



# Consequences of analysing complex survey data using inappropriate analysis and software computing packages

S-T Wang, M-L Yu and L-Y Lin

Department of Public Health, National Cheng Kung University Medical Center, 138 Sheng Li Road, Tainan, Taiwan

In the analysis of complex survey data such as stratified multi-stage cluster samples, ignoring the design effects such as clustering and stratification usually will lead to erroneous conclusions. In this paper, we will demonstrate the consequences in the estimation of means and proportions by two examples from a stratified two-stage cluster sample. A brief review of methodology will be presented, and some suggestions on computational issues will be provided.

**Keywords:** complex survey data; clustering; stratification

## Introduction

Consider a simple study where a sample of size one is drawn from a population containing four elementary units. Suppose the values of a population characteristics  $Y$  for these units are 1, 2, 5, and 10. With equal probability sampling, an unbiased estimator of the mean of  $Y$  is  $\bar{Y}$ , the outcome of the drawn unit. With unequal probability sampling,  $\bar{Y}$  is not an unbiased estimator of the mean. For instance,  $E(\bar{Y}) = 7.8$  when the sampling probabilities are 0.10, 0.10, 0.10, and 0.70, respectively, which is larger than the actual population average, 4.5. In other words, the sampling probability associated with each drawn unit is an important consideration in the estimation of population parameters.

In sample surveys of human populations, it is often not feasible—perhaps not even possible—to compile sampling frames that list all the elementary units for the entire population. On the other hand, sampling frames can often be constructed that identify groups or clusters of elementary units without listing explicitly the elementary units. These sample designs are known as cluster samples and are widely used in practice. Typical examples are household and school surveys (Table 1). Stratification of the population with respect to another variable (or variables) that is thought to be associated with the variable of interest is also widely applied for administrative reasons, and statistical efficiency considerations.

It is a well-known fact that in the analysis of data collected by means of a survey of complex design, ignoring the design effects such as stratification and/or clustering usually leads to seriously misleading results.<sup>1</sup> In spite of the fact, it is not uncommon to see that such survey data be analysed in many published results using inappropriate statistical computing packages such as SAS, SPSS, and BMDP. In this paper we will discuss this problem in the estimation of population totals or means, and the comparison of subpopulation totals or means. First, a brief review of methodology for estimation in both simple random sampling and complex sampling design will be presented. And then two examples from a survey including stratification and clustering in its design will be used to

demonstrate our points. Finally, some suggestions on computational issues will be provided.

## Methods

To illustrate our points, a sample survey using stratified two-stage cluster sampling scheme is used. This sampling scheme includes, in its first stage, the sampling of some fixed number or proportion of primary sampling units (PSU), essentially clusters of elementary units, from each stratum using either simple random sampling (SRS) or probabilities-proportional-to-sizes sampling with replacement (PPS).<sup>2</sup> In its second stage, some fixed number or proportion of elementary units are sampled from each selected PSU using primarily SRS. With this survey design, a population total, say  $Y$ , can be expressed as

$$Y = \sum_{h=1}^L \sum_{i=1}^{M_h} \sum_{j=1}^{N_{hi}} y_{hij} \quad (1)$$

where  $L$  is the number of strata,  $M_h$  is the number of PSU in stratum- $h$ ,  $N_{hi}$  is the number of elementary units in the PSU- $i$  of stratum  $h$ , and  $y_{hij}$  is the value for the  $j$ th elementary unit. Let  $m_h$  denote the number of sampled PSUs in stratum- $h$ ,  $\pi_{hi}$  denote the selection probability for the PSU- $i$  of stratum  $h$ , and  $n_{hi}$  denote the number of sampled elementary units in the PSU- $i$  of stratum  $h$ . An unbiased estimator of  $Y$  is

$$\begin{aligned} \hat{Y} &= \sum_{h=1}^L \sum_{i=1}^{m_h} (1/\pi_{hi}) \sum_{j=1}^{n_{hi}} (N_{hi}/n_{hi}) \times y_{hij} \\ &= \sum_{h=1}^L \sum_{i=1}^{m_h} \sum_{j=1}^{n_{hi}} w_{hi} \times y_{hij} \end{aligned} \quad (2)$$

where  $w_{hi} = (1/\pi_{hi}) (N_{hi}/n_{hi})$ . We can see that  $\hat{Y}$  is a weighted sum of observed data, and the weight  $w_{hi}$  is simply the number of people in the population who are represented by a person in the PSU- $i$  of stratum  $h$ . Using these analysis weights, the total number of people in the population can be estimated by

$$\hat{W} = \sum_{h=1}^L \sum_{i=1}^{m_h} \sum_{j=1}^{n_{hi}} w_{hi} \quad (3)$$

It seems natural then to estimate the population mean, say  $\bar{Y}$ , by the ratio estimator

$$\hat{\bar{Y}} = \hat{Y}/\hat{W} \quad (4)$$

**Table 1** Two practical examples of cluster samples

Cluster	Listing unit	Elementary unit	Application
City block	Household	People	Estimation of mean serum uric acid level among the elderly in a city
School	Classroom	Student	Estimation of the prevalence of nearsightedness among elementary students in a school district

According to the suggestion by Cochran [Eq. 11.35],<sup>2</sup> we can compute the sampling error of  $\hat{Y}$  by a two-stage procedure described below.

*Step 1.* Compute the weighted sum of the data observed in each PSU. For the PSU- $i$  of stratum  $h$ , the weighted sum is

$$Z_{hi} = \sum_{j=1}^{n_{hi}} w_{hi} \times y_{hij}. \tag{5}$$

*Step 2.* Substitute  $Z_{hi}$  into

$$\text{Var}(\hat{Y}) = \sum_{h=1}^L m_h \times \hat{\sigma}_h^2 \tag{6}$$

where  $\hat{\sigma}_h^2$  is the sample variance based on  $Z_{hi}$  in stratum  $h$ . Using Taylor-series linearization of  $\hat{Y}$ , we can show that the sampling error of  $\hat{Y}$  can be computed by the above procedure where

$$Z_{hi} = \sum_{j=1}^{n_{hi}} w_{hi}(y_{hij} - \hat{Y})/\hat{W}. \tag{7}$$

A common but erroneous practice in the estimation of  $Y$  or  $\bar{Y}$  is to assume  $w_{hi}$  were equal when, in fact, they are not. In addition, design effects are usually ignored in variance estimation. It is not hard to find in the literature that many authors do all their variance calculation using formulas valid for SRS, which ignore the positive correlation between the data collected from the same cluster. As a result, the sampling variances of  $\hat{Y}$  and  $\bar{Y}$  are underestimated, and, in turn, the statistical inferences, including interval estimation and hypothesis testing, about  $Y$  and  $\bar{Y}$  are simply misleading. Specifically, the confidence intervals are narrower than they should be or the null hypotheses are rejected more often than they should be.

The same kind of erroneous practices also exist in the estimation of subpopulation totals or means, and the differences between them. Inferences on these parameters are often erroneously conducted. See Appendix for details on associated variance estimation.

In the following section, we will demonstrate the consequences of ignoring analysis weights, and design effects in the estimation of totals or means by two examples from a regional health survey using stratified two-stage cluster sampling. We first analysed the data ignoring analysis weights and design effects (Method I), and then analysed the same data considering analysis weights, stratification, and clustering (Method II). We first analysed the data ignoring analysis weights and design effects (Method I). Specifically, the data were analysed as though they had been sampled by simple random sampling. Then we analysed the same data considering analysis weights, stratification, and clustering (Method II) by the use of equations (1)–(7) in the text, and (8)–(9) in the Appendix.

**Results**

In the stratified two-stage cluster sampling scheme, area code was used as a stratification variable (stratum), clusters of telephone numbers as the primary sampling units, and individual telephone numbers as the elementary units.

The first example is to estimate the average number of cigarettes smoked per day in male and female smokers. For this example,  $y_{hij}$  is the number of cigarettes smoked per day by individual  $j$  in the PSU- $i$  of stratum  $h$ , and  $I_{hij(k)}$  (see Appendix), where  $k=1, 2$ , an indicator of gender groups; for instance,  $k=1$  for male and  $k=2$  for female.

The sample sizes of male and female smokers were 203 and 235, respectively, and the total number of cigarettes smoked per day by the male and female smokers were 3931 and 3361, respectively (Table 2). Using the unweighted sizes and totals (that is assuming  $w_{hi(k)}=1$  in  $\hat{W}_k$  and  $\hat{Y}_k$ ), the average number of cigarettes smoked per day ( $\bar{Y}_k$ ) in male smokers was 19.36, and 15.58 in female smokers. By weighting the data (that is, using the actual  $w_{hi(k)}$  in  $\hat{W}_k$  and  $\hat{Y}_k$ ), estimates of the total number of male smokers ( $\hat{W}_1$ ) and the total number of cigarettes smoked by male smokers per day ( $\hat{Y}_1$ ) were 53,966 and 985,142, and the corresponding values for female smokers ( $\hat{W}_2$  and  $\hat{Y}_2$ ) were 51,094 and 814,622. Using the weighted sizes and totals, the average number of cigarettes smoked per day in male and female smokers were 18.25 and 15.94, respectively. The weighted and unweighted estimates of subpopulation means and their standard errors are listed in Table 3. As expected, the standard errors of the unweighted estimates were smaller than those of corresponding weighted estimates. As a

**Table 2** Sample (or unweighted) and weighted total numbers of smokers and total numbers of cigarettes smoked per day by gender

	Gender		Total
	Male	Female	
Sample size	203	235	438
Sample total*	3931	3661	7592
Weighted size	53966	51094	105059
Weighted total*	985142	814622	1799763

\* total numbers of cigarettes smoked per day

**Table 3** The average number of cigarettes smoked per day in male and female smokers, and their difference obtained by Method I and Method II

Method	Overall	Gender		Difference
		Male	Female	
I	17.33	19.36	15.58	3.78
	(0.52) <sup>a</sup>	(0.72)	(0.73)	(1.02)
II	17.13	18.25	15.94	2.31
	(0.73)	(1.13)	(0.90)	(1.44)

<sup>a</sup> All the entries in parentheses are standard errors.

**Table 4** Sample (or unweighted) and weighted total numbers of women, and numbers of women who had a mammogram by age group

	Age (in years)			Total
	18-34	35-54	≥ 55	
Sample size	301	374	129	804
Sample total*	34	243	90	367
Weighted size	78830	71417	24310	174556
Weighted total*	10036	49743	19335	79113

\* number of women who had a mammogram

**Table 5** The percentage of women who had a mammogram in each one of the three age groups obtained by Method I and Method II

Method	Overall	Age (in years)		
		18-34	35-54	≥ 55
I	45.65 (1.76) <sup>a</sup>	11.30 (1.83)	64.97 (2.47)	69.77 (4.06)
II	45.32 (2.55)	12.73 (2.64)	69.65 (3.32)	79.54 (4.30)

<sup>a</sup> All the entries in parentheses are standard errors.

result, the 95% confidence interval on the difference between male and female smokers in the average number of cigarettes smoked per day obtained by Method I, (1.77, 5.79), was narrower than that obtained by Method II, (-0.53, 5.15). If Method I had been used, we would have erroneously concluded that there was gender difference in the average number of cigarettes smoked per day when, in fact, there was no gender difference as indicated by Method II.

The second example is to estimate the percentage of women who had a mammogram in each of the three age groups, 18-34, 35-54, and ≥ 55 y. For this example,  $y_{hij}$  is a dichotomous variable that has a value of 1 if the woman  $j$  in the PSU- $i$  of stratum  $h$  had a mammogram; otherwise, it has a value of 0, and  $I_{hij(k)}$ , where  $k = 1, 2, 3$ , an indicator of age groups.

The sample size of female participants was 804, 301 of age 18-34 y, 374 of age 35-54 y, and 129 of age greater than or equal to 55 y, and 367 of the sampled women had a mammogram (Table 4). Using the unweighted analysis, 45.65% of the study women had a mammogram, and the corresponding proportions were 11.30%, 64.97%, and 69.77% in the groups of 18-34, 35-54, and ≥ 55 y of age, respectively (Table 5). By weighting the data, the estimated proportions in the three age groups were 12.73%,

69.65%, and 79.54%, respectively. The estimates between unweighted and weighted data were different by nearly 1.5% in the 18-34 age group, nearly 5% in the 35-54 age group, and nearly 10% in the ≥ 55 age group. In addition, the standard errors of the unweighted estimates were smaller than those of the corresponding weighted estimates. As a result, the 95% confidence intervals on age contrasts of the proportion of women who had a mammogram obtained by Method I were narrower than those obtained by Method II, and more importantly, were biased (Table 6).

### Discussion

Our examples demonstrate that ignoring analysis weights and design effects can lead to misleading results. It is clear that clustering induces larger and positive correlations between element values which can not be ignored in variance estimation. Most of the commonly-used statistical computing packages (SAS, SPSS, BMDP) assume data were obtained from a simple random sample; that is, the observations are independent and identically distributed. When data have been collected from a survey which has a complex sample design, the simple random sample assumption can often lead to an underestimate of the variance, which can therefore lead to artificially small confidence intervals and anticonservative hypothesis testing; namely rejecting the null hypothesis when it is in fact true. It is also clear that had we ignored analysis weights in the estimation of means or proportions, the resulting estimates would be biased. More specifically, when unequal probability samples are drawn, the results will be biased by unweighted analysis. Most of the statistical computing packages actually have the feature to compute the weighted means and proportions of equation (4). For example, by specifying a WEIGHT statement for the sampling weights in the SAS procedure PROC UNIVARIATE and in the procedure PROC FREQ, we can compute the weighted estimates.<sup>3</sup> However, the sampling variances are still computed in these packages as though the sample was selected by means of SRS giving

$$\text{Var}(\hat{Y}) = \hat{\sigma}^2/n$$

where  $\hat{\sigma}^2$  is the sample variance based on weighted data, and  $n$  is the total number of elementary units drawn. These will not be the same as the sampling variances given by equation (6). The same bias occurs in employing similar routines in SPSS and BMDP, and therefore the inferential statistics on population means or proportions generated by SAS, SPSS, or BMDP in the analysis of complex survey data may lead to erroneous conclusions even if sampling weights are accounted for. So may the inferential statistics on the comparisons of subgroup means or proportions as

**Table 6** Ninety-five percent confidence intervals on age contrasts of the percentage of women who had a mammogram obtained by Method I and Method II

Method	95% C.I. <sup>a</sup> on age contrasts		
	A vs B <sup>b</sup>	A vs C	B vs C
I	(-59.69, -47.65)	(-67.19, -49.75)	(-14.11, 4.51)
II	(-65.25, -48.59)	(-76.71, -56.91)	(-20.57, 0.79)

<sup>a</sup> C.I.: Confidence intervals.

<sup>b</sup> A: 18-34 y, B: 35-54 y, and C: ≥ 55 y.

shown in Table 5. In carrying out a regression analysis of complex survey data, ignoring stratification and/or clustering in complex survey designs may also yield seriously misleading results.<sup>4-6</sup> The SAS procedure PROC REG<sup>7</sup> and similar routines in SPSS and BMDP all fail to provide usable options for taking into account the design effect in variance estimation.

For readers familiar with the derivation of equations (1)–(7) in the text, and (8)–(9) in the Appendix, the calculation of these formulas is just simple mathematics and can be easily programmed by a computer language such as FORTRAN 77.<sup>8</sup> For readers not familiar with the derivation, several computing resources for variance estimation with complex survey data are also available.<sup>9,10</sup> For personal computer (PC) users, PC SUDAAN<sup>11</sup> and PC CARP,<sup>12</sup> are two widely used software packages designed for this purpose. Carlson *et al*<sup>13</sup> evaluated various features and flaws of the two packages, and concluded that using a PC for complex survey data analysis is certainly feasible, and may be desirable in many circumstances. Compared to PC CARP, PC SUDAAN has two desirable features that might be appealing to SAS users; one is that its command syntax is similar to that used by SAS, and the other is that it accepts data files of SAS format.

### Conclusions

Using inappropriate methods and statistical computing packages in the analysis of complex survey data may yield misleading results that could bear serious consequences on health policy and clinical practice. These consequences can be avoided by applying appropriate computer software packages and procedures suggested by the authors.

### Acknowledgements

We would like to thank Donald Brogan, Ph.D. for her enjoyable teaching of sample survey that helped us greatly in writing this paper. This study was in part supported by the National Science Council (NSC 82-0412-B006-079T), Taiwan, R.O.C.

### Appendix

The total for a subpopulation, say  $k$ , can be expressed as

$$Y_k = \sum_{h=1}^L \sum_{i=1}^{M_h} \sum_{j=1}^{N_{hi}} y_{hij(k)}$$

where

$$y_{hij(k)} = y_{hij} \times I_{hij(k)}$$

with

$$I_{hij(k)} = \begin{cases} 1 & \text{if; the elementary unit } j \text{ in the PSU-}i \text{ of} \\ & \text{stratum } h \text{ is an element of domain } k \\ 0 & \text{otherwise.} \end{cases}$$

$Y_k$  clearly has the form of equation (1) in the text. So it can be estimated by an unbiased estimator,  $\hat{Y}_k$ , of the form (2) where  $y_{hij}$  is replaced with  $y_{hij(k)}$ . Its sampling error can be computed using the variance formula (6) where  $Z_{hi}$  is obtained by substituting  $y_{hij(k)}$  in (5) for  $y_{hij}$ . Similarly, the total number of people in domain  $k$  can be estimated by

an estimator,  $\hat{W}_k$ , of the form (3) where  $w_{hi}$  is replaced with  $w_{hi(k)} = w_{hi} \times I_{hij(k)}$ . Replacing the numerator and denominator of (4) with  $\hat{Y}_k$  and  $\hat{W}_k$ , respectively yields an estimate,  $\hat{Y}_k$ , of the subpopulation mean. Its sampling error can be computed using the variance formula (6) where

$$Z_{hi} = \sum_{j=1}^{N_{hi}} w_{hi} (y_{hij(k)} - \hat{Y}_k \times I_{hij(k)}) / \hat{W}_k. \quad (8)$$

The procedures for the estimation of subpopulation totals or means are basically the same as those for the estimation of population totals or means. Therefore, if the analysis weights were assumed to be equal or sampling errors were computed ignoring the design effects, the estimates of subpopulation totals or means would be biased and associated statistical inferences would be misleading.

Simple algebra can show that the difference between two subpopulation totals,  $Y_k$  and  $Y_{k'}$ , also has the form of (1), and can be estimated by an unbiased estimator,  $\hat{Y}_k - \hat{Y}_{k'}$ , of the form (2) where  $y_{hij}$  is replaced with  $y_{hij(k)} - y_{hij(k')}$ . The sampling error of  $\hat{Y}_k - \hat{Y}_{k'}$  can be computed using the variance formula (6) where  $Z_{hi}$  is obtained by substituting  $y_{hij(k)} - y_{hij(k')}$  in (5) for  $y_{hij}$ . By analogy, the difference between  $Y_k$  and  $Y_{k'}$  can be estimated by  $\hat{Y}_k - \hat{Y}_{k'}$ . Using Taylor-series linearization of  $\hat{Y}_k - \hat{Y}_{k'}$ , the sampling error of  $\hat{Y}_k - \hat{Y}_{k'}$  can be computed using the variance formula (6) where

$$Z_{hi} = \sum_{j=1}^{N_{hi}} w_{hi} (y_{hij(k)} - \hat{Y}_k \times I_{hij(k)}) / \hat{W}_k \\ - \sum_{j=1}^{N_{hi}} w_{hi} (y_{hij(k')} - \hat{Y}_{k'} \times I_{hij(k')}) / \hat{W}_{k'}. \quad (9)$$

### References

- 1 Kish L, Frankel MR. Inference from complex samples (with Discussion). *J R Stat Soc* 1974; **B36**: 1–37.
- 2 Cochran WG. *Sampling Techniques* (3rd edn). John Wiley & Sons: New York, 1977.
- 3 SAS Institute Inc. *SAS User's Guide: Basics* (5th edn). SAS Institute Inc.: Cary, NC, 1985.
- 4 Konijn HS. Regression in sample surveys. *J Am Stat Ass* 1962; **51**: 590–606.
- 5 Holt D, Smith TMF, Winter PD. Regression analysis of data from complex surveys. *J R Stat Soc* 1980 (Series A); **A143**: 474–487.
- 6 Nathan G, Holt D. The effect of survey design on regression analysis. *J R Stat Soc* 1980; **B42**: 377–386.
- 7 SAS Institute Inc. *SAS/STAT User's Guide* (6.03 edn). SAS Institute Inc.: Cary, NC, 1988.
- 8 Mojena R, Ageloff R. *Fortran 77*. Wadsworth, Inc.: Belmont, CA, 1990.
- 9 Lee ES, Forthofer RN, Lorimor RJ. *Analyzing Complex Survey Data*. Sage Publications Inc.: Newbury Park, CA, 1989.
- 10 Wolter KW. *Introduction to Variance Estimation*. Springer-Verlag New York Inc.: New York, NY, 1985.
- 11 Research Triangle Institute. *SUDAAN User's Manual: Software for Analysis of Correlated Data* (6.04 edn). Research Triangle Institute: Research Triangle Park, NC, 1988.
- 12 Fuller WA *et al*. *PC CARP*. Statistical Laboratory, Iowa State University: Ames, IA, 1986.
- 13 Carlson BL, Johnson AE, Cohen SB. An evaluation of the use of personal computers for variance estimation with complex survey data. *J Official Stat* 1993; **9(4)**: 795–814.