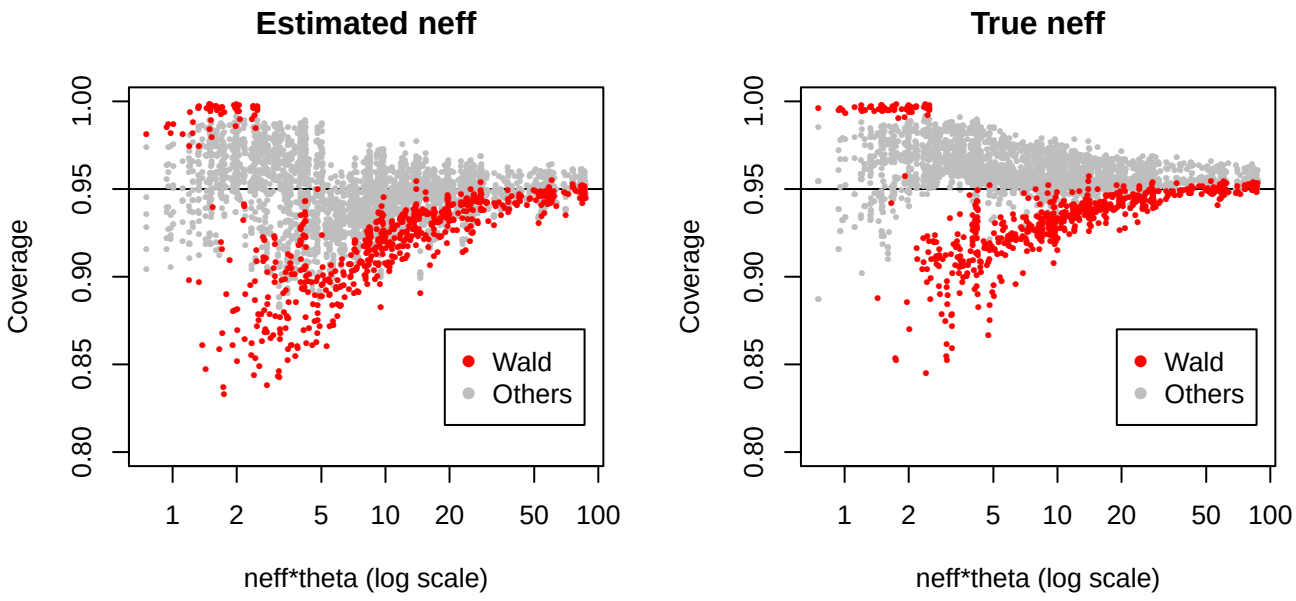

Figure 1: Left Panel: Coverage of seven CIs (all of those in Section 2 except Logit) using effective sample size estimate (14), plotted against effective expected number of successes ($n_{\text{eff}} \cdot \theta$, plotted on the log scale), for each simulation configuration, where the red points correspond to the Wald interval and the gray points correspond to all other intervals. Right Panel: Analogous to left panel, using the true effective sample size (12) instead of (14).

coverage of the Wald CI relative to others is evident (Fig. 1, left panel). For CIs other than Wald, undercoverage is rarely a major problem when the true design effect is known. The Wald interval does very poorly even when the design effect is known (in the right panel), and Figure 1 sufficiently justifies eliminating it from consideration.

The format of plots in Figure 2 is the same as that of Figure 1, but with coverage plotted for only one CI method in each row, and the Wald interval excluded. All 6 CIs tend to be conservative when based on the true design effect (right panel of Fig. 2). The Clopper-Pearson interval with estimated n_{eff} tends to over-cover, at the expense of very large width (see Section 5.3). For all methods, coverage tends to the nominal as $n_{\text{eff}} \cdot \theta$ increases, but convergence can be slow. (Note the log scale on the horizontal axis.)

In practice the sampling variance is unknown, and comparison of the left and right panels of Figure 2 suggests that variance estimation is the primary source of undercoverage in CIs from complex surveys, so that improving the estimate of variance (and hence of the effective sample size) will improve CI coverage.

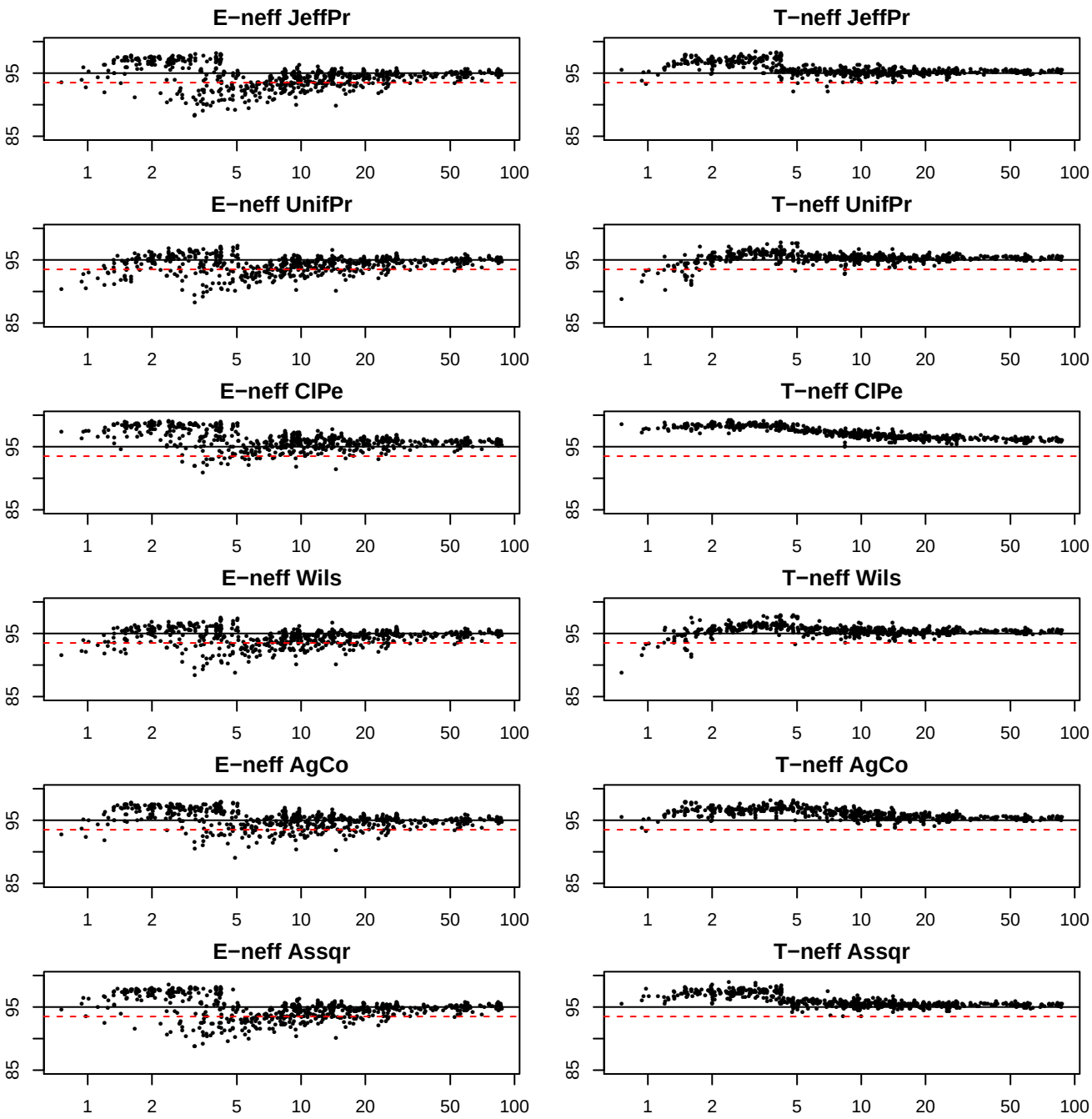

Figure 2: Left Panels: Coverage of 6 CIs using design-based effective sample size estimate, plotted against effective expected number of successes ($n_{\text{eff}} \cdot \theta$, plotted on the log scale), for each simulation configuration. `JeffPr` refers to Jeffreys interval, `UniPr` to Uniform, `ClPe` to Clopper-Pearson, `Wils` to Wilson, `AgCo` to Agresti-Coull, and `Assqr` to Arcsine Square Root interval. Solid line at 95 represents nominal coverage, and dashed line at 93.5 undercoverage by 1.5%. Right panels: Analogous to left panels, using the true effective sample size (12) instead of (14)

5.2 Adjustments & Alternatives to Design-Based Estimates

Motivated by the good coverage properties of all CIs other than Wald with true effective sample size in the right panels of Figure 2, the next subsection directly examines the improved mean-squared error (MSE) achieved by estimating sampling variance with (15) in place of (13). We compare in subsection 5.2.2 the performance of non-Wald intervals using the Dean-Pagano modification to the estimated effective sample size versus CIs based on formula (15).

5.2.1 Properties of the alternative design effect estimator

The Kish-type formula (15) exploits a superpopulation model. Although the design-effect estimator in (14) is essentially unbiased, the corresponding effective sample size estimator is not. The effective sample size estimator (18) corresponding to the model-based variance estimator (15) has some biases that vary systematically with cluster size θ and ICC ρ. Table 1 displays within the simulation design of Section 4, for $n = 200$, the simulation-averaged ratio of the estimated effective sample size (18) divided by the true design effect (12) averaged across scenarios.

The biases of estimated effective sample size (12) turn out to be very slight when $\rho = 0.1$ (not shown) or when $c = 1$, are largest when $\theta = 0.1$, are negative by up to 22% when $\rho = .001$ and $c \geq 3$, and can be quite positive when $\rho = 0.25$ and $\theta = 0.1$. These biases are tolerable because the MSE of the effective sample size estimator (18) is low compared to that of (14).

The biases in estimated effective sample size illustrated in Table 1, as well as those not shown, are generally associated with upward bias in the corresponding design-effect estimates (18). However, these biases in estimating design effect and effective sample size in the Kish method are generally accompanied by a notable decrease in RMSE by comparison with the purely design-based estimators. Table 2 shows the ratio of RMSE for estimated effective sample size (18) over the RMSE of the corresponding design-based estimate in (14) for ICC $\rho = 0.25$ and $n = 200$, by Scenario. Parameter combinations with $c = 1$ are not shown, since in those cases the Kish and design-based estimators are algebraically equivalent. The table shows the considerable improvements in RMSEs when using the Kish method relative to the design-based method. For other values of n, the pattern is the same as that shown, with RMSE ratios often even smaller

	$\rho = .001$			$\rho = 0.25$		
Clus-Size	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.3$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.3$
c=1	1.07	1.02	1.01	1.07	1.02	1.01
c=3	0.98	0.93	0.90	1.16	1.03	1.00
c=5	0.89	0.86	0.86	1.27	1.02	1.02
c=7	0.82	0.78	0.79	1.33	1.10	1.00

Table 1: Average ratio of estimated effective sample size (18) divided by true effective sample size, based on $10,000$ replications and averaged across Scenarios, with $N = 10^4, n = 200$.

(many in the range 0.3–0.7). The favorable performance of the Kish method in confidence interval construction appears to be due to its reduced RMSE in combination with its positive biases in estimated design effect which increase with cluster size.

	$\theta = 0.1$			$\theta = 0.2$			$\theta = 0.3$		
Scen.	c=3	c=5	c=7	c=3	c=5	c=7	c=3	c=5	c=7
C	0.92	0.88	0.85	0.94	0.90	0.86	0.94	0.90	0.87
I	0.89	0.83	0.83	0.84	0.83	0.77	0.82	0.78	0.72
D	0.83	0.83	0.73	0.68	0.54	0.58	0.58	0.49	0.44
H	0.88	0.85	0.81	0.88	0.81	0.79	0.91	0.82	0.81

Table 2: Ratio of RMSE for estimated effective sample size (18) over RMSE for design-based estimated effective sample size(14) for ICC $\rho = 0.25$ and $n = 200$, by Scenario, based on $N = 10^4$ and 10,000 replications.

It should be noted that in the exhibits of this subsection, as elsewhere in the paper's displays of simulation results, a new and independent random population of size N is generated for each simulation configuration. Accordingly, each cell in the tables and point in the figures has inherent variability in repeated runs due to finite population differences. Nevertheless, the patterns described in the paper are fairly consistent and stable and support general conclusions.

5.2.2 Comparison of Kish-type formula CIs to Dean-Pagano CIs

We now discuss CI results for the Kish-formula (15)–(18) method of estimating effective sample size – which we refer to as the Kish n_{eff} method – versus the Dean and Pagano (DP) method applying the modification (20) to the design-based effective sample size (14).

Briefly, the two methods are broadly similar in their coverage rates, although the Kish method tends to have slightly higher coverage. When there is no clustering (i.e., $c = 1$), undercoverage is not a big problem and the Kish method is essentially the same as the design-based method. For $c = 3$, undercoverage is frequent when using the design-based method, and both the DP modification and the Kish methods reduce it to a similar extent. In configurations with $c \geq 3$, there are slightly more configurations aggregated across the six non-Wald intervals considered in this subsection in which DP coverage falls below 93.5% , 94% or 94.5% as compared with Kish, and this comparison holds for almost every combination of θ and ρ when $c > 1$ and $n > 50$ (tables shown in the Online Supplement).

For large cluster-size c and ICC ρ, undercoverage for either the DP or Kish method is common. The most problematic setting is $c = 7$, and Figure 3 contrasts the methods in this case. In each panel labeled by a CI method, the ratio of average interval lengths with effective sample size estimated by the Kish method over the DP method is plotted against the non-coverage ratio under the two n_{eff} methods. For all CI types, most points have one ratio > 1 and one < 1. Among points with width ratios > 1 and non-coverage ratios < 1, the cyan ones for which DP coverage was below nominal can be viewed as favorable for the Kish method, and perhaps so are the black points with width-ratios < 1 and noncoverage ratios > 1 and above-nominal DP coverage. The points in the lower-left quadrant in each panel are very favorable to the Kish method because they reflect settings in which Kish-method coverage is larger than for DP while width is smaller. Notably, none of the CI types show any upper-right quadrant configurations in which the DP method would have smaller average width but larger coverage. Similar pictures for $c = 3$ and $c = 5$, included in the Online Supplement, show a similar pattern but not quite so strikingly favorable to the Kish n_{eff} method over the DP method.

Figure 4 presents another view of the coverage of the same six CI types as Figure 3 computed with the Kish versus DP n_{eff} method. In each of these 6 panels, which correspond to the case $c = 7$, the middle range of DP points with near-nominal (94%–96%) coverage correspond to a range 94%–97.5% of Kish-method coverage. This observation is consistent with the width-comparisons from Figure 3. Analogous figures for the cases $c = 3, 5$ are presented in the Online Supplement.

Both the DP and the Kish methods have increased width over all (non-Wald) types compared to the design-based method of estimating sampling variance and n_{eff}. In fact, for the six intervals considered in

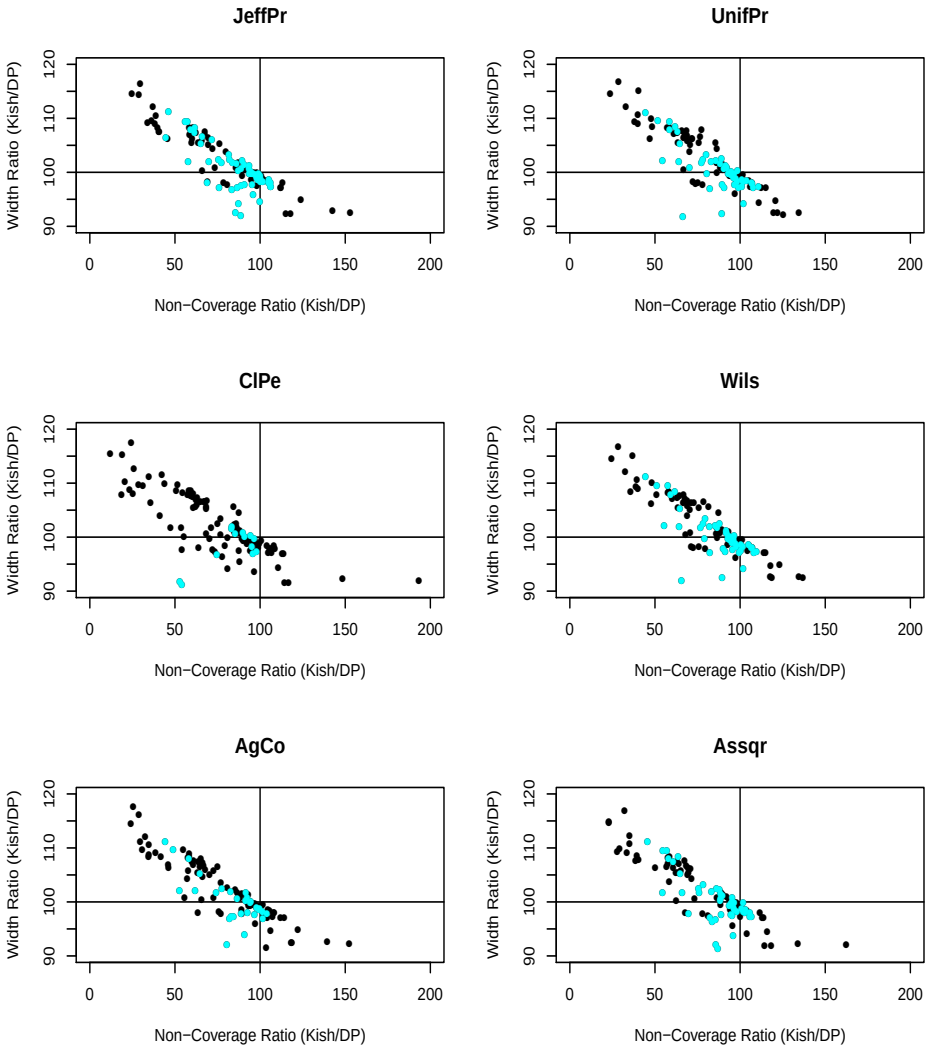

Figure 3: Comparison of metrics for Kish versus DP n_{eff} methods, for each of six CIs under 108 simulation settings with $c = 7$. Plotted points are: y = 100 times ratio of widths for Kish n_{eff} method over DP, versus x = 100 times ratio of non-coverage for Kish n_{eff} method over non-coverage for DP. Points with below-nominal DP coverage plotted in cyan. Vertical line indicates coverage ratio 1, horizontal line width-ratio 1.

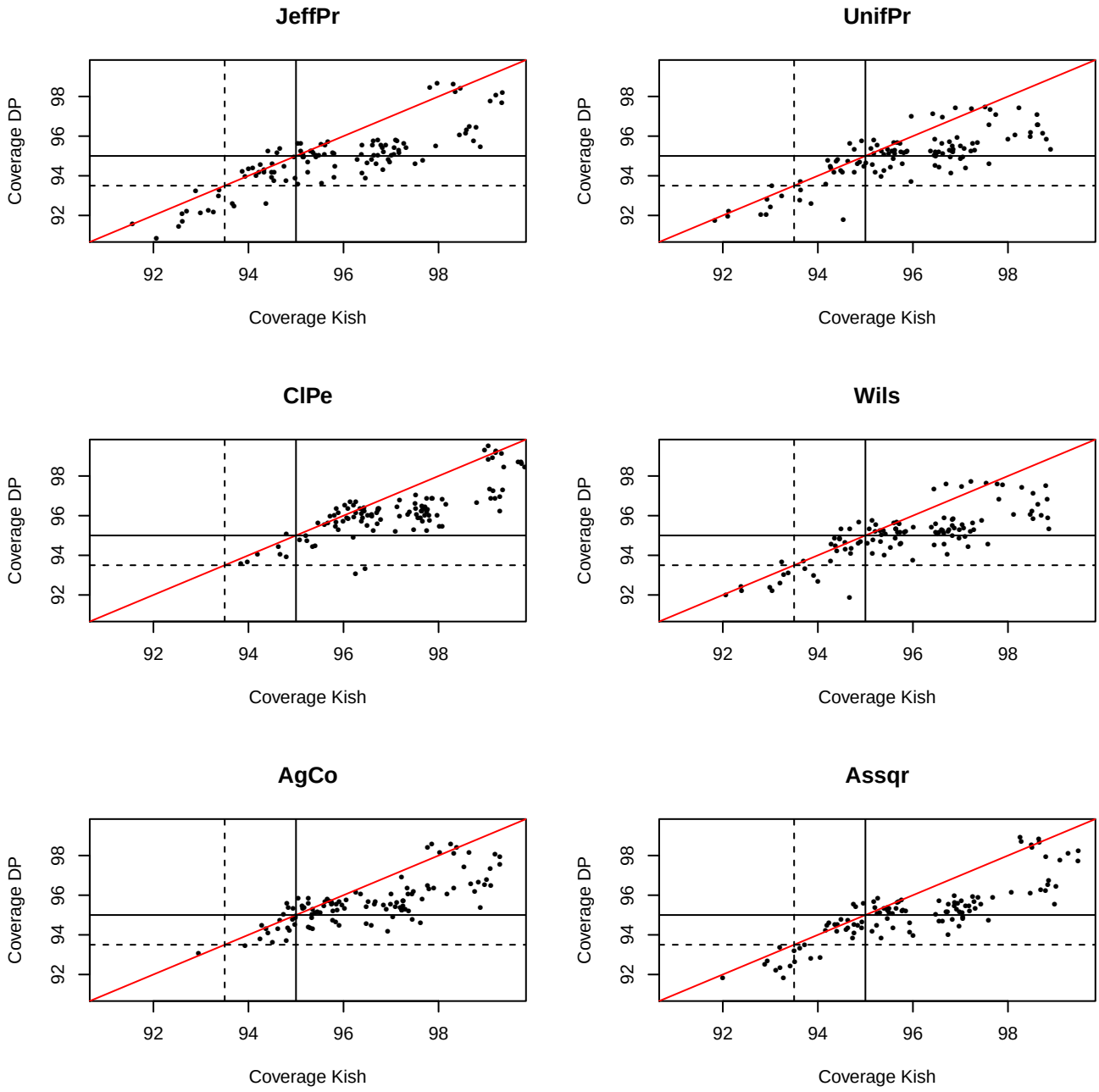

Figure 4: Coverage for 6 CI types, scaled by 100, for the Kish and DP n_{eff} methods in 108 simulation configurations with $c = 7$. Equal coverage is indicated by red $45°$ line, nominal (95%) coverage by black solid lines, and extreme (93.5%) undercoverage by black dashed lines.

this subsection, the Kish method increases width 0–41.1% with a mean increase of 6.9%, and the DP modification increases width 0.4–50.6% with a mean of 6.8%. Figure 4 indicates the somewhat higher coverage for the Kish versus DP method for each interval type. The increased coverage is acceptable because the overall message from the Kish versus DP comparisons is that the Kish method makes more effective use than DP of CI widths, with slightly better success at mitigating undercoverage in the presence of clustering.

5.3 Comparison of Alternative Intervals

We move now to highlight relative advantages among the non-Wald CI types of Sec. 2. In this comparison, we examine results using the Kish method (18) of n_{eff} estimation. Considering first the rather good coverage properties of these CI types based on true n_{eff} in the rightmost panels of Figure 2, the coverage performance of the Jeffreys- and Uniform-prior and Wilson intervals seem most favorable to us: Clopper-Pearson is excessively conservative, with systematically above-nominal coverage also for Agresti-Coull and Arcsine Square Root. When n_{eff} is estimated, the leftmost panels in Figure 2 show Clopper-Pearson to be overly conservative, and Arcsine Square Root erratic and dominated across the range of $\theta \cdot n_{\text{eff}}$ by the Jeffreys-prior interval, but it is rather hard to choose among the Jeffreys, Uniform, Wilson and Agresti-Coull alternatives. In the presence of extensive clustering $(c = 7)$, Figure 4 shows somewhat more detail, but also does not provide a compelling reason to prefer any of the Jeffreys, Uniform, Wilson and Agresti-Coull to the others, although Jeffreys has a slightly wider range of coverage and Agresti-Coull a more systematically conservative tendency than the others in this group.

Width and coverage are shown simultaneously for the six CI types in a further pictorial display in Figures 5 and 6. In these, we plot the ratio of each interval width to that of Clopper-Pearson (since that CI is typically the widest and has highest coverage) versus the non-coverage, plotting a separate panel for each level of clustering and within each c, for each overall proportion θ. Points in the lower left of each panel have high coverage and small widths. The definition of relative width removes much of the dependence on θ and n. Figure 5 contains the three θ panels for $c = 1$, and Figure 6 for $c = 5$. The other cases $c = 3$, $c = 7$ can be found in the Supplement.

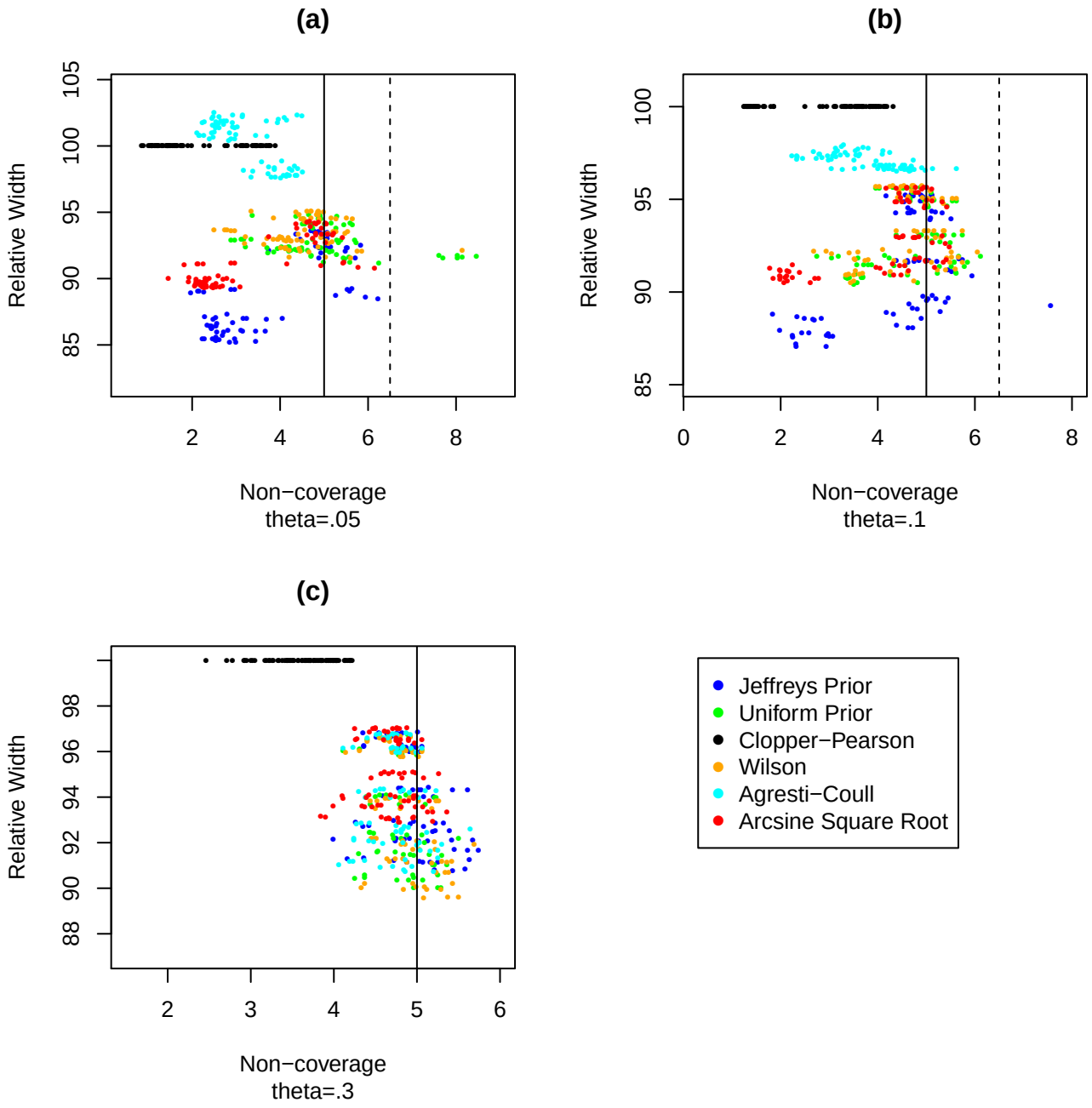

Figure 5: Relative width (ratio of width of Jeffreys, Uniform, Clopper-Pearson, Wilson, and Agresti-Coull interval to that of the Clopper-Pearson, multiplied by 100) vs non-coverage for 72 configurations with no clustering ($c = 1$) and (a) $\theta = 0.05$, (b) $\theta = 0.1$, (c) $\theta = 0.3$. The solid vertical line represents nominal non-coverage. The dotted line, given for reference, represents undercoverage of 1.5 percentage points.

Some patterns are common to all these plots: the Agresti-Coull and Clopper-Pearson tend to be the widest for $\theta = 0.05, 0.1$, and also tend to have higher coverage, typically overcovering. The other intervals are

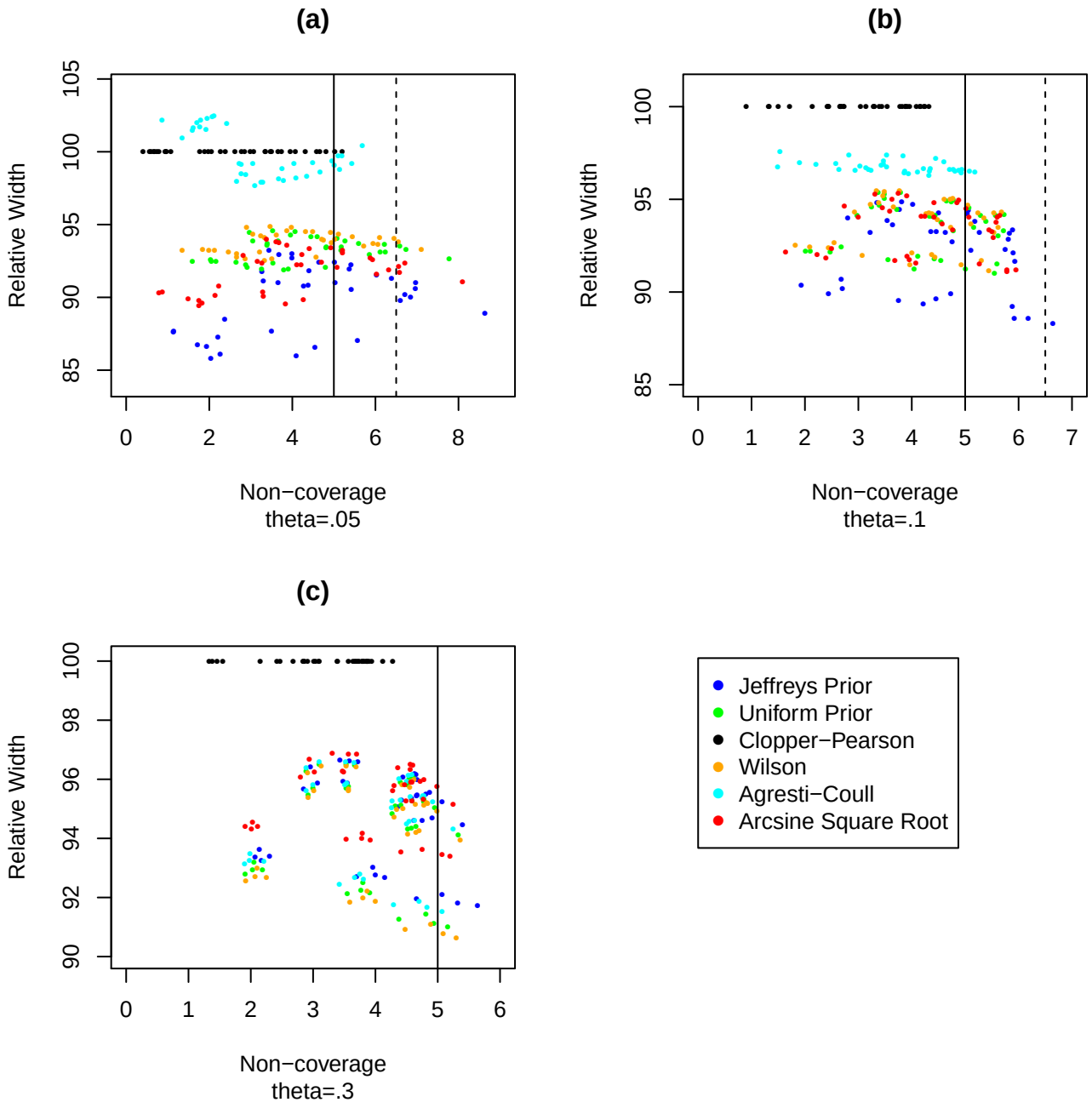

Figure 6: Ratio of width of Jeffreys, Uniform, Clopper-Pearson, Wilson, Agresti-Coull and Arcsine Square Root to that of the Clopper-Pearson, multiplied by 100, plotted vs non-coverage for 36 simulation configurations with $c = 5$ and (a) $\theta = 0.05$, (b) $\theta = 0.1$, (c) $\theta = 0.3$. The solid vertical line represents nominal non-coverage. The dotted line, given for reference, represents undercoverage of 1.5 percentage points.

comparable to each other in coverage, with the Jeffreys the shortest, though having slightly more cases of marked undercoverage, especially for ICC ρ of 0.25. For $\theta = 0.3$, the Clopper Pearson is the widest and

most conservative. The other intervals are comparable in coverage, with the Wilson and Uniform tending to be the shortest.

In Figure 5 there is a general tendency towards over-coverage and few cases of pronounced undercoverage. Panels (a) and (b), respectively for $\theta = .05$ and 0.1, show that the Clopper-Pearson (CP), Agresti-Coull (AC), Jeffreys, and Arcsine Square Root intervals all have many over-coverage events in the region where non-coverage is less than 4%, but CP and AC are widest in this region, achieving over-coverage at the cost of increased width. AC can be wider than CP for $\theta = 0.05$. Among other CIs, Jeffreys strikes a good width/coverage trade-off, but not in all situations (showing marked undercoverage $\theta = 0.1$ and $n = 50$). The Wilson and Uniform CIs are relatively well-calibrated, though there are a few instances of non-coverage beyond 7% for $\theta = 0.05$ and $n = 30$, most of them for the Uniform. For $\theta = 0.3$, the Wilson and the Uniform intervals tend to be shortest, with Jeffreys shortest for $\theta = 0.05$ and $\theta = 0.1$.

In Figure 6, many of the same comments apply, including the excessive width of Agresti-Coull and Clopper-Pearson for $\theta = .05$, the well-calibrated Uniform and Wilson, the short but occasionally severely undercovering Jeffreys, and somewhat erratic Arcsine Square Root.

The most pronounced undercoverage for all intervals occurs when $c = 7$ and $\rho = 0.25$. Even the Clopper-Pearson undercovers there, although not more than 1.5%.

Some of the width relationships among interval widths can be proved analytically. Specifically, both the Uniform and the Jeffreys intervals are contained in the Clopper-Pearson interval (see Brown et al., 2001). A proof is supplied in Appendix A. Though the proof does not cover the relation between the Jeffreys and the Uniform, we have verified numerically that for $\alpha = 0.10, 0.05, 0.01$ and $n = 2, \ldots, 10,000$ the Jeffreys interval's lower endpoint is always smaller than that of the Uniform's when the binomial count $Y < n/2$. That is, qbeta$(\alpha/2, y + 1/2, n - y + 1/2) <$ qbeta$(\alpha/2, y + 1, n - y + 1)$ for $y < n/2$.

5.3.1 The Logit interval (see the Online Supplement)

In our simulations, the Logit interval shows a similar performance to the Agresti-Coull, but in some cases was extremely wide, as shown in Figures 7-11 of the Online Supplement. Figure 7 in the Supplement is

analogous to Figure 2 in the paper but includes also the Logit interval. Figures 8-11 in the Supplement are analogous to Figure 5-6 in the main paper, covering the cases $c = 1, 3, 5$ and 7, but plotting points corresponding to the Logit in cyan, the Agresti-Coull in blue, the Clopper-Pearson in black, for reference, and all others in gray, to highlight the similar pattern of behavior of the Agresti-Coull and Logit. The strikingly high widths in some cases, seen mostly for $\theta = 0.05$ and sometimes for $\theta = 0.01$, are consistent with the Brown et al. (2001) finding that the Logit interval is "unnecessarily long" in the binomial case.

6 Conclusions and Future Work

We have seen that the Wald CI is badly flawed for estimating proportions in complex surveys due to its severe undercoverage in a variety of situations. Improving the estimation of sampling variance will not salvage the Wald interval, which performs poorly even when the true sampling variance is known. Since the alternative methods studied are straightforward to implement and clearly superior, the Wald approach should not be used, especially not in complex surveys.

For the other intervals considered, notable undercoverage can also occur when there is clustering. Improving the estimation of sampling variance by using simple superpopulation model assumptions can greatly enhance the performance of these intervals. This approach worked well throughout our factorial design, better than the modification of effective sample size by Dean and Pagano (2015), and can be applied more generally. This approach to improving coverage by improving estimation of the effective sample size is perhaps our main contribution.

Among the CI methods studied, there was no clear winner with respect to coverage or length. Our comparisons of coverage and lengths suggest the Wilson, Uniform, and Jeffreys intervals tended to have shorter lengths (the former two especially for larger θ such as $\theta = 0.3$ and the latter for smaller θ), and coverage closest to nominal. The Clopper-Pearson interval, recommended by Korn and Graubard (1999) and by Dean and Pagano (2015) in cases with high clustering and extreme proportions, tends to be much longer, and should only be used if conservative coverage is paramount.

Our method of estimating the effective sample size has been developed and tested for the

Horvitz-Thompson estimator under stratified one stage sampling of clusters of equal sizes. The appendix extends the method to unequal cluster sizes (equation 24), and to the case where the weights might not be inverse inclusion probabilities, but design-consistent variance estimates of the stratum totals are available (equation 25) and may come, for instance, from random groups, Balanced Repeated Replication (Wolter, 1985), jackknife or bootstrap (Shao and Tu 1995). Future research will extend and test the method using other designs and other types of estimators. The ratio estimator or combined ratio estimator will be particularly relevant, as these are frequently used in surveys to achieve gains in precision in estimating proportions when the cluster sizes are not equal or when a good auxiliary variable is available (see for instance, Lohr, 2010). We expect that under moderate misspecification of the sampling design or the model the method will still perform well. Further research about the impact of model misspecification is recommended. In particular, our simulations do not test the performance of the method in settings with cross-cluster correlation. Sampling variance estimators can also be developed using the same ideas under other super-population model assumptions, e.g., allowing for other correlation structures, but care must be taken that the number of parameters to be estimated is not too large given the sample size.

In the case where a data user is only provided with replicate weights in a public-use data file, with no information about clustering, our method of estimating sampling variances will not apply. Even when our method cannot be used, a strong recommendation still emerges from our simulations–that the Wald interval not be used, and be replaced by the preferred non-Wald method, where the best available sampling variance estimate is used to compute the effective sample size and effective sample count as described in Section 2.8, and the effective sample count and effective sample size are then used in the confidence interval formulas (2.2)- (2.6). Possible variance estimators include those based directly on supplied weight-replicates, or random-group or jackknife estimators in which weights and replicates are used to define the groups, or others such as bootstrap variances in complex surveys when those can be justified as consistent (see, for instance, Rust and Rao 1996 for a review of replication techniques for variance estimation in complex surveys).

Several other lines of investigation of CI performance for proportions based on complex survey data deserve attention. Coverage of all of the intervals tends to fall below nominal as cluster sizes increase, and variants of these intervals, or more urgently of the underlying estimation of effective sample size, which mitigate this tendency are needed. In particular, a fully Bayesian approach with weakly informative prior

distributions, which incorporates complex design features like clustering and stratification through a hierarchical Bayes model appropriate for a binary outcome, deserves consideration. We did not assess this approach, since we confined attention to simple computational approaches that are more readily implemented in ACS-type settings. Indeed, further research is needed to confirm that any method performs well consistently across designs with widely varying (non-constant) cluster sizes and other sorts of inhomogeneity, and it is in such settings where we believe the model-based approach introduced here shows greatest promise.

References

[1] Agresti, A. and Coull, B. (1998), "Approximate is better than 'exact' for interval estimation of binomial proportions," *American Statistician*, 52, 119-126.

[2] Blyth, C. and Still, H. (1983), "Binomial Confidence Intervals," *Journal of the American Statistical Association*, 78, 108-116.

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

[4] Brown, L., Cai, T., and DasGupta, A. (2002), "Confidence Intervals for a Binomial Proportion and Asymptotic Expansions," *Annals of Statistics*, 30, 160-201.

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

[6] Carlin, B. and Louis, T. (2009), *Bayesian Methods for Data Analysis, 3^{rd} ed.* Chapman & Hall/CRC Press, Boca Raton, FL.

[7] Casella, G. and Berger, R. (2002), *Statistical Inference, 2nd ed.* Pacific Grove, CA: Duxbury.

[8] Casella, G. (1986), "Refining Binomial Confidence Intervals," *Canadian Journal of Statistics*, 78, 107-116.

[9] Chen, S. and Rust, K. (2017), "An extension of Kish's Formula for Design Effects to Two and Three-Stage Designs with Stratification," *Jour. Survey Statist. & Methodol.* 5, 111-130.

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

[11] Dean, N. and Pagano, M., "Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys." (2015), *Journal of Survey Statistics and Methodology*, 3, no. 4, 484-503.

[12] Fay, R. and Train, G. (1995), "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics and States and Counties," *Proc. Government Statist. Section*, Alexandria, VA: American Statistical Association, 154-159.

[13] Franco, C., Little, R., Louis, T. and Slud, E. (2014), "Coverage Properties of Confidence Intervals for Proportions in Complex Sample Surveys," *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 1799-1813.

[14] Frankel, M. (1971). *Inference from Survey Samples: An Empirical Investigation.* Ann Arbor, MI: Institute for Social Research, The University of Michigan.

[15] Gabler, S., Haeder, S. and Lahiri,P. (1999), "A Model-Based Justification of Kish's Formula for Design Effects for Weighting and Clustering," *Survey Methodology*, 25, 105-106.

[16] Gilary, A., Maples, J., and Slud E. (2012) "Small Area Confidence Bounds on Small Cell Proportions in Survey Populations," *JSM Proceedings*, Survey Research Section, 3541-3555.

[17] Isaki, C. T., Fuller W. A. (1982)., "Survey Design Under the Regression Superpopulation Model," *Journal of th American Statistical Society*, 77, 377, 89-96

[18] Kalton, G. (1979), "Ultimate Cluster Sampling," *Journal of the Royal Statistical Society–Series A* 142 (2), 210-222.

[19] Kish, L. (1965) *Survey Sampling*, New York: Wiley.

[20] Kish, L. (1987), "Weighting in Deft^2," *The Survey Statistician*, June 1987.

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

[22] Korn, E. and Graubard, B. (1990), "Simultaneous Testing of Regression Coefficients with Complex Survey Data: Use of Bonferroni t-statistics," *Statistica Sinica*, 8, 1131-1151.

[23] Kott, P.S., Andersson, P. G. and Nerman, O. (2001) ,"Two-sided Coverage Intervals for Small Proportions Based on Survey Data," *Proceedings of the FCSM Research Conference*, Washington DC.

[24] Kott, P.S., and Liu, Y. K. (2009), "One-Sided Coverage Intervals for a Proportion Estimated from a Stratified Simple Random Sample." *International Statistical Review*, 77, 2, 251-265

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

[26] R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. URL https://www.R-project.org/

[27] Rust, K. F, and Rao, J. N. K (1996). "Variance Estimation for Complex Surveys Using Replication Techniques." *Statistical Methods in Medical Research*, 5, 283-310.

[28] Lohr, S. (2010), *Sampling: Design and Analysis, 2nd ed.* Boston: Brooks-Cole.

[29] Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*, New York: Springer.

[30] Slud, E. (2012), "Assessment of Zeroes in Survey-Estimated Tables via Small Area Confidence Bounds," *Jour. Indian Soc. Agricultural Statistics*, 66, 157-169.

[31] Sokal, R. and Rohlf, F. (1994), *Biometry: the Principles and Practice of Statistics in Biological Research, 2nd ed.*, New York: W. H. Freeman.

[32] U.S. Census Bureau (2014). *American Community Survey Design and Methodology* [online]. `http://www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/acs_design_methodology_report_2014.pdf`

[33] Wilson, E. (1927), "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association* 22, 209-212.

[34] Wolter, K. (2001) *Variance Estimation, 2nd ed.* New York: Springer.

[35] Wright, T. (1997), "A Simple Algorithm for Tighter Exact Upper Confidence Bounds with Rare Attributes in Finite Universes," *Statistics and Probability Letters*, 36, 59-67.

APPENDIX

A Ordering of Clopper-Pearson, Jeffreys, and Uniform Intervals

Theorem: Let $[L_J(X,n), U_J(X,n)]$, $[L_{UNI}(X,n), U_{UNI}(X,n)]$, $[L_{CP}(X,n), U_{CP}(X,n)]$ be the lower and upper bounds of the Jeffreys, Uniform, and Clopper-Pearson Intervals, respectively. Then

$$[L_J(X,n), U_J(X,n)] \subset [L_{CP}(X,n), U_{CP}(X,n)]$$

$$[L_{UNI}(X,n), U_{UNI}(X,n)] \subset [L_{CP}(X,n), U_{CP}(X,n)]$$

Proof: The result follows from the fact that if f and g are densities for random variables W and Z with cdf's F and G, respectively, such that f/g is increasing, then W is stochastically bigger then Z. To show this, write

$$\frac{d}{du}[F(u) - G(u)] = g(u)(\frac{f(u)}{g(u)} - 1).$$

Because the ratio $f(u)/g(u)$ is increasing, the derivative on the left hand side can only change from negative to positive. Hence the function $F(u) - G(u)$ can have only one local minimum, and its value is zero at ∞ and $-\infty$. Hence $F(u) - G(u) \leq 0$. The Jeffreys, and Clopper-Pearson endpoints can all be expressed from the quantiles of the beta distribution, as described in Section 2. The result follows by taking the ratios of the beta densities for each of the interval endpoints and showing that each is decreasing or increasing. For instance, in terms of the Beta function $\beta(a,b)$

$$f(u) = \frac{\text{dbeta}(u, X+1/2, n-X+1/2)}{\text{dbeta}(u, X, n-X+1)} = \frac{\beta(X+1/2, n-X+1/2)}{\beta(X, n-X+1)} \{u/(1-u)\}^{1/2}$$

It is easy to check that $f'(u) > 0$, so $f(u)$ is increasing. This implies

$$L_J(X,n) > L_{CP}(X,n).$$

The relations between the other endpoints are proved analogously. □

B Justification of Model-based Estimation Formulas

Consider the survey-weighted estimator of Y applicable in survey settings with stratification and single-stage cluster sampling, but not necessarily equal-sized clusters or stratumwise SRS single-stage cluster sampling. In terms of single and joint inclusion probabilities $\pi_{hki},\ \pi_{hki,hk'j}$, the general survey-weighted (Horvitz-Thompson) estimator of Y becomes

$$\hat{Y} \ = \ \hat{Y}^{\mathrm{HT}} \ = \ \sum_{h=1}^{H} \sum_{k \in S_h} \sum_{i \in C_{kh}} Y_{khi}/\pi_{hki} \tag{22}$$

and a general expression for the *anticipated variance* (Isaki and Fuller, 1982), where the expectation is taken with respect to both the sampling design and the super population model, is

$$\mathrm{AV}(\hat{Y}^{HT}) \ = \ \sum_{h=1}^{H} \sum_{k,k' \in \mathcal{U}_h} \sum_{i \in C_{kh},\, j \in C_{k'h}} \frac{\pi_{hki,hk'j} - \pi_{hki}\pi_{hk'j}}{\pi_{hki}\,\pi_{hk'j}} \Big[E(Y_{hki})\, E(Y_{hk'j}) \ + \ \mathrm{Cov}(Y_{hki},\, Y_{hk'j}) \Big] \tag{23}$$

where $\mathcal{U}_h$ is the set of clusters in stratum h of the population.

Our blanket assumption is that

(A.o) clusters are sampled all-or-none

which implies that the single and double inclusion probabilities are constant over clusters, so we drop the indices i, j from their notation. Let M_{kh} denote the number of units i in cluster C_{kh}. Then the extra assumptions underlying the formula simplifications in Sections 3.1 and 3.2 are that all M_{kh} are equal to c and that

(A.iv) the single and pairwise inclusion probabilities are equal to those of stratified SRS cluster sampling:

$$\pi_{hk} \ = \ \frac{n_h^C}{K_h} \quad , \qquad \pi_{hk,hk'} \ = \ \frac{n_h^C\,(n_h^C - 1)}{K_h(K_h - 1)} \quad \text{for} \qquad k \in U_h, \quad k' \neq k$$

Under assumptions (A.o)-(A.ii) and (A.iii), the mean and variance of the cluster-attribute $Y_{kh+} \ = \ \sum_{i \in C_{kh}} Y_{khi}$ are given by

$$E(Y_{kh+}) \ = \ M_{kh}\,\tau_h \ , \qquad V(Y_{kh+}) \ = \ \{M_{kh} \ + \ M_{kh}(M_{kh} - 1)\rho\}\,\sigma_h^2$$

Then formula (23) simplifies to

$$\mathrm{AV}(\hat{Y}^{HT}) \;=\; \sum_{h=1}^{H} \tau_h^2 \sum_{k,k'=1}^{K_h} \frac{\pi_{hk,hk} \;-\; \pi_{hk}\,\pi_{hk'}}{\pi_{hk}\,\pi_{hk'}}\, M_{kh}\, M_{k'h} + \sum_{h=1}^{H} \sum_{k=1}^{K_h} \frac{1-\pi_{hk}}{\pi_{hk}} \left\{ M_{kh} + M_{kh}(M_{kh}{-}1)\rho \right\} \sigma_h^2 \tag{24}$$

However, in many real surveys where the weights w_i are based on calibration, raking, nonresponse adjustment and/or weight-trimming steps, the fiction that these weights are inverse inclusion probabilities cannot be maintained, and therefore anticipated-variance formulas like (23) must be replaced by some off-the-shelf design-consistent variance estimation method such as random-groups or Balanced Repeated Replication (Wolter 1985), jackknife or bootstrap (Shao and Tu 1995). Let $\hat{V}_h(\mathbf{z}) \;=\; \hat{V}_h(\{z_k\}_k)$ denote the estimated variance in stratum h by any of these methods applicable to the total of a cluster-level attribute z_k for $k \in \{1, \ldots, K_h\}$. Then, under assumptions (A.i),(A.ii) and (A.iii), the anticipated variance AV in (23) or (24) is estimated by

$$\widehat{\mathrm{AV}}(\hat{Y}^{HT}) \;=\; \sum_{h=1}^{H} \hat{\tau}_h^2\; \hat{V}_h(\{M_{kh}\}_k) \;+\; \sum_{h=1}^{H} \hat{\sigma}_h^2 \sum_{k \in S_h} \frac{M_{kh}}{\pi_{kh}} \left(\frac{1-\pi_{kh}}{\pi_{kh}} \right) (1 + (M_{kh}-1)\hat{\rho}) \tag{25}$$

where $\hat{\tau}_h,\, \hat{\sigma}_h^2,\, \hat{\rho}$ are design-based estimators derived from sample-weighted moments. Natural formulas for such estimators are:

$$1 - \hat{\rho} \;=\; \frac{n-H}{n} \;\cdot\; \frac{\sum_{h=1}^{H} \sum_{k \in S_h} \sum_{i,j \in C_{kh}} (Y_{hki} - Y_{hkj})^2/2}{\sum_{h=1}^{H} \hat{\tau}_h(1-\hat{\tau}_h) \sum_{k \in S_h} M_{kh}\,(M_{kh}-1)}$$

$$\hat{\sigma}_h^2 \;=\; \frac{1}{n_h - 1} \sum_{k \in S_h} \sum_{i \in C_{kh}} (Y_{hki} - \hat{\tau}_h)^2 \quad , \qquad \hat{\tau}_h \;=\; \frac{1}{n_h} \sum_{k \in S_h} Y_{hk+}$$

where $n_h = \sum_{k \in S_h} M_{kh}$. In a complex survey with stratification and take-all clusters, formula (25) provides a general variance estimator as part of our proposed design-effect and CI estimators. If assumption (A.iv) also holds, then these variance-estimation formulas take the explicit form:

$$\widehat{\mathrm{AV}}^{*}(\hat{Y}) \;=\; \sum_{h=1}^{H} \frac{K_h - n_h^C}{n_h^C} \Big\{ \hat{\tau}_h^2\, \frac{K_h}{n_h^C - 1} \sum_{k \in S_h} (M_{kh} - \frac{n_h}{n_h^C})^2 \;+\; \hat{\sigma}_h^2\, \frac{K_h}{n_h^C} \Big(n_h\,(1-\hat{\rho}) + \hat{\rho} \sum_{k \in S_h} M_{kh}^2 \Big) \Big\} \tag{26}$$

where $\hat{\tau}_h, \hat{\sigma}_h^2$ and $\hat{\rho}$ are as above. Then the expressions for estimated design effect and effective sample size are as in (18).

Note that in the further restricted setting of Section 3.2, where $M_{kh} = c$ is constant, formula (26) reduces to (15), which in the main text we denote as $\widehat{\text{Var}}^*(\hat{Y})$ for simplicity of exposition. This formula would imply the usual Kish formula Deff$= 1 + \hat{\rho}(c-1)$ under the further assumption of proportional allocation with cluster sample sizes n_h^C to stratum h, if $\hat{\sigma}_h^2$ were changed to a variance assumed not to depend on stratum. With proportional allocation, the design-based estimated sampling variance $\widehat{\text{Var}}(\hat{Y})$ in (13) is essentially equal to $\widehat{\text{Var}}^*(\hat{Y})$, the only difference being the $K_h - 1$ terms instead of K_h in the denominators of stratum sample variances in (13).

The idea of this Section has been that approximately design-based estimators of the variance of survey-weighted estimates of proportions, with anticipated variance calculated on the basis of slightly misspecified models expressed in terms of a few unknown parameters to be estimated, can be an effective and numerically stable way to produce estimated design effects.

C Additional Simulation Details

Patterns of joint variation of stratum sampling fractions $f_h = n_h/N_h$ and stratum proportions θ_h.

$$\theta_h = \theta\left(1 + b_0\,(h - \frac{H+1}{2})\right) \quad , \quad \bar{f}_h \propto \theta_h^{a_1}\,(2\theta - \theta_h)^{a_2} \quad , \quad 1 \le h \le H \tag{27}$$

Initially we consider four patterns of parameters for the pair of H-vectors:

C:	$b_0 = a_1 = a_2 = 0$	level θ's, $\bar{f}$'s
I:	$b_0 = 3/(4H-4)$, $a_1 = 1$, $a_2 = 0$	$\nearrow$ θ's, $\nearrow$ $\bar{f}$'s
D:	$b_0 = 3/(4H-4)$, $a_1 = 0$, $a_2 = 1$	$\nearrow$ θ's, $\searrow$ $\bar{f}$'s
H:	$b_0 = 3/(4H-4)$, $a_1 = 1$, $a_2 = 1$	$\nearrow$ θ's, $\cap$ shaped $\bar{f}$'s

The purpose of the choice $b_0 = 3/(4H-4)$ in the scenarios **I**, **D**, and **H** with $H = 4$ strata is to fix the smallest of the ratios θ_h/θ as $5/8$, and the largest as $11/8$. Of course, other choices are possible: to set the minimum value θ_1/θ of these ratios to $\beta_0 < 1/2$, fix $b_0 = 2(1-\beta_0)/(H-1)$. The intent was to choose values that would allow for differences between the scenarios while avoiding extreme values that would

lead to poorly behaved and unusual results.

Rounding of Stratum Sample Sizes

Stratum sample sizes to be calculated following the rules given above need to be rounded to become integers and the corresponding proportions are changed to reflect the integer population and sample sizes. Thus: with `round` denoting the operation of rounding a number to the nearest integer,
$n = \texttt{round}(fN/c)$, then $f \equiv cn/N$; then $N_h = \texttt{round}(N\lambda_h)$, $\lambda_h \equiv N_h/N$;
then from the values $\bar{f}_h$ defined by (27), $n_h^C = \texttt{round}(\bar{f}_h N_h/c)$, $f_h \equiv n_h^C c/N_h$.

D R Code for Design Effect and CI Calculations

In Section 1 of the Online Supplement, we describe the use of two `R` (R Core Team, 2017) functions adapted from those used in the simulations of Section 4, `VarKish` and `CIarrFcn`, and provide an illustrated example based on simulated data. The `R` function `VarKish` calculates the design-based and 'Kish-type' variances for a survey-weighted total of a binary attribute Y_{hki} in the setting of a stratified single-stage cluster-sample (in which all units are taken from each sampled cluster). The `R` function `CIarrFcn` encodes the calculation of all 8 types of confidence intervals studied in this paper. These functions along with parameter values and data objects used in the illustration are contained in the supplementary `R` workspace `RSupp.RData`, where function listings and the data objects can be found. After explaining the inputs and outputs of the functions in successive subsections, we present a a detailed example of the use of these functions similar to the way they were applied in the simulations.