

2 Complex Sampling Concepts

2.1 Introduction

A common theme in alcohol abuse research is that data are usually obtained from a multi-stage or complex sample design. An example of a typical complex sampling scheme is:

- Stratify the geographical area under study according to census geography and census socio-economic variables.
- Form meaningful clusters of population elements, called primary sampling units (PSUs), for example schools, in each stratum.
- Draw a predetermined number of PSUs from each stratum, using probability sampling proportional to size.
- Do one or more stages of subsampling within each PSU.
- Draw a simple random sample of ultimate sampling units (USUs) at the last stage.

The main advantages of a complex sample (CS) in comparison with a simple random sample (SRS) are:

- CS does not require a complete sampling frame of the population elements.
- CS is more economical and practical.
- CS guarantees a representative sample of the population.
- CS makes a step-by-step design of the sample possible.

The main disadvantage of CS is that it is generally less efficient than SRS, *i.e.*, it yields estimates of lower precision for a fixed sample size.

In the application of CS, the design effect ($deff$) and sampling weights play an important role. The design effect is defined as

$$deff = \frac{\text{Variance of an estimate using complex sampling}}{\text{Variance of an estimate under SRS}}$$

The design effect ($deff$) provides a rough and ready method of estimating the variance of survey statistics and of adjusting the output of standard statistical software packages for the complex sampling design. This aspect of $deff$ derives from its assumed portability. See Kish (1965) for a discussion of design effects.

The $deff$ is used not only to produce estimates of variance, but also to adjust the output of standard analyses. For example, the practitioner may utilize standard statistical software packages to conduct a regression analysis of a hypothesized linear relationship between survey variables, or to formulate a multi-way table and conduct a χ^2 test of independence between survey variables. The output of standard statistical software packages gives wrong answers for such problems (because

the underlying assumptions of the methods are not satisfied for complex survey designs). A first-order correction may be obtained by dividing the corresponding test statistic by the estimated deff. See Rao & Scott (1981) and Skinner, Holt, & Smith (1989).

In this chapter we start with some known results for the sum of random variables and for multiple linear regression to illustrate adjustments that must be made to accommodate complex sampling properly. In Sections 2.2 and 2.3 we provide a brief summary of important concepts in complex sampling and in Sections 2.4 to 2.7 discuss how these concepts are currently applied to fit regression models to survey data. We will show that standard software packages for regression analysis allow for a weight variable, but do not yield the correct standard error estimates and measures of fit.

2.2 Indicator variables and t-estimators

Consider a finite population of identifiable units $U = \{u_1, u_2, \dots, u_N\}$ where the size N of the population is assumed known. The inclusion of a given element u_i in a sample s is a random event indicated by the random variable I_i (sample membership indicator of element i), $i = 1, 2, \dots, N$ defined as

$$I_i = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{else.} \end{cases}$$

The probability that u_i will be included in the sample, denoted by π_i , is

$$\pi_i = P(u_i \in s) = P(I_i = 1).$$

The probability that both u_i and u_j will be included in the sample, denoted by π_{ij} , is

$$\pi_{ij} = P(u_i \in s \text{ and } u_j \in s) = P(I_i \cdot I_j = 1)$$

By definition (see *e.g.* Traat, Meister, & Söstra, 2001). Therefore, $E(I_i) = \pi_i$, $Var(I_i) = \pi_i(1 - \pi_i)$ and $Cov(I_i, I_j) = \pi_{ij} - \pi_i\pi_j$.

The selected sample $s = \{u_{(1)}, u_{(2)}, \dots, u_{(n)}\}$, is an unordered set of population units, where n denotes the sample size. A sampling weight w_i for the i -th USU is usually calculated as $1/\pi_i$, where π_i

denotes the inclusion probability. If $\pi_1 = \pi_2 = \dots = \pi$ the sample is called a self-weighting sample. Sometimes w_i is called the base weight.

Let \mathbf{y}_N and \mathbf{z}_N be $N \times 1$ vectors of finite population values with typical elements y_j and z_j , $j = 1, 2, \dots, N$ respectively. Denote the values drawn from a multi-stage sample of size n by \mathbf{y}_s and \mathbf{z}_s , where

$$\mathbf{y}_s = (y_{(1)}, y_{(2)}, \dots, y_{(j)}, \dots, y_{(n)})',$$

$$\mathbf{z}_s = (z_{(1)}, z_{(2)}, \dots, z_{(j)}, \dots, z_{(n)})'.$$

Here $z_{(j)}$ denotes the j -th sample element, $z_{(j)} \in \mathbf{z}_s$.

Consider the population totals $t_1 = \sum_{i=1}^N y_i$, $t_2 = \sum_{i=1}^N y_i^2$, $t_3 = \sum_{i=1}^N z_i$, $t_4 = \sum_{i=1}^N z_i^2$, and $t_5 = \sum_{i=1}^N z_i y_i$. Each population total t_j can be estimated by the corresponding π -estimator $\hat{t}_{j\pi}$ (Horvitz & Thompson, 1952). For example,

$$\hat{t}_{1\pi} = \sum_{i=1}^n y_{(i)} / \pi_i, \quad \hat{t}_{5\pi} = \sum_{i=1}^n z_i y_i / \pi_i.$$

Each estimated total can be written as a linear function of the sample membership indicators $I_i, i = 1, 2, \dots, N$. For example,

$$\hat{t}_{1\pi} = \sum_{i=1}^N I_i y_i / \pi_i.$$

This estimator (Horvitz-Thompson) is an unbiased estimator of t_1 since

$$E(\hat{t}_{1\pi}) = \sum_{i=1}^N E(I_i) y_i / \pi_i = \sum_{i=1}^N y_i = t_1.$$

Use of $Cov(I_i, I_j) = \pi_{ij} - \pi_i \pi_j$ gives

$$Var(\hat{t}_{1\pi}) = \sum_{i=1}^N \sum_{j=1}^N \frac{1}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j.$$

Hence,

$$\widehat{Var}(\hat{t}_{1\pi}) = \sum_{i=1}^n \sum_{j=1}^n \frac{1}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_{(i)} y_{(j)}.$$

This simple example shows that the expected value and variance of the sum are rather different under complex sampling than for the corresponding case of simple random sampling.

Standard methods are available for the point estimation of sample functions of the population totals, such as means, ratios, and differences of ratios. These methods are based (see *e.g.* Särndal, *et. al.*, 1992) on the following result. Given that a population parameter θ can be expressed as a function of several population totals, *i.e.* $\theta = f(t_1, t_2, \dots, t_q)$, then an estimator $\hat{\theta}$ of θ is obtained from $\hat{\theta} = f(\hat{t}_{1\pi}, \dots, \hat{t}_{j\pi}, \dots, \hat{t}_{q\pi})$ where $\hat{t}_{j\pi}$ is the corresponding π -estimator of t_j . Additionally, consistent estimators of the sample variances of the estimators are available and have been implemented in various programs for the analysis of survey data (*cf.* Section 2.2). One method of estimating the variance of $\hat{\theta}$ if θ is a nonlinear function of the totals is by a first-order Taylor approximation of $f(\hat{t}_{1\pi}, \dots, \hat{t}_{j\pi}, \dots, \hat{t}_{q\pi})$ (see *e.g.* Wolter, 1985).

To estimate the variance of a survey estimator in the case of a single-stage sampling design, there are typically two alternatives: (1) the variance estimator based upon pps sampling with replacement, and (2) the Yates-Grundy estimator of variance (Yates & Grundy, 1953; Biyani, 1980) for pps without replacement sampling. Many survey practitioners will find the first estimator to be a satisfactory approximation to the variance given the actual survey design. For instances in which it is important to reflect the without replacement sampling design (*e.g.*, important sampling fraction) and where it is feasible to calculate the joint inclusion probabilities (*e.g.*, Durbin's two-per-stratum design; see, Durbin, 1967, and Shapiro & Bateman, 1978), you would have the opportunity to specify the Yates-Grundy estimator. Applied to a multi-stage sampling design, these estimators usually provide a very good approximation to the total variance.

2.3 Additional weight adjustments

It was previously mentioned that the sampling weight w_i is usually calculated as $w_i = 1/\pi_i$, the so-called base weight.

In a practical application each weight w_i usually undergoes additional adjustments,

- such as $w_i = \frac{1}{\pi_i} \frac{1}{\widehat{R}_i}$ (nonresponse adjusted weight), where \widehat{R}_i is the response rate in a cell containing the i -th unit
- $w_i = \frac{1}{\pi_i} \frac{1}{\widehat{R}_i} F_i$ (post stratification weight), where $F_i = \frac{T_k}{\widehat{T}_k}$, and where T_k denotes the total number of units belonging to the k -th cell, $k = 1, 2, \dots, K$ of a contingency table formed from a set of categorical variables. For example, consider the variables age (5 categories), gender (2 categories), and ethnic group (3 categories). In this case $K = 5 \times 2 \times 3 = 30$. From the sample one can establish, for each of these variables, which category (*e.g.* male/female) is assigned to the i -th USU, and hence determine the cell number for that specific combination. The estimated total \widehat{T}_k is obtained using

$$w_i = \frac{1}{\pi_i} \frac{1}{\widehat{R}_i}.$$

It is evident that non-response and post stratification adjusted weights will have an impact on the estimation of population totals and functions of population totals.

2.4 Linear regression

In this section we show the effect of sampling on the estimates and variability of regression coefficients. Again, the results are rather different from those that can be derived under the more familiar case of simple random sampling.

Suppose \mathbf{Y}_N is an $N \times p$ matrix defined as $\mathbf{Y}_N' = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$, where the elements of \mathbf{y}_i are values of p variables of interest.

2.4.1 Example 1

Let y_{ij} denote a typical element of \mathbf{y}_i , where y_{ij} is the number of occasions alcohol was consumed by student i ($i = 1, 2, \dots, N$) in the prior 30 days, and where $j = 1$ denotes grade 8, $j = 2$ denotes grade 9, ..., $j = 5$ denotes grade 12.

2.4.2 Example 2

Let the subscript i denote the student i ($i = 1, 2, \dots, N$). Suppose y_{i1} equals the number of times this student was under the influence of alcohol in the prior year; y_{i2} is a language score, and y_{i3} is a math score.

Example 1 above describes a longitudinal study, often referred to in the literature as a repeated measurements study, since measurements are made on the same individual on successive occasions. Note that, in general, measurement occasions are not necessarily equally spaced over time.

Example 2 describes a typical cross-sectional study. Note, however, that this study may have been carried out in 1998, and subsequently repeated in 2000 and 2002. It is evident that the finite populations U_{1998} , U_{2000} , and U_{2002} will overlap if, for example, 8th to 12th graders in the state of Texas are defined as the population elements. Hence, the samples s_{1998} , s_{2000} , and s_{2002} may also have overlapping units. A cross-sectional study, repeated over time, is often referred to as a panel study, but the statistical treatment usually treats the data as multiple-group data. In this examples the year of study defines the group. A typical multiple group application is to test for differences in the means of latent variables under the assumption of factor invariance.

Consider the case $p = 1$ (univariate regression) so that $\mathbf{Y}_N = \mathbf{y}_N$, an N -dimensional vector. Suppose further that \mathbf{X}_N is an $N \times r$ matrix defined as $\mathbf{X}'_N = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$, where the elements of \mathbf{x}_i are values of r auxiliary variables, for example x_{i1} = gender, x_{i2} = socio-economic status, and x_{i3} = age.

The finite population regression coefficient vector $\boldsymbol{\beta}$ is a function of \mathbf{y} and is defined as

$$\boldsymbol{\beta} = (\mathbf{X}'_N \mathbf{X}_N)^{-1} \mathbf{X}'_N \mathbf{y}_N. \quad (2.1)$$

Under a so-called design-based approach, $\boldsymbol{\beta}$ is an obvious choice for the parameter of interest when regression is based on sample survey data. In the estimation of $\boldsymbol{\beta}$ we assume an underlying homoscedastic model

$$E(\mathbf{y}_N | \mathbf{X}_N) = \mathbf{X}_N \boldsymbol{\beta}; \text{Cov}(\mathbf{y}_N | \mathbf{X}_N) = \sigma^2 \mathbf{I}_N. \quad (2.2)$$

Let \mathbf{X}_s denote an $n \times r$ matrix of rows of \mathbf{X}_N selected according to some sampling design s .

The ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}_s' \mathbf{X}_s)^{-1} \mathbf{X}_s' \mathbf{y}_s$ of $\boldsymbol{\beta}$ is not, in general, a design-consistent estimator of $\boldsymbol{\beta}$.

An equivalent expression for (2.1) is

$$\boldsymbol{\beta} = \mathbf{T}^{-1} \mathbf{t} \quad (2.3)$$

where $\mathbf{T} = \sum_{\alpha=1}^N \mathbf{x}_\alpha \mathbf{x}_\alpha'$, and $\mathbf{t} = \sum_{\alpha=1}^N \mathbf{x}_\alpha y_\alpha$.

Let t_{ij} and t_{i0} denote typical elements of \mathbf{T} and \mathbf{t} respectively, then

$$t_{ij} = \sum_{\alpha=1}^N x_{i\alpha} x_{j\alpha} \quad \text{and} \quad t_{i0} = \sum_{\alpha=1}^N x_{i\alpha} y_\alpha.$$

Each total (*cf.* Section 2.2) can be estimated by their unbiased π -estimators. For example,

$$\hat{t}_{ij,\pi} = \sum_{\alpha=1}^n x_{(i\alpha)} x_{(j\alpha)} / \pi_\alpha,$$

$$\hat{t}_{i0,\pi} = \sum_{\alpha=1}^n x_{(i\alpha)} y_{(\alpha)} / \pi_\alpha, \quad i, j = 1, 2, \dots, r.$$

In matrix notation,

$$\hat{\mathbf{T}} = \sum_{\alpha=1}^n \frac{\mathbf{X}_{(\alpha)} \mathbf{X}_{(\alpha)'}}{\pi_\alpha} = \mathbf{X}_s' \mathbf{W}_s \mathbf{X}_s, \quad (2.4)$$

and

$$\hat{\mathbf{t}} = \sum_{\alpha=1}^n \frac{\mathbf{X}_{(\alpha)} \mathcal{Y}_{(\alpha)}}{\pi_{\alpha}} = \mathbf{X}'_s \mathbf{W}_s \mathbf{y}_s, \quad (2.5)$$

where $\mathbf{x}_{(\alpha)}$ denotes the α -th column of \mathbf{X}'_s .

This yields the design-weighted estimator

$$\hat{\boldsymbol{\beta}}_W = \hat{\mathbf{T}}^{-1} \hat{\mathbf{t}} = (\mathbf{X}'_s \mathbf{W}_s \mathbf{X}_s)^{-1} \mathbf{X}'_s \mathbf{W}_s \mathbf{y}_s, \quad (2.6)$$

which is a design-consistent estimator of $\boldsymbol{\beta}$.

The weighting matrix is defined as $\mathbf{W}_s = \text{diag}(w_1, w_2, \dots, w_n)$, where $w_i = 1/\pi_i$ denotes the sampling weight for the i -th USU. In the case of a self-weighting design, *i.e.*, $\pi_i = n/N, i = 1, 2, \dots, N$, the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_W$ are identical.

2.5 Standard error estimation

For most sample designs used in practice, the sampling variance of $\hat{\boldsymbol{\beta}}_W$ cannot be estimated using standard computer packages, and a variance estimating technique has to be used.

The basic methods available (see *e.g.* Rao, 1975) are:

(a) Linearization or Taylor expansion methods (Wolter, 1985, Binder, 1983):

Suppose that the parameter θ is a nonlinear function $f(t_1, t_2, \dots, t_q)$ of population totals, then θ is consistently estimated by $\hat{\theta} = f(\hat{t}_{1,\pi}, \hat{t}_{2,\pi}, \dots, \hat{t}_{q,\pi})$. By Taylor linearization it follows that

$$f(\hat{t}_{1,\pi}, \hat{t}_{2,\pi}, \dots, \hat{t}_{q,\pi}) \approx f(t_1, t_2, \dots, t_q) + \sum_{j=1}^q a_j (\hat{t}_{j,\pi} - t_j),$$

where

$$a_j = \frac{\partial f(\hat{t}_{1,\pi}, \hat{t}_{2,\pi}, \dots, \hat{t}_{q,\pi})}{\partial \hat{t}_{j,\pi}} \Big|_{\hat{t}_{1,\pi}=t_1, \dots, \hat{t}_{q,\pi}=t_q}$$

From (1.3) it follows that $\boldsymbol{\beta}$ is a nonlinear function of $r(r+1)/2+r$ population totals, with corresponding estimator $\hat{\boldsymbol{\beta}}_w$ as defined in (2.6).

Using a first-order Taylor approximation, it can be shown that (cf. (2.3))

$$\hat{\boldsymbol{\beta}}_w \approx \boldsymbol{\beta} + \mathbf{T}^{-1}(\hat{t} - \hat{\mathbf{T}}\boldsymbol{\beta}). \quad (2.7)$$

From (2.7) it follows that

$$\text{Cov}(\hat{\boldsymbol{\beta}}_w) \approx \mathbf{T}^{-1}\mathbf{V}\mathbf{T}^{-1}.$$

An approximate expression for $\mathbf{T}^{-1}\mathbf{V}\mathbf{T}^{-1}$ is $\hat{\mathbf{T}}^{-1}\hat{\mathbf{V}}\hat{\mathbf{T}}^{-1}$. Typical elements of \mathbf{V} and $\hat{\mathbf{V}}$ are given in, for example, Särndal *et. al.* (1992, page 194).

- (b) Balanced repeated replication (McCarthy, 1969): Statistics based on half-samples, which are selected so as to ensure an orthogonal balanced set, are computed and the empirical covariances of these statistics are used as the appropriate estimator.

In longitudinal studies an increase in precision is obtained if allowance is made for the fact that units sampled over time are correlated. Replication methods provide a simple means for incorporating this correlation.

- (c) Jackknife (Miller, 1974): The sample is first split into subsamples, each of which reflects the original complex design. Statistics based on the sample data without one of the subsamples are computed and the empirical covariances of these statistics serve as covariance estimators. A more detailed account is given in Wolter (1985).
- (d) Bootstrap (Efron, 1981, 1982; Kovar, Rao & Wu, 1988): The sample data is used to construct an artificial population U^* which is assumed to mimic the real, but unknown, population U . The original design is used to draw a series of K samples (with replacement) from U^* . For each “bootstrap” sample, i , an estimate $\hat{\theta}_i^*$ of the population parameter θ is computed and subsequently $\hat{\theta}$ and $\text{var}(\hat{\theta})$ are estimated from $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_K^*$.

While the standard statistical package computer programs do not in general deal with the complex sample design situation, several special purpose programs for covariance estimation have been developed for use with complex sample designs. Lepkowski and Bowles (1996) give a list of eight software packages that are available for use by the general survey analyst. The eight catalogued in their paper are CENVAR, CLUSTERS, EpiInfo, PC CARP, STATA, SUDAAN, VPLX and WesVar. The

first six programs use the Taylor series expansion for variance estimation and the remaining programs use replication methods.

Theoretical comparisons of the different methods of covariance estimation by Krewski & Rao (1981) and empirical comparisons by Kish & Frankel (1974) and by Richards & Freeman (1980) indicate their performance is very similar in many cases.

2.6 Heteroscedastic model

Some predictors may exhibit heteroscedasticity, in which changes in variance occur with changes in the values of the predictor. For example, the variance of income across individuals is systematically higher for higher-income individuals.

In the case of a heteroscedastic model

$$E(\mathbf{y}_N | \mathbf{X}_N) = \mathbf{X}_N \mathbf{B} \quad \text{Cov}(\mathbf{y}_N | \mathbf{X}_N) = \sigma^2 \mathbf{V}, \quad (2.8)$$

the finite population parameter

$$\boldsymbol{\beta}^* = (\mathbf{X}_N' \mathbf{V}^{-1} \mathbf{X}_N)^{-1} \mathbf{X}_N' \mathbf{V}^{-1} \mathbf{y}_N \quad (2.9)$$

is a more suitable parameter for inference.

If \mathbf{V} is diagonal and the inclusion probabilities are proportional to the variances, then $\hat{\boldsymbol{\beta}}_w$ (*cf.* (2.6)) coincides with the weighted least squares estimator

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{y}_s \quad (2.10)$$

where \mathbf{V}_s is the appropriate $n \times n$ submatrix of \mathbf{V} .

Standard programs compute the OLS estimator, $\hat{\boldsymbol{\beta}}$, and can often also compute the generalized OLS estimator, $\hat{\boldsymbol{\beta}}^*$, together with unbiased estimators of their model-variances $\sigma^2 (\mathbf{X}_s' \mathbf{X}_s)^{-1}$ under (2.2) and $\sigma^2 (\mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{X}_s)^{-1}$ under (2.8) respectively. The design-weighted estimator, $\hat{\boldsymbol{\beta}}_w$, can also be obtained by the weighted regression options of standard statistical packages (*e.g.* LISREL or SPSS) by using the weights $1/\pi_i$. Alternatively, $\hat{\boldsymbol{\beta}}_w$ can be obtained by unweighted regression on the transformed variables $y_i / \sqrt{\pi_i}$ and $\mathbf{x}_i / \sqrt{\pi_i}$. Nathan (1988), however, has pointed out that the

reported variances and covariances will be incorrect. This implies that the standard significance tests (*e.g.* F-tests) will be invalid and can result in misleading conclusions.

The programs that use weighted regression, with weights $1/\pi_i$, report the estimator of the variance-covariance matrix as $\hat{\sigma}^2 (\mathbf{X}'_s \mathbf{W}_s \mathbf{X}_s)^{-1}$. The model-variance of $\hat{\boldsymbol{\beta}}_W$, under the homoscedastic model (2.2), is

$$\hat{\sigma}^2 (\mathbf{X}'_s \mathbf{W}_s \mathbf{X}_s)^{-1} (\mathbf{X}'_s \mathbf{W}_s \mathbf{V}_s \mathbf{W}_s \mathbf{X}_s) (\mathbf{X}'_s \mathbf{W}_s \mathbf{X}_s)^{-1}$$

which simplifies to $\hat{\sigma}^2 (\mathbf{X}'_s \mathbf{W}_s \mathbf{X}_s)^{-1}$ under the homoscedastic model ($\mathbf{V} = \mathbf{I}$) only for self-weighting designs and under the heteroscedastic model (2.8) only if \mathbf{V} is diagonal and the inclusion probabilities are proportional to the variances.

2.7 Covariance matrix of vector of totals

2.7.1 Introduction

In this section formulae for the estimation of the covariance matrix of a vector of totals are given for single-stage, two-stage, and three-stage sampling designs.

For a multi-stage sampling design we assume the following general sampling methods at each stage:

- First stage: random sampling with replacement (WR), random sampling without replacement and equal probability of selection (WOR), and random sampling without replacement and unequal probabilities (UWOR).
- Second stage: if the first stage is not WR, then WR, WOR, or UWOR.
- Third stage: if second stage is not WR, then WR, WOR, or systematic.

From the above it follows that all specifications other than weights are ignored for subsequent stages if a multi-stage sample contains a WR, or an approximation to WR, stage.

Overall weights for each ultimate sampling unit can be obtained as a product of weights for corresponding units computed in each sampling stage.

2.7.2 Notation

N	: Total number of elements in the population
n	: Total number of elements in the sample
H	: Number of strata
n_h	: Sampled number of primary sampling units (PSU) per stratum
m_{hi}	: Number of elements in the i -th sampled PSU in stratum h , $i=1, \dots, n_h$
w_{hij}	: Overall sampling weight for j -th element in the i -th sampled PSU in stratum h
\mathbf{y}_{hij}	: Values of vector \mathbf{y} for the j -th element in the i -th sampled PSU in stratum h
\mathbf{y}_T	: Population total sum for vector of variables \mathbf{y}

2.7.3 Total covariances

To simplify the expressions for the estimated covariance matrix of a vector of totals, let

$$\mathbf{z}_{hij} = w_{hij} \mathbf{y}_{hij} \quad (2.11)$$

where the index h denotes a stratum within a given sampling stage, i denotes the i -th sampled unit within stratum h in the same sampling stage and j denotes all final stage units contained within hi .

Let

$$\mathbf{z}_{hi} = \sum_{j=1}^{m_{hij}} \mathbf{z}_{hij} \quad (2.12)$$

$$\bar{\mathbf{z}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \mathbf{z}_{hi} \quad (2.13)$$

and

$$\mathbf{S}_h^2(\mathbf{y}) = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (\mathbf{z}_{hi} - \bar{\mathbf{z}}_h) (\mathbf{z}_{hi} - \bar{\mathbf{z}}_h)' \quad (2.14)$$

Single stage sample

The covariance of the total for vector \mathbf{y} in a single-stage sample is estimated by:

$$\widehat{\mathbf{V}}(\widehat{\mathbf{y}}_T) = \widehat{\mathbf{V}}_1(\widehat{\mathbf{y}}_T) = \sum_{h=1}^H \mathbf{U}_h(\widehat{\mathbf{y}}_T) \quad (2.15)$$

where $\mathbf{U}_h(\widehat{\mathbf{y}}_T)$ is an estimated contribution from stratum $h = 1, \dots, H$ and depends on the sampling method used:

- For WR, $\mathbf{U}_h(\widehat{\mathbf{y}}_T) = n_h \mathbf{S}_h^2(\mathbf{y})$,
- For simple random sampling, $\mathbf{U}_h(\widehat{\mathbf{y}}_T) = (1 - f_h) n_h \mathbf{S}_h^2(\mathbf{y})$,
- and for sampling WOR and unequal probabilities,

$$\mathbf{U}_h(\widehat{\mathbf{y}}_T) = \sum_{i=1}^{n_h} \sum_{i>j}^{n_h} \left(\frac{\pi_{hi} \pi_{hj}}{\pi_{hij}} - 1 \right) (\mathbf{z}_{hi} - \bar{\mathbf{z}}_h)(\mathbf{z}_{hj} - \bar{\mathbf{z}}_h)'$$

In the variance estimator, π_{hi} and π_{hj} are the inclusion probability for units i and j in stratum h , and π_{hij} is the joint inclusion probability for the same units (Yates & Grundy, 1953; Sen, 1953). In some situations it may yield a negative estimate and is treated as undefined.

Currently, for each stratum h containing a single element, the covariance contribution $\mathbf{U}_h(\widehat{\mathbf{y}}_T)$ is set to zero. An alternative procedure is to collapse strata. Presently, we leave it to the discretion of the user to collapse strata prior to any further statistical analysis.

Two-stage sample

When two-stage sampling is used and sampling WOR is applied in the first stage, the following estimate of the covariance of the total for vector \mathbf{y} may be used:

$$\widehat{\mathbf{V}}(\widehat{\mathbf{y}}_T) = \widehat{\mathbf{V}}_2(\widehat{\mathbf{y}}_T) = \widehat{\mathbf{V}}_1(\widehat{\mathbf{y}}_T) + \sum_{h=1}^H \sum_{i=1}^{n_h} \pi_{hi} \sum_{k=1}^{K_{hi}} \mathbf{U}_{hik}(\widehat{\mathbf{y}}_T). \quad (2.16)$$

- Here π_{hi} represents the first stage inclusion probability for the primary sampling unit i from stratum h .

- If simple random sampling is used, the inclusion probability is equal to the sampling rate f_h for stratum h .
- The number of second stage strata in the primary sampling unit i within the first stage stratum h is denoted by K_{hi} .
- $\mathbf{U}_{hik}(\hat{\mathbf{y}}_T)$ is the covariance contribution from the second stage stratum k from the primary sampling unit hi . It depends on the sampling method used in the second stage (see formulae above).

Three-stage sample

For a three-stage sample where first stage sampling is done without replacement, and simple random sampling is applied in the second stage, the following estimate of the covariance of the total for vector \mathbf{y} may be used:

$$\widehat{\mathbf{V}}(\hat{\mathbf{y}}_T) = \widehat{\mathbf{V}}_2(\hat{\mathbf{y}}_T) + \sum_{h=1}^H \sum_{i=1}^{n_h} \pi_{hi} \sum_{k=1}^{K_{hi}} f_{hik} \sum_{j=1}^{n_{hik}} \sum_{l=1}^{L_{hikj}} \mathbf{U}_{hikjl}(\hat{\mathbf{y}}_T), \quad (2.17)$$

where

- f_{hik} represents the sampling rate for the secondary sampling units in the second-stage stratum hik ,
- L_{hikj} indicates the number of third-stage strata in the secondary sampling unit $hikj$, and
- $\mathbf{U}_{hikjl}(\hat{\mathbf{y}}_T)$ denotes the covariance contribution from the third-stage stratum l , which is contained in the secondary sampling unit $hikj$. Again, this depends on the third-stage sample method (see formulae above).

2.8 Approximate covariance matrix of estimators

In this section we provide a general procedure for the estimation of the approximate covariance matrix of estimators. The results derived are based on Binder (1983) and use a first-order Taylor linearization.

Assume that L is the likelihood function or any other appropriate function of the vector $\boldsymbol{\gamma}$ of unknown parameters, and that an estimate $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$ is obtained by solving the set of simultaneous equations

$$\left. \frac{\partial \ln L}{\partial \boldsymbol{\gamma}} \right|_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}} = \mathbf{0} \quad (2.18)$$

In general, no closed-form solution to the set of equations (2.18) exists, and therefore parameter estimates are obtained iteratively using the Fisher scoring algorithm, for example,

$$\hat{\boldsymbol{\gamma}}^{(t+1)} = \hat{\boldsymbol{\gamma}}^{(t)} + \mathbf{I}_n^{-1}(\hat{\boldsymbol{\gamma}}^{(t)}) \mathbf{g}(\hat{\boldsymbol{\gamma}}^{(t)}) \quad (2.19)$$

where $\hat{\boldsymbol{\gamma}}^{(t)}$ denotes the parameter values at iteration t , $t = 1, 2, \dots$; $\mathbf{g}(\cdot)$ denotes the gradient vector; and $\mathbf{I}_n(\cdot)$ denotes the information matrix. In other words,

$$\mathbf{g}(\boldsymbol{\gamma}) = \frac{\partial \ln L}{\partial \boldsymbol{\gamma}} \quad (2.20)$$

and

$$\mathbf{I}_n(\boldsymbol{\gamma}) = -E \left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \right] \quad (2.21)$$

Denote the contribution to the gradient vector of each first-stage element for a given sampling stage by \mathbf{g}_{hij} , where h denotes stratum, and i the i -th unit within this stratum. The index j denotes a typical final stage element contained within the PSU hi , then

$$[\mathbf{g}(\boldsymbol{\gamma})]_r = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} [\mathbf{g}_{hij}(\boldsymbol{\gamma})]_r \quad (2.22)$$

From (2.18), (2.20), and (2.22) it follows that $\hat{\boldsymbol{\gamma}}$ is the solution to the set of equations

$$\hat{\mathbf{w}}(\hat{\boldsymbol{\gamma}}) = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \mathbf{g}_{hij}(\hat{\boldsymbol{\gamma}}) = \mathbf{0} \quad (2.23)$$

Using a first-order Taylor expansion of $\hat{\mathbf{w}}(\hat{\boldsymbol{\gamma}})$ at $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}$, it follows that

$$\mathbf{0} = \hat{\mathbf{w}}(\hat{\boldsymbol{\gamma}}) \approx \hat{\mathbf{w}}(\boldsymbol{\gamma}) + \frac{\partial \hat{\mathbf{w}}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \quad (2.24)$$

Taking variances on both sides, it further follows that

$$Cov(\hat{\mathbf{w}}(\hat{\boldsymbol{\gamma}})) \approx \frac{\partial \hat{\mathbf{w}}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'} Cov(\hat{\boldsymbol{\gamma}}) \left(\frac{\partial \hat{\mathbf{w}}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'} \right)' \quad (2.25)$$

Thus, provided that (cf. (2.23)) $\frac{\partial^2 \ln L}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = \frac{\partial}{\partial \boldsymbol{\gamma}} \left[\frac{\partial \mathbf{g}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'} \right]$ is a non-singular matrix,

$$Cov(\hat{\boldsymbol{\gamma}}) \approx \left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \right]^{-1} Cov(\hat{\mathbf{w}}(\hat{\boldsymbol{\gamma}})) \left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \right],$$

where $E \left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \right] = -\mathbf{I}_n(\boldsymbol{\gamma})$.

Therefore, an approximate expression for the asymptotic covariance matrix of $\hat{\boldsymbol{\gamma}}$ is given by

$$Cov(\hat{\boldsymbol{\gamma}}) \approx \mathbf{I}_n^{-1}(\boldsymbol{\gamma}) \mathbf{G} \mathbf{I}_n^{-1}(\boldsymbol{\gamma}) \quad (2.26)$$

where $\mathbf{G} = Cov(\hat{\mathbf{w}}(\hat{\boldsymbol{\gamma}}))$.

Using results derived by Fuller (1975) (see also Section 2.7), it follows that, under single stage sampling with replacement (WR) or without replacement (WOR),

$$\mathbf{G} = \sum_{h=1}^H \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (\mathbf{t}_{hi.} - \bar{\mathbf{t}}_{h..}) (\mathbf{t}_{hi.} - \bar{\mathbf{t}}_{h..})' \quad (2.27)$$

where:

- $n_h = \sum_{j=1}^{n_{hj}} m_{hij}$, with m_{hij} the number of cases with identical response patterns within stratum h , cluster i , and USU j . If $f_{hij} = 1$ for all h , then $m_{hij} = 1$ for all h, i and j .
- $f_h = \frac{n_h}{N_h}$, the sampling rate for stratum h .
- $\mathbf{t}_{hij} = \mathbf{g}_{hij}(\hat{\boldsymbol{\gamma}})$ where $\mathbf{g}_{hij}(\hat{\boldsymbol{\gamma}})$ is the h_{ij} -th contribution to the gradient vector $\mathbf{g}(\boldsymbol{\gamma})$ as defined by (2.19).

- $\mathbf{t}_{hi.} = \sum_{j=1}^{m_{hj}} \mathbf{t}_{hij} .$
- $\mathbf{t}_{h..} = \sum_{j=1}^{n_h} \mathbf{t}_{hi.} \quad \bar{\mathbf{t}}_{h..} = \frac{1}{n_h} \sum_{i=1}^{n_h} \mathbf{t}_{hi.}$

Currently, we assume a zero contribution to \mathbf{G} for strata that contain a single PSU (cluster). Alternatively, the collapsing of strata or PSUs is left to the user's discretion (see Section 2.7.3). Additionally, if there is no variable to define clusters, the observations within each stratum are treated as being the primary sampling units.

2.9 References

- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review*, **51**, 279-292.
- Binder, D.A. & Hidirolou, M.A. (1988). Sampling in time. In: P.R. Krishnaiah & C.R. Rao (Eds.). *Handbook of Statistics*, Vol. **6**. Amsterdam: North-Holland, pp. 187-211.
- Biyani, S.H. (1980). On variance estimator in unequal probability sampling, *Proceedings of the Survey Research Methods American Statistical Association*, 634-637.
- Durbin, J. (1967). Design of Multi-Stage Surveys for the Estimation of Sampling Errors, *Applied Statistics*, **XVI**, 152-164.
- Efron, B. (1981). Nonparametric standard errors and confidence intervals, *Canadian Journal of Statistics*, **9**, 139-172.
- Efron, B. (1982). The jackknife, the bootstrap and other resampling plans, CBMS-NSF Regional Conf. *Series in Applied Mathematics*, no. 38.
- Fuller, W.A. (1975). Regression Analysis for Sample Survey. *Sankhya*, Series **C**, **37**, 117-132.
- Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663-685.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley.
- Kish, L., & Frankel, M.R. (1974). Inference from Complex Samples, *Journal of Royal Statistical Society Ser. B*, **36**, 1-37.
- Kovar, J., Rao, J.N.K., & Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates, *Canadian Journal of Statistics*, **16** (Supplement), 25-45.
- Krewski, D., & 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.

Lepkowski, J., & Bowles, J. (1996). Sampling error software for personal computers, *Survey Statistician*, **35**, 10-17.

McCarthy, P.J. (1969). Pseudo-replication: Half samples, *Internat. Stat. Rev.*, **37**, 239-264.

Miller, R.G. (1974). The jackknife: A review, *Biometrika*, **61**, 1-15.

Rao, J.N.K. (1975). Unbiased variance estimation for multistage designs, *Sankhya*, **C37**, 133-139.

Rao, J.N.K., & Scott, A.J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables, *Journal of the American Statistical Association*, **76**, 221-230.

Richards, V., & Freeman, D.H. (1980). A comparison of replicated and pseudo-replicated covariance matrix estimators for the analysis of contingency tables, *Proceedings of the Survey Research Methods, American Statistical Association*, 209-211.

Särndal, C.E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 55-77.

Shapiro, G.M., & Bateman, D.V. (1978). A better alternative to the collapsed stratum variance estimate, *Proceedings of the Survey Research Methods, American Statistical Association*, 451-456.

Skinner, C.J., Holt, D., & Smith, T.M.F. (1989). *Analysis of Complex Surveys*. Chichester: Wiley.

Traat, I., Meister, K., & Söstra, K. (2001). Statistical inference in sampling theory, *Theory of stochastic processes*, Vol. **7(23)**, no. 1-2, 301-316.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Yates, F., & Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size, *Journal of the Royal Statistical Society, Series B*, **15**, 253-261.