

## CONFIDENCE INTERVALS FOR PROPORTIONS AND QUANTILES UNDER TWO-STAGE SAMPLING DESIGNS: AN EMPIRICAL STUDY

Cindy X. Feng<sup>1</sup> and Randy R. Sitter

### ABSTRACT

It has been well known that the conventional confidence interval for population proportions does not perform well for large or small values of proportions. Several alternative methods have been proposed in the literature, where the sample data are independent and identically distributed. For finite populations the problem is further complicated due to the use of complex sampling designs and issues related to effective sample sizes and effective degrees of freedom. In this paper we investigate the performance of several confidence intervals for proportions and quantiles under two-stage sampling designs through simulation studies. An application to the U.S. National Health and Nutrition Examination Surveys (NHANES) is briefly discussed.

KEY WORDS: Complex Surveys; Design Effect; Domain Estimation, Effective Sample Size; Effective Degrees of Freedom

### RÉSUMÉ

C'est connu que l'intervalle de confiance classique calculé pour les proportions ne donne pas de bons résultats en cas de proportions grandes ou petites. Plusieurs méthodes alternatives ont été proposées dans la littérature dans le contexte de données indépendantes et identiquement distribuées. Pour des populations finies, le problème se complique en raison des plans d'échantillonnage complexes et des questions autour des tailles effectives d'échantillon et du nombre effectif de degrés de liberté. Dans cet article, nous étudions par simulation la performance de plusieurs intervalles de confiance pour les estimations de proportions et de quantiles dans un plan de sondage à deux degrés. Nous discutons aussi brièvement d'une application à l'enquête américaine « National Health and Nutrition Examination Survey » (NHANES).

MOTS CLÉS : degrés de liberté effectifs; effet du plan de sondage; enquêtes complexes; estimation par domaines; taille d'échantillon effectif.

### 1. INTRODUCTION

Interval Estimation of a binomial proportion is a very basic practical problem in statistics. The popular standard interval in an independent and identically distributed (IID) setting,  $\hat{p} \pm z_{\alpha/2} n^{-1/2} \sqrt{\hat{p}(1-\hat{p})}$ , is in nearly universal use. Recent literature has highlighted the "chaotic" behaviour of the standard interval for the binomial proportion and proposed some alternatives. Brown, Cai and Dasgupta (2001, 2002) made many constructive remarks and suggestions concerning this issue and recommended some of the alternative intervals proposed in the literature in the IID case. However, many large complex surveys are conducted using a stratified multi-stage sampling design. It is often that the sampling design induces a non-IID structure on the data (e.g., stratification, clustering and unequal probability sampling). Therefore, in complex survey settings, confidence intervals for the proportion based on the binomial distribution, normal approximation to binomial distribution and Poisson approximation to the binomial distribution are not directly applicable. Different approaches have been suggested, such as a modified binomial approach (Korn and Graubard, 1998), an interval based on a Poisson approximation (Breeze, 1990) and variance stabilization approaches. We also extend the Wilson (Wilson, 1927) and the likelihood ratio intervals from the IID case to the complex survey context to incorporate the complex survey data structure. We evaluate all these methods under some preliminary complex survey settings and then reconsider these methods with regards to the National Health and Nutrition Examination Surveys (NHANES), where small proportion is of particular concern.

---

<sup>1</sup> Cindy X. Feng, PhD Candidate, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, B.C., V5A1S6, Canada Email: [xfeng@stat.sfu.ca](mailto:xfeng@stat.sfu.ca) and Randy R. Sitter, Professor, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, B.C., V5A1S6, Canada

Another concern in large scale surveys is the confidence intervals for small quantiles, which are usually obtained by inverting an estimated distribution function. Sitter and Wu (2001) show that despite the poor performance of the standard confidence interval for small or large  $p$ , Woodruff intervals for small or large quantiles by inverting the badly behaved intervals perform very well. They also argue that improving the confidence interval for  $p$  should negatively impact intervals for small and large quantiles using Woodruff's method. However, they do not consider weighting, multi-stage sampling, design effects, effective sample size, etc. Thus, we extend the Woodruff intervals by inverting all of the intervals discussed for the confidence interval for  $p$  and investigate its performance via simulation in a similar fashion to our study of intervals for  $p$ .

The paper is organized as follows. Section 2 introduces notation for multi-stage stratified sampling and describes the confidence intervals for proportions in complex surveys. Section 3 investigates properties of the methods via simulation. Section 4 describes the sampling design of NHANES. Section 5 introduces the confidence intervals for quantiles. Concluding remarks are made in section 6.

## 2. CONFIDENCE INTERVALS FOR PROPORTIONS IN COMPLEX SURVEYS

### 2.1 Stratified Two-Stage Sampling

Suppose a finite population consisting of  $N = \sum_{i=1}^{N_h} N_h$  primary sampling units (PSU's) is partitioned into  $H$  non-overlapping strata of  $N_1, N_2, \dots, N_H$  PSU's, respectively. Each PSU consists of secondary sampling units (SSU's), and so on through possibly many stages. The additional stages of sampling within PSU's can be quite complicated, possibly involving not only multiple additional stages, but also further stratification, clustering, etc. The total number of ultimate units is  $M = \sum_{h=1}^H \sum_{i=1}^{N_h} M_{hi}$ , where  $M_{hi}$  is the number of ultimate units in the  $i$ th PSU within stratum  $h$ . Suppose  $n_h$  PSU's are selected from the  $N_h$  PSU's in stratum  $h$  under some sampling scheme. The total number of sampled PSU's is  $n = \sum_{h=1}^H n_h$ . A typical technique developed by Hansen and Hurwitz (1943) is to select the PSU's with probabilities proportion to their sizes  $M_{hi}$  with corresponding probability of selection  $z_{hi} = M_{hi} / M_h$  provided all the  $M_{hi}$  are known. In some applications, the sizes  $M_{hi}$  are unknown or 'size' is not the number of elements in the PSU but a measure of size that is known to be highly correlated with the PSU total. Within each selected cluster, say the  $hi$ th PSU,  $m_{hi}$  ultimate units are selected from the  $M_{hi}$  ultimate units according to certain sampling plan.

### 2.2 Confidence Intervals for Proportions in Complex Surveys

Suppose the total number of sampled ultimate units is  $m = \sum_{h=1}^H \sum_{i=1}^{n_h} m_{hi}$ . The set of all ultimate units in the sample is denoted as  $S = \{(hij) : h = 1, \dots, H; i = 1, \dots, n_h; j = 1, \dots, m_{hi}\}$ . An approximately unbiased ratio estimator for the proportion is given by

$$\hat{p} = \frac{\sum_{(hij) \in S} w_{hij} y_{hij}}{\sum_{(hij) \in S} w_{hij}} = \sum_{(hij) \in S} w_{hij}^* y_{hij}, \quad (1)$$

where  $w_{hij}^* = \frac{w_{hij}}{\sum_{(hij) \in S} w_{hij}}$ . The property that the estimator is consistent is a fundamental one. It requires that the estimator

converges in probability to the true value of the parameter as the sample size goes to infinite. Several authors extended the notion of consistency to complex surveys under sufficient conditions (Krewski and Rao, 1981; Isakl and Fuller, 1982). Most of the proposed intervals in complex surveys, including those introduced here, amount to replacing the  $n$  in the IID interval by the estimated effective sample size,

$$m_e = m / deff(\hat{p}) = m / [v(\hat{p}) / v_{SRS}(\hat{p})] = \hat{p}(1 - \hat{p}) / v(\hat{p}), \quad (2)$$

where  $deff(\hat{p}) = v(\hat{p}) / v_{SRS}(\hat{p})$  indicates the estimated design effect of  $\hat{p}$  and also replacing degrees of freedom for the IID interval with the effective degrees of freedom ( $df_e$ ) to account for the variability of the variance estimator due to the complex design (Rust, 1986),

$$df_e = 2[E[v(\hat{p})]]^2 / Var[v(\hat{p})] \quad (3)$$

Korn and Graubard (1998) refer to  $d = n - H$  as the *nominal* degrees of freedom under certain assumptions.

- **Interval #1: Standard (SD) Interval**

In complex surveys, the approximate  $100(1 - \alpha)\%$  standard interval for  $p$  can be easily obtained by replacing  $n$  for the standard interval in the IID case by the effective sample size and replacing  $z_{\alpha/2}$  in the interval by  $\kappa^* = t_d(1 - \alpha/2)\sqrt{m/m_e}$ , which gives  $\hat{p} \pm t_d(1 - \alpha/2)[v(\hat{p})]^{1/2}$ .

- **Interval #2: Wilson (WS) Interval**

Design-consistent estimator  $\hat{p}$  converges to  $p$  in probability. According to Central Limit Theorem and Slutsky's Theorem, it can then be shown that

$$\frac{\sqrt{m_e}(\hat{p} - p)}{\sqrt{p(1-p)}} = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{p(1-p)}} \left[ \frac{\sqrt{m_e}(\hat{p} - p)}{\sqrt{\hat{p}(1-\hat{p})}} \right] \xrightarrow{d} N(0,1),$$

which yields the lower and upper limits for the Wilson interval,

$$[p_{WS}^L, p_{WS}^U] = \left[ \frac{\hat{p} + t_d^2(1 - \alpha/2)/(2m_e)}{1 + t_d^2(1 - \alpha/2)/(m_e)} - \frac{t_d(1 - \alpha/2)m_e^{1/2}}{m_e + t_d(1 - \alpha/2)^2} \sqrt{\hat{p}(1 - \hat{p}) + t_d^2(1 - \alpha/2)/(4m_e)}, \right. \\ \left. \frac{\hat{p} + t_d^2(1 - \alpha/2)/(2m_e)}{1 + t_d^2(1 - \alpha/2)/(m_e)} + \frac{t_d(1 - \alpha/2)m_e^{1/2}}{m_e + t_d(1 - \alpha/2)^2} \sqrt{\hat{p}(1 - \hat{p}) + t_d^2(1 - \alpha/2)/(4m_e)} \right] \quad (4)$$

- **Interval #3: Logit (LG) Interval**

To stabilize the variance, logit transformation,  $\hat{\lambda} = \log[\hat{p}/(1 - \hat{p})]$  can be applied to the standard interval, which yields

$$[p_{LG}^L, p_{LG}^U] = \left[ \frac{e^{\lambda_{LG}^L}}{1 + e^{\lambda_{LG}^L}}, \frac{e^{\lambda_{LG}^U}}{1 + e^{\lambda_{LG}^U}} \right], \text{ where}$$

$$[\lambda_{LG}^L, \lambda_{LG}^U] = \left\{ \log[\hat{p}/(1 - \hat{p})] - t_d(1 - \alpha/2)[m_e \hat{p}(1 - \hat{p})]^{-1/2}, \log[\hat{p}/(1 - \hat{p})] + t_d(1 - \alpha/2)[m_e \hat{p}(1 - \hat{p})]^{-1/2} \right\}. \quad (5)$$

- **Interval #4: Arcsine (AR) Interval**

Arcsine transformation is another variance stabilization technique by considering  $\hat{\delta} = \arcsin(\hat{p}^{1/2})$ , which gives the interval limits,

$$[p_{AR}^L, p_{AR}^U] = \left\{ \sin^2[\arcsin(\hat{p}^{1/2}) - t_d(1 - \alpha/2)m_e^{-1/2}/2], \sin^2[\arcsin(\hat{p}^{1/2}) + t_d(1 - \alpha/2)m_e^{-1/2}/2] \right\}. \quad (6)$$

- **Interval #5: Clopper-Pearson (KG) Interval**

A simple modification of the Clopper-Pearson interval to be used in complex surveys is proposed by Korn and Graubard (1998). The degrees-of-freedom adjusted effective sample size is defined by

$$m_{df}^* = \frac{\hat{p}(1 - \hat{p})}{\text{var}(\hat{p})} \left\{ \frac{t_{n-1}(1 - \alpha/2)}{t_d(1 - \alpha/2)} \right\}^2. \quad (7)$$

Korn and Graubard (1998) gave some heuristic justification for the second fraction in (7) by considering the issue that  $v(\hat{p})$  will be typically more variable than a variance estimator that would be used for simple random sampling. Both  $m_e$  and  $m_{df}^*$  are set equal to  $m$  when  $\hat{p} = 0$ . The modified Clopper-Pearson interval (KG) is

$$[p_{KG}^L, p_{KG}^U] = \left[ \frac{v_1 F_{v_1, v_2}(\alpha/2)}{v_2 + v_1 F_{v_1, v_2}(\alpha/2)}, \frac{v_3 F_{v_3, v_4}(\alpha/2)}{v_4 + v_3 F_{v_3, v_4}(\alpha/2)} \right], \quad (8)$$

where  $v_1 = 2m_{df}^* \hat{p}$ ,  $v_2 = 2(m_{df}^* - m_{df}^* \hat{p} + 1)$ ,  $v_3 = 2(m_{df}^* \hat{p} + 1)$ ,  $v_4 = 2(m_{df}^* - m_{df}^* \hat{p})$  and  $F_{v_i, v_j}(\gamma)$  is the  $\gamma$  th quantile of an F distribution.

- **Interval #6: Breeze (BZ) Interval**

Breeze (1990) adapted the Poisson confidence interval in the IID case to incorporate the complex survey structures, which gives,

$$[p_{BZ}^L, p_{BZ}^U] = \left[ \frac{1}{2m_e} \chi_{v_1}^2(\alpha/2), \frac{1}{2m_e} \chi_{v_2}^2(1 - \alpha/2) \right], \quad (9)$$

where  $v_1 = 2m_e \hat{p}$ ,  $v_2 = 2(m_e \hat{p} + 1)$  and  $\chi_v^2(\gamma)$  is the  $\gamma$  th quantile of  $\chi^2$  distribution.

- **Interval #7: Likelihood Ratio (LR) Interval**

For the IID case, let

$$Y_n(p) = 2n \left\{ \hat{p} [\log \hat{p} - \log p] + (1 - \hat{p}) [\log(1 - \hat{p}) - \log(1 - p)] \right\}, \quad (10)$$

then the LR interval is given by

$$\left\{ p : Y_n(p) \leq z_{\alpha/2}^2 \right\}. \quad (11)$$

one can justify the resulting interval without resorting to maximum likelihood theory. This is important if we wish to extend to surveys, since there it would be difficult to justify a likelihood interpretation. To do so, note that by Taylor series expansion

$$\log \hat{p} \approx \log p + (\hat{p} - p)(1/p) - (\hat{p} - p)^2(1/2p^2), \quad (12)$$

which implies  $Y_n(p)$  is approximately equals to  $\frac{n(\hat{p} - p)^2}{p(1-p)} \xrightarrow{d} N(0,1)$  by the Central Limit Theorem and Slutsky's Theorem. Therefore, we could, as with other intervals, replace  $n$  by the effective sample size  $m_e$  for the LR interval in the IID case to derive the LR interval in complex surveys. In addition, we can replace  $z_{\alpha/2}$  by  $t_d(1 - \alpha/2)$  with  $d$  nominal degrees of freedom, if desired.

### 3. SIMULATION STUDIES

In this section, we perform some preliminary simulations to compare the intervals for proportions. For simplicity, we consider two-stage *pps* with-replacement sampling design.

#### 3.1 General Description of Simulation

One can generate a finite population of  $y$ 's taking values 0 and 1 based on a latent population of continuous  $x$ 's, which can be treated as an unobservable variable that has a relationship to clusters and strata. This can be done by denoting the desired overall proportion of  $y$ 's equal to 1 as  $p$  and then letting  $y_{hij} = 1$  if  $x_{hij} \leq \zeta_p$  and 0 otherwise, where  $\zeta_p$  is the  $p$ th quantile of the latent variable  $x_{hij}$ . This then ensures the proportion of 1's differs between strata and clusters within strata in the population in some reasonable way. We constrain ourselves to  $H = 15$  and  $N_h$  to be a sequence of values ranging from 29 to 47 and  $n_h = 2$ , so that  $N_h$  are of manageable size and the first stage sampling fraction ( $f_h = n_h / N_h$ ) are small. Next, the total number of the ultimate units in the  $i$ th PSU within the  $h$ th stratum,  $M_{hi}$  are generated from a Gamma distribution ranging from 50 to 2338 and the sample size within  $hi$ th PSU is  $m_{hi} = m_0 = 50$ . The choice of parameters for the Gamma distribution keeps the PSU size manageable and the within PSU sampling fraction reasonably small. The variances of  $x_{hij}$  are generated as  $v_h^2 \stackrel{IID}{\sim} N(\mu_v, \sigma_v^2)$  with the choice of the parameters arbitrary to ensure a reasonable range of  $v_h$  from 15.7 to 22.7. Once the values of  $N_h$ ,  $M_{hi}$  and  $v_h$  were generated, they remained fixed and were used to generate finite populations of the latent variable,  $x$ , with differing characteristics. To do this, the ultimate population units  $x$  were defined as

$$x_{hij} = \mu_{hi} + \varepsilon_{hij} \quad (h = 1, \dots, H; i = 1, \dots, N_h; j = 1, \dots, M_{hi}). \quad (13)$$

The average of the  $x_{hij}$  in cluster  $hi$  is generated by  $\mu_{hi} \stackrel{IID}{\sim} N(\mu_h, \sigma_h^2)$ , where  $\sigma_h^2 = \rho_h v_h^2$  with  $\rho_h$  denotes the intra-cluster correlation. The variation among SSU's within the  $hi$ th PSU is incorporated into the model via  $\varepsilon_{hij} \stackrel{IID}{\sim} N(0, \sigma_{hi}^2)$ , where  $\sigma_{hi}^2$  is generated from gamma distribution with mean  $(1 - \rho_h)v_h^2$  and variance  $v_h^2/10$ . Two-stage sampling with probability proportional to the size (*pps*) at the PSU level and SRS at SSU level is used throughout the simulation study.

#### 3.2 Point Estimation for Proportion

The estimated proportion is given by  $\hat{p} = [\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_0} w_{hij} y_{hij}] / M$ , where  $w_{hij} = M_{hi} / (n_h z_{hi} m_{hi})$ . Result for *pps* with  $z_{hi} = M_{hi} / M_h$  follows as a special case. If the sample sizes within all the PSU's are the same  $m_{hi} = m_0$ , the sampling weight  $w_{hij} = M_h / (nm_0)$ . In this case, the variance estimator takes the simpler form

$$v(\hat{p}) = \sum_{h=1}^H \frac{W_h^2}{n_h(n_h - 1)} \sum_{i=1}^{n_h} (\bar{y}_{hi} - \bar{y}_h)^2, \quad (14)$$

where  $W_h = M_h / M$ ,  $\bar{y}_{hi} = \sum_{j=1}^{m_0} y_{hij} / m_0$  is the sample mean for  $hi$  th PSU and  $\bar{y}_h = \sum_{i=1}^{n_h} \bar{y}_{hi} / n_h$  is the sample mean for stratum  $h$ .

In the simulation studies, 10,000 independent samples were taken under a stratified two-stage *pps* with-replacement sampling design. To compare the intervals in a broad range of population structures, setting I, II, and III are generated by setting  $\rho_h = 0.05, 0.08$  and  $0.2$ . It can be shown that the variation among clusters gets larger as  $\rho_h$  increases. Figure 1 shows how the variation among clusters varies as the intra-cluster correlation coefficient increases.

Table 1 summarizes the simulation results for the confidence intervals with  $t_d(1-\alpha/2)$  as critical value ( $d$ : nominal degrees of freedom) for SETTING I, II and III. With regards to the same proportion, the coverage probabilities for all the intervals tend to become lower than the nominal level as the design effect gets larger. The KG interval can maintain the nominal level better than the other intervals for nearly all the settings, but when the cluster means tend to be homogeneous (approximately equivalent to SRS), it is more conservative compared to the other intervals, which is consistent with the discussion in the IID case. In this case, the WS and LR intervals are preferred. The BZ interval performs well for small proportions in SETTING I, but tends to become conservative as the proportion gets larger and does not perform as well as the KG interval in SETTING II and III. Poor performance of the BZ interval when  $p$  is not small is to be expected as a Poisson approximation will be poor in these cases. In terms of interval balance, the WS and the LR intervals perform relatively better than the other intervals for SETTING I, but the KG interval performs relatively better than the other intervals for SETTING II and III. The left error rates for the WS and the LG intervals are closer to the nominal level, 2.5%, than the other intervals. In terms of interval length, the BZ and the KG intervals are relatively longer than the other intervals. The LG interval is slightly longer than the LR, the AR and the WS intervals. The AR interval tends to be less balanced compared to the KG, the LR and the WS intervals. In general, as the clusters get more heterogeneous, the coverage probabilities for the intervals tend to fall below the nominal level and the intervals are less balanced for the small proportions.

#### 4. APPLICATION TO NHANES

The National Health and Nutrition Examination Surveys (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States, combining household interviews and medical examinations.

In our simulation studies, the sampling design will be simplified into a stratified two-stage design, where two PSU's are sampled *pps* within stratum and SSU's (persons) are sampled SRS within obtained PSU's. There will be three race/ethnicity domains under consideration and the sampling design will be structured so that the design is approximately self-weighting within domain, with approximately equal workload for each PSU, where workload means the sample size per PSU. This is a simplified version of NHANES where they use a stratified four-stage sampling design, with age, sex, and race/ethnicity domains with *pps* sampling at the first three stages and SRS at the final stage. The primary reason for achieving self-weighting within domain is to stabilize variance estimators, since the variation of weights can increase the variance. The purpose for approximately equal workload for each PSU is to make sure the field work and respondent burden is kept under control.

In recent years, RTI (Research Triangle Institute) has used a composite size measure procedure for achieving self-weighting samples and equal workload per PSU for multiple domains in multistage designs, which was developed by Folsom in early 1978. We will utilize this procedure to construct the sampling design. In our case, we will ignore strata size and create the self-weighting within stratum. This will mean the sampling weights will depend on stratum. In addition, if one creates estimates over multiple domains the design will not be self-weighting within stratum. By then making the true proportions depend on domain, which they typically do, and the domain sizes depend upon strata, the weights should be correlated with the proportion in a reasonable way. This is in contrast to the previous sections, where the weights and proportions were essentially unrelated.

The above method for creating a self-weighting design is an extension of the procedure proposed by Folsom, Potter and Williams (1987). The procedure provides an efficient approach for obtaining domain estimates with controlled precision.

The rare domains tend to be oversampled by selecting PSU's with probabilities proportional to a specifically structured composite size measure to gain roughly the same precision over multiple domains for the stratified multi-stage design. Self-weighting samples and equal workload can theoretically be obtained for each domain. Detailed procedure for generating and analyzing the domains (subpopulations) based on the sample provided by NCHS are provided in Feng (2006). We will not elaborate the results here, which are essentially consistent with results from the preliminary simulation studies.

## 5. CONFIDENCE INTERVALS FOR QUANTILES

Another related issue is the use of Woodruff's method for obtaining confidence intervals for small and large quantiles based upon inverting the standard interval for the distribution function. Sitter and Wu (2001) show that despite the poor performance of the standard interval for small  $p$ , Woodruff's method still performs well for the associated quantiles. They also argue that improving the confidence interval for  $p$  negatively impacts intervals for small and large quantiles using Woodruff's method. Thus, one must take care in deciding what to advise to end-users depending upon their goals. Sitter and Wu (2001) perform a very limited simulation study and do not consider weighting, multi-stage sampling, design effects, effective sample size, etc. Thus, we extend our investigation to consideration of Woodruff's method.

In the stratified multistage survey design, the sample estimate of the cumulative distribution function (CDF) at some fixed value  $x$  is

$$\hat{F}(x) = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} I\{x_{hij} \leq x\} / \hat{M}, \quad (15)$$

where  $w_{hij}$  denotes the sample weight of the unit  $hij$  and  $\hat{M}$  is the estimated total number of population units. Thus, for fixed  $x$ , the estimated cumulative distribution function,  $\hat{F}(x)$  is the estimated binomial proportion  $\hat{p}$ . Thus, the confidence intervals for  $\hat{F}(x)$  are exactly the confidence intervals for the  $p$  discussed in Section 3. It can be shown that the sample estimate of the  $p$  th quantile is given by,

$$\hat{\zeta}_p = \hat{F}^{-1}(p) = \inf_{(hij) \in S} \{x_{hij} : \hat{F}^{-1}(x_{hij}) \geq p\}. \quad (16)$$

The sample estimator of  $\hat{\zeta}_p$  can be found by first sorting the sampled units in order and then cumulating their normalized sample weights  $w_{hij} / \hat{M}$  until  $p$  is first exceeded.

### 5.1 Woodruff Interval for Quantiles

Woodruff interval (Woodruff, 1952) for quantiles is obtained by inverting the standard interval for the distribution function and replacing the unknown  $\zeta_p$  with an estimator  $\hat{\zeta}_p$  in the variance estimator to yield the approximately  $100(1 - \alpha)\%$  confidence interval for  $\hat{\zeta}_p$ ,

$$\hat{F}^{-1}\left[p - t_d(1 - \alpha/2)\sqrt{v[\hat{F}(\hat{\zeta}_p)]}\right] \leq \zeta_p \leq \hat{F}^{-1}\left[p + t_d(1 - \alpha/2)\sqrt{v[\hat{F}(\hat{\zeta}_p)]}\right] \quad (17)$$

Francisco and Fuller (1991) have justified (17) under certain assumptions, such as  $\hat{F}(\cdot)$  and  $\hat{F}^{-1}(\cdot)$  are good approximations to  $F(\cdot)$  and  $F^{-1}(\cdot)$ .

Sitter and Wu (2001) use some heuristic arguments together with limited simulations to show that Woodruff interval can perform quite well even in the case of small or large values of  $p$ , even though the standard interval for  $p$  it is based on does not. In fact, they argue that the poor performance of the standard interval for  $p$  is in some sense in the opposite direction to the impact of replacing  $\zeta_p$  by  $\hat{\zeta}_p$  above and "two wrongs, make a right" in this case.

## 5.2 Modified Woodruff Confidence Intervals for Quantiles

We extend the idea of Woodruff interval, so as to apply it to the confidence intervals mentioned in section 2. Assume that  $\hat{F}^L(\hat{\zeta}_p)$  and  $\hat{F}^U(\hat{\zeta}_p)$  are lower and upper limits for  $F(\zeta_p)$ . Then,

$$P[\hat{F}^L(\hat{\zeta}_p) \leq F(\zeta_p) \leq \hat{F}^U(\hat{\zeta}_p)] \approx 1 - \alpha. \quad (18)$$

By replacing  $\zeta_p$  with  $\hat{\zeta}_p$ , we have a  $1 - \alpha$  level confidence interval for  $\zeta_p$ ,

$$\hat{F}^{-1}[\hat{F}^L(\hat{\zeta}_p)] \leq \zeta_p \leq \hat{F}^{-1}[\hat{F}^U(\hat{\zeta}_p)]. \quad (19)$$

Alternatively, the  $1 - \alpha$  confidence interval for quantile  $\zeta_p$  can be generally expressed as

$$[\hat{\zeta}_p^L, \hat{\zeta}_p^U] = \left[ \inf\{x_{hij} : \hat{F}(x_{hij}) \geq \hat{F}^L(\hat{\zeta}_p)\}, \inf\{x_{hij} : \hat{F}(x_{hij}) \geq \hat{F}^U(\hat{\zeta}_p)\} \right]. \quad (20)$$

Table 2 reports the simulation results for the Woodruff ( $W_{SD}$ ) interval and the *modified* Woodruff intervals for quantile  $\zeta_{0.005}$ ,  $\zeta_{0.05}$  and  $\zeta_{0.1}$  with  $t_d(1 - \alpha/2)$  as critical value in finding  $\hat{F}^L(\hat{\zeta}_p)$  and  $\hat{F}^U(\hat{\zeta}_p)$  ( $d$ : nominal degrees of freedom) in the same SETTINGS I, II and III as in section 3.1. It is evident that not only does the Woodruff interval maintain the nominal level better, but also it is much more balanced than the *modified* Woodruff intervals. The lower error rates for all the *modified* Woodruff intervals for the small quantiles tend to be much higher than the nominal error rate, 2.5% and the upper error rates tend to be lower than 2.5%. The coverage probability for the  $W_{KG}$  interval tends to be close to the nominal level; however, the interval is less balanced than the  $W_{SD}$  interval and the coverage probability for the  $W_{BZ}$  interval is severely below the nominal level for the small quantiles. In addition, as the clusters become more dispersed, the  $W_{SD}$  interval for the small quantiles keeps its good performance surprisingly better than the standard interval for  $F(\zeta_p)$  for the small proportions. Besides, as the number of strata increases  $H=30, 60$  and  $90$  and in turn increase the sample size from 3000 to 6000 to 9000, both the standard interval for  $F(\zeta_p)$  and the  $W_{SD}$  interval for  $\zeta_p$  improve, verifying their asymptotic properties. In fact, the appealing performance of the  $W_{SD}$  interval for the small quantiles is partially due to the inflated length of the interval (Sitter and Wu, 2001). Thus, using a "better" interval for  $p$  in combination with Woodruff's method is not a good idea. It is much better to use the Woodruff interval based on inverting the standard interval for the distribution function.

## 6. CONCLUDING REMARKS

In this paper, we reviewed one of the most basic and methodologically important problems in statistical practice, namely, the confidence intervals for small proportions in the complex surveys in terms of coverage probability, interval length, and interval balance. We extended the WS and the AR intervals in the IID case to the complex surveys theoretically and we also extended the LR interval in the IID case to complex surveys based on some heuristic justification. Then, we investigate and compare all the alternative intervals including the intervals suggested in the literature under various complex survey settings induced by varying the intra-cluster correlation coefficient of the artificial population. For the small proportions, the standard interval performs badly with coverage probability much lower than the nominal level and the two tail errors severely unbalanced, since the sampling distribution for the estimated proportion is not symmetric for small  $p$  with moderate sample size. By contrast, the other intervals perform uniformly better than the standard interval in terms of coverage probability and interval balance for small or large proportions. However, as the clusters become more heterogeneous with respect to the same proportion, the coverage probabilities for all the intervals tend to be below the nominal level and the intervals are less balanced with the interval length inflated. In general, the KG interval maintains the nominal level the best, especially when the clusters tend to be quite heterogeneous and it is a little bit conservative when the clusters are homogeneous. The WS interval is less conservative than the KG interval and it performs quite well in

terms of interval balance with the lower error rate close to 2.5% for small proportions. Therefore, in the case that the clusters are homogeneous, either the WS or the LR interval can be recommended. The choice depends on the user's personal preferences. The BZ interval can be suggested when the proportions are quite small with large number of sampled clusters, since it does not account for the degrees of freedom for the variance estimator.

In the application to NHANES, we utilized a composite size measure procedure for achieving approximately self-weighting samples for multiple domains and approximately equal workload per PSU to generate the subpopulations for the stratified two stage design. Then, we compare the alternative intervals for each domain proportion and the overall proportion. In general, the results are consistent with the situations without considering domain.

Under the two-stage sampling design considered in the simulation studies, the Woodruff interval by inverting the SD interval for the cumulative distribution function performs better than the *modified* Woodruff intervals, especially for the small proportions. As the clusters become more dispersed with respect to the same quantile, the coverage probabilities for all the intervals tend to be below the nominal level and the intervals are less balanced with the inflated interval length. However, the Woodruff interval works surprisingly well for small quantiles even when the design effect tends to be quite large. There are some other methods for constructing confidence intervals for quantiles, such as the Francisco-Fuller (FF) or "test inversion" procedure (Francisco and Fuller, 1991), which is closely related to Woodruff interval. The Woodruff interval uses only the estimated variance at the estimated quantile, while the FF interval uses information about the variance at points close to the estimated quantile, so it is computationally far more intensive. The literature tends to suggest FF will outperform Woodruff. However, the Woodruff procedure is simple and faster to compute than the other methods.

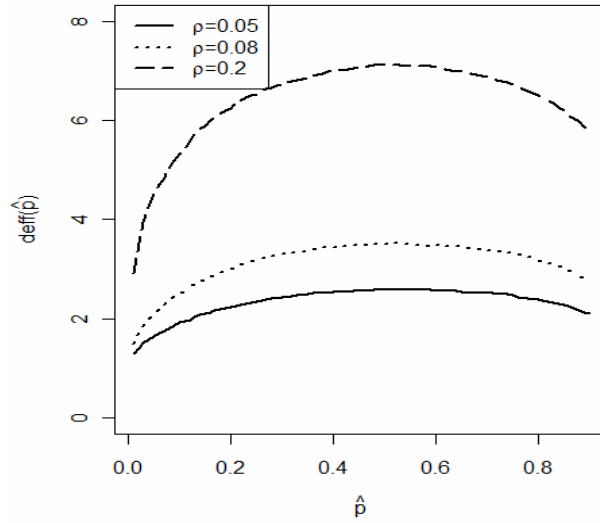
## ACKNOWLEDGEMENTS

We would like to thank Westat, the Mathematics of Information Technology and Complex Systems (MITACS) and the National Program for Complex Data Structures (NPCDS) for the support. I thank Professor Changbao Wu for helpful comments and suggestions.

## REFERENCES

- Breeze, E. (1990). "General Household Survey: Report on Sampling Error". London: Her Majesty's Stationery Office (Office of Population Censuses and Surveys).
- 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.
- Cochran, W.G. (1997). *Sampling Techniques*. Third Edition. New York: Wiley.
- Feng (2006). "On Confidence Intervals for Proportions with Focus on the U.S. National Health and Nutrition Examination Surveys". *M.Sc. thesis*, Simon Fraser University.
- Francisco, C.A. and Fuller, W.A.(1991). "Quantile Estimation with a Complex Survey Design", *the Annals of Statistics*, **2**, 568-571.
- Hansen, M.H., and Hurwitz, W.N. (1943). "On the theory of sampling from finite populations", *Ann. Math. Stat.*, **14**, 333-362.
- Isaki, C.T. and Fuller, W.A. (1982). "Survey design under a regression superpopulation model", *Journal of the American Statistical Association*, **77**, 89-96.
- 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.

- Korn, L. Edward and Graubard, I. Barry (1998). "Confidence Intervals for Proportions With Small Expected Number of Positive Counts Estimated from Survey Data", *Survey Methodology*, **24**, 193-201.
- R.E. Folsom, F.J. Potter, and S.R. Williams (1987). "Note on A Composite Size Measure for Self-Weighting Samples in Multiple Domains", Research Triangle Institute.
- Rust, K.F. (1986). "Efficient Formation of Replicates for Replicated Variance Estimation", *Proceedings of the American Statistical Association Section on Survey Research Methods*, pp. 81-87.
- Sitter, R.R. and Wu, C. (2001). "A Note on Woodruff Confidence Intervals for Quantiles", *Statistics and Probability Letter*, **52**, 353-358.
- Wu, C. (1999). "The Effective Use of Complete Auxiliary Information from Survey Data". *Ph.D. Dissertation*, Simon Fraser University.
- Wilson, E.B. (1927). "Probable Inference, the Law of Succession, and Statistical Inference", *Journal of the American Statistical Association*, **22**, 209-212.
- Woodruff, R.S. (1952). "Confidence Intervals for Median and Other Position Measures", *Journal of American Statistical Association*, **47**, 635-646.



**Figure 1** Illustration of the estimated design effect,  $deff(\hat{p})$ , varies as the estimated proportions,  $\hat{p}$ , for different intra-cluster correlation coefficient

**Table 1 Coverage probability ( $CP$  %), lower ( $L$  %) and upper ( $U$  %) tail error rates and average interval length ( $AL$  %) for the 95% confidence interval for binomial proportion ( $p = 0.005$ ,  $p = 0.05$  and  $p = 0.1$ ) with  $t_d(0.975)$  as critical value ( $d$  : nominal degrees of freedom) for SETTING I, II and III. Note that all values in the table are in percentage.**

	$p = 0.005$				$p = 0.05$				$p = 0.1$			
	$CP$	$L$	$U$	$AL$	$CP$	$L$	$U$	$AL$	$CP$	$L$	$U$	$AL$
	<b>SETTING I <math>\rho_h = 0.05</math></b>											
SD	92.51	0.40	7.09	0.85	94.87	1.24	2.89	3.04	94.82	1.52	3.66	4.50
LG	97.36	2.47	0.17	0.98	95.78	2.02	2.20	3.08	95.14	2.12	2.74	4.53
AR	95.11	1.14	3.75	0.85	95.55	1.53	2.92	3.04	94.97	1.82	3.21	4.50
WS	96.51	2.64	0.85	0.94	95.67	2.07	2.26	3.07	95.11	2.13	2.76	4.51
LR	95.86	1.51	2.63	0.87	95.60	1.68	2.72	3.05	94.98	1.94	3.08	4.50
KG	97.42	1.38	1.20	0.96	95.60	1.66	2.53	3.15	95.29	1.85	2.86	4.63
BZ	96.25	1.84	1.91	0.89	95.14	1.99	2.87	2.99	94.95	2.00	3.05	4.51
	<b>SETTING II <math>\rho_h = 0.08</math></b>											
SD	91.67	0.33	8.00	0.88	93.96	1.34	4.70	3.38	94.33	1.44	4.23	5.12
LG	97.51	2.30	0.19	1.02	95.13	2.15	2.72	3.44	95.05	2.04	2.91	5.15
AR	94.54	1.12	4.34	0.88	94.65	1.73	3.62	3.38	94.81	1.69	3.50	5.11
WS	96.51	2.45	1.04	0.98	95.04	2.18	2.78	3.42	95.00	2.04	2.96	5.13
LR	95.43	1.44	3.13	0.90	94.70	1.98	3.32	3.39	94.78	1.83	3.39	5.11
KG	97.15	1.34	1.51	1.00	95.28	1.76	2.96	3.54	95.33	1.70	2.97	5.30
BZ	95.99	1.66	2.35	0.93	94.41	2.13	3.46	3.34	94.75	1.91	3.34	5.14
	<b>SETTING III <math>\rho_h = 0.2</math></b>											
SD	85.66	0.48	13.86	1.06	90.69	1.07	8.24	4.78	92.61	1.12	6.27	7.29
LG	97.33	2.01	0.66	1.32	93.10	2.00	4.90	4.95	93.93	1.92	4.15	7.39
AR	90.44	0.97	8.59	1.06	91.85	1.39	6.76	4.77	93.32	1.42	5.26	7.27
WS	94.91	2.32	2.77	1.22	92.86	2.07	5.07	4.89	93.78	1.97	4.25	7.28
LR	92.25	1.22	6.53	1.10	92.22	1.60	6.18	4.79	93.40	1.63	4.97	7.28
KG	95.29	1.06	3.65	1.26	93.60	1.37	5.03	5.13	94.46	1.39	4.15	7.68
BZ	93.46	1.31	5.23	1.16	92.08	1.70	6.22	4.85	93.52	1.61	4.87	7.47

**Table 2 Coverage probability ( $CP$  %), lower ( $L$  %) and upper ( $U$  %) tail error rates and average interval length ( $AL$  %) for the 95% Woodruff and modified Woodruff confidence intervals for quantiles ( $\zeta_{0.005}$ ,  $\zeta_{0.05}$  and  $\zeta_{0.1}$ ) with  $t_d(0.975)$  as critical value ( $d$  : nominal degrees of freedom) for SETTING I, II and III. Note that all values in the table are in percentage.**

	$\zeta_{0.005}$				$\zeta_{0.05}$				$\zeta_{0.1}$			
	<i>CP</i>	<i>L</i>	<i>U</i>	<i>AL</i>	<i>CP</i>	<i>L</i>	<i>U</i>	<i>AL</i>	<i>CP</i>	<i>L</i>	<i>U</i>	<i>AL</i>
	<b>SETTING I <math>\rho_h = 0.05</math></b>											
$W_{SD}$	97.53	1.17	1.30	20.92	96.23	2.18	1.59	6.07	95.59	2.65	1.76	5.23
$W_{LG}$	94.63	5.22	0.15	13.12	95.31	3.80	0.89	5.89	95.38	3.54	1.08	5.16
$W_{AR}$	96.63	2.89	0.48	14.32	95.88	2.94	1.18	5.92	95.51	3.11	1.38	5.17
$W_{WS}$	94.38	5.49	0.16	12.63	95.28	3.82	0.90	5.86	95.36	3.55	1.09	5.14
$W_{LR}$	96.06	3.64	0.30	13.47	95.61	3.29	1.10	5.89	95.46	3.23	1.31	5.16
$W_{KG}$	97.09	2.76	0.15	15.55	96.44	2.79	0.17	10.85	96.02	2.97	1.01	5.40
$W_{BZ}$	92.92	6.36	0.72	11.53	91.50	6.16	2.34	4.88	91.79	5.51	2.70	4.24
	<b>SETTING II <math>\rho_h = 0.08</math></b>											
$W_{SD}$	97.47	1.37	1.16	22.86	96.05	2.48	1.47	6.92	95.69	2.68	1.63	6.03
$W_{LG}$	94.33	5.58	0.09	13.81	95.25	4.02	0.73	6.64	95.38	3.74	0.88	5.92
$W_{AR}$	96.29	3.25	0.46	15.12	95.72	3.21	1.07	6.69	95.62	3.21	1.17	5.94
$W_{WS}$	94.10	5.78	0.12	13.23	95.19	4.08	0.73	6.60	95.35	3.76	0.89	5.90
$W_{LR}$	95.71	4.02	0.27	14.19	95.66	3.43	0.91	6.65	95.55	3.38	1.07	5.92
$W_{KG}$	96.79	3.13	0.08	16.41	96.32	3.05	0.63	7.09	96.10	3.08	0.82	6.23
$W_{BZ}$	92.88	6.48	0.64	12.23	91.35	6.68	1.97	5.54	91.74	6.04	2.22	5.02
	<b>SETTING III <math>\rho_h = 0.2</math></b>											
$W_{SD}$	96.76	2.36	0.88	33.81	95.18	3.70	1.12	11.75	95.75	3.11	1.14	9.07
$W_{LG}$	90.53	9.25	0.22	17.83	92.75	6.71	0.54	9.83	94.15	5.16	0.69	8.62
$W_{AR}$	93.83	5.77	0.40	18.22	94.07	5.14	0.79	10.03	95.02	4.14	0.84	8.71
$W_{WS}$	89.99	9.74	0.27	16.51	92.62	6.82	0.56	9.67	94.04	5.24	0.72	8.54
$W_{LR}$	92.46	7.19	0.35	17.74	93.73	5.60	0.67	9.86	94.69	4.51	0.80	8.64
$W_{KG}$	94.19	5.59	0.22	20.47	94.66	4.84	0.50	10.85	95.53	3.80	0.67	9.32
$W_{BZ}$	89.17	10.4	0.42	15.97	89.54	9.35	1.11	8.40	91.12	7.53	1.35	7.44