properties that cover the most common types of complex sampling designs and nonlinear estimators.

Approximative variance estimators can be used for variance estimation of a nonlinear estimator. These variance estimators are not sampling-design-specific, unlike those for linear estimators. Approximative variance estimators are flexible so that they can be applied for different kinds of nonlinear estimators, including the ratio estimator, under a variety of multi-stage designs covering all the different real sampling designs selected for this book. We use the *linearization method* as the basic approximation method. Alternative methods are based on *sample reuse techniques* such as *balanced half-samples, jackknife* and *bootstrap*. Approximative techniques for variance estimation are available in statistical software products for variance estimation in complex surveys.

Certain simplifying assumptions are often made when using approximative variance estimators. In variance estimation under a multi-stage design, each sampling stage contributes to the total variance. For example, under a two-stage design, an analytical variance estimator of a population total is composed of a sum of the between-cluster and within-cluster variance components as shown in Section 3.2. In the simplest use of the approximation methods, a possible multi-stage design is reduced to a one-stage design, and the clusters are assumed to be drawn with replacement. Variances are then estimated using the between-cluster variation only. In more advanced uses of the approximation techniques, the variation of all the sampling stages can be properly accounted for.

5.3 LINEARIZATION METHOD

Linearization Method for a Nonlinear Estimator

In estimating the variance of a general nonlinear estimator, denoted by $\hat{\theta}$, we adopt a method based on the so-called *Taylor series expansion*. The method is usually called the *linearization* method because we first reduce the original nonlinear quantity to an approximate linear quantity by using the linear terms of the corresponding Taylor series expansion, and then construct the variance formula and an estimator of the variance of this linearized quantity.

Let an s-dimensional parameter vector be denoted by $\mathbf{Y} = (Y_1, \ldots, Y_s)'$ where Y_j are population totals or means. The corresponding estimator vector is denoted by $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_s)'$ where $\hat{Y}_j$ are estimators of Y_j. We consider a nonlinear parameter $\theta = f(\mathbf{Y})$ with a consistent estimator denoted by $\hat{\theta} = f(\hat{\mathbf{Y}})$. A simple example is a subpopulation mean parameter $\theta = \overline{Y} = Y_1/Y_2$ with a ratio estimator $\hat{\theta} = \overline{y} = \hat{Y}_1/\hat{Y}_2 = y/x$, where $y = \sum_{h=1}^{H} \sum_{i=1}^{m_h} y_{hi}$ is the subgroup sample sum of the response variable and $x = \sum_{h=1}^{H} \sum_{i=1}^{m_h} x_{hi}$ is the subgroup sample size, both regarded as random quantities.

Suppose that for the function $f(\mathbf{y})$, continuous second-order derivatives exist in an open sphere containing $\mathbf{Y}$ and $\hat{\mathbf{Y}}$. Using the linear terms of the Taylor series expansion, we have an approximative linearized expression,

$$\hat{\theta} - \theta \doteq \sum_{j=1}^{s} \frac{\partial f(\mathbf{Y})}{\partial y_j}(\hat{Y}_j - Y_j), \tag{5.3}$$

where $\partial f(\mathbf{Y})/\partial y_j$ refers to partial derivation. Using the linearized equation (5.3), the variance approximation of $\hat{\theta}$ can be expressed by

$$V(\hat{\theta}) \doteq V\left(\sum_{j=1}^{s} \frac{\partial f(\mathbf{Y})}{\partial y_j}(\hat{Y}_j - Y_j)\right) = \sum_{j=1}^{s}\sum_{l=1}^{s} \frac{\partial f(\mathbf{Y})}{\partial y_j}\frac{\partial f(\mathbf{Y})}{\partial y_l} V(\hat{Y}_j, \hat{Y}_l), \tag{5.4}$$

where $V(\hat{Y}_j, \hat{Y}_l)$ denote variances and covariances of the estimators $\hat{Y}_j$ and $\hat{Y}_l$. We have hence reduced the variance of a nonlinear estimator $\hat{\theta}$ to a function of variances and covariances of s linear estimators $\hat{Y}_j$. A variance estimator $\hat{v}(\hat{\theta})$ is obtained from (5.4) by substituting the variance and covariance estimators $\hat{v}(\hat{Y}_j, \hat{Y}_l)$ for the corresponding parameters $V(\hat{Y}_j, \hat{Y}_l)$. The resulting variance estimator is a first-order Taylor series approximation where justification for ignoring the remaining higher-order terms is essentially based on practical experience derived from various complex surveys in which the sample sizes have been sufficiently large.

As an example of the linearization method, let us consider further a ratio estimator. The parameter vector is $\mathbf{Y} = (Y_1, Y_2)'$ with the corresponding estimator vector $\hat{\mathbf{Y}} = (\hat{Y}_1, \hat{Y}_2)'$. The nonlinear parameter to be estimated is $\theta = f(\mathbf{Y}) = Y_1/Y_2$, and the corresponding ratio estimator is $\hat{\theta} = f(\hat{\mathbf{Y}}) = \hat{Y}_1/\hat{Y}_2$. The partial derivatives are

$$\partial f(\mathbf{Y})/\partial y_1 = 1/Y_2 \quad \text{and} \quad \partial f(\mathbf{Y})/\partial y_2 = -Y_1/Y_2^2.$$

Hence we have

$$\begin{aligned} V(\hat{\theta}) &\doteq \sum_{j=1}^{2}\sum_{l=1}^{2} \frac{\partial f(\mathbf{Y})}{\partial y_j}\frac{\partial f(\mathbf{Y})}{\partial y_l} V(\hat{Y}_j, \hat{Y}_l) \\ &= \frac{1}{Y_2}\frac{1}{Y_2}V(\hat{Y}_1) + \frac{1}{Y_2}\left(-\frac{Y_1}{Y_2^2}\right)V(\hat{Y}_1, \hat{Y}_2) \\ &\quad + \left(-\frac{Y_1}{Y_2^2}\right)\frac{1}{Y_2}V(\hat{Y}_2, \hat{Y}_1) + \left(-\frac{Y_1}{Y_2^2}\right)\left(-\frac{Y_1}{Y_2^2}\right)V(\hat{Y}_2) \\ &= (1/Y_2^2)(V(\hat{Y}_1) + \theta^2 V(\hat{Y}_2) - 2\theta V(\hat{Y}_1, \hat{Y}_2)) \\ &= \theta^2(Y_1^{-2}V(\hat{Y}_1) + Y_2^{-2}V(\hat{Y}_2) - 2(Y_1Y_2)^{-1}V(\hat{Y}_1, \hat{Y}_2)). \end{aligned} \tag{5.5}$$

Basic principles of the linearization method for variance estimation of a nonlinear estimator under complex sampling are due to Keyfitz (1957) and Tepping (1968). Woodruff (1971) suggested simplified computational algorithms for the approximation by transforming an s-dimensional situation to a one-dimensional case. A good reference for the method is Wolter (1985). The linearization method can also be used for more complex nonlinear estimators such as correlation and regression coefficients. The linearization method is used in most survey analysis software products for variance estimation of ratio estimators and for more complicated nonlinear estimators. We next consider the estimation of the approximative variance of a ratio estimator using the linearization method.

Linearization Method for a Combined Ratio Estimator

A variance estimator of the ratio estimator $\hat{r} = y/x = \sum_{h=1}^{H} \sum_{i=1}^{m_h} y_{hi} / \sum_{h=1}^{H} \sum_{i=1}^{m_h} x_{hi}$ given by (5.1) should, according to equation (5.5), include the following terms: first, a term accounting for cluster-wise variation of the subgroup sample sums y_{hi}, second, a term accounting for cluster-wise variation of the subgroup sample sizes x_{hi}, and finally, a term accounting for joint cluster-wise variation of the sample sums y_{hi} and x_{hi}, i.e. their covariance. A variance estimator of $\hat{r}$ can thus be obtained from equation (5.5) by substituting the estimators $\hat{v}(y)$, $\hat{v}(x)$ and $\hat{v}(y, x)$ for the corresponding variance and covariance terms $V(y)$, $V(x)$ and $V(y, x)$. Hence we have

$$\hat{v}_{des}(\hat{r}) = \hat{r}^2(y^{-2}\hat{v}(y) + x^{-2}\hat{v}(x) - 2(yx)^{-1}\hat{v}(y, x)), \tag{5.6}$$

as the *design-based* variance estimator of $\hat{r}$ based on the linearization method, where $\hat{v}(y)$ is the variance estimator of the subgroup sample sum y, $\hat{v}(x)$ is the variance estimator of the subgroup sample size x, and $\hat{v}(y, x)$ is the covariance estimator of y and x.

The variance estimator (5.6) is consistent if the estimators $\hat{v}(y)$, $\hat{v}(x)$ and $\hat{v}(y, x)$ are consistent. The cluster sample sizes x_{hi} should not vary too much for the reliable performance of the approximation based on the Taylor series expansion. The method can be safely used if the coefficient of variation of x_{hi} is less than 0.2. If the cluster sample sizes are equal, the variance and covariance terms $\hat{v}(x)$ and $\hat{v}(y, x)$ are zero and the variance approximation reduces to $\hat{v}_{des}(\hat{r}) = \hat{v}(y)/x^2$. And for a binary response from simple random sampling with replacement, this variance estimator reduces to the binomial variance estimator $\hat{v}_{des}(\hat{p}) = \hat{v}_{bin}(\hat{p}) = \hat{p}(1 - \hat{p})/x$, where $x = n$, the size of the available sample data set.

The variance estimator (5.6) is a large-sample approximation in that a good variance estimate can be expected if not only a large element-level sample is available but a large number of sample clusters is also present. In the case of a small number of sample clusters, the variance estimator can be unstable; this will be examined in Section 5.7.

Strictly speaking, the variance and covariance estimators in (5.6) depend on the actual sampling design. But assuming that at least two sample clusters are drawn from each stratum and by using the with-replacement assumption, i.e. assuming that clusters are drawn independently of each other, we obtain relatively simple variance and covariance estimators, which can be generally applied for multi-stage stratified epsem samples:

$$\hat{v}(y) = \sum_{h=1}^{H} m_h \hat{s}_{yh}^2, \quad \hat{v}(x) = \sum_{h=1}^{H} m_h \hat{s}_{xh}^2$$

and

$$\hat{v}(y, x) = \sum_{h=1}^{H} m_h \hat{s}_{yxh},$$

where

$$\hat{s}_{yh}^2 = \sum_{i=1}^{m_h} (y_{hi} - y_h/m_h)^2/(m_h - 1),$$

$$\hat{s}_{xh}^2 = \sum_{i=1}^{m_h} (x_{hi} - x_h/m_h)^2/(m_h - 1),$$

and

$$\hat{s}_{yxh} = \sum_{i=1}^{m_h} (y_{hi} - y_h/m_h)(x_{hi} - x_h/m_h)/(m_h - 1). \tag{5.7}$$

Note that by using the with-replacement approximation, only the between-cluster variation is accounted for. Therefore, the corresponding variance estimators underestimate the true variance. This bias is negligible if the stratum-wise first-stage sampling fractions are small, which is the case when there are a large number of population clusters in each stratum (see Section 3.2).

For the estimation of the between-cluster variance, at least two sample clusters are needed. If the sampling design is such that exactly two clusters are drawn from each stratum, the estimators (5.7) can be further simplified:

$$\hat{v}(y) = \sum_{h=1}^{H} (y_{h1} - y_{h2})^2, \quad \hat{v}(x) = \sum_{h=1}^{H} (x_{h1} - x_{h2})^2$$

and

$$\hat{v}(y, x) = \sum_{h=1}^{H} (y_{h1} - y_{h2})(x_{h1} - x_{h2}). \tag{5.8}$$

Covariance-matrix Estimation

The unknown population covariance matrix $\mathbf{V}/n$ of the ratio estimator vector $\hat{\mathbf{p}}$ has u rows and u columns, thus it is a $u \times u$ matrix. $\mathbf{V}/n$ is symmetric such that the lower and upper triangles of the matrix are identical. Variances of the domain ratio estimators are placed on the main diagonal of $\mathbf{V}/n$ and covariances of the corresponding domain ratio estimators on the off-diagonal part of the matrix. There is a total of $u \times (u+1)/2$ distinct parameters in $\mathbf{V}/n$ that need to be estimated.

The variance and covariance estimators $\hat{v}_{des}(\hat{r}_j)$ and $\hat{v}_{des}(\hat{r}_j, \hat{r}_l)$, being respectively the diagonal and off-diagonal elements of a consistent covariance-matrix estimator $\hat{\mathbf{V}}_{des}$ of the asymptotic covariance matrix $\mathbf{V}/n$ of the ratio estimator vector $\hat{\mathbf{r}} = (\hat{r}_1, \ldots, \hat{r}_u)'$, are derived using the linearization method considered in Section 5.3. The variance and covariance estimators of the sample sums y_j and x_j in a variance estimator $\hat{v}_{des}(\hat{r}_j)$ of $\hat{r}_j = y_j/x_j$, and the covariance estimators of the sample sums y_j, y_l, x_j and x_l in the covariance estimators $\hat{v}_{des}(\hat{r}_j, \hat{r}_l)$ of $\hat{r}_j$ and $\hat{r}_l$ in separate domains, are straightforward generalizations of the corresponding variance and covariance estimators given in Section 5.3 for the variance estimator of a single ratio estimator $\hat{r}$. We therefore do not show these formulae.

Like the scalar case, the variance and covariance estimators of $\hat{r}_j$ and $\hat{r}_l$ are based on the with-replacement assumption and the variation accounted for is the between-cluster variation. This causes bias in the estimates, but the bias can be assumed to be negligible if the first-stage sampling fraction is small.

The variance and covariance estimators of y_j, x_j, y_l and x_l are finally collected into the corresponding $u \times u$ covariance-matrix estimators $\hat{\mathbf{V}}_{yy}$, $\hat{\mathbf{V}}_{xx}$ and $\hat{\mathbf{V}}_{yx}$. Using these estimators, the design-based covariance-matrix estimator of $\hat{\mathbf{r}}$ based on the linearization method is given by

$$\begin{aligned}\hat{\mathbf{V}}_{des} = \operatorname{diag}(\hat{\mathbf{r}})(&\mathbf{Y}^{-1}\hat{\mathbf{V}}_{yy}\mathbf{Y}^{-1} + \mathbf{X}^{-1}\hat{\mathbf{V}}_{xx}\mathbf{X}^{-1} \\ &- \mathbf{Y}^{-1}\hat{\mathbf{V}}_{yx}\mathbf{X}^{-1} - \mathbf{X}^{-1}\hat{\mathbf{V}}_{xy}\mathbf{Y}^{-1})\operatorname{diag}(\hat{\mathbf{r}}),\end{aligned} \tag{5.35}$$

where

$\operatorname{diag}(\hat{\mathbf{r}}) = \operatorname{diag}(\hat{r}_1, \ldots, \hat{r}_u) = \operatorname{diag}(y_1/x_1, \ldots, y_u/x_u)$
$\mathbf{Y} = \operatorname{diag}(\mathbf{y}) = \operatorname{diag}(y_1, \ldots, y_u)$
$\mathbf{X} = \operatorname{diag}(\mathbf{x}) = \operatorname{diag}(x_1, \ldots, x_u)$
$\hat{\mathbf{V}}_{yy}$ is the covariance-matrix estimator of the sample sums y_j and y_l
$\hat{\mathbf{V}}_{xx}$ is the covariance-matrix estimator of the sample sums x_j and x_l
$\hat{\mathbf{V}}_{yx}$ is the covariance-matrix estimator of the sums y_j and x_l, and
$\hat{\mathbf{V}}_{xy} = \hat{\mathbf{V}}'_{yx}$

and the operator 'diag' generates a diagonal matrix with the elements of the corresponding vector as the diagonal elements and with off-diagonal elements equal to zero. Note that in a linear case, all elements of the covariance-matrix estimators $\hat{\mathbf{V}}_{xx}$, $\hat{\mathbf{V}}_{yx}$ and $\hat{\mathbf{V}}_{xy}$ are zero.

In the estimation of the elements of $\hat{\mathbf{V}}_{des}$, at least two clusters are assumed to be drawn with replacement from each of the H strata. In the special case of $m_h = 2$ clusters routinely used in survey sampling, the estimators can be simplified in a manner similar to that done in Section 5.3.

As a simple example, let the number of domains be $u = 2$. The elements of the covariance-matrix estimator

$$\hat{\mathbf{V}}_{des} = \begin{bmatrix} \hat{v}_{des}(\hat{r}_1) & \hat{v}_{des}(\hat{r}_1, \hat{r}_2) \\ \hat{v}_{des}(\hat{r}_2, \hat{r}_1) & \hat{v}_{des}(\hat{r}_2) \end{bmatrix}$$

are the following:
Variance estimator:

$$\hat{v}_{des}(\hat{r}_j) = \hat{r}_j^2(y_j^{-2}\hat{v}(y_j) + x_j^{-2}\hat{v}(x_j) - 2(y_j x_j)^{-1}\hat{v}(y_j, x_j)), \quad j = 1, 2.$$

Covariance estimator:

$$\begin{aligned}\hat{v}_{des}(\hat{r}_1, \hat{r}_2) = \hat{r}_1\hat{r}_2((y_1y_2)^{-1}\hat{v}(y_1, y_2) + (x_1x_2)^{-1}\hat{v}(x_1, x_2) \\ - (y_1x_2)^{-1}\hat{v}(y_1, x_2) - (y_2x_1)^{-1}\hat{v}(y_2, x_1)).\end{aligned}$$

The estimator $\hat{v}_{des}(\hat{r}_2, \hat{r}_1)$ is equal to $\hat{v}_{des}(\hat{r}_1, \hat{r}_2)$ because of symmetry of $\hat{\mathbf{V}}_{des}$. If the estimators $\hat{r}_j$ are taken as linear estimators, then the denominators x_j are assumed fixed. In this case, the variance and covariance estimates $\hat{v}(x_j)$ and $\hat{v}(y_j, x_j)$ are zero, and $\hat{v}_{des}(\hat{r}_j) = \hat{v}(y_j)/x_j^2$. And for a binary response in the binomial case, this estimator reduces to $\hat{v}_{bin}(\hat{p}_j) = \hat{p}_j(1 - \hat{p}_j)/n_j$.

It is important to note that $\hat{\mathbf{V}}_{des}$ is distribution-free so that it requires no specific distributional assumptions about the sampled observations. This allows an estimate $\hat{\mathbf{V}}_{des}$ to be nondiagonal. The nondiagonality of $\hat{\mathbf{V}}_{des}$ is because the ratio estimators $\hat{r}_j$ and $\hat{r}_l$ from distinct domains can have nonzero correlations. In contrast, the binomial covariance-matrix estimators considered in this section have zero correlation by definition.

One source of nonzero correlation of the estimators $\hat{r}_j$ and $\hat{r}_l$ from separate domains comes from the clustering of the sample. Varying degrees of correlation can be expected depending on the type of the domains. If the domains cut smoothly across the sample clusters, distinct members in a given sample cluster may fall in separate domains j and l such as cross-classes like demographic or related factors. Large correlations can then be expected if the clustering effect is noticeable. In contrast, if the domains are totally segregated in such a way that all members of a given sample cluster fall in the same domain, zero correlations of distinct estimates $\hat{r}_j$ and $\hat{r}_l$ are obtained. This happens if the predictors used in forming the domains are cluster-specific unlike cross-classes where factors are essentially individual-specific. If, for example, households are clusters, typical cluster-specific factors are net income of the household and family size, whereas age and sex of a family member are individual-specific. Mixed-type domains, often met in practice, are intermediate, so that nonzero correlations are present in some dimensions of the table with zero correlations in the others.

another. Obviously, in self-weighting samples both approaches should yield equal design-effect estimates.

It should be noted that, in the design-effects matrix estimator (5.37) only the contribution of the clustering is accounted for, because a binomial covariance-matrix estimator of the consistent weighted proportion estimator vector is used. By using in (5.37) a binomial covariance-matrix estimator of the unweighted proportion estimator vector instead of that of the weighted proportion estimator vector, all the contributions of complex sampling on covariance-matrix estimation are reflected, such as unequal inclusion probabilities, clustering and adjustment for nonresponse. Obviously, both approaches give similar design-effect matrix estimates when working with self-weighting samples. If adopting as a rule the use of a consistent proportion estimator $\hat{\mathbf{p}}$, then working with weighted observations, and thus with (5.37) would be reasonable. Then, the crucial role of adjusting for the clustering effect in the analysis of complex surveys would also be emphasized. However, the calculation of the deff matrix estimate by using both versions of the binomial covariance-matrix estimate can be useful in assessing the contribution of weighting to the design effects.

Example 5.5

Covariance-matrix and design-effects matrix estimation with the linearization method. Using the OHC Survey data we carry out a detailed calculation of the covariance-matrix estimate $\hat{\mathbf{V}}_{des}$ of a proportion estimate $\hat{\mathbf{p}}$ of the binary response PHYS (physical health hazards of work), and of a mean estimate $\overline{\mathbf{y}}$ of the continuous response PSYCH (the first standardized principal component of nine psychic symptoms), in the simple case of $u = 2$ domains formed by the variable sex. $\hat{\mathbf{V}}_{des}$ is thus a 2×2 matrix, and the domains are of a cross-class type. A part of the data set needed for the covariance-matrix estimation is displayed in Table 5.9. Note that these data are cluster-level, consisting of $m = 250$ clusters in five strata. Thus, the degrees of freedom $f = 245$. The employee-level sample size is $n = 7841$.

The ratio estimator is $\hat{\mathbf{r}} = (\hat{r}_1, \hat{r}_2)' = (y_1/x_1, y_2/x_2)'$, where $\hat{r}_1$ and $\hat{r}_2$ are given by (5.34). For the binary response PHYS, we denote the ratio estimator as $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2)'$, and for the continuous response PSYCH $\overline{\mathbf{y}} = (\overline{y}_1, \overline{y}_2)'$. The following figures for PHYS are calculated from Table 5.9.

Sums of the cluster-level sample sums $y_{jhi} (= y_{ji})$ and $x_{jhi} (= x_{ji})$:

$$\hat{n}_{11} = y_1 = 2061 \quad \text{and} \quad \hat{n}_1 = x_1 = 4485 \text{ (males)},$$
$$\hat{n}_{21} = y_2 = 650 \quad \text{and} \quad \hat{n}_2 = x_2 = 3356 \text{ (females)}.$$

Proportion estimates for PHYS, i.e. the elements of $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2)'$:

$$\hat{p}_1 = y_1/x_1 = 2061/4485 = 0.4595 \text{ (males)},$$

Table 5.9 Cluster-level sample sums y_{1i} (males) and y_{2i} (females) of the response variables PHYS and PSYCH with the corresponding cluster sample sizes x_{1i} (males) and x_{2i} (females) in sample clusters $i = 1, \ldots, 250$ in two domains formed by sex (the OHC Survey).

Stratum	Cluster	PHYS		PSYCH			
h	i	y_{1i}	y_{2i}	y_{1i}	y_{2i}	x_{1i}	x_{2i}
2	1	11	3	−0.1434	−0.0322	36	22
2	2	18	4	−0.1925	0.1867	57	21
2	3	4	5	0.0045	0.3674	9	15
2	4	2	2	0.7135	−0.3679	12	15
2	5	1	0	−0.1681	0.1235	27	8
2	6	1	0	−0.2673	0.1504	19	21
2	7	9	4	0.0099	0.2099	23	27
2	8	4	2	0.3681	0.0155	16	31
2	9	0	0	−0.5033	0.0755	6	6
2	10	3	0	−0.3176	−0.2516	8	8
2	11	2	7	0.9746	0.1903	6	67
2	12	7	3	−0.3361	0.5572	22	31
2	13	4	1	−0.2329	−0.2181	9	7
2	14	0	0	−0.2032	0.5893	13	16
2	15	1	23	0.4137	0.2565	4	56
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
6	245	14	2	0.1984	−0.4271	23	7
6	246	2	1	−0.1049	0.3905	7	7
6	247	4	7	−0.2961	0.5018	7	13
6	248	0	1	−0.8073	0.9278	3	9
6	249	2	0	0.0006	−0.3484	16	13
6	250	13	1	−0.1273	−0.1466	26	4
Total sample		2061	650	−26.7501	33.7983	4485	3356

and

$$\hat{p}_2 = y_2/x_2 = 650/3356 = 0.1937 \text{ (females)}.$$

We next construct the diagonal 2×2 matrices $\text{diag}(\hat{\mathbf{p}})$, $\mathbf{Y}$ and $\mathbf{X}$ for the calculation of the estimate $\hat{\mathbf{V}}_{des}$ for the PHYS proportion estimator $\hat{\mathbf{p}}$:

$$\text{diag}(\hat{\mathbf{p}}) = \begin{bmatrix} 0.4595 & 0 \\ 0 & 0.1937 \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} 2061 & 0 \\ 0 & 650 \end{bmatrix}$$

and

$$\mathbf{X} = \begin{bmatrix} 4485 & 0 \\ 0 & 3356 \end{bmatrix}.$$

The covariance-matrix estimates $\hat{\mathbf{V}}_{yy}$, $\hat{\mathbf{V}}_{xx}$ and $\hat{\mathbf{V}}_{yx}$, also obtained from the cluster-level data displayed in Table 5.9, are the following:

$$\hat{\mathbf{V}}_{yy} = \begin{bmatrix} 15\,722.50 & -130.45 \\ -130.45 & 3261.71 \end{bmatrix},$$

$$\hat{\mathbf{V}}_{xx} = \begin{bmatrix} 34\,560.23 & -7315.43 \\ -7315.43 & 34\,099.04 \end{bmatrix},$$

and

$$\hat{\mathbf{V}}_{yx} = \begin{bmatrix} 18\,973.88 & -5907.69 \\ -1098.11 & 6051.14 \end{bmatrix} = \hat{\mathbf{V}}'_{xy}.$$

By using these matrices we finally calculate for PHYS proportions the covariance-matrix estimate $\hat{\mathbf{V}}_{des}$ given by (5.35). Hence we have

$$\hat{\mathbf{V}}_{des} = \begin{bmatrix} \hat{v}_{des}(\hat{p}_1) & \hat{v}_{des}(\hat{p}_1, \hat{p}_2) \\ \hat{v}_{des}(\hat{p}_2, \hat{p}_1) & \hat{v}_{des}(\hat{p}_2) \end{bmatrix} = 10^{-4} \begin{bmatrix} 2.775 & 0.576 \\ 0.576 & 1.951 \end{bmatrix}.$$

For example, using the estimates calculated, the variance estimate $\hat{v}_{des}(\hat{p}_1)$ is obtained as

$$\begin{aligned} \hat{v}_{des}(\hat{p}_1) = 0.4595^2 \times (2061^{-2} \times 15\,722.50 + 4485^{-2} \times 34\,560.23 \\ - 2 \times (2061 \times 4485)^{-1} \times 18\,973.88) = 0.2775 \times 10^{-3}. \end{aligned}$$

Correlation of $\hat{p}_1$ and $\hat{p}_2$ is 0.25, which is quite large and indicates that the domains actually constitute cross-classes. The condition number of $\hat{\mathbf{V}}_{des}$ is $\text{cond}(\hat{\mathbf{V}}_{des}) = 1.9$, indicating stability of the estimate owing to a large f and small u.

For PSYCH, the following figures are calculated from Table 5.9.
Sums of the cluster-level sample sums y_{jhi} and x_{jhi}:

$$y_1 = -26.7501 \quad \text{and} \quad x_1 = 4485 \text{ (males)},$$
$$y_2 = \quad 33.7983 \quad \text{and} \quad x_2 = 3356 \text{ (females)}.$$

Mean estimates for PSYCH, i.e. the elements of $\overline{\mathbf{y}} = (\overline{y}_1, \overline{y}_2)'$:

$$\overline{y}_1 = y_1/x_1 = -0.1008 \text{ (males)},$$

and

$$\overline{y}_2 = y_2/x_2 = 0.1347 \text{ (females)}.$$

The diagonal 2×2 matrices diag($\overline{\mathbf{y}}$), $\mathbf{Y}$ and $\mathbf{X}$ are constructed in the same way as for PHYS. The covariance-matrix estimate $\hat{\mathbf{V}}_{xx}$ is equal to that for PHYS, and the covariance-matrix estimates $\hat{\mathbf{V}}_{yy}$ and $\hat{\mathbf{V}}_{yx}$ are:

$$\hat{\mathbf{V}}_{yy} = \begin{bmatrix} 6765.34 & 1036.34 \\ 1036.34 & 6585.20 \end{bmatrix},$$

$$\hat{\mathbf{V}}_{yx} = \begin{bmatrix} -3139.98 & 2129.01 \\ -2051.46 & 2259.73 \end{bmatrix} = \hat{\mathbf{V}}'_{xy}.$$

By using these matrices we calculate for PSYCH means the covariance-matrix estimate $\hat{\mathbf{V}}_{des}$:

$$\hat{\mathbf{V}}_{des} = \begin{bmatrix} \hat{v}_{des}(\overline{y}_1) & \hat{v}_{des}(\overline{y}_1, \overline{y}_2) \\ \hat{v}_{des}(\overline{y}_2, \overline{y}_1) & \hat{v}_{des}(\overline{y}_2) \end{bmatrix} = 10^{-4} \begin{bmatrix} 3.223 & 0.427 \\ 0.427 & 5.856 \end{bmatrix}.$$

Results from the design-based covariance-matrix estimation for PHYS proportions and PSYCH means including the standard-error estimates $\text{s.e}_{des}(\hat{r}_j)$ are displayed below.

		PHYS		PSYCH		
j	Domain	$\hat{p}_j$	$\text{s.e}_{des}(\hat{p}_j)$	$\overline{y}_j$	$\text{s.e}_{des}(\overline{y}_j)$	$\hat{n}_j$
1	Males	0.460	0.0167	−0.1008	0.0180	4485
2	Females	0.194	0.0140	0.1347	0.0242	3356
Total sample		0.346	0.0144	0.0000	0.0158	7841

Variance and covariance estimates $\hat{\mathbf{V}}_{yy}$, $\hat{\mathbf{V}}_{xx}$ and $\hat{\mathbf{V}}_{yx}$ can be calculated using the cluster-level data set displayed in Table 5.9 by suitable software for correlation analysis. The matrix operations in the formula of $\hat{\mathbf{V}}_{des}$ can be executed by any suitable software for matrix algebra. In practice, however, it is convenient to estimate $\hat{\mathbf{V}}_{des}$ using an element-level data set using appropriate software for survey analysis. Generally, in the case of u domains formed by several categorical predictors, a linear ANOVA model can be used by fitting, with an appropriate sampling design option, for the response variable, a full-interaction model excluding the intercept. The model coefficients are then equal to the domain proportion or mean estimates, and the covariance-matrix estimate of the model coefficients provides the covariance-matrix estimate $\hat{\mathbf{V}}_{des}$ of the proportions or means.

We next calculate the design-effects matrix. For this, a binomial covariance-matrix estimate is needed.

For PHYS, by computing the elements of the binomial covariance-matrix estimate

$$\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}}) = \begin{bmatrix} \hat{v}_{bin}(\hat{p}_1) & 0 \\ 0 & \hat{v}_{bin}(\hat{p}_2) \end{bmatrix} = \begin{bmatrix} \hat{p}_1(1-\hat{p}_1)/\hat{n}_1 & 0 \\ 0 & \hat{p}_2(1-\hat{p}_2)/\hat{n}_2 \end{bmatrix}$$

of the proportion vector $\hat{\mathbf{p}}$ we obtain

$$\hat{p}_1(1-\hat{p}_1)/\hat{n}_1 = 0.4595(1-0.4595)/4485 = 0.0000554 \text{ (males)},$$

and

$$\hat{p}_2(1-\hat{p}_2)/\hat{n}_2 = 0.1937(1-0.1937)/3356 = 0.0000465 \text{ (females)}.$$

Inserting these variance estimates in $\hat{\mathbf{V}}_{bin}$ we have

$$\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}}) = 10^{-4}\begin{bmatrix} 0.554 & 0 \\ 0 & 0.465 \end{bmatrix}.$$

It is important to note that the covariance-matrix estimate $\hat{\mathbf{V}}_{bin}$ is diagonal because the proportion estimates $\hat{p}_1$ and $\hat{p}_2$ are assumed to be uncorrelated. The effect of clustering is not accounted for, even in the variance estimates, in the estimate $\hat{\mathbf{V}}_{bin}$. Therefore, with positive intra-cluster correlation, the binomial variance estimates $\hat{v}_{bin}(\hat{p}_j)$ tend to be underestimates of the corresponding variances. This appears when calculating the design-effects matrix estimate $\hat{\mathbf{D}} = \hat{\mathbf{V}}_{bin}^{-1}\hat{\mathbf{V}}_{des}$ of the estimate $\hat{\mathbf{p}}$:

$$\begin{aligned}\hat{\mathbf{D}}(\hat{\mathbf{p}}) &= \begin{bmatrix} 18\,058.295 & 0 \\ 0 & 21\,489.421 \end{bmatrix} \times 10^{-4}\begin{bmatrix} 2.775 & 0.576 \\ 0.576 & 1.951 \end{bmatrix} \\ &= \begin{bmatrix} 5.01 & 1.04 \\ 1.24 & 4.19 \end{bmatrix}.\end{aligned}$$

The design-effect estimates $\hat{d}_j$ on the diagonal of $\hat{\mathbf{D}}$ are thus

$$\hat{d}(\hat{p}_1) = \hat{v}_{des}(\hat{p}_1)/\hat{v}_{bin}(\hat{p}_1) = 0.0002775/0.0000554 = 5.01 \text{ (males)},$$

and

$$\hat{d}(\hat{p}_2) = \hat{v}_{des}(\hat{p}_2)/\hat{v}_{bin}(\hat{p}_2) = 0.0001951/0.0000465 = 4.19 \text{ (females)}.$$

These estimates are quite large, indicating a strong clustering effect for the response PHYS. This results in severe underestimation of standard errors of the estimates $\hat{p}_j$ when the binomial covariance-matrix estimate $\hat{\mathbf{V}}_{bin}$ is used. In addition to the design-effect estimates, the eigenvalues of the design-effect matrix, i.e. the generalized design effects, can be calculated. These are $\hat{\delta}_1 = 5.81$ and $\hat{\delta}_2 = 3.39$. It may be noted that the sum of the design-effect estimates is 9.20, which is equal to the sum of the eigenvalues. The mean of the design-effect estimates is 4.60, which indicates a strong average clustering effect over the sex groups. However, the mean is noticeably smaller than the overall design-effect estimate $\hat{d} = 7.2$ for the proportion estimate $\hat{p}$ calculated from the whole sample. This is due to

the property of design-effect estimates that, when compared against the overall design-effect estimate, they tend to get smaller in cross-class-type domains.

Estimation results for PHYS proportions are collected below.

j	Domain	$\hat{p}_j$	s.e_{des}	s.e_{bin}	$\hat{d}_j$	$\hat{n}_j$
1	Males	0.460	0.0167	0.0074	5.01	4485
2	Females	0.194	0.0140	0.0068	4.19	3356
Total sample		0.346	0.0144	0.0054	7.17	7841

5.8 CHAPTER SUMMARY AND FURTHER READING

Summary

Proper estimation of the variance of a ratio estimator is important in the analysis of complex surveys. First, variance estimates are needed to derive standard errors and confidence intervals for nonlinear estimators such as a ratio estimator. The estimation of the variance of ratio mean and ratio proportion estimators was carried out under an epsem two-stage stratified cluster-sampling design, where the sample data set was assumed self-weighting so that adjustment for nonresponse was not necessary. The demonstration data set from the modified sampling design of the Mini-Finland Health Survey (MFH Survey) fulfilled these conditions.

A ratio-type estimator $\hat{r} = y/x$ was examined for the estimation of the subpopulation mean and proportion in the important case of a subgroup of the sample whose size x was not fixed by the sampling design. Therefore, the denominator quantity x in $\hat{r}$ is a random variable, involving its own variance and covariance with the numerator quantity y. In addition to the variance of y, these variance and covariance terms contributed to the variance estimator of a ratio estimator calculated with the linearization method. This method was considered in depth because of its wide applicability in practice and popularity in software products for survey analysis.

We also introduced alternative methods for variance estimation of a ratio estimator based on sample reuse methods. The techniques of balanced half-samples (BRR) and jackknife (JRR) are traditional sample reuse methods, but the bootstrap (BOOT) has been applied for complex surveys only recently. Being computer-intensive, they differ from the linearization technique but are, as such, readily applicable for different kinds of nonlinear estimators. With-replacement sampling of clusters was assumed for all the approximation methods. With this assumption, the variability of a ratio estimate was evaluated using the between-cluster variation only, leading to relatively simple variance estimators. The design effect was used extensively as a measure of the contribution of the clustering on