Journal of Official Statistics, Vol. 25, No. 4, 2009, pp. 569–588

# Evaluating Alternative One-Sided Coverage Intervals for a Proportion

*Yan K. Liu*[1] *and Phillip S. Kott*[2]

The construction of coverage intervals for a proportion is difficult, especially when the proportion is very small or very large. Most of the methods treated in the literature implicitly assume simple random sampling. These interval-construction methods are not immediately applicable to data derived from a complex sample design. Some recent papers have addressed this problem, proposing modifications for complex samples. Matters are further complicated when a one-sided coverage interval is desired. This article provides an extensive review of existing methods for constructing coverage intervals for a proportion under both simple random and complex sample designs. It also evaluates the empirical performances of different one-sided coverage intervals under both a simple random and a stratified random sample design.

*Key words:* Coverage probability; effective sample size; stratified random sample.

## 1. Introduction

In survey practice, we are often interested in constructing a coverage interval for a proportion of successes in the population that represents one of two outcomes. The binomial distribution is frequently used to model the number of successes in a simple random sample of size $n$ from a population of size $N$. If the sample selections are not independent (i.e., sampling is *without* replacement), the resulting distribution is a hypergeometric distribution, not a binomial one. However, for $N$ much larger than $n$, the binomial distribution is a good approximation and widely used. Therefore, the proportion is also referred as the binomial proportion when the population size is large enough. Throughout this article, we use the term "coverage interval" instead of "confidence interval," which is explained at the end of Section 2.

It is well-known that the standard Wald method for constructing coverage intervals around a proportion behaves erratically, especially when the proportion is near 0 or 1. Its coverage probability can be severely under or over the nominal level even when sample size is large. Because of the poor performance of the standard Wald method, the literature contains a series of modifications, alternative methods, and comparisons for a two-sided coverage interval under a simple random sample design (Brown et al. 2001; Agresti

[1] Statistics of Income Division, IRS, P.O. Box 2608, Washington, DC 20013, U.S.A. Email: yan.k.liu@irs.gov
[2] RTI International, 6110 Executive Blvd., Rockville, MD 20852, U.S.A. Email: pkott@rti.org
**Acknowledgments:** The authors express a special note of thanks to Kevin Cecco at IRS for his encouragement and direction at the beginning of this project. The authors are very grateful to five anonymous referees and the associate editor for their valuable comments and suggestions.

and Coull 1998; Vollset 1993; Clopper and Pearson 1934). Some recent papers have addressed this problem under more complex sample designs (Feng 2006; Sukasih and Jang 2005; Kott et al. 2001; Korn and Graubard 1998).

The construction of coverage intervals for a proportion is difficult because the sampling distribution of the proportion does not closely follow its normal approximation and because the binomial distribution has the lattice structure. Constructing empirically effective one-sided coverage intervals can be an even more difficult task because of the skewness of the sampling distribution of the proportion. Cai (2004) and Hall (1982) use an Edgeworth expansion to develop one-sided coverage intervals under a simple random sample. Kott and Liu (2009) modify the Hall method and extend it to handle data from a complex sample design with a particular emphasis on stratified (simple) random sampling.

We are particularly interested here in constructing one-sided coverage intervals for proportions that are either very small (less than 20%) or very large (more than 80%). This is a case with many useful applications. For example, when auditing insurance claims, the proportion of underpaid claims and its upper bound are often of interest. Because of the large number of claims in the population, only a small statistical sample can be reviewed. Stratified simple random sampling is often used to select the sample because different sampling rates are needed and different error rates are believed in different strata. The stratifiers can be geographic location, type of insurance, characteristic of insurer, and so forth. The coverage interval of a proportion is also often used in quality control. The Statistics of Income division in the IRS has been reviewing the quality of IRS customer services using statistical samples. One of such projects is the National Quality Review System (NQRS) that reviews telephones and paper cases. A statistical sample is selected for reviewing E-mails, Account calls, Tax Law calls and so on. A fixed number of sampling units are randomly selected from each service location every day. Because the total numbers of units are different on different days and at different locations, the sample is considered a stratified random sample with day and location as stratifiers. After the sampled units are reviewed each month, the data collected is used to estimate the accuracy rates as well as the coverage limits. In some circumstances, these estimated accuracy rates can be close to 1, where the lower bounds may be desired. The applications in health care are often carried out using complex sample designs. The 2006/2007 Northern Ireland Drug Prevalence Survey (UK Department of HSSPS, Technical Report 2008) published a series of proportions of drug use and their coverage intervals. The survey uses a multistage sample design. In the first stage, a stratified sampling design is used to select primary sampling units of electoral districts. Within each electoral district, residential households are randomly selected. One member of each sampled household is selected as the final sampling unit. The effective-sample-size-adjusted Clopper-Pearson method is used for interval estimates of prevalence rates. Only two-sided coverage intervals are published in this report, but one-sided upper limits may also be of interest.

Organizationally, the article is divided up into four parts. This introduction is Section 1. Section 2 provides an extensive list of coverage-interval methods under simple random sampling and then compares them. Section 3 looks at interval methods modified to handle complex sample data and evaluates their performances under stratified random sampling. Section 4 contains a summary and discussion of our results.

## 2. Interval Construction Under a Simple Random Sample

Let $X$ follow a binomial distribution with parameters $n$ and $p$. The parameter $p$ is sometimes called a "binomial proportion." In the survey sampling setting, $n$ is the sample size of a simple random sample. Let $k$ denote a sampled element and $x_k$ be either 0 or 1. Assuming that $x_k$ follows the Bernoulli distribution with parameter $p$, the estimator for $p$ from the sample is $\hat{p} = x/n$, where $x = \sum^n x_k$.

This section contains a summary of many of the interval-construction methods under simple random sampling that have appeared in the literature. All the methods assume that the population size is large enough not to need a finite population correction. The symbol $z$ is used to denote the $z$-score of a standard normal distribution associated with one-sided $(1 - \alpha)\%$ coverage intervals. For 95% coverage intervals, $\alpha = 0.05$, and the $z$-score is 1.645.

### *2.1. The Methods*

#### 2.1.1. Standard Wald Interval

This is the best known and most commonly used interval. It is based on the limiting distribution (as $n$ grows arbitrarily large): $(\hat{p} - p)/\sqrt{v(\hat{p})} \rightarrow N(0, 1)$, where $v(\hat{p}) = \hat{p}(1 - \hat{p})/(n - 1)$. The lower and upper bounds are

$$L_S = \hat{p} - z\sqrt{\hat{p}\big(1 - \hat{p}\big)/(n - 1)}, \text{ and } \quad U_S = \hat{p} + z\sqrt{\hat{p}\big(1 - \hat{p}\big)/(n - 1)} \tag{1}$$

That is to say, the two one-sided Wald intervals for $p$ are $p \geq L_S$, and $p \leq U_S$.

#### 2.1.2. Wilson (Score) Interval

Instead of using the variance estimator for $\hat{p}$, this interval employs the true variance $V(\hat{p}) = p(1 - p)/n$. It is based on the limit: $(\hat{p} - p)/\sqrt{V(\hat{p})} \rightarrow N(0, 1)$. The lower and upper bounds are

$$L_W = \tilde{p} - \frac{z\sqrt{n}}{n + z^2}\sqrt{\hat{p}\big(1 - \hat{p}\big) + \frac{z^2}{4n}}, \text{ and } \quad U_W = \tilde{p} + \frac{z\sqrt{n}}{n + z^2}\sqrt{\hat{p}\big(1 - \hat{p}\big) + \frac{z^2}{4n}} \tag{2}$$

$$\text{where } \tilde{p} = \frac{\hat{p} + z^2/2n}{1 + z^2/n}$$

#### 2.1.3. Logit Interval

A logistic transformation $\hat{\lambda} = \log[\hat{p}/(1 - \hat{p})]$ stabilizes the variance of $\hat{p}$. The logit interval is based on the limit: $(\hat{\lambda} - \lambda)/\sqrt{v(\hat{\lambda})} \rightarrow N(0, 1)$, where $v(\hat{\lambda}) = 1/[n\hat{p}(1 - \hat{p})]$. The lower and upper bounds are

$$\begin{aligned} L_L &= \frac{e^{\lambda_L}}{1 + e^{\lambda_L}}, \text{ where } \lambda_L = \hat{\lambda} - z\sqrt{v(\hat{\lambda})}, \text{ and} \\ U_L &= \frac{e^{\lambda_U}}{1 + e^{\lambda_U}}, \text{ where } \lambda_U = \hat{\lambda} + z\sqrt{v(\hat{\lambda})} \end{aligned} \tag{3}$$

2.1.4. Angular (Arcsine of Square Root) Interval

Another variance-stabilizing transformation is the angular transformation, $\delta = \arcsin(\sqrt{p})$. The interval for δ is based on the limit: $(\hat{\delta} - \delta)/\sqrt{v(\hat{\delta})} \rightarrow N(0,1)$, where $\hat{\delta} = \arcsin(\sqrt{\hat{p}})$ and $v(\hat{\delta}) = 1/(4n)$. This results in these lower and upper bounds for $p$:

$$\begin{aligned} L_A &= \sin^2(\delta_L) = \sin^2\left[\arcsin(\hat{\delta}) - z/(2\sqrt{n})\right], \text{ and} \\ U_A &= \sin^2(\delta_L) = \sin^2\left[\arcsin(\hat{\delta}) + z/(2\sqrt{n})\right] \end{aligned} \tag{4}$$

2.1.5. Jeffreys Interval

The Bayesian Posterior interval under a Jeffreys prior of the beta distribution Beta (1/2,1/2) is

$$\begin{aligned} L_J &= \text{Beta}(\alpha; x + 1/2, n - x + 1/2), \text{ and} \\ U_J &= \text{Beta}(1 - \alpha; x + 1/2, n - x + 1/2) \end{aligned} \tag{5}$$

2.1.6. Clopper-Pearson Exact Interval

This interval is based on inverting the equal-tailed binomial tests of the null hypothesis $H_0 : p = p_0$ against the alternative hypothesis $H_1 : p \neq p_0$. The lower and upper bounds can be obtained by solving the polynomial equations:

$$L_{CP} = \left\{ p : \sum_{t=0}^{x-1} \binom{n}{t} p^t (1-p)^{n-t} = 1 - \alpha \right\}, \text{ and}$$
$$U_{CP} = \left\{ p : \sum_{t=0}^{x} \binom{n}{t} p^t (1-p)^{n-t} = \alpha \right\}$$

They can be expressed in terms of a beta distribution as

$$L_{CP} = \text{Beta}(\alpha; x, n - x + 1), \text{ and } \quad U_{CP} = \text{Beta}(1 - \alpha; x + 1, n - x) \tag{6}$$

2.1.7. Mid-P Clopper-Pearson Interval

One way to reduce the perceived over-conservativeness of the Clopper-Pearson method obtains by solving the polynomial equations:

$$p_L = \left\{ p : \frac{1}{2}\binom{n}{x} p^x (1-p)^{n-x} + \sum_{t=0}^{x-1} \binom{n}{t} p^t (1-p)^{n-t} = 1 - \alpha \right\}$$
$$p_U = \left\{ p : \frac{1}{2}\binom{n}{x} p^x (1-p)^{n-x} + \sum_{t=0}^{x-1} \binom{n}{t} p^t (1-p)^{n-t} = \alpha \right\}$$

The interval can be expressed in terms of a beta distribution as

$$\begin{aligned} L_{MP} &= \frac{1}{2}\{\text{Beta}(\alpha; x, n-x+1) + \text{Beta}(\alpha; x+1, n-x)\}, \text{ and} \\ U_{MP} &= \frac{1}{2}\{\text{Beta}(1-\alpha; x, n-x+1) + \text{Beta}(1-\alpha; x+1, n-x)\} \end{aligned} \tag{7}$$

Note its similarity to the Jeffreys interval in Equation (5).

Brown et al. (2001) evaluate the properties of these seven methods for constructing two-sided intervals (replacing $\alpha$ by $\alpha/2$ and $z$ by the $z$-score of $1 - \alpha/2$). Unfortunately, an effective two-sided-interval method may not work as well in constructing a one-sided interval. This is because a two-sided interval can have compensating one-sided errors due to the sampling distribution of $\hat{p}$ being asymmetric.

The following methods were developed specifically to construct one-sided intervals based on an Edgeworth expansion that explicitly adjusts for the skewness in $\hat{p}$.

### 2.1.8. Hall Interval

The bounds for this interval translate the Wald bounds in Equation (1) towards $\frac{1}{2}$. They are

$$\begin{aligned} &L_H = \hat{p} + \delta - z\sqrt{v(\hat{p})}, \text{ and} \quad U_H = \hat{p} + \delta + z\sqrt{v(\hat{p})} \\ &\text{where } v(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1} \text{ and} \quad \delta = \left(\frac{z^2}{3} + \frac{1}{6}\right)\frac{(1-2\hat{p})}{n} \end{aligned} \tag{8}$$

The translation term, $\delta$, is $O_P(1/n)$. Terms of smaller asymptotic order have been dropped. Hall (1982) has $n$ in the denominator of $v(\hat{p})$ rather than $n - 1$. This difference has no practical consequence when $n \geq 30$.

### 2.1.9. Cai Interval

Cai (2004) goes further than Hall in correcting for the skewness in $\hat{p}$ by keeping $O_P(1/n^2)$ terms producing the bounds:

$$\begin{aligned} L_{Cai} &= \breve{p} - \frac{z}{\sqrt{n}}\sqrt{\hat{p}(1-\hat{p}) + \frac{\gamma_1\hat{p}(1-\hat{p}) + \gamma_2}{n}}, \text{ and} \\ U_{Cai} &= \breve{p} + \frac{z}{\sqrt{n}}\sqrt{\hat{p}(1-\hat{p}) + \frac{\gamma_1\hat{p}(1-\hat{p}) + \gamma_2}{n}} \\ &\text{where } \breve{p} = \frac{\hat{p} + \eta/n}{1 + 2\eta/n},\ \eta = \frac{z^2}{3} + \frac{1}{6},\ \gamma_1 = -\frac{13}{18}z^2 - \frac{17}{18} \text{ and} \quad \gamma_2 = \frac{1}{18}z^2 + \frac{7}{36} \end{aligned} \tag{9}$$

### 2.1.10. Kott-Liu Interval

Under simple random sampling, Kott and Liu (2009) propose a slight modification of the Hall interval that better handles samples with small $\hat{p}(1 - \hat{p})$ values:

$$L_{KL} = \hat{p} + \delta - \sqrt{z^2 v(\hat{p}) + \delta^2}, \text{ and} \quad U_{KL} = \hat{p} + \delta + \sqrt{z^2 v(\hat{p}) + \delta^2} \tag{10}$$

where $v(\hat{p})$ and $\delta$ are unchanged from those in Hall. Notice that the lower bound attains its minimum value, 0, when $\hat{p} = 0$, and the upper bound attains its maximum value, 1, when $\hat{p} = 1$. This method will be described further in the following section.

#### 2.1.11. Other Intervals

There are also various continuity-correction approaches (Vollset 1993) that are not included in this article. Two other methods not treated here are the Wilson-logit and the likelihood-ratio interval (Brown et al. 2002; Feng 2006). These methods employ an iteration algorithm to obtain the interval end-points and therefore harder to compute. Finally, when $n$ is large and $p$ is close to 0, the binomial distribution Bin($n$,$p$) can be approximated by a Poisson distribution $P(X = x) = \lambda^x e^{-\lambda}/x!$, where $\lambda = np$ (Newcombe 1998; Feng 2006). The lower and upper bounds for $p$ are

$$L_P = \chi^2_{2x,\alpha}/(2n), \text{ and } \quad U_P = \chi^2_{2(x+1),1-\alpha}/(2n)$$

This method has to be redefined for $p$ near 1 to be effective and is not useful when $p$ is not very near either 0 or 1.

### 2.2. *Comparison of One-Sided Intervals Under Simple Random Sampling*

In this subsection, the methods defined in Equations (1) through (10) are used to construct one-sided 95% coverage intervals. They are then compared in terms of their coverage probabilities and the average distances from their endpoints to the true value of $p$.

The *coverage probability* for the given $p$ and $n$ is defined as the probability of $p$ falling within the coverage interval $CI$, that is,

$$P(p \in CI) = \sum_{x=0}^{n} I(x)P(x)$$

$$\text{where } CI = \begin{cases} (L, 1), \text{ for lower bound} \\ (0, U), \text{ for upper bound} \end{cases}$$

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x},\ 0 < p < 1 \text{ and } \quad I(x) = \begin{cases} 1, \text{ if } p \in CI \\ 0, \text{ if } p \notin CI \end{cases}$$

The *average distance* for the given $p$ and $n$ is defined here as the mean of the absolute distance of lower or upper bound from the true value of $p$, that is,

$$AD = \sum_{x=0}^{n} D(x)P(x)$$

$$\text{where } D(x) = \begin{cases} |L(x) - p|, \text{ for the lower bound} \\ |U(x) - p|, \text{ for the upper bound} \end{cases}$$

We are interested in a setting where the sample size $n$ is relatively small but large enough for the asymptotic theory supporting some of the methods to be effective. Therefore, we evaluate a sample of size 30. Coverages perform differently for different sample sizes and

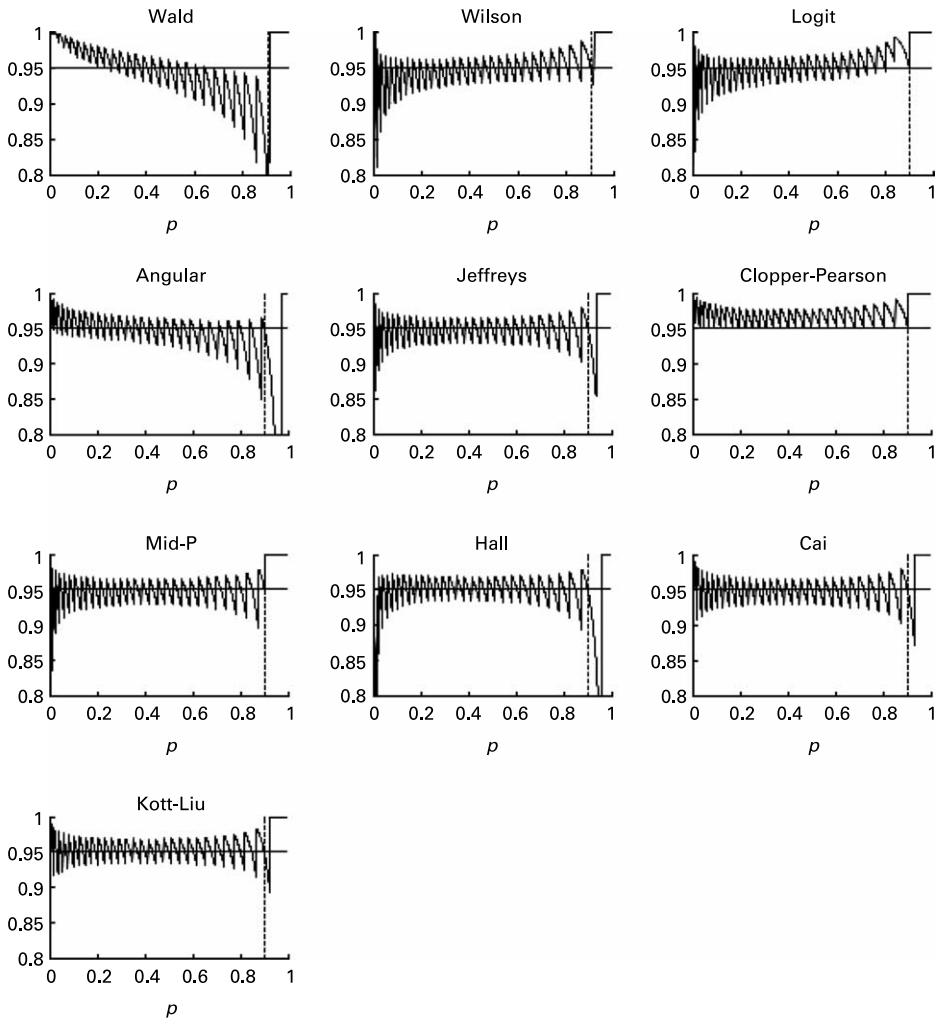


*Fig. 1. Coverage probabilities of lower bound at 95% nominal level: simple random sample with n = 30*

different values of $p$ (Brown et al. 2001, 2002, discuss this at length for two-sided intervals). Thus, we evaluate one-sided coverages over the entire range of potential $p$-values.

We make a few sensible modifications of the methods when $x = 0$ or $n$. We force the lower bound to be 0 at $x = 0$ and the upper bound to be 1 at $x = n$. We also force the lower bound to be 0 if it falls below 0 and the upper bound to be 1 if it falls above 1. In addition, when a bound is not defined at $x = 0$ or $n$ for a method (the Wald, Logit and Mid-P), we take a conservative stance and replace it with the Clopper-Pearson.

The coverage probabilities and average distances for all the methods are symmetric or very nearly so in the range $0 \leq p \leq 1$. Consequently, conclusions drawn about lower bounds for, say, $p < .2$ also apply to upper bounds for $p > .8$, and conclusions about lower bounds for $p > .8$ apply to upper bounds for $p < .2$. These values are calculated at $p = .001, .002, .003, \ldots, .998, .999$.

The plots of coverage probabilities for $n = 30$ are displayed in Figure 1. The vertical line at $p = .905$ represents the $p$-value where $p^{30} = .05$. So, when $p \geq .905$, $\hat{p}$ has at least a 5% probability of being 1.

The following conclusions can be drawn from the plots in Figure 1:

- All methods have 100% coverages as $p$ gets very close to 1. The region of the 100% coverage is called "lip" in this article. All methods can sometimes experience a downward spike before the "lip," we call it "dip."
- The Wald and Angular methods are systematically biased, sometimes in one direction sometimes in the other.
- The Clopper-Pearson method always has at least the nominal coverage (95%), but often over-covers. It has 100% coverage when $p \geq .905$.
- The Wilson and Logit methods are systematically biased in the opposite direction of the Wald but to a lesser degree. They tend to under-cover for small $p$ and over-cover for large $p$. The over-coverage for the Wilson near $p = 1$ is not as pronounced as for the Clopper-Pearson.
- The Jeffreys and Hall methods have large downward spikes (under-coverages) near the two boundaries.
- The Mid-P has some large downward spikes near $p = 0$, but performs reasonably well for large $p$.
- The Kott-Liu and Cai methods provide good coverages almost everywhere. Both have 100% coverages as $p$ gets very close to 1, but this "lip" begins for the Kott-Liu (at 0.929) while the Cai is still experiencing its worst downward spike or "dip" (it reaches a minimum coverage of 88% before beginning its lip at 0.935; the Kott-Liu minimum coverage is 90.1%). Before then, the two methods have identical coverages for large $p$-values ($\geq .873$).

The above analysis of coverage shows that when the proportion is in the middle range (between 0.2 and 0.8), there are many good methods: Jeffreys, Hall, Mid-P, Cai, and Kott-Liu. When the proportion is either very small (less than 20%) or very large (more than 80%), Mid-P, Cai, and Kott-Liu are the better methods. Analogous graphs for a few other sample sizes $n = 20, 60,$ and $120$ (not shown) behave similarly.

We plot the average distances of lower bounds versus the tail values of $p$ for the better methods (Mid-P, Cai, and Kott-Liu) and for the conservative Clopper-Pearson in Figure 2. In general, the average distance is longer when the coverage probability is larger. The Clopper-Pearson has a much longer average distance than the other methods, not surprisingly since it tends to be conservative. For small $p$, the Kott-Liu and Cai behave very similarly. For large $p$, the Kott-Liu tends to be slightly longer than the Cai. The Mid-P becomes longer than Kott-Liu and Cai when $p$ gets near 1 but not before.

In summary, the Kott-Liu and Cai methods are the best in terms of having coverages almost always close to the nominal and a reasonable average distance. For Mid-P method, when $p$ is near the end (larger than 0.95), it is more conservative than the Kott-Liu and Cai methods but less conservative than the Clopper-Pearson. Mid-P has large downward spikes when $p$ is near 0. The Clopper-Pearson never under-covers, but has longer average distances. Many view the property of never providing less than nominal coverage as very desirable, if not absolutely required (see the discussions in Brown et al. 2001). They argue

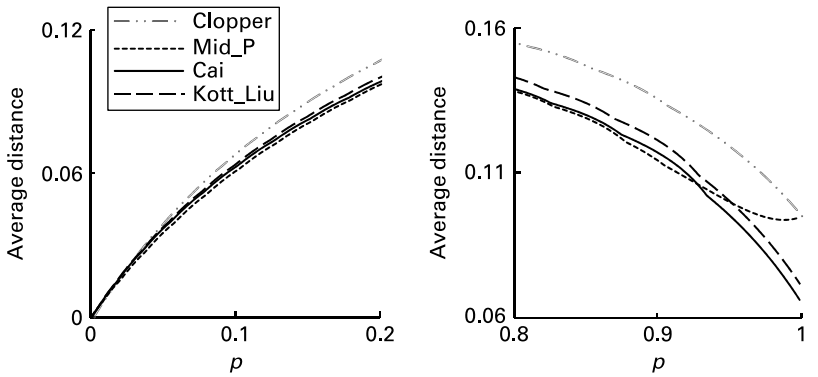


*Fig. 2. Average distance of lower bound at 95% nominal level: simple random sample with n = 30*

that users should have confidence that their intervals always cover at least as well as advertised, hence the popular term "confidence interval." However, the confidence interval can be too conservative for particular values of $p$. In addition, such confidence is not always justified with some complex sample data, as we shall see. Therefore, we use the term "coverage interval" instead of "confidence interval" in this article. Our goal is to determine coverage intervals for particular values of $p$.

## 3. Interval Construction Methods Under Stratified Random Sampling

Let $s$ denote elements of the whole sample, $k$ (again) denote an element, and $w_k$ the weight of element $k$. Let $x_k$ be either 0 or 1. The estimated proportion is $\hat{p} = \sum_s w_k x_k / \sum_s w_k$.

### *3.1. The Methods*

The most common way of extending interval-construction methods to handle sample data from a complex design is by replacing the sample size $n$ with the (estimated) *effective sample size* $n^*$ and replacing $x$ with $x^* = n^*\hat{p}$. When $v(\hat{p}) > 0$, where $v(\hat{p})$ is the estimated variance of $\hat{p}$ under the complex sample design, the effective sample size $n^*$ can be defined as

$$n^* = \frac{\hat{p}(1-\hat{p})}{v(\hat{p})} \tag{11}$$

Sometimes, $n^*$ is defined as 1 plus the left-hand side of Equation (11). The distinction is usually trivial when $n \geq 30$. The ratio $n/n^*$ is called "the (estimated) design effect."

The *idealized effective sample size* $\tilde{n}$ features the true variance $v(\hat{p})$ in the denominator of Equation (11) in place of the estimated variance $v(\hat{p})$. Unfortunately, $V(\hat{p})$ is unknown and needs to be estimated from the sample in practice.

The *ad hoc* procedure of replacing $n$ by $n^*$ and $x$ by $x^*$ is used and discussed in Breeze (1990, cited in Feng 2006) for modifying the Poisson interval, in Kott and Carr (1997) for modifying the Wilson interval and in Korn and Graubard (1998) for modifying the Clopper-Pearson interval. Using the same procedure, Feng (2006) treats a few other two-sided intervals: Wald, Logit, Angular, and Likelihood Ratio intervals. We also apply

this procedure to the one-sided Hall interval and Cai interval. The estimated variance $v(\hat{p})$ in the Hall interval is calculated under the complex sample design, that is, $v(\hat{p}) = \hat{p}(1-\hat{p})/n^*$.

We focus in this section on an empirical evaluation of one-sided interval methods under stratified random sampling. We apply the effective sample size procedure to all the methods from Section 2 except the Kott-Liu, which was designed especially to handle data from stratified random samples. We follow Korn and Graubard and set $n^* = n$ when $v(\hat{p}) = 0$.

Let $W_h = N_h/N$ for a stratified random sample with $H$ strata. The estimated overall proportion is $\hat{p} = \sum^H W_h \hat{p}_h$, where $\hat{p}_h$ is the observed stratum proportion of stratum $h$. Adapting the Edgeworth expansions in Hall (1982) and Cai (2004) under a simple random sampling, Kott and Liu (2009) discuss three different coverage intervals for data from a stratified random sample.

### 3.1.1. Basic Kott-Liu Interval

$$L_{KL1} = \hat{p} + \delta_1 - \sqrt{z^2 v_1(\hat{p}) + \delta_1^2}, \text{ and } \quad U_{KL1} = \hat{p} + \delta_1 + \sqrt{z^2 v_1(\hat{p}) + \delta_1^2} \tag{12}$$

$$\text{where } v_1(\hat{p}) = \sum\nolimits_{h=1}^{H} W_h^2 \hat{p}_h(1-\hat{p}_h)/(n_h - 1), \text{ and}$$

$$\delta_1 = \left(\frac{z^2}{3} + \frac{1}{6}\right) \frac{\sum^H W_h^3 \hat{p}_h(1-\hat{p}_h)(1-2\hat{p}_h)/[(n_h-1)(n_h-2)]}{\sum^H W_h^2 \hat{p}_h(1-\hat{p}_h)/(n_h-1)} \tag{13}$$

The variance of $\hat{p}$ is not a simple function of the true $p$ and $n$ under stratified random sampling as it is under simple random sampling. As a result, $V(\hat{p})$ must be estimated from the sample. The estimation has its own random error, which cannot be completely eliminated from the Edgeworth expansion (moreover, following Cai and keeping $O_P(1/n^2)$ terms becomes impossible).

### 3.1.2. DF-adjusted Kott-Liu Interval

One way to adjust for the error in the implicit estimator for $V(\hat{p})$ in the basic Kott-Liu method is by replacing the $z$-score in Equation (12) with a $t$-score from a Student $t$. A $t$-distribution needs a degrees-of-freedom calculation. Kott and Liu (2009) discuss a number of ways of estimating the *effective degrees of freedom*. When each stratum has at least ten observations, a nearly unbiased estimator for this quantity is

$$df_1 = \frac{2a_1^2}{a_3 - a_2^2/a_1}$$

where

$$a_1 = \sum^H W_h^2 \hat{p}_h(1-\hat{p}_h)/n_h,\ a_2 = \sum^H W_h^3 \hat{p}_h(1-\hat{p}_h)(1-2\hat{p}_h)/n_h^2, \text{ and}$$
$$a_3 = \sum^H W_h^4 \hat{p}_h(1-\hat{p}_h)(1-2\hat{p}_h)^2/n_h^3$$

An asymptotically biased, but more stable, effective-degrees-of-freedom estimator treats $p_h$ as if they were equal:

$$df_2 = \frac{2\left(\sum_{h=1}^{H} W_h^2/n_h\right)^2 \hat{p}(1-\hat{p})}{\left\{\sum_{h=1}^{H} \frac{W_h^4}{n_h^3} - \left(\sum_{h=1}^{H} W_h^3/n_h^2\right)^2 \Big/ \sum_{h=1}^{H} \frac{W_h^2}{n_h}\right\}(1-2\hat{p})^2}$$

A slightly conservative policy, followed here, sets the estimated effective degrees of freedom at $df = \min(df_1, df_2)$ and uses $t(df, 1-\alpha)$ in place of $z$ in the lower and upper bounds defined in Equation (12).

#### 3.1.3. Kott-Liu iid Interval

If an independent and identically distributed (*iid*) Bernoulli model is assumed, then a different way to generalize Equation (10) is with

$$L_{KL2} = \hat{p} + \delta_2 - \sqrt{z^2 v_2(\hat{p}) + \delta_2^2}, \text{ and } \quad U_{KL2} = \hat{p} + \delta_2 + \sqrt{z^2 v_2(\hat{p}) + \delta_2^2} \tag{14}$$

where, $v_2(\hat{p}) = \sum^{H} W_h^2 \hat{p}(1-\hat{p})/n_h$, and

$$\delta_2 = \left(\frac{1-z^2}{6}\frac{\sum^{H} W_h^3/n_h^2}{\sum^{H} W_h^2/n_h} + \frac{z^2}{2}\sum^{H}\frac{W_h^2}{n_h}\right)(1-2\hat{p}) \tag{15}$$

Since both the basic and DF-adusted Kott-Liu intervals are undefined when $\hat{p} = 0$ or 1, Kott and Liu (2009) suggest using the *iid* method in Equation (14) in this situation.

### *3.2. Comparison of One-Sided Intervals Under Stratified Random Sampling*

All the methods described in the text are evaluated under the following stratified random sampling designs using simulations. A population of 6,000 is divided into 3 equal strata, that is, $N_h = 2,000$, $h = 1, 2, 3$. The overall proportion $p$ takes the values of 0.001, 0.002, 0.003, . . ., 0.998, 0.999. We consider these six settings for the stratum sample sizes and the comparative values of $p_h$. They are shown in Table 1. One sample size allocation – 10, 10, 10 – has a total sample size of 30, our minimum. The other allows one stratum to be big enough to stand alone, $n_h = 30$, while the other two strata contain 10 units. As for the

*Table 1. Simulation settings*

| | Stratum proportions ($p_1$, $p_2$, $p_3$) | |
|---|---|---|
| Stratum sample sizes $n_1$, $n_2$, $n_3$ | Equal ($p$, $p$, $p$) | Unequal ($p - pq$, $p$, $p + pq$) |
| 10, 10, 10 | A | B |
| 10, 30, 10 | C | D |
| 10, 10, 30 | Same as C | E |
| 30, 10, 10 | Same as C | F |

$q = 1 - p$.

$p_h$-values, either they are all equal or their spread is, in some sense, maximized while allowing all the $p_h$-values fall into the 0 to 1 range.

For the simulations, we first generate a finite population of 2,000 units in each stratum $h$, denoted as $x_{hi} = 1, 2, \ldots, 2,000$. We then draw 1,000 stratified random samples for each stratum sample size allocation. For each stratum proportion $p_k$, we set

$$y_{hi} = \begin{cases} 1, \text{ if } x_{hi} \leq 2,000\, p_{hi} \\ 0, \text{ otherwise} \end{cases}$$

The weighted estimate for the proportion of $y = 1$ is calculated for each value of $p$ and for each sample. The coverage intervals are constructed using the methods described earlier in the text with the coverage probabilities and the average distances calculated from the 1,000 samples for each $p$.

Analogously with the simple random sampling case, only the simulation results for a lower bound need be considered. Given space limitation, we only display the lower-bound coverage plots using the Mid-P, Cai, three Kott-Liu methods, and Clopper-Pearson. As discussed in Section 2.2, Mid-P, Cai, and Kott-Liu are the better methods than others; and Clopper-Pearson method is the conservative benchmark to compare with.

The plots for setting A (not displayed) mirror those in Figure 1 with the three Kott-Liu methods being virtually identical. This is not surprising since the $p_h$ are equal, the idealized effective sample size is 30, and the effective degrees of freedom are nearly infinite (as in simple random sampling).

Figure 3 displays the coverage plots for Setting B that has same stratum sampling rates and unequal $p_h$. Despite the variability in the $p_h$, not much changes from Setting A. The Clopper-Pearson still has coverage above the nominal level. Its lip again begins at 0.905, which is marked by a vertical dash in all the plots. The basic and DF-adjusted Kott-Liu methods remain virtually identical everywhere, while the *iid* version is slightly more variable than the others when $p$ is roughly between 0.2 and 0.8 but matches their behavior

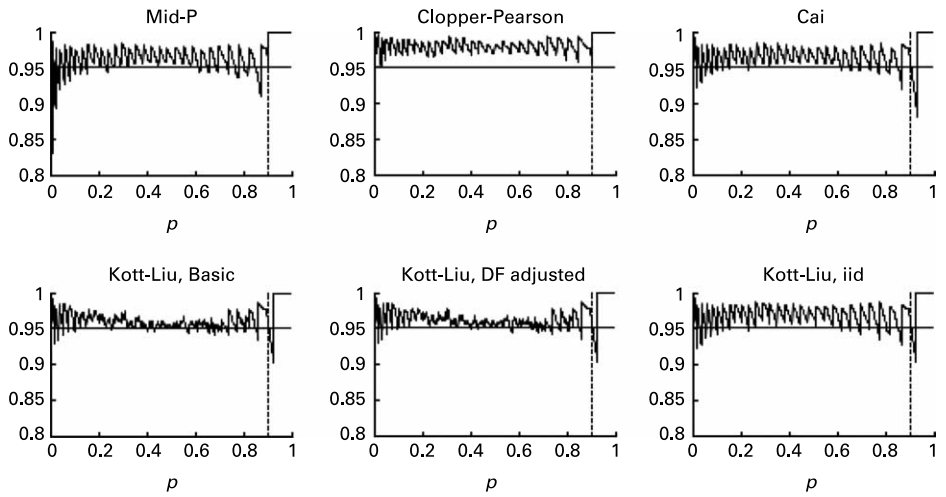


*Fig. 3. Coverage probabilities of lower bound at 95% nominal level for setting B: stratified random sample with* $(n_1, n_2, n_3) = (10, 10, 10)$ *and* $(p_1, p_2, p_3) = (p - pq, p, p + pq)$

in the tails. The Mid-P is similar to the three Kott-Liu methods when $p > .8$ for both settings A and B, but continues to be plagued by downward spikes for some very small $p$. Cai is similar to the Kott-Liu *iid* methods in the whole range of $p$.

Figure 4 displays the coverage plots for Setting C that has different stratum sampling rates and same stratum proportion $p_h$. The vertical dash line at $p = .942$ in all the plots corresponds to the $p$-value at which the Clopper-Pearson starts to lip. The basic Kott-Liu has a very final deep dip just before its lip. The DF-adjusted version is only slightly better. Its lip starts at 0.951 rather than at 0.956 (the basic has a minimum coverage of 82.0%, the DF-adjusted 84.1%). The Kott-Liu *iid* method hardly dips at all. Its lip starts at 0.948. Its coverage is close to nominal level almost everywhere. The lip for the Mid-P starts at 0.942, just like the lip of the Clopper-Pearson. The Cai's lip does not begin until 0.961, while its dip (bottoming at 87.9%) is not as great as those of the basic and DF-adjusted Kott-Liu methods.

For settings D, E, and F where the stratum proportions $p_h$ are not equal and stratum sampling rates vary, the coverage plots are displayed in Figures 5–7. For Setting D (Figure 5), except that the Kott-Liu *iid* method has a much higher coverage level, other methods behave similarly to those in Setting C. This suggests that Kott-Liu *iid* method may be sensitive to the assumption of equal stratum proportions $p_h$.

In Setting E (Figure 6), all the methods suffer from a deep dip before the final lip. Here, there is no advantage of the DF-adjusted Kott-Liu over the basic. Its lip starts slightly earlier, but by then the basic's dip has ended. The Clopper-Pearson has the slightest dip and longest lip among the methods, but its dip is well below the nominal (87.8% at 0.941 as opposed to *iid* Kott-Liu's 84.4% at 0.947). The Cai has the deepest dip (74.1% as opposed to the basic and DF-adjusted Kott-Liu's 83.1%). Both the Clopper-Pearson and the Kott-Liu *iid* method consistently over-cover when $p$ is less than 0.5.

In Setting F (Figure 7), only the basic and DF-adjusted Kott-Liu methods have final dips, and these are modest (the basic's bottom is 88.8% at 0.955, while the DF-adjusted's

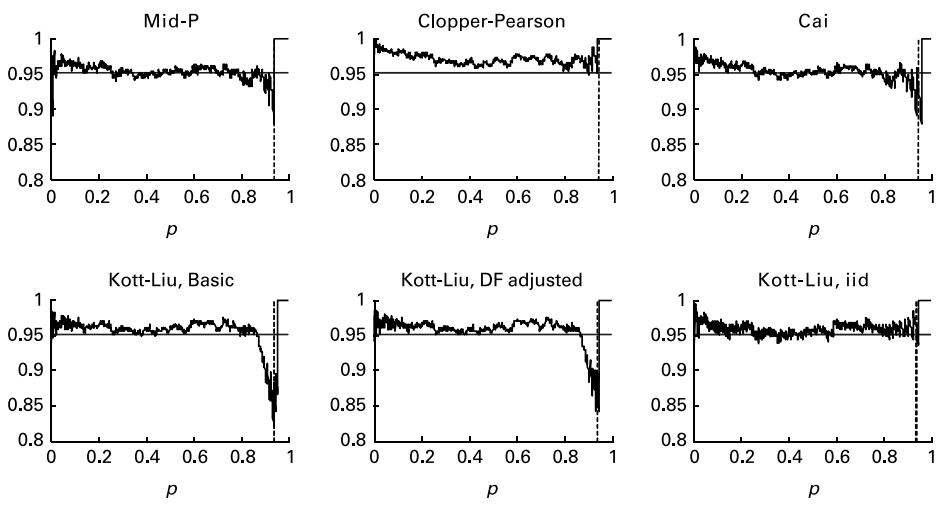


*Fig. 4. Coverage probabilities of lower bound at 95% nominal level for setting C: stratified random sample with* $(n_1, n_2, n_3) = (10, 30, 10)$ *and* $(p_1, p_2, p_3) = (p, p, p)$

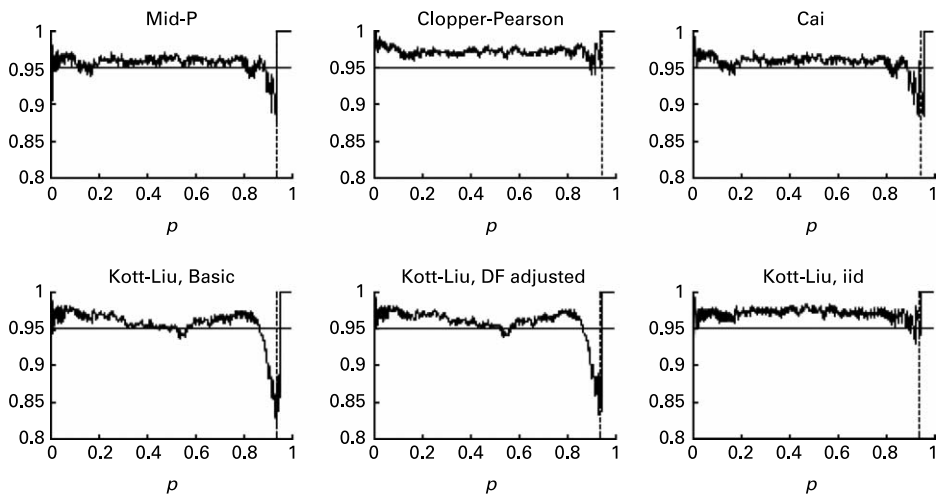


*Fig. 5. Coverage probabilities of lower bound at 95% nominal level for setting D: stratified random sample with* $(n_1, n_2, n_3) = (10, 30, 10)$ *and* $(p_1, p_2, p_3) = (p - pq, p, p + pq)$

is 91.1% at 0.951). The Clopper-Pearson consistently over-covers for all values of $p$. The Kott-Liu *iid* method consistently over-covers when $p$ is greater than 0.5 and suffers downward spikes for very low values of $p$, but not as severely as the Mid-P. The Mid-P and Cai tend to over-cover for $p > .6$, but by not as much as the Clopper-Pearson and Kott-Liu *iid* methods.

The average distances for tail $p$-values in Settings B, C, D, E, and F are displayed in Figure 8 for $p \leq .2$ and $p \geq .8$. Because the average distances of the basic and DF-adjusted Kott-Liu methods are so close, only the DF-adjusted version is displayed in the graphs. The conservative Clopper-Pearson method exhibits the longest average distances, while the Cai method tends to have the smallest average distances, but not by much.

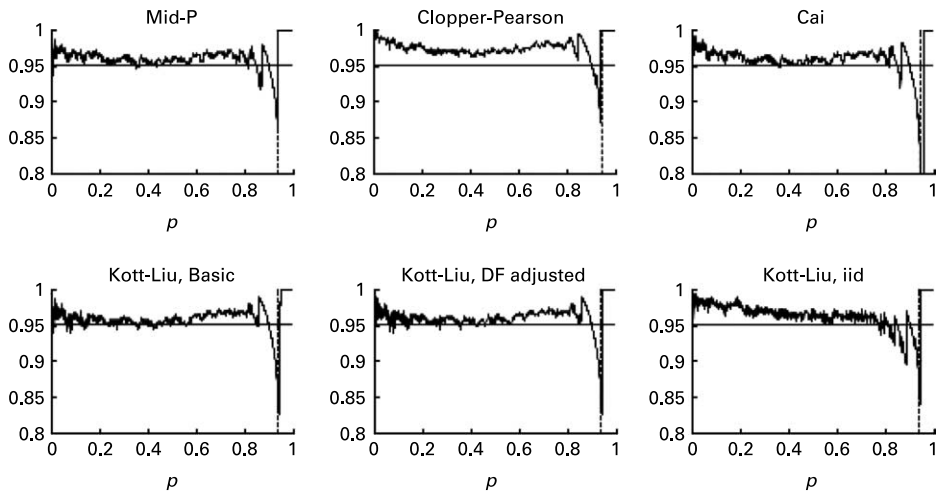


*Fig. 6. Coverage probabilities of lower bound at 95% nominal level for setting E: stratified random sample with* $(n_1, n_2, n_3) = (10, 10, 30)$ *and* $(p_1, p_2, p_3) = (p - pq, p, p + pq)$

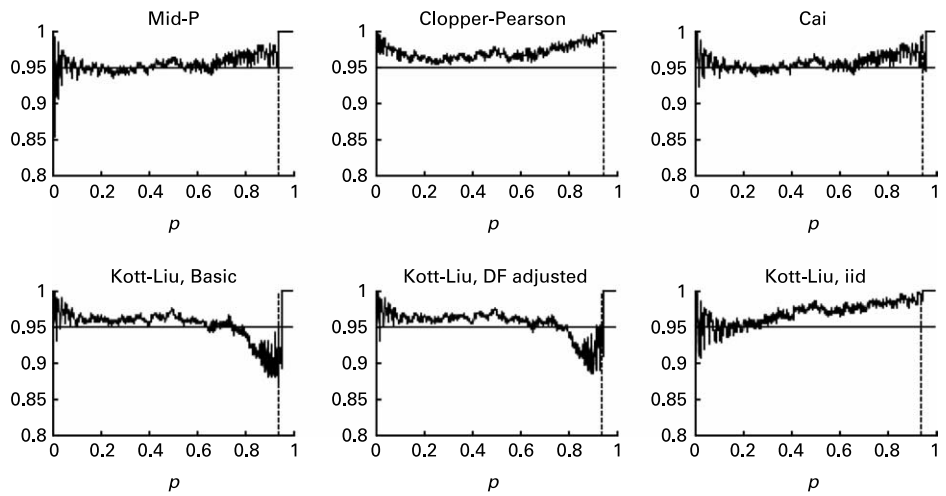


*Fig. 7. Coverage Probabilities of Lower Bound at 95% Nominal Level for Setting F: Stratified Random Sample with* $(n_1, n_2, n_3) = (30, 10, 10)$ *and* $(p_1, p_2, p_3) = (p - pq, p, p + pq)$

An important scenario in survey practice posed by an anonymous referee is when one wishes to estimate the upper bound of a rare trait and knows in advance that the trait is concentrated in a particular stratum. The example given by the referee is that Americans who become addicted to smoking cigarettes are more likely to be from families of lower socioeconomic status. Since one usually wants to study the relationship of the trait to other characteristics, it is often efficient for that purpose to oversample the stratum where the trait is most prevalent. Figure 6 shows that a few methods are conservative when the highest stratum proportion is paired with the highest stratum sampling fraction. In order to give more direct details to support this conclusion, we have added simulations for the stratum proportions of ($p$/3, 2$p$/3, 2$p$) that allows the prevalent rate in Stratum 3 ($p_3 = 2p$) is much higher than those in the other two strata ($p_1 = p/3$ and $p_2 = 2p/3$). We consider a sample size $n = 60$ with stratum allocations of (10, 20, 30), (20, 20, 20), and (30, 20, 10), and a larger sample size $n = 180$ with stratum allocations of (30, 60, 90), (60, 60, 60), and (90, 60, 30). Figures 9 and 10 give the coverage levels of the *upper* bound for the overall proportion $p$ in the range of [0, 0.3]. Because of the large sample size, the basic and DF-adjusted Kott-Liu methods remain identical everywhere. The DF-adjusted Kott-Liu method is not included in the graphs. A vertical dash line in each graph represents the $p$-value at which the Clopper-Pearson method starts to have 100% coverage. It is at $p = .048$ in Figure 9 and $p = .016$ in Figure 10. As shown in Figure 9 and 10, when the highest proportion is paired with the highest sampling fraction (settings (10, 20, 30) and (30, 60, 90)), the Mid-P and Cai methods are conservative, but not as much as the Clopper-Pearson. When the stratum sample sizes are reasonably large, the less conservative Kott-Liu (basic or DF-adjusted) is preferred as the coverage level is very close to the nominal. On the other hand, when the lowest proportion is paired with the highest sampling fraction (settings (30, 20, 10) and (90, 60, 30)), there is a dip when the proportion is close to 0 in all methods. When the proportion is not near 0, Kott-Liu, Cai, and Mid-P have good coverage level. Figure 10 confirm that the equal stratum sampling rates for a given total sample size always give a coverage level closer to the nominal level than unequal allocations.

## 4. Summary and Discussion

After reviewing much of the literature on constructing one-sided coverage intervals under simple random sampling, we conducted our own empirical evaluation and found that, among the methods reviewed, Cai and Kott-Liu produced one-sided interval coverages closest to nominal. We also confirmed that the Clopper-Pearson method always provided at least the nominal coverage, which many find a singularly desirable property.

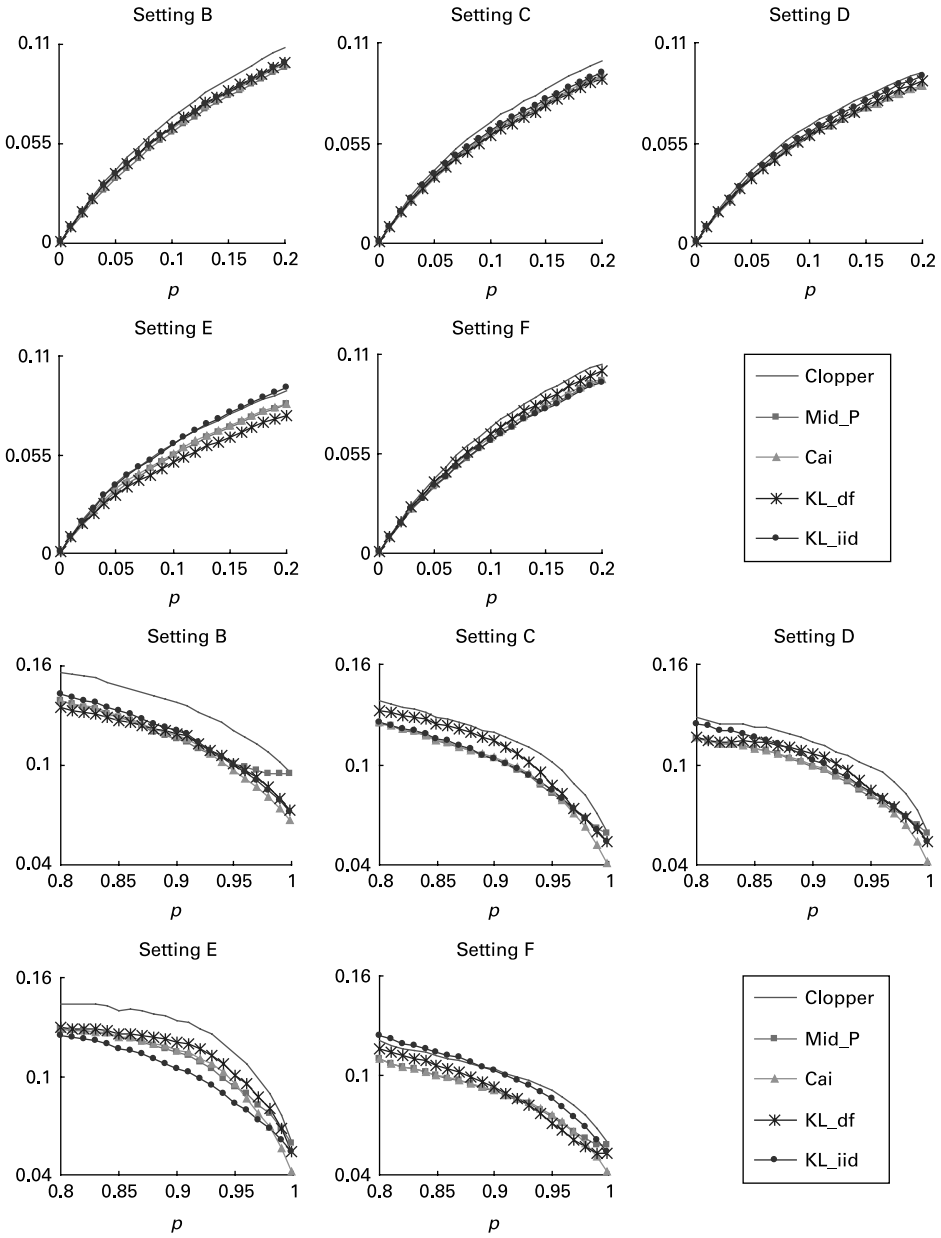


*Fig. 8. Average distances of lower bound at 95% nominal level for settings B–F and P in the range of [0, 0.2] and [0.8, 1]*

We then turned to stratified random sampling. We adjusted all the non-Kott-Liu methods by replacing the sample size with an estimate for the effective sample size. The Clopper-Pearson was still the most conservative method with coverage probabilities usually, *but not always*, at or above the nominal level.

For a given sample size, the setting of equal stratum sampling rates gave a better coverage than settings of unequal stratum sampling rates. The potential for under-coverage was larger when the sampling fraction varied across the strata.

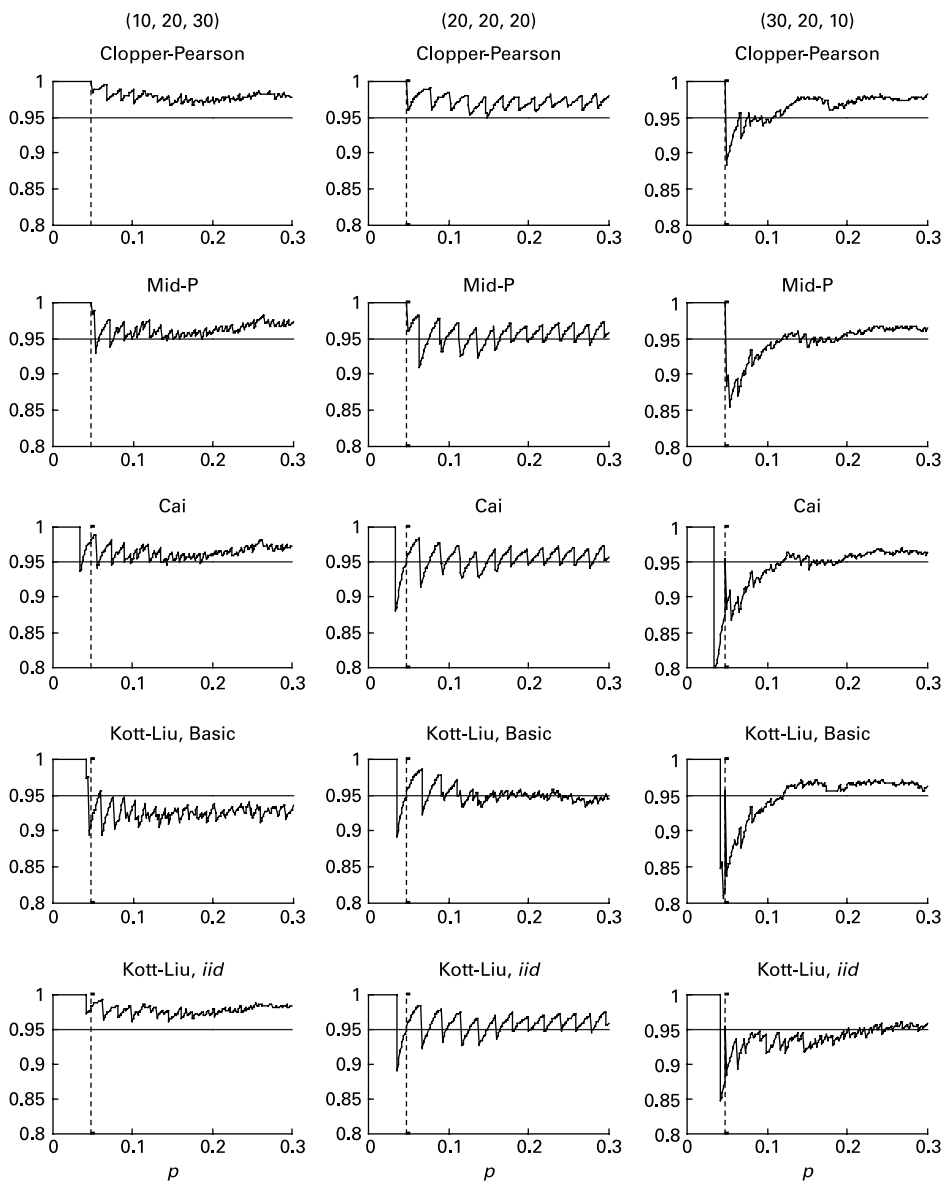


*Fig. 9. Coverage probabilities of upper bound at 95% nominal level for settings:* $(n_1, n_2, n_3) = (10, 20, 30)$, $(20, 20, 20)$, $(30, 20, 10)$ *and* $(p_1, p_2, p_3) = (p/3, 2p/3, 2p)$

The Cai and Mid-P methods appeared to be more conservative than Kott-Liu (basic). Forcing the lower bound to be zero when $\hat{p} = 0$ removed what would have been sharp downward spikes for small $p$-values. The Cai and Mid-P methods outperformed Kott-Liu for unequal stratum proportions and small sample size. When the sample size is reasonably large, which is often the case in practice, Kott-Liu is better in terms of the coverage and average distance.

The *iid* version of the Kott-Liu provides the best coverage when the assumption of same stratum proportions $p_h$ is reasonable. The basic Kott-Liu method worked well

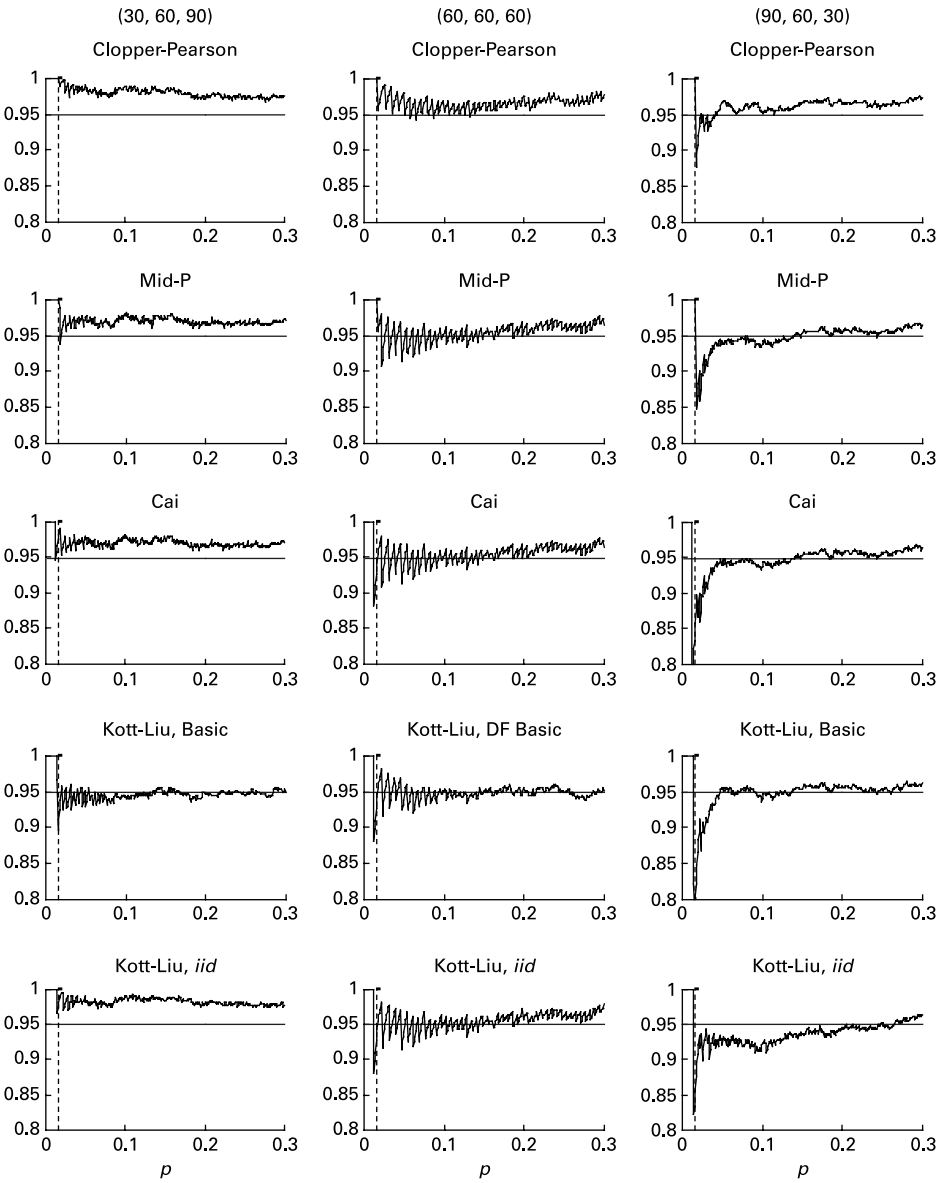


*Fig. 10. Coverage probabilities of upper bound at 95% nominal level for settings:* $(n_1, n_2, n_3) = (30, 60, 90)$, $(60, 60, 60)$, $(90, 60, 30)$ *and* $(p_1, p_2, p_3) = (p/3, 2p/3, 2p)$

when the sample size was reasonably large. Adjusting the basic Kott-Liu method for its effective degrees of freedom improved the coverage for extremely large and small $p$, but not by much.

Finally, we suggest another alternative when the value of $p$ is extreme. The lower bounds constructed using any of the methods have "lips" very near 1, that is, a region in which coverage is 100%. It is easy to see that this region includes all $p > 1 - 2\delta_1$ (see Equations (12) and (13)) using the basic Kott-Liu method and all $p > 1 - 2\delta_2$ using the *iid* method (see Equations (14) and (15)).

Using the Clopper-Pearson method, the lip begins in general at $p = p_L$ for the lower bound and at $p = 1 - p_L$ for the upper bound, where $p_L^n = \alpha$, or equivalently,

$$p_L = p_L(\alpha, n) = \exp[\log(\alpha)/n] \tag{16}$$

Suppose all the $p_h$ were equal to, say, $r$. If $r$ were greater or equal to $p_L$, and thus in the Clopper-Pearson lip, then $\hat{p}$ would have at least a probability $\alpha$ of being 1. No matter how large $\hat{p}$ was, $r$ would have to be in the lower one-sided interval to assure at least $(1 - \alpha)\%$ coverage. As a consequence, finding a lower bound producing close to the nominal $(1 - \alpha)\%$ coverage when $p = r$ can be an impossible task. Nevertheless, it would be a prudent rule not to let the lower bound for an interval be any higher than $p_L$ (and, symmetrically, not let the upper bound be any lower than $p_U = 1 - p_L$). The size of the lip from using this rule is of asymptotic order $1/n$: it decreases as the sample size increases.

We have marked where $p_L$ falls in our coverage plots. Notice that not allowing the lower bound to be higher than $p_L$ reduces the size of dips that would result from using the Cai or one of the Kott-Liu methods in the settings displayed in Figures 1 and 3. There remain deep dips using all the methods in Setting E (Figure 6), even the Clopper-Pearson. This may be because the $p_h$ are not all equal and neither are the sampling fractions.

Observe that when the sampling fractions are the same across the three strata –

$$\log(p_1^{n_1} p_2^{n_2} p_3^{n_3}) = \sum n_h \log\{p[1 + (p_h - p)/p)]\} \approx \sum n_h\{\log(p) + (p_h - p)/p\}$$
$$= \log(p) = \log(p^n)$$

–the impact of the variability of the $p_h$ is muted. This suggests the following policy when the sampling fractions are not all equal: setting the maximum value of the lower bound at $p_{L2} = p_L(\alpha, N \min\{n_h/N_h\})$ with $p_L(.,.)$ defined in Equation (16) (and setting the minimum value of the upper bound at $1 - p_L(\alpha, N \min\{n_h/N_h\})$. Such a policy will often be very conservative, extending the region where coverage will be 100%. This is a reflection of the difficulty of constructing a lower bound at all when $p > p_L(\alpha, N \min\{n_h/N_h\})$, and the variability among the $p_h$ is unknown. *There is no parallel difficulty constructing an upper bound for large* p *or a lower bound for small* p. In any event, when one-sided coverage intervals for a small or a large proportion is a survey goal, it would be wise to avoid stratification schemes with widely varying sampling fractions if possible.

Constructing one-sided coverage intervals with the Kott-Liu method under sampling designs with multiple stages has not been addressed in this article, but these methods (perhaps modified in the tails) can be extended to cover such samples. Two components in Equation (13) are replaced by the counterparts under complex sample design. Replacing the estimated third central moment of $\hat{p}$, $\sum^H W_h^3 \hat{p}_h(1 - \hat{p}_h)(1 - 2\hat{p}_h)/[(n_h - 1)(n_h - 2)]$,

is discussed in Kott et al. (2001). Similarly, $\sum^{H} W_h^2 \hat{p}_h(1 - \hat{p}_h)/(n_h - 1)$ can be replaced by a standard randomization-based variance estimator for $\hat{p}$ under complex sample design. This variance may be estimated using repeated sampling methods or the Taylor linearization method. More work on data from such designs will have to wait for another time. For the other methods, one need only replace $n$ by the effective sample size $n^*$ and $x$ by $x^* = n^*\hat{p}$, which was explained in the text.

## 5. References

Agresti, A. and Coull, B.A. (1998). Approximate Is Better Than "Exact" for Interval Estimation of Binomial Proportions. The American Statistician, 52, 119–126.

Brown, L.D., Cai, T., and Dasgupta, A. (2001). Interval Estimation for a Binomial Proportion. Statistical Science, 16, 101–133.

Brown, L.D., Cai, T., and Dasgupta, A. (2002). Confidence Intervals for a Binomial Proportion and Asymptotic Expansions. The Annals of Statistics, 30, 160–201.

Cai, T. (2004). One-sided Confidence Intervals in Discrete Distributions. Journal of Statistical Planning and Inference, 131, 63–88.

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.

Feng, X. (2006). Confidence Intervals for Proportions With Focus on the U.S. National Health and Nutrition Examination Survey. Master's Thesis, Simon Fraser University.

Hall, P. (1982). Improving the Normal Approximation When Constructing One-sided Confidence Intervals for Binomial or Poisson Parameters. Biometrika, 69, 647–652.

Korn, E.L. and Graubard, B.I. (1998). Confidence Intervals for Proportions With Small Expected Number of Positive Counts Estimated From Survey Data. Survey Methodology, 24, 193–201.

Kott, P.S., Anderson, P.G., and Nerman, O. (2001). Two-Sided Coverage Intervals for Small Proportion Based on Survey Data. Presented at Federal Committee on Statistical Methodology Research Conference, Washington, DC.

Kott, P.S. and Carr, D.A. (1997). Developing an Estimation Strategy for a Pesticide Data Program. Journal of Official Statistics, 13, 367–383.

Kott, P.S. and Liu, Y.K. (2009). One-sided Coverage Intervals for a Proportion Estimated From a Stratified Simple Random Sample. Internal Statistical Review, 77, 251–265.

Newcombe, R.G. (1998). Two-sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. Statistics in Medicine, 17, 857–872.

Sukasih, A. and Jang, D. (2005). An Application of Confidence Interval Methods for Small Proportions in the Health Care Survey of DoD Beneficiaries. Proceedings of the American Statistical Association, Survey Methodology Section, 3608–3612.

UK Department of HSSPS (2008). Northern Ireland Drug Prevalence Survey 2006/2007. Technical Report.

Vollset, S.E. (1993). Confidence Intervals for a Binomial Proportion. Statistics in Medicine, 12, 809–827.

Received November 2007
Revised May 2009