Survey Methodology, December 1998
Vol. 24, No.2, pp. 193-201
Statistics Canada



# Confidence Intervals for Proportions With Small Expected Number of Positive Counts Estimated From Survey Data

EDWARD L. KORN and BARRY I. GRAUBARD[1]

ABSTRACT

In the nonsurvey setting, "exact" confidence intervals for proportions calculated using the binomial distribution are frequently used instead of intervals based on approximate normality when the number of positive counts is small. With complex survey data, the binomial intervals are not applicable, so intervals based on the assumed approximate normality of the sample-weighted proportion are used, even if the number of positive counts is small. We propose a simple modification of the binomial intervals to be used in this situation. Limited simulations are presented that show the coverage probability of the proposed intervals is superior to that of the normality-based intervals, logit-transform intervals, and intervals based on a Poisson approximation. Applications are given involving the prevalence of Human Immunodeficiency Virus (HIV) based on data from the third National Health and Nutrition Examination Survey, and the proportion of users of cocaine based on data from the Hispanic Health and Nutrition Examination Survey.

KEY WORDS: Binomial confidence interval; Exact confidence interval; Logit transformation; Poisson confidence interval.

## 1. INTRODUCTION

With complex survey data, the typical construction of a $1 - \alpha$ level confidence interval for a proportion of positive counts for a 0-1 variable is

$$\hat{p} \pm t_d\,(1 - \alpha/2)\,[\hat{\text{var}}(\hat{p})]^{1/2} \qquad (1.1)$$

where $\hat{p}$ is the sample-weighted estimator of the proportion, $\hat{\text{var}}(\hat{p})$ is the variance estimator of $\hat{p}$, and $t_d(1 - \alpha/2)$ is the $1 - \alpha/2$ quantile of a $t$ distribution with $d$ degrees of freedom. The estimator $\hat{\text{var}}(\hat{p})$ is computed using linearization or a replication method to reflect the sample design, including the fact that $\hat{p}$ is a sample-weighted estimator. By complex survey data, we mean data obtained from a multistage design with stratified selection of clusters at the first stage. For such a sample design, $d$ is usually taken to be equal to the number of sampled clusters minus the number of strata (Korn and Graubard 1990). The confidence interval (1.1), which we shall refer to as the "linear interval", is based on the assumption that $\hat{p}$ is approximately normally distributed. Under various reasonable asymptotics, this is known to be true (Krewski and Rao 1981). The use of the $t$ quantile rather than a normal-distribution quantile in (1.1) is based on empirical evidence (Frankel 1971, ch. 7), and it can also be formally justified using strong assumptions (Korn and Graubard 1990).

When the expected number of positive counts is small, the approximate normality of $\hat{p}$ breaks down (Cochran 1977, p. 58). For a simple random sample (or in the nonsurvey setting), one can avoid the normality assumption by using the Clopper and Pearson (1934) confidence interval based on the binomial distribution; see Vollset (1993) for a complete discussion of confidence intervals for proportions in the nonsurvey setting. When $x$ positive responses are seen in a simple random sample of size $n$, the Clopper-Pearson $1 - \alpha$ level confidence interval $(p_L(x, n), p_U(x, n))$ can be expressed as (Johnson, Kotz and Kemp 1993, p. 130):

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

$$p_U(x, n) = \frac{v_3 F_{v_3, v_4}(1 - \alpha/2)}{v_4 + v_3 F_{v_3, v_4}(1 - \alpha/2)} \qquad (1.2)$$

where $v_1 = 2x, v_2 = 2\,(n - x + 1), v_3 = 2(x + 1), v_4 = 2\,(n - x)$ and $F_{d_1, d_2}(\beta)$ is the $\beta$ quantile of an $F$ distribution with $d_1$ and $d_2$ degrees of freedom. For one-sided confidence bounds, $\alpha$ is used instead of $\alpha/2$ in the above expressions. For a simple random sample, these intervals are known to have coverage probability greater than or equal to their nominal level, regardless of the expected number of positive counts. They are sometimes referred to as "exact" confidence intervals; we shall refer to them as the "binomial intervals".

In this paper we suggest a simple modification to the binomial intervals to make them applicable for a proportion estimated from complex survey data. We are especially interested in the situation when the expected number of positive counts is small. Many survey analysts would not

[1] Edward L. Korn, Biometric Research Branch, EPN-739, National Cancer Institute, Bethesda, MD 20892, U.S.A.; Barry I. Graubard, Biostatistics Branch, National Cancer Institute, Bethesda, MD 20892, U.S.A.

present estimated proportions in this situation, since they are unreliable. For example, applying the relative-standard-error criterion for presenting proportions in the 1996 National Household Survey on Drug Abuse (SAMHSA 1998), the estimated proportion of women using cocaine in Table 7 would not be presented. We believe such proportions can provide valuable information, but that their lack of precision needs to be explicitly stated by presenting confidence intervals. In section 2, we define our proposed confidence intervals and define intervals based on a logit transformation and the Poisson distribution that have been suggested in the literature. Simulation results are presented in section 3 that compare the intervals. We find that the proposed intervals behave well in terms of coverage probability of the true proportion and in terms of their average width. Two applications are given in section 4 involving large surveys, but where the number of positive counts is expected to be small. We end with a discussion of some related work that constructs confidence intervals that are guaranteed to attain their nominal coverage probability regardless of the population configuration of counts.

## 2. PROPOSED AND OTHER CONFIDENCE LIMITS

For a $1-\alpha$ level confidence interval based on a sample of size $n$, first define the effective sample size by

$$n^* = \frac{\hat{p}(1-\hat{p})}{\hat{\text{var}}(\hat{p})} \tag{2.1}$$

and the degrees-of-freedom adjusted effective sample size by

$$n^*_{df} = \frac{\hat{p}(1-\hat{p})}{\hat{\text{var}}(\hat{p})}\left(\frac{t_{n-1}(1-\alpha/2)}{t_d(1-\alpha/2)}\right)^2 \tag{2.2}$$

Both $n^*$ and $n^*_{df}$ are set equal to $n$ when $\hat{p}=0$. The proposed limits substitute $n^*_{df}$ for $n$, and $\hat{p}n^*_{df}$ for $x$ in (1.2), *viz.* $p_L(\hat{p}n^*_{df}, n^*_{df})$ and $p_U(\hat{p}n^*_{df}, n^*_{df})$. (When $n$ is large, the $1-\alpha/2$ quantile of a normal distribution can be used in place of $t_{n-1}(1-\alpha/2)$ in (2.2).) For estimating a confidence interval for a proportion on a subdomain of the population, the sample size $n$ is taken to be equal to the sample size restricted to the subdomain.

A heuristic justification for this procedure is as follows. The effective sample size (2.1) is $n$ divided by an estimator of the design effect of the survey. This seems to be a reasonable way to incorporate the additional variability of $\hat{p}$ due to the complex sampling. For confidence interval construction, the variability of the variance estimator is also important. The second fraction in (2.2) takes into account the fact that $\hat{\text{var}}(\hat{p})$ will typically be more variable than a variance estimator that would be used for simple random sampling. If $d$ is large, then this factor is close to one and unneeded. For small $d$ and large $n$ and $\hat{p}n^*_{df}$, we would like the proposed interval to be close to the interval (1.1), which is appropriate in this situation. Using the fact that $F_{u,w}(\beta) \cong 1 + z(\beta)\sqrt{2(1/u + 1/w)}$ for large $u$ and $w$ (Johnson and Kotz 1970, p. 81), this is true, *i.e.*, $\hat{p} - p_L(\hat{p}n^*_{df}, n^*_{df}) \cong p_U(\hat{p}n^*_{df}, n^*_{df}) - \hat{p} \cong t_d(1-\alpha/2)[\hat{\text{var}}(\hat{p})]^{1/2}$.

A procedure closely related to the proposed procedure was developed by Breeze (1990) for use in the U.K. General Household Survey. This procedure is based on the simple-random-sampling $1-\alpha$ confidence interval $(po_L(x), po_U(x))$ for a Poisson random variable $x$, which can be expressed as (Johnson *et al.* 1993, p. 171):

$$po_L(x) = 0.5\,\chi^2_{\nu_1}(\alpha/2) \text{ and } po_U(x) = 0.5\,\chi^2_{\nu_2}(1-\alpha/2)$$

where $\nu_1 = 2x$, $\nu_2 = 2(x+1)$, and $\chi^2_\nu(\beta)$ is the $\beta$ quantile of a $\chi^2$ distribution with $\nu$ degrees of freedom. With complex survey data, the confidence interval is taken to be $(po_L(\hat{p}n^*)/n^*, po_U(\hat{p}n^*)/n^*)$.

A third procedure for confidence interval construction is based on a logit transform. For a $1-\alpha$ level confidence interval, the interval is

$$\left(\frac{1}{1+\exp(-LLOGIT)}, \frac{1}{1+\exp(-ULOGIT)}\right)$$

where

$$LLOGIT = \log\frac{\hat{p}}{1-\hat{p}} - t_d(1-\alpha/2)\frac{[\hat{\text{var}}(\hat{p})]^{1/2}}{\hat{p}(1-\hat{p})} \tag{2.3}$$

and

$$ULOGIT = \log\frac{\hat{p}}{1-\hat{p}} + t_d(1-\alpha/2)\frac{[\hat{\text{var}}(\hat{p})]^{1/2}}{\hat{p}(1-\hat{p})} \tag{2.4}$$

These intervals, with a normal-distribution quantile instead of a $t$ distribution quantile, were suggested for use with the 1996 National Household Survey on Drug Abuse (SAMHSA 1998). When $\hat{p}=0$, in the nonsurvey setting one might add a small constant to the observed number of events and nonevents, *e.g.*, 1/2, to be able to calculate the logit-transform confidence interval (Agresti 1990, pp. 249-250). In the present setting, when $\hat{p}=0$, we set the confidence interval equal to the binomial interval $(p_L(0,n), p_U(0,n))$.

In applications where it is known before sampling that the (true) design effect will be greater than 1, various modifications of the above procedures are possible. For our proposal, we recommend in this situation truncating the degrees-of-freedom adjusted effective sample size at $n$. That is, if $n^*_{df}$ is greater than $n$, we set its value to $n$, and define the lower and upper confidence limits to be $p_L(\hat{p}n, n)$ and $p_U(\hat{p}n, n)$. For the Breeze intervals, one could set $n^*$ to be $n$ if $n^* > n$. For the linear or logit intervals, one can use the simple-random-sampling variance estimator $\hat{p}(1-\hat{p})/n$ in place of $\hat{\text{var}}(\hat{p})$ in (1.1), (2.3) and (2.4) if $n^* > n$; see SAMHSA (1998) for additional truncation suggestions. The justification of these truncation

procedures is that the design effect may be estimated to be less than one because of instability of the variance estimator $\hat{\text{var}}(\hat{p})$. This type of instability may be especially large because $\hat{p}$ is small (SAMHSA 1998). The effect of these truncation procedures is to make the confidence intervals wider and more conservative. In theory, one could also adjust the estimated effective sample sizes when it is known before the sampling that the (true) design effect is less than one. However, to be conservative, we do not recommend doing this.

Our focus in this paper is on confidence intervals for the "superpopulation" probability that the outcome $Y = 1$ rather that the finite-population proportion. That is, the target parameter is $p = \sum_{u=1}^{N} p_u / N$ rather than $P = \sum_{u=1}^{N} Y_u / N$, where $Y_u$ has a Bernoulli distribution with parameter $p_u$, and $N$ is the population size. The simulated coverage probabilities given in the next section therefore refer to coverage of $p$. With this target parameter in mind, we do not use finite-population correction factors when estimating $\hat{\text{var}}(\hat{p})$ for use in (2.2); additional adjustments to the design-based variance $\hat{\text{var}}(\hat{p})$ for superpopulation inference are not pursued here (Korn and Graubard 1998). A referee suggests the possibility of a model-based approach to estimating a confidence interval for $p$. However, in our limited experience, such approaches yield estimators similar to weighted estimators and offer no advantages for inference (Pfeffermann and LaVange 1989; Graubard and Korn 1996).

If one were interested in a confidence interval for $P$, we would recommend using the proposed intervals but with $\hat{\text{var}}(\hat{p})$ in (2.2) containing the finite-population correction factors. A confidence interval for $\sum_{u=1}^{N} Y_u$ could be obtained by multiplying the ends of the confidence interval for $P$ by $N$, if known, or by an estimator $\hat{N}$ of $N$, if not known. (In theory, one could account for the variability of $\hat{N}$, but this additional variability will be small.) An alternative approach for estimating a confidence interval for $P$ would be to modify the usual limits (Guenther 1983) appropriate for a simple random sample (based on the hypergeometric distribution) similarly to the way the proposed intervals modify the binomial intervals.

## 3. SIMULATIONS

The main simulation results are presented in Tables 1-5. Table 1 presents the results of simulations in which datasets of 32 clusters, each with sample size 100, were simulated. Within cluster $i$, the number of positive events was simulated with a binomial distribution with probability parameter $p_i$. In Table 1, we refer to the $\{p_i, i = 1, ..., 32\}$ as the cluster probabilities. For the top third of the table, the cluster probabilities are taken to be the constant $p = .1, .02,$

**Table 1**
Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32 Clusters and 100 Observations Per Cluster; Sample Weights are 1 Or 10 with Probability 1/2 (Noninformative Sample Weights)

| Distribution of cluster proportions[a] | Overall proportion | Expected number positive | Method of calculating confidence bounds | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Linear | | Logit | | Breeze | | Proposed | |
| | | | Lower | Upper | Lower | Upper | Lower | Upper | Lower | Upper |
| (1) | | | | | | | | | | |
| .1 | .1 | 320 | 4.6 | 5.5 | 5.3 | 4.6 | 4.5 | 4.1 | 4.8 | 4.4 |
| .02 | .02 | 64 | 3.4 | 7.1 | 5.2 | 4.6 | 4.5 | 4.7 | 4.2 | 4.4 |
| .01 | .01 | 32 | 2.9 | 8.0 | 5.4 | 4.5 | 4.4 | 4.5 | 4.0 | 4.1 |
| .0025 | .0025 | 8 | 1.6 | 9.5 | 5.5 | 1.8 | 3.6 | 2.2 | 3.3 | 1.8 |
| (1/2, 1/2) | | | | | | | | | | |
| .05, .15 | .1 | 320 | 4.3 | 5.8 | 5.5 | 4.3 | 4.3 | 3.8 | 4.7 | 4.1 |
| .01, .03 | .02 | 64 | 3.1 | 7.5 | 5.2 | 4.8 | 4.3 | 4.8 | 4.0 | 4.5 |
| .005, .015 | .01 | 32 | 2.7 | 8.6 | 5.2 | 4.7 | 4.1 | 4.9 | 3.7 | 4.4 |
| .00125, .00375 | .0025 | 8 | 1.5 | 9.9 | 5.4 | 2.0 | 3.4 | 2.3 | 3.1 | 2.0 |
| (3/4, 1/4) | | | | | | | | | | |
| .05, .25 | .1 | 320 | 3.1 | 7.8 | 4.7 | 5.6 | 3.4 | 5.0 | 3.6 | 5.3 |
| .01, .05 | .02 | 64 | 2.7 | 8.6 | 5.1 | 5.3 | 4.0 | 5.4 | 3.7 | 5.0 |
| .005, .025 | .01 | 32 | 2.2 | 9.8 | 5.0 | 5.3 | 3.7 | 5.5 | 3.3 | 5.0 |
| .00125, .00625 | .0025 | 8 | 1.3 | 10.7 | 5.3 | 2.2 | 3.3 | 2.5 | 3.0 | 2.2 |

(a) Fractions in parentheses are the probabilities that the cluster proportions have the stated value.

**Table 2**
Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32 Clusters and 100 Observations Per Cluster; Informative Sample Weights are 1 or 10 (See Text)

| Distribution of cluster proportions[a] | Overall weighted proportion | Expected number positive | Method of calculating confidence bounds | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Linear | | Logit | | Breeze | | Proposed | |
| | | | Lower | Upper | Lower | Upper | Lower | Upper | Lower | Upper |
| (1) | | | | | | | | | | |
| .1 | .1 | 191.0 | 4.3 | 5.9 | 5.1 | 4.9 | 4.2 | 4.4 | 4.6 | 4.6 |
| .02 | .02 | 36.9 | 3.3 | 7.3 | 5.3 | 4.3 | 4.4 | 4.4 | 4.1 | 4.1 |
| .01 | .01 | 18.4 | 2.8 | 8.7 | 5.5 | 4.0 | 4.3 | 4.3 | 3.9 | 3.7 |
| .0025 | .0025 | 4.6 | 1.3 | 18.7 | 6.1 | 4.8 | 3.2 | 4.8 | 2.8 | 4.8 |
| (1/2, 1/2) | | | | | | | | | | |
| .05, .15 | .1 | 191.0 | 5.0 | 5.0 | 6.4 | 3.7 | 5.1 | 3.2 | 5.4 | 3.4 |
| .01, .03 | .02 | 36.9 | 3.0 | 7.9 | 5.4 | 4.5 | 4.3 | 4.6 | 4.0 | 4.3 |
| .005, .015 | .01 | 18.4 | 2.5 | 9.2 | 5.4 | 4.2 | 4.1 | 4.4 | 3.7 | 3.9 |
| .00125, .00375 | .0025 | 4.6 | 1.3 | 19.0 | 6.1 | 4.9 | 3.2 | 4.9 | 2.8 | 4.9 |
| (3/4, 1/4) | | | | | | | | | | |
| .05, .25 | .1 | 191.0 | 4.7 | 5.7 | 7.1 | 4.1 | 5.1 | 3.6 | 5.5 | 3.8 |
| .01, .05 | .02 | 36.9 | 2.6 | 8.9 | 5.2 | 5.2 | 4.0 | 5.3 | 3.7 | 4.9 |
| .005, .025 | .01 | 18.4 | 2.1 | 10.1 | 5.3 | 4.8 | 3.8 | 5.1 | 3.4 | 4.5 |
| .00125, .00625 | .0025 | 4.6 | 1.2 | 19.8 | 5.9 | 5.3 | 3.2 | 5.3 | 2.8 | 5.3 |

(a) Fractions in parentheses are the probabilities that the cluster weighted proportions have the stated value.

**Table 3**
Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32 Clusters and 100 Observations Per Cluster; Unweighted Analyses

| Distribution of cluster proportions[a] | Overall proportion | Expected number positive | Method of calculating confidence bounds | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Linear | | Logit | | Breeze | | Proposed | |
| | | | Lower | Upper | Lower | Upper | Lower | Upper | Lower | Upper |
| (1) | | | | | | | | | | |
| .1 | .1 | 320 | 5.0 | 4.9 | 5.7 | 4.2 | 4.9 | 3.8 | 5.2 | 4.1 |
| .02 | .02 | 64 | 3.8 | 6.3 | 5.2 | 4.5 | 4.7 | 4.8 | 4.4 | 4.4 |
| .01 | .01 | 32 | 3.5 | 6.8 | 5.6 | 4.4 | 4.7 | 4.4 | 4.3 | 4.0 |
| .0025 | .0025 | 8 | 2.5 | 8.8 | 5.6 | 3.8 | 4.1 | 3.9 | 3.9 | 3.9 |
| (1/2, 1/2) | | | | | | | | | | |
| .05, .15 | .1 | 320 | 4.5 | 5.6 | 5.6 | 4.2 | 4.5 | 3.7 | 4.8 | 4.0 |
| .01, .03 | .02 | 64 | 3.4 | 7.0 | 5.1 | 4.8 | 4.5 | 4.9 | 4.1 | 4.6 |
| .005, .015 | .01 | 32 | 3.0 | 7.6 | 5.2 | 4.8 | 4.4 | 4.8 | 3.9 | 4.4 |
| .00125, .00375 | .0025 | 8 | 2.2 | 9.2 | 5.4 | 4.3 | 3.8 | 4.3 | 3.5 | 4.3 |
| (3/4, 1/4) | | | | | | | | | | |
| .05, .25 | .1 | 320 | 3.3 | 7.7 | 4.8 | 5.6 | 3.5 | 5.1 | 3.7 | 5.3 |
| .01, .05 | .02 | 64 | 2.9 | 8.1 | 5.1 | 5.2 | 4.1 | 5.3 | 3.8 | 4.9 |
| .005, .025 | .01 | 32 | 2.5 | 9.2 | 4.9 | 5.6 | 3.9 | 5.6 | 3.5 | 5.2 |
| .00125, .00625 | .0025 | 8 | 2.0 | 10.4 | 5.3 | 5.1 | 3.8 | 5.1 | 3.3 | 5.1 |

(a) Fractions in parentheses are the probabilities that the cluster proportions have the stated value.

**Table 4**
Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32 Clusters and 10 Observations Per Cluster; Sample Weights are 1 or 10 with Probability 1/2 (Noninformative Sample Weights)

| Distribution of cluster proportions[a] | Overall proportion | Expected number positive | Method of calculating confidence bounds | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Linear | | Logit | | Breeze | | Proposed | |
| | | | Lower | Upper | Lower | Upper | Lower | Upper | Lower | Upper |
| (1) | | | | | | | | | | |
| .2 | .2 | 64 | 4.0 | 6.6 | 5.2 | 4.7 | 3.1 | 4.7 | 4.2 | 4.3 |
| .1 | .1 | 32 | 3.2 | 7.8 | 5.3 | 4.4 | 3.6 | 3.8 | 3.9 | 4.0 |
| .025 | .025 | 8 | 1.7 | 10.2 | 5.5 | 2.1 | 3.4 | 2.1 | 3.2 | 2.4 |
| (1/2, 1/2) | | | | | | | | | | |
| .1, .3 | .2 | 64 | 3.6 | 7.0 | 5.0 | 4.9 | 2.8 | 3.4 | 3.9 | 4.4 |
| .05, .15 | .1 | 32 | 3.0 | 8.1 | 5.1 | 4.6 | 3.4 | 4.0 | 3.7 | 4.2 |
| .0125, .0375 | .025 | 8 | 1.6 | 10.6 | 5.4 | 2.1 | 3.3 | 2.1 | 3.1 | 2.5 |
| (3/4, 1/4) | | | | | | | | | | |
| .1, .5 | .2 | 64 | 3.1 | 7.8 | 4.6 | 5.3 | 2.4 | 3.9 | 3.3 | 4.8 |
| .05, .25 | .1 | 32 | 2.5 | 9.2 | 4.8 | 5.2 | 3.0 | 4.6 | 3.3 | 4.8 |
| .0125, .0625 | .025 | 8 | 1.5 | 11.5 | 5.3 | 2.4 | 3.2 | 3.5 | 3.0 | 2.8 |

(a) Fractions in parentheses are the probabilities that the cluster proportions have the stated value.

**Table 5**
Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32 Clusters and 10 or 100 Observations Per Cluster with Probability 1/2; Sample Weights are 1 or 10 with Probability 1/2 (Noninformative Sample Weights)

| Distribution of cluster proportions[a] | Overall proportion | Expected number positive | Method of calculating confidence bounds | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Linear | | Logit | | Breeze | | Proposed | |
| | | | Lower | Upper | Lower | Upper | Lower | Upper | Lower | Upper |
| (1) | | | | | | | | | | |
| .1818 | .1818 | 320 | 5.1 | 6.0 | 5.7 | 5.2 | 4.2 | 4.1 | 5.2 | 5.0 |
| .0364 | .0364 | 64 | 4.1 | 7.6 | 5.7 | 5.2 | 5.0 | 5.2 | 4.8 | 4.9 |
| .0182 | .0182 | 32 | 3.4 | 8.5 | 5.7 | 5.0 | 4.7 | 5.1 | 4.4 | 4.7 |
| .0045 | .0045 | 8 | 2.0 | 12.7 | 5.9 | 3.4 | 4.0 | 4.3 | 3.6 | 3.8 |
| (1/2, 1/2) | | | | | | | | | | |
| .0909, .2727 | .1818 | 320 | 5.0 | 6.4 | 6.1 | 4.8 | 4.2 | 3.6 | 5.2 | 4.4 |
| .0182, .0545 | .0364 | 64 | 3.9 | 8.1 | 6.0 | 5.1 | 4.9 | 5.0 | 4.7 | 4.8 |
| .0091, .0273 | .0182 | 32 | 3.1 | 9.3 | 5.8 | 5.2 | 4.5 | 5.3 | 4.2 | 4.9 |
| .0023, .0068 | .0045 | 8 | 1.8 | 13.2 | 5.9 | 3.6 | 3.9 | 4.5 | 3.5 | 4.0 |
| (3/4, 1/4) | | | | | | | | | | |
| .0909, .4545 | .1818 | 320 | 3.1 | 9.9 | 4.6 | 7.6 | 2.5 | 6.3 | 3.3 | 7.1 |
| .0182, .0909 | .0364 | 64 | 2.8 | 10.9 | 5.3 | 7.3 | 3.9 | 7.3 | 3.7 | 7.0 |
| .0091, .0455 | .0182 | 32 | 2.4 | 11.5 | 5.4 | 6.8 | 3.9 | 6.9 | 3.6 | 6.5 |
| .0023, .0114 | .0045 | 8 | 1.6 | 14.5 | 5.7 | 4.0 | 3.7 | 5.0 | 3.3 | 4.4 |

(a) Fractions in parentheses are the probabilities that the cluster weighted proportions have the stated value.

.01, or .0025, corresponding to an expected number of positive events equal to 320, 64, 32, or 8 out of the sample size of 3200. For the middle third of the table, the cluster probabilities are taken to be $p/2$ with probability 1/2 or $3p/2$ with probability 1/2, with $p$ as in the first third of the table. Varying the $p_i$ across the clusters induces an intracluster correlation among the observations. For the middle third of the table, these correlations (ignoring the sample weights) are .00278, .0051, .0025 and .0006 corresponding to the expected number of positive events being 320, 64. 32, or 8, respectively. For the bottom third of the table, the cluster probabilities are taken to be $p/2$ with probability 3/4 or $5p/2$ with probability 1/4, corresponding to intraclass correlations of .0833, .0153, .0076 and .0019. For all simulations in Table 1, sample weights of 1 or 10 are randomly assigned with probability 1/2 to each observation (noninformative weights).

The results presented in Table 1 are appropriate for one-sided 95% upper and lower confidence limits; ideally the lack-of-coverage percentages in the table should be less than or equal to the nominal value of 5.0. The results are also relevant for two-sided 90% confidence intervals, for which ideally both the upper and lower values in the table should both be $\leq 5.0$ (Jennings 1987). For each line of the table, 100,000 datasets were simulated using the random number generator in SAS (1990, p. 631) to estimate the probabilities of noncoverage of the confidence limits.

For the linear confidence bounds, the upper confidence limit falls below the true value more than the 5% nominal level. Somewhat surprisingly, this is true even with as large as an expected 320 positive counts, especially with positive intracluster correlation (middle and bottom third of the table). For the logit-transform confidence bounds, the noncoverage appears slightly higher than the nominal level, especially for the lower limits. Both the Breeze and proposed confidence bounds appear generally conservative. Simple-random-sampling binomial limits are not appropriate for the cases simulated in Table 1 because of the sample weights and the intracluster correlation (in the bottom two-thirds of the table). This can be demonstrated by noting that the lack of coverage for both the upper and lower binomial bounds are greater than 8% for all the cases considered in the table (results not shown).

As it is slightly complicated to discuss confidence interval "lengths" for one-sided bounds, we restrict discussion to the lengths of the two-sided 90% confidence intervals. Over all the simulations presented in Table 1, the Breeze and proposed intervals are 3.3% and 4.9% wider on average than the logit-transform intervals.

Table 2 presents simulation results for the same setup as Table 1 except that the sample weights were taken to be informative. This was done by setting the sample weight to be 10 with probability 2/3 if the event was positive and with probability 1/3 if the event was not positive, otherwise the weight was set to 1. The probability that an event was positive in each cluster was adjusted downwards so that the overall weighted proportions were the same as in Table 1. The results in Table 2 look similar to those in Table 1 except the linear and logit intervals tend to have worse coverage probabilities.

Table 3 presents simulation results for the same setup as Table 1 except the analysis is unweighted. The results are very similar to the Table 1 results. Since the top third of Table 3 corresponds to no intracluster correlation, one could also use the simple-random-sampling binomial limits there. Averaging over the four situations in this third of the table, the proposed limits are 2.5% wider that the binomial limits (results not shown). As the true design effect is 1.0 in the top third of Table 3, these simulations can be used to examine the effect of truncation of $n^*_{df}$ in the proposed procedure. (Truncation is uncommon in the simulations in Table 1, since the true design effects there are all >1.) Simulation using the proposed procedure with truncation lead to wider more conservative intervals than for the proposed intervals in the top third of Table 3. Averaging over the four situations considered, the proposed limits with truncation are 4.0% wider than the proposed limits (results not shown for truncated limits).

Table 4 presents simulation results for the same setup as Table 1 except 10 rather than 100 observations are simulated within each cluster. The results are very similar to Table 1 when one compares simulations with the same expected number of positive events. The one exception is the increased conservativeness of the Breeze intervals as compared to the proposed method. This is because the overall proportions are higher in Table 4 than Table 1 for a given expected number of positive events (since the sample size is smaller in Table 4). The Poisson intervals of Breeze do not work well with proportions that are not small. For example, we performed a simulation corresponding to the top third of Table 1 except that the overall proportion $p = .5$ with 1600 expected number of positive events. The simulated lower and upper lack-of-coverage percentages for the Breeze bounds were 1.2% and 1.3%, compared to 4.6% and 4.7% for the proposed method. The Breeze intervals were on average 37% wider than the proposed intervals.

The Breeze intervals also do not work well when the number of clusters is very small, since they do not account for degrees of freedom of the variance estimation. For example, we performed a simulation corresponding to the top third of Table 1 except that data from only 8 clusters were simulated (with 100 observations per cluster), and $p_i \equiv .1$ so that the expected number of positive events was 80. The simulated lower and upper lack-of-coverage percentages for the Breeze bounds were 6.1% and 5.4%, compared to 4.7% and 4.0% for the proposed method.

Table 5 presents simulation results for the same setup as Table 1 except the cluster sizes were taken to be 10 or 100 with probability 1/2. The lack-of-coverage probabilities are larger than the nominal 5% in the bottom third of the table for all the methods. The logit intervals also do not behave as well as in Table 1 for the top two-thirds of the table.

An additional set of simulations was done in which two clusters (each of sample size 50) were simulated from each 32 strata. The expected numbers of positive event were taken as in Table 1, the weights were randomly set to 1 or 10, and the probability of a positive event was taken to be different in the different strata to simulate an intracluster correlation. The results (not shown) were very similar to the results given in Table 1.

## 4. APPLICATIONS

In this section we consider two applications in which the numbers of positive counts are small. In the first application, involving estimating HIV positivity in an unselected population, the numbers of positive counts are small because the rates of HIV infection are small. In the second application, involving estimating whether individuals have ever used cocaine, the rates are not small but the numbers of positive counts are small because we restrict the analyses to relatively small subdomains. For both applications, SUDAAN (Shah, Barnwell and Bieler 1995) was used to calculate the (design-based) standard errors of the proportions, and the function "FINV" in SAS (1990, p. 547) was used to calculate the quantiles of the $F$ distribution in (1.2).

### 4.1 Seroprevalence of HIV Estimated From the Third National Health and Nutrition Examination Survey (NHANES III)

NHANES III was a survey conducted in 1988-1994 of the civilian noninstitutionalized population ages 2 months or older of the United States (National Center for Health Statistics 1994). An HIV test was performed on participating individuals 18 years of age or older. McQuillan, Khare, Karon, Schable and Vlahov (1997) studied the seroprevalence of HIV for individuals under the age of 60 years and various subgroups, some of which are displayed in Table 6. Of the 11,202 individuals tested, 59 were infected. The estimated prevalence in Table 6, 0.32%, is far from the unweighted proportion, 0.53% = 59/11202, because the estimated prevalence is a weighted proportion utilizing the sample weights. Because the testing for HIV was anonymous, for these analyses the sample weights were derived from the original NHANES III sample weights of all individuals in the same stand (survey location), race/ethnicity group, sex, and age group (18-39 *vs.* 40-59) of the tested individual (M. Khare, personal communication). The pseudo-design for variance estimation was the sampling of 2 pseudo-PSU's from each of 23 strata (M. Khare, personal communication), which is not the pseudo-design typically used for NHANES III variance estimation.

The linear 90% confidence intervals for prevalence for the various groups in Table 6 are shifted to the left and shorter than the other intervals, which are similar to each other. The proposed intervals are very slightly wider than the Breeze or logit intervals. The effective sample sizes calculated in (2.1) are markedly smaller than the sample sizes because of the design effects of the survey; the confidence intervals based on the truncated procedures will therefore be identical to the ones given in Table 6. The differences between $n^*$ and $n^*_{df}$ are relatively minor. For this relatively rare outcome, the simulations given in section 3 suggest that the Breeze and proposed confidence intervals may maintain their nominal 90% coverage probabilities better than the other intervals.

**Table 6**
Seroprevalence of HIV Among Adults Aged 18-59 Years Based on the Third National Health and Nutrition Examination Survey

| | Total | Sex | | Race/ethnicity | | |
|---|---|---|---|---|---|---|
| | | Male | Female | White | Black | Mex. - Amer. |
| Sample size | 11202 | 5142 | 6060 | 4128 | 3579 | 3495 |
| Number infected | 59 | 44 | 15 | 9 | 38 | 12 |
| Prevalence (%) ± SE | 0.320 ± 0.076 | 0.519 ± 0.130 | 0.127 ± 0.053 | 0.203 ± 0.071 | 1.100 ± 0.247 | 0.368 ± 0.134 |
| Effective sample size | | | | | | |
| $n^*$ | 5588 | 3056 | 4433 | 3976 | 1779 | 2039 |
| $n^*_{df}$ | 5148 | 2816 | 4084 | 3664 | 1640 | 1880 |
| Linear 90% con. int. | (0.19, 0.45) | (0.30, 0.74) | (0.04, 0.22) | (0.08, 0.33) | (0.68, 1.52) | (0.14, 0.60) |
| Logit 90% con. int. | (0.21, 0.48) | (0.34, 0.80) | (0.06, 0.26) | (0.11, 0.37) | (0.75, 1.62) | (0.20, 0.69) |
| Breeze 90% con. int. | (0.21, 0.48) | (0.32, 0.79) | (0.05, 0.26) | (0.10, 0.37) | (0.73, 1.61) | (0.18, 0.68) |
| Proposed 90% con. int. | (0.20, 0.48) | (0.32, 0.80) | (0.05, 0.26) | (0.10, 0.37) | (0.71, 1.63) | (0.17, 0.69) |

**Table 7**
"Ever Users" of Cocaine Among Adults Ages 12-44 Years Based on Individuals with 16 or More Years of Education Sampled in Hispanic Health and Nutrition Examination Survey

| | Total | Sex | |
|---|---|---|---|
| | | Male | Female |
| Sample size | 123 | 69 | 54 |
| Ever-users | 13 | 10 | 3 |
| Proportion (%) ± SE | 11.6 ± 2.5 | 14.3 ± 3.4 | 7.0 ± 4.8 |
| Effective sample size | | | |
| $n^*$ | 167.1 | 105.0 | 28.2 |
| $n^*_{df}$ | 132.8 | 84.4 | 22.9 |
| Linear 90% confidence int. | (7.0, 16.2) | (8.0, 20.7) | (-1.9[a], 15.9) |
| Logit 90% confidence int. | (7.8, 17.1) | (9.1, 21.9) | (1.9, 22.8) |
| Breeze 90% confidence int. | (7.7, 17.0) | (8.3, 23.2) | (0.9, 24.8) |
| Proposed 90% confidence int. | (7.4, 17.2) | (8.5, 22.1) | (0.9, 22.7) |
| Truncated Procedures | | | |
| Linear 90% confidence int. | (6.3, 17.0) | (6.5, 22.2) | same as above |
| Logit 90% confidence int. | (7.2, 18.2) | (8.1, 24.1) | " |
| Breeze 90% confidence int. | (7.1, 18.1) | (7.7, 24.4) | " |
| Proposed 90% confidence int. | (7.2, 17.5) | (8.0, 23.2) | " |

(a) In practice, this interval would be presented as (0, 15.9) since negative proportions are impossible.

### 4.2 Use of Cocaine Among College-educated Individuals Sampled in the Hispanic Health and Nutrition Examination Survey (HHANES)

HHANES was a survey conducted in 1982-1983 of three Hispanic groups living in the United States (National Center for Health Statistics 1985). We restrict attention here to the Mexican-American sample. Individuals ages 12-44 years were asked "About how old were you the first time you tried cocaine?". The possible answers were the age of the individual (in years) when he first tried cocaine, a "never used" category, and a "don't know" category. We consider estimating the proportion of "ever-users" among individuals who completed 16 or more years of education (for which there were no "don't know" responses).

There were 13 ever-users among 123 sampled individuals, with the sample-weighted proportion being 11.6% (Table 7). The design-based standard error, 2.5%, is estimated with only 8 degrees of freedom since the sampling design of HHANES can be approximated by the sampling of 2 PSU's from each of 8 strata (Kovar and Johnson 1986). The effective sample sizes are $n^* = 167.1$ and $n^*_{df} = 132.8$, which are both greater than the sample size. This is because the estimated design effect is .736, so that $n^* = 123/.736 = 167.1$. (The second factor in (2.2) is 0.794.) Despite the stratification, we think that the true design effect is greater than 1 for this survey because of the clustering and the sample weighting. (The estimated design effect is estimated poorly because of the limited degrees of freedom.) We therefore think that the truncated procedures are reasonable for this application.

Because of the limited degrees of freedom, and because the outcome is not rare, there are more differences between the logit, Breeze and proposed confidence intervals displayed in Table 7. Based on the simulations given in section 3, we recommend the proposed (truncated) confidence intervals.

Our approach may appear slightly inconsistent for this survey in that we accept poorly-estimated effective sample sizes less than the sample size but truncate those greater. We believe that this is a reasonable conservative approach to use when it is thought that the true design effect is probably greater than 1.

## 5. DISCUSSION

Although the confidence intervals proposed here had adequate coverage probability for almost all the simulations performed, this is not guaranteed for all possible configurations of the population, *e.g.*, see the bottom third of Table 5. An example with a more serious lack of coverage can also easily be constructed: Suppose that the population consists of clusters of size 100, and that 10% of the clusters have all positive units and the remaining 90% have all zero units. If we sample 10 clusters as a simple random sample, and subsample all the units in the sampled clusters, then 35% (= $(1-.1)^{10}$) of the time we will observe no positive units in the sample size of 1000. In this situation, our proposed intervals reduce to the usual binomial ones, so that, *e.g.*, the upper 95% confidence limit for the population proportion is given by .003 ( = $1-.05^{1/1000}$). This implies that

the upper 95% confidence interval is less that the true value of .10 at least 35% of the time, a serious undercoverage.

It is possible in simple sampling situations to construct confidence intervals that are guaranteed to have at least their nominal coverage probability by considering all possible configurations of the population, and using a least-favorable configuration for the coverage probability. For the hypothetical single-stage cluster sample mentioned above, for example, an upper 95% confidence limit could be given by the binomial limit based on 0 positive units out of 10, *i.e.*, .26 (=1-.05$^{1/10}$). Such confidence intervals, which can become computationally intensive to calculate, have been studied by Gross and Frankel (1991), who also suggest some less computationally intensive approximations.

The advantages of our proposed intervals over such approaches are (1) they are easy to calculate, (2) they accommodate any complex sampling design, including nonresponse and postsratification adjustments to the sample weights, (3) they will generally maintain their nominal coverage probability, (4) they will be less conservative than intervals that are guaranteed to maintain their nominal coverage probability for all population configurations, and (5) they have better properties than the linear intervals, logit-transform or Breeze intervals. Conclusions (2) and (5) are based on our simulation results, which of course do not cover all possible situations. More research would be useful in this regard.

## ACKNOWLEDGEMENTS

The authors like to thank M. Khare for providing the prevalence estimates and their design-based standard errors given in Table 6, and the Associate Editor and referees for their helpful comments.

## REFERENCES

AGRESTI, A. (1990). *Categorical Data Anlaysis*. New York: Wiley.

BREEZE, E. (1990). *General Household Survey: Report on Sampling Error*. London: Her Majesty's Stationery Office (Office of Population Censuses and Surveys).

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.

COCHRAN, W.G. (1977). *Sampling Techniques*. Third Edition. New York: Wiley.

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

GRAUBARD, B.I., and KORN, E.L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*, 5, 263-281.

GROSS, S.T., and FRANKEL, M.R. (1991). Confidence limits for small proportions in complex samples. *Communications in Statistics – Theory and Methods*, 20, 951-975.

GUENTHER, W.C. (1983). Hypergeometric distributions. In *Encyclopedia of Statistical Sciences*, Volume 3, (Eds. S. Kotz and N.L. Johnson). New York: Wiley, 707-712.

JENNINGS, D.E. (1987). How do we judge confidence-interval adequacy? *American Statistician*, 41, 335-337.

JOHNSON, N.L., and KOTZ, S. (1970). *Continuous Univariate Distributions – 2*. New York: Wiley.

JOHNSON, N.L., KOTZ, S., and KEMP, A.W. (1993). *Univariate Discrete Distributions*. Second Edition. New York: Wiley.

KORN, E.L., and GRAUBARD, B.I. (1990). Simultaneous testing or regression coefficients with complex survey data: use of Bonferroni *t*-statistics. *American Statistician*, 44, 270-276.

KORN, E.L., and GRAUBARD, B.I. (1998). Variance estimation for superpopulation parameters. *Statistica Sinica*, 8, 1131-1151.

KOVAR, M.G., and JOHNSON, C. (1986). Design effects from the Mexican American portion of the Hispanic Health and Nutrition Examination Survey: a strategy for analysts. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-399.

KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

McQUILLAN, G.M., KHARE, M., KARON, J.M., SCHABLE, C.A., and VLAHOV, D. (1997). Update on the seroepidemiology of Human Immunodeficiency Virus in the United States household population: NHANES III, 1988-94. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 14, 355-360.

NATIONAL CENTER FOR HEALTH STATISTICS (1985). Plan and operation of the Hispanic Health and Nutrition Examination Survey, 1982-84. *Vital and Health Statistics* 1(19). Hyattsville, MD: National Center for Health Statistics.

NATIONAL CENTER FOR HEALTH STATISTICS (1994). Plan and operation of the Third National Health and Nutrition Examination Survey, 1988-94. *Vital and Health Statistics* 1(32). Hyattsville, MD: National Center for Health Statistics.

PFEFFERMANN, D., and LAVANGE, L. (1989). Regression models for stratified multi-stage cluster samples. In *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt, and T.M.F. Smith). New York: Wiley, 237-260.

SAMHSA (1998). National Household Survey on Drug Abuse: Main Findings 1996. (DHHS Publication No. (SMA) 98-3200). Rockville, MD: SAMHSA.

SAS (1990). *SAS Language: Reference, Version 6*, First Edition. Cary, NC: SAS Institute Inc.

SHAH, B.V., BARNWELL, B.G., and BIELER, G.S. (1995). *SUDAAN User's Manual, Release 6.40*. Research Triangle Park, NC: Research Triangle Institute.

VOLLSET, S.E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine*, 12, 809-824.