VARIANCE ESTIMATION FOR EXPERIMENTS EMBEDDED IN COMPLEX SAMPLING SCHEMES

Jan van den Brakel, Statistics Netherlands, David Binder, Statistics Canada
Jan van den Brakel, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands

Keywords: completely randomized designs, design-based analysis, imputation, survey methodology.

1 Introduction

Field experiments embedded in ongoing sample surveys are highly appropriate to investigate possible improvements of a sample survey process. In many practical situations such experiments are aimed to test effects of alternative survey methodologies on the outcomes of a current survey. In Van den Brakel and Renssen (1998) a series of such field experiments are described. Fienberg and Tanur (1988) discussed how to take advantage of the parallels between the principals of design and analysis of experiments and sampling theory, in order to improve the efficiency of experiments embedded in sample surveys.

The typical situation considered here, is an experiment designed to compare the impact of K different survey implementations, or treatments, on the estimates of the finite population parameters of a current survey. To this end a sample, drawn from a finite population, is randomly divided into K subsamples according to some experimental design. Each subsample is assigned to one of the K treatments. The analysis of such embedded experiments generally serves two purposes. First is the estimation of finite population parameters obtained under the alternative survey implementations. This enables the measurement of the effect of alternative approaches on the main estimates of the survey. Secondly, we may wish to test hypotheses concerning the differences between the estimated population parameters obtained under the different survey implementations. Statistical methods traditionally used in the analysis of experiments are model dependent and typically require identically and independently distributed (IID) observations. Since in embedded experiments, experimental units are selected by some complex sampling design from a finite population the assumption of IID data is generally violated. Moreover, if an experiment is embedded in a complex sampling scheme, it is not always obvious how the analysis results concerning treatment effects obtained in a model-based analysis procedure are related to the finite population parameters as defined in the sample survey. This complicates the interpretation of the results obtained in a model-based procedure. A natural approach for the analysis of such embedded experiments is to formulate a hypothesis, which concerns the differences between the finite population parameters observed under the different survey implementations. Based on the K subsamples, a design unbiased estimator for these K population parameters and the covariance matrix of these K population parameter estimates can be derived under both the randomization mechanism of the sampling design and the experimental design. As a result, a design-based Wald statistic is obtained to test hypotheses. Van den Brakel and Renssen (1998, 2000) developed such a design-based theory for the analysis of embedded completely randomized designs (CRD's) and randomized block designs (RBD's).

Since the K subsamples are drawn without replacement from a finite population, there is a nonzero design covariance between the different subsample estimates. Design-unbiased estimators for these covariance terms require paired observations of the target parameter under the different treatments obtained at each experimental unit. Since each individual is assigned to either one of the K treatments, such paired observations are not available. In Van den Brakel and Renssen (2000), an estimator is derived for the covariance matrix of the $K-1$ contrasts between the K parameter estimates which is design-unbiased under specific measurement error models. In this paper an alternative approximately design-unbiased estimator for the covariance matrix of the K parameter estimates is derived for a CRD, using an imputation technique for the missing paired observations. For more technical details and derivations of the results presented in this paper, we refer to Van den Brakel and Binder (2000).

2 Embedding experiments in ongoing surveys

Assume we have a finite population U with N units. Consider an experiment, embedded in an ongoing survey aimed to compare the impact of $K-1$ alternative survey methodologies (treatments) with respect to the standard approach of the current survey on the parameter estimates of this survey. Let $y_{i(k)}$ denote the value of the $i-$th unit under treatment k, for $i = 1, \ldots, N$ and $k = 1, \ldots, K$. Then $\bar{Y}_{(k)}$ and $Y_{(k)}$ can be defined as the population mean and total observed under treatment k. Let $\bar{\mathbf{Y}} = (\bar{Y}_{(1)}, \ldots, \bar{Y}_{(K)})^t$, the K vector containing the population treatment means. The objective of this experiment is to test the hypothesis

$$\begin{aligned} H_0: &\quad \mathbf{C}\bar{\mathbf{Y}} = \mathbf{0}, \\ H_1: &\quad \mathbf{C}\bar{\mathbf{Y}} \neq \mathbf{0}. \end{aligned} \tag{1}$$

Here $\mathbf{0}$ denotes the $K-1$ vector with each element equal to zero and $\mathbf{C}$ a $(K-1)\times K$ contrast matrix, for example $(\mathbf{j} \mid -\mathbf{I})$ where $\mathbf{j}$ denotes a $K-1$ vector with each element one and $\mathbf{I}$ a $(K-1)\times(K-1)$ identity matrix. Let $\hat{\bar{\mathbf{Y}}}$ denote a design unbiased estimator for $\bar{\mathbf{Y}}$, $\boldsymbol{\Sigma}$ the covariance matrix of $\hat{\bar{\mathbf{Y}}}$, and $\hat{\boldsymbol{\Sigma}}$ a design unbiased consistent estimator for $\boldsymbol{\Sigma}$. Hypothesis (1) can be tested with the design-based Wald statistic

$$W = \hat{\bar{\mathbf{Y}}}^t \mathbf{C}^t \left(\mathbf{C}\hat{\boldsymbol{\Sigma}}\mathbf{C}^t\right)^{-1} \mathbf{C}\hat{\bar{\mathbf{Y}}}. \tag{2}$$

If a limit theorem holds such that $\hat{\bar{\mathbf{Y}}}$ is asymptotically multivariate normally distributed with mean $\bar{\mathbf{Y}}$ and covariance matrix $\boldsymbol{\Sigma}$, then the Wald statistic is, under the null hypothesis, asymptotically chi-squared distributed with $K-1$ degrees of freedom. If the sample s is drawn by means of simple random sampling without replacement and the randomization of the n elements of s to the K treatments is accomplished by means of simple random sampling without replacement, i.e. by means of a CRD, then Lehmann (1975, appendix 8) gives sufficient conditions under which $\hat{\bar{\mathbf{Y}}} \rightarrow \mathcal{N}(\bar{\mathbf{Y}}, \boldsymbol{\Sigma})$. Under more complex sampling schemes, the limit distribution of $\hat{\bar{\mathbf{Y}}}$ will generally be unknown. In such situations, it is usually assumed that a limit theorem holds such that $\hat{\bar{\mathbf{Y}}} \rightarrow \mathcal{N}(\bar{\mathbf{Y}}, \boldsymbol{\Sigma})$.

3 Estimation of treatment effects

To test hypothesis (1) a sample s of size n is drawn from a finite population U of size N under a possibly complex sampling design with first and second order inclusion probabilities π_i and π_{ij}. According to the experimental design, this sample is randomly divided into K subsamples s_k of size n_k. All the elements of the $k-$th subsample undergo treatment k, so that we observe only the values of $y_{i(k)}$ for units in the $k-$th group. Also the randomization mechanism of the experimental design can be described by means of first and second order inclusion probabilities. Let $\pi_{i|s}^{(k)}$ denote the conditional probability that the $i-$th unit is in treatment group k, given that the sample s is selected and $\pi_{ij}^{(kl)}$ the conditional joint inclusion probability that the $i-$th unit is in treatment group k and the $j-$th unit in treatment group l. Consider for example a CRD. Since the randomization mechanism of a CRD comes down to simple random sampling without replacement of K subsamples of size n_k from the sample s of size n, we have the following conditional first and second order inclusion probabilities: $\pi_i^{(k)} = \pi_{ii}^{(kk)} = n_k/n$, $\pi_{ii}^{(kl)} = 0$, $\pi_{ij}^{(kk)} = (n_k(n_k-1))/(n(n-1))$, and $\pi_{ij}^{(kl)} = (n_k n_l)/(n(n-1))$, $i \neq j$. Since each subsample can be considered as a two-phase survey sample a design-unbiased estimator for $\bar{Y}_{(k)}$ is given by the two-phase estimator

$$\hat{\bar{Y}}_{(k)} = \frac{1}{N} \sum_{i=1}^{n_k} \frac{y_{i(k)}}{\pi_i \pi_{i|s}^{(k)}}. \tag{3}$$

For the special case of a CRD, $\pi_{i|s}^{(k)} = n_k/n$.

4 Covariance matrix of $\hat{\bar{\mathbf{Y}}}$

In this section the covariance matrix, $\boldsymbol{\Sigma}$, of the estimated population treatment means $\hat{\bar{\mathbf{Y}}}$ is derived. The following expression for the variance of $\hat{\bar{Y}}_{(k)}$ can be derived by conditioning on the realization of the sample s:

$$\begin{aligned} \mathrm{Var}(\hat{\bar{Y}}_{(k)}) = &\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \triangle_{ij} \check{y}_{i(k)} \check{y}_{j(k)} \\ &+ \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} \triangle_{ij|s}^{(kk)} \check{y}_{i(k)} \check{y}_{j(k)}}{\pi_{i|s}^{(k)} \pi_{j|s}^{(k)}}, \end{aligned} \tag{4}$$

where $\check{y}_{i(k)} = y_{i(k)}/\pi_i$, $\triangle_{ij} = \pi_{ij} - \pi_i \pi_j$ and, $\triangle_{ij|s}^{(kl)} = \pi_{ij|s}^{(kl)} - \pi_{i|s}^{(k)} \pi_{j|s}^{(l)}$. The first component is the variance of the sampling scheme of the conditional expectation of the experimental design and can be interpreted as the design variance of a one phase sampling scheme. The second component is the expectation with respect to the sampling design of the conditional variance of the experimental design. In an equivalent

way, it can be shown that the design covariance between $\hat{\bar{Y}}_{(k)}$ and $\hat{\bar{Y}}_{(l)}$ is given by:

$$\mathrm{Cov}(\hat{\bar{Y}}_{(k)},\hat{\bar{Y}}_{(l)}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\triangle_{ij}\check{y}_{i(k)}\check{y}_{j(l)} + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}\triangle_{ij|s}^{(kl)}\check{y}_{i(k)}\check{y}_{j(l)}}{\pi_{i|s}^{(k)}\pi_{j|s}^{(l)}}. \quad (5)$$

In the special case of a CRD, it follows from (4) that

$$\mathrm{Var}(\hat{\bar{Y}}_{(k)}) = \frac{n}{N^2(n-1)n_k}\Big[(n_k-1)\sum_{i=1}^{N}\sum_{j=1}^{N}\triangle_{ij}\check{y}_{i(k)}\check{y}_{j(k)} + (n-n_k)\left(\sum_{i=1}^{N}\pi_i\check{y}_{i(k)}^2 - \frac{1}{n}Y_{(k)}^2\right)\Big]. \quad (6)$$

and from (5) that

$$\mathrm{Cov}(\hat{\bar{Y}}_{(k)},\hat{\bar{Y}}_{(l)}) = \frac{n}{N^2(n-1)}\Big[\sum_{i=1}^{N}\sum_{j=1}^{N}\triangle_{ij}\check{y}_{i(k)}\check{y}_{j(l)} - \sum_{i=1}^{N}\pi_i\check{y}_{i(k)}\check{y}_{i(l)} + \frac{1}{n}Y_{(k)}Y_{(l)}\Big]. \quad (7)$$

5 Estimation of the covariance matrix

In this section an estimator for the covariance matrix, $\boldsymbol{\Sigma}$, is derived. Since there are only n_k observations obtained under treatment k, a design unbiased estimator for $\mathrm{Var}(\hat{\bar{Y}}_{(k)})$, is provided by

$$\hat{\mathrm{Var}}(\hat{\bar{Y}}_{(k)}) = \frac{1}{N^2}\sum_{i=1}^{n_k}\sum_{j=1}^{n_k}\frac{\check{\triangle}_{ij}\check{y}_{i(k)}\check{y}_{j(k)}}{\pi_{ij|s}^{(kk)}} + \frac{1}{N^2}\sum_{i=1}^{n_k}\sum_{j=1}^{n_k}\frac{\check{\triangle}_{ij|s}^{(kk)}\check{y}_{i(k)}\check{y}_{j(k)}}{\pi_{i|s}^{(k)}\pi_{j|s}^{(k)}}, \quad (8)$$

where $\check{\triangle}_{ij} = \triangle_{ij}/\pi_{ij}$ and $\check{\triangle}_{ij|s}^{(kk)} = \triangle_{ij|s}^{(kk)}/\pi_{ij|s}^{(kk)}$. For the special case of a CRD we have

$$\hat{\mathrm{Var}}(\hat{\bar{Y}}_{(k)}) = \frac{n(n-1)}{N^2n_k(n_k-1)}\sum_{i=1}^{n_k}\sum_{j=1}^{n_k}\check{\triangle}_{ij}\check{y}_{i(k)}\check{y}_{j(k)} + \frac{n(n-n_k)}{N^2n_k(n_k-1)}\left(\sum_{i=1}^{n_k}\pi_i\check{y}_{i(k)}^2 - \frac{1}{n_k}\left(\sum_{i=1}^{n_k}\check{y}_{i(k)}\right)^2\right). \quad (9)$$

We now turn our attention to the derivation of an estimator for the covariance between $\hat{\bar{Y}}_{(k)}$ and $\hat{\bar{Y}}_{(l)}$. Since $\pi_{ii|s}^{(kl)} = 0$ for $k \neq l$, we cannot observe the values $y_{i(k)}$ and $y_{i(l)}$ on the same unit i, which complicates the estimation of $\mathrm{Cov}(\hat{\bar{Y}}_{(k)},\hat{\bar{Y}}_{(l)})$ considerably. To cope with this problem, we derive an estimator where we impute for the unobserved or missing values of the paired observations. We restrict ourselves to the covariance of a CRD, given by (7). First note that (7) can be expressed as

$$\mathrm{Cov}(\hat{\bar{Y}}_{(k)},\hat{\bar{Y}}_{(l)}) = \sum_{i=1}^{N}\sum_{j\neq i}^{N}\beta_{ij}y_{i(k)}y_{j(l)} + \sum_{i=1}^{N}\beta_{ii}y_{i(k)}y_{i(l)}, \quad (10)$$

where $\beta_{ij} = [(n\pi_{ij})/(N^2(n-1)\pi_i\pi_j) - 1/N^2]$ and $\beta_{ii} = -1/N^2$. The first term of (10) doesn't contain unobserved values and therefore it can be estimated directly by

$$\sum_{i=1}^{n_k}\sum_{j=1}^{n_l}\frac{n(n-1)}{n_kn_l\pi_{ij}}\beta_{ij}y_{i(k)}y_{j(l)} = \hat{\bar{Y}}_{(k)}\hat{\bar{Y}}_{(l)} - \sum_{i=1}^{n_k}\sum_{j=1}^{n_l}\frac{n(n-1)y_{i(k)}y_{j(l)}}{N^2n_kn_l\pi_{ij}}. \quad (11)$$

To find an estimator for the second term in (10), we impute for the unobserved values. Without loss of generality, it is assumed that the $i-$th unit has been assigned to the $k-$th treatment group. For $k \neq l$, the potential donor deck for imputing $y_{i(l)}$ is the set of units allocated to treatment l. Let $\delta_{ij}^{(l)}$ denote the indicator variable taking the value 1 when the $j-$th unit is selected to be in the imputation group for imputing $y_{i(l)}$. The imputed value for $y_{i(l)}$ can be defined as:

$$\hat{y}_{i(l)} = \sum_{j=1}^{n_l}\delta_{ij}^{(l)}w_{ij}^{(kl)}y_{j(l)}, \quad (12)$$

where the $w_{ij}^{(kl)}$'s are weights to be determined. Consider the quantity

$$\sum_{i=1}^{n_k}\beta_{ii}\frac{ny_{i(k)}\hat{y}_{i(l)}}{n_k\pi_i}, \quad (13)$$

as an estimator for the second term in (10). We consider the simplest case, where all units allocated to treatment group l are included in the imputation group. If (12) is substituted into (13), then we can derive $w_{ij}^{(kl)}$ such that (13) is an estimate for the second term in (10). If we take

$$w_{ij}^{(kl)} = \frac{(n-1)\pi_i}{(N-1)n_l\pi_{ij}}, \quad (14)$$

then it follows that the imputed values for the unobserved $y_{i(l)}$'s are

$$\hat{y}_{i(l)} = \sum_{j=1}^{n_l} \frac{(n-1)\pi_i}{(N-1)n_l\pi_{ij}} y_{j(l)}. \quad (15)$$

An estimate of $\sum_{i=1}^{N} \beta_{ii} y_{i(k)} y_{i(l)}$ is given by

$$-\sum_{i=1}^{n_k}\sum_{j=1}^{n_l} \frac{n(n-1)}{N^2(N-1)n_k n_l \pi_{ij}} y_{i(k)} y_{j(l)}. \quad (16)$$

The design expectation of (16) is given by

$$-\frac{1}{N^2(N-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} y_{i(k)}y_{j(l)} - \sum_{i=1}^{N} y_{i(k)}y_{i(l)}\right). \quad (17)$$

Using imputed values for $y_{i(l)}$ will generally lead to a biased estimate for the covariance term. To examine this bias, consider the basic model $y_{i(k)} = u_k + \varepsilon_{i(k)}$, with model assumptions

$$\begin{aligned} &\mathrm{E}(\varepsilon_{i(k)}) = 0, \quad \mathrm{Var}(\varepsilon_{i(k)}) = \sigma_k^2, \\ &\mathrm{Cov}(\varepsilon_{i(k)}, \varepsilon_{i(l)}) = \sigma_{kl}, \quad \mathrm{Cov}(\varepsilon_{i(k)}, \varepsilon_{j(l)}) = 0. \end{aligned} \quad (18)$$

The expected value of $\sum_{i=1}^{N} \beta_{ii} y_{i(k)} y_{i(l)}$ under this model equals $-(u_k u_l + \sigma_{kl})/N$. The model expectation of (17) equals $-(u_k u_l)/N$. Since we cannot estimate σ_{kl} directly, the approximation of (16) for the second term in (10) is only unbiased under the strong assumption that the treatment effects are not correlated, i.e. $\sigma_{kl} = 0$. Now, using (11) and (16) an approximation for $\mathrm{Cov}(\hat{\bar{Y}}_{(k)}, \hat{\bar{Y}}_{(l)})$ under a CRD, given by (7), is

$$\begin{aligned} &\hat{\mathrm{Cov}}(\hat{\bar{Y}}_{(k)}, \hat{\bar{Y}}_{(l)}) = \\ &\hat{\bar{Y}}_{(k)}\hat{\bar{Y}}_{(l)} - \sum_{i=1}^{n_k}\sum_{j=1}^{n_l} \frac{n(n-1)}{N(N-1)n_k n_l \pi_{ij}} y_{i(k)} y_{j(l)}. \end{aligned} \quad (19)$$

The bias of this covariance estimate is

$$\frac{1}{N(N-1)} \sum_{i=1}^{N} (y_{i(k)} - \bar{Y}_{(k)})(y_{i(l)} - \bar{Y}_{(l)}). \quad (20)$$

If s is drawn by means of simple random sampling without replacement, then it follows from (15) that the imputed values equals the subsample mean $\hat{y}_{i(l)} = (1/n_l)\sum_{j=1}^{n_l} y_{j(l)}$ for all i. Since we impute a fixed value for each individual i, the covariance approximation (19) equals zero and consequently the bias in (20) equals the covariance between two subsamples drawn by means of simple random sampling without replacement, (i.e. $-S^2/N$).

In an attempt to reduce this bias the imputation method can be improved by forming more or less homogeneous imputation groups. Potential groups are for example pre- and post strata, primary sampling units and clusters. To this end, the finite population is divided into H groups U_h of size N_h. Since only elements from the same group U_h are used as imputation donor, it follows that $\delta_{ij}^{(l)} = 1$ if $i \in U_h$ and $j \in U_h$, $\delta_{ij}^{(l)} = 0$ if $i \in U_h$ and $j \in U_{h'}$. Quantity (13) is still our estimator for the second term in (10). If we take

$$w_{ij}^{(kl)} = \frac{(n-1)\pi_i}{(N_h - 1)n_l\pi_{ij}}, \quad (21)$$

then it follows that the imputed value for $y_{i(l)}$ equals

$$\hat{y}_{i(l)} = \sum_{j=1}^{n_{hl}} \frac{(n-1)\pi_i}{(N_h-1)n_l\pi_{ij}} y_{j(l)}, \quad (22)$$

and that an estimate of $\sum_{i=1}^{N} \beta_{ii} y_{i(k)} y_{i(l)}$ is given by

$$-\sum_{h=1}^{H}\sum_{i=1}^{n_{hk}}\sum_{j=1}^{n_{hl}} \frac{n(n-1)}{N^2(N_h-1)n_k n_l \pi_{ij}} y_{i(k)} y_{j(l)}, \quad (23)$$

where n_{hk} and n_{hl} denotes the number of individuals in imputation group h assigned to respectively treatment group k and l. The design expectation of (23) is given by

$$-\frac{1}{N^2}\sum_{h=1}^{H}\frac{1}{(N_h-1)}\left(\sum_{i=1}^{N_h}\sum_{j=1}^{N_h} y_{i(k)}y_{j(l)} - \sum_{i=1}^{N_h} y_{i(k)}y_{i(l)}\right). \quad (24)$$

The model expectation of (24) under the basic model equals $-(u_k u_l)/N$. An approximation for $\mathrm{Cov}(\hat{\bar{Y}}_{(k)}, \hat{\bar{Y}}_{(l)})$ under a CRD,(7), is obtained by the sum of (11) and (23). The bias of this covariance estimate equals

$$\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 \frac{1}{N_h(N_h-1)} \sum_{i=1}^{N_h} (y_{i(k)} - \bar{Y}_{(hk)})(y_{i(l)} - \bar{Y}_{(hl)}),$$

where $\bar{Y}_{(hk)}$ and $\bar{Y}_{(hl)}$ denotes the population means in under both treatments. It follows that the bias in the covariance approximation is reduced with the covariance between the imputation groups. Consider the most extreme case where there is no covariance within the imputation groups. Then the bias of our approximation would be zero. Under this extreme situation, we actually have paired observations.

6 Variance estimation of contrasts

In the preceding section we saw that the derivation of a design unbiased estimator for the covariance between $\hat{\bar{Y}}_{(k)}$ and $\hat{\bar{Y}}_{(l)}$ is not possible, since for $k \neq l$, we have $\pi_{ii|s}^{(kl)} = 0$. However, for the Wald statistic it is sufficient to have a design unbiased estimator for the covariance matrix of the $K-1$ contrasts $\mathbf{C\Sigma C}^t$. For a CRD the problem of the missing observations can be avoided by concentrating on the variance of the contrasts between $\hat{\bar{Y}}_{(k)}$ and $\hat{\bar{Y}}_{(l)}$. Let $\hat{\bar{Y}}_{I(k)}$ denote the Horvitz-Thompson estimator for $\bar{Y}_{(k)}$ based on the n elements of sample s. Then

$$\mathrm{Var}(\hat{\bar{Y}}_{I(k)}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\triangle_{ij}\check{y}_{i(k)}\check{y}_{j(k)}, \qquad (25)$$

denotes the design variance of $\hat{\bar{Y}}_{I(k)}$ according to the first phase, i.e. the sampling design used to draw s. Furthermore

$$\tilde{\mathrm{Var}}(\hat{\bar{Y}}_{I(k)}) = \frac{1}{n}\left(\sum_{i=1}^{N}\frac{n y_{i(k)}^2}{N^2\pi_i} - \bar{Y}_{(k)}^2\right), \qquad (26)$$

denotes the variance of $\hat{\bar{Y}}_{I(k)}$ as if s, in the first phase, has been drawn with replacement with selection probabilities π_i/n. It follows that (6) can be expressed as

$$\mathrm{Var}(\hat{\bar{Y}}_{(k)}) = \frac{n}{(n-1)}\frac{n}{n_k}\left(\tilde{\mathrm{Var}}(\hat{\bar{Y}}_{I(k)}) - \frac{1}{n}\mathrm{Var}(\hat{\bar{Y}}_{I(k)})\right) - \frac{n}{(n-1)}\left(\tilde{\mathrm{Var}}(\hat{\bar{Y}}_{I(k)}) - \mathrm{Var}(\hat{\bar{Y}}_{I(k)})\right). \qquad (27)$$

In an equivalent way

$$\mathrm{Cov}(\hat{\bar{Y}}_{I(k)}, \hat{\bar{Y}}_{I(l)}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\triangle_{ij}\check{y}_{i(k)}\check{y}_{j(l)}, \qquad (28)$$

denotes the design covariance between $\hat{\bar{Y}}_{I(k)}$ and $\hat{\bar{Y}}_{I(l)}$ according to the of the first phase and

$$\tilde{\mathrm{Cov}}(\hat{\bar{Y}}_{I(k)}, \hat{\bar{Y}}_{I(l)}) = \frac{1}{n}\left(\sum_{i=1}^{N}\frac{n y_{i(k)} y_{i(l)}}{N^2\pi_i} - \bar{Y}_{(k)}\bar{Y}_{k(l)}\right), \qquad (29)$$

the covariance according to the first phase, as if s has been drawn with replacement with selection probabilities π_i/n. Now we can express (7) as

$$\mathrm{Cov}(\hat{\bar{Y}}_{(k)}, \hat{\bar{Y}}_{(l)}) = -\frac{n}{(n-1)}\left[\tilde{\mathrm{Cov}}(\hat{\bar{Y}}_{I(k)}, \hat{\bar{Y}}_{I(l)}) - \mathrm{Cov}(\hat{\bar{Y}}_{I(k)}, \hat{\bar{Y}}_{I(l)})\right]. \qquad (30)$$

Now the covariance matrix of $\hat{\bar{\mathbf{Y}}}$ can be expressed as $\mathbf{\Sigma} = \mathbf{D} + \mathbf{\Lambda}$, where $\mathbf{D}$ denotes a diagonal matrix with elements

$$d_k = \frac{n}{(n-1)}\frac{n}{n_k}\left(\tilde{\mathrm{Var}}(\hat{\bar{Y}}_{I(k)}) - \frac{1}{n}\mathrm{Var}(\hat{\bar{Y}}_{I(k)})\right) \qquad (31)$$

and $\mathbf{\Lambda}$ a matrix with diagonal elements

$$\lambda_{kk} = -\frac{n}{(n-1)}\left(\tilde{\mathrm{Var}}(\hat{\bar{Y}}_{I(k)}) - \mathrm{Var}(\hat{\bar{Y}}_{I(k)})\right) \qquad (32)$$

and the off-diagonal elements λ_{kl} the covariance terms defined by (30). A design unbiased estimator for d_k is given by

$$\hat{d}_k = \frac{1}{n_k(n_k-1)}\sum_{i=1}^{n_k}\left(\frac{n y_{i(k)}}{N\pi_i} - \frac{1}{n_k}\sum_{j=1}^{n_k}\frac{n y_{i(k)}}{N\pi_i}\right)^2. \qquad (33)$$

The covariance matrix of $\mathbf{C}\hat{\bar{\mathbf{Y}}}$ equals $\mathbf{C\Sigma C}^t = \mathbf{CDC}^t + \mathbf{C\Lambda C}^t$. Van den Brakel and Renssen (1996) showed that under the null hypothesis, $\mathbf{C\Lambda C}^t$ is zero and under the alternative hypothesis at least negligible with respect to the leading term $\mathbf{CDC}^t$. As a result an estimator for $\mathbf{C\Sigma C}^t$ is given by $\mathbf{C\hat{D}C}^t$ where $\hat{\mathbf{D}}$ is a diagonal matrix with elements $\hat{d}_k$. This estimator is design unbiased under the null hypothesis but slightly over estimate the variance under the alternative hypothesis. Consider the measurement error model $y_{i(k)} = u_i + b_k + \varepsilon_{i(k)}$, where u_i is the unobservable, intrinsic value of individual i, b_k an additive treatment effect and $\varepsilon_{i(k)}$ the measurement errors with model assumptions (18). Let $\hat{\bar{Y}}^*_{i(k)} = \hat{\bar{Y}}_{i(k)} / \sum_{i=1}^{n_k}\pi_i\pi_{i|s}^{(k)}$, denotes the extended Horvitz-Thompson estimator of $\bar{Y}_{(k)}$. Van den Brakel and Renssen (2000) showed that if the extended Horvitz-Thompson estimator is used, an estimator for the covariance matrix of the contrasts which is approximately design unbiased under the null- as well as the alternative hypothesis, is obtained by (33), where $y_{i(k)}$ is replaced by $(y_{i(k)} - \hat{\bar{Y}}^*_{(k)})$.

Due to the diagonal structure of $\hat{\mathbf{D}}$, the Wald statistic can be further simplified to

$$W = \sum_{k=1}^{K}\frac{1}{\hat{d}_k}\left(\hat{\bar{Y}}_{(k)} - \hat{\bar{Y}}_w\right)^2, \qquad (34)$$

where

$$\hat{\bar{Y}}_w = \left(\sum_{k=1}^{K}\frac{1}{\hat{d}_k}\right)^{-1}\left(\sum_{k=1}^{K}\frac{\hat{\bar{Y}}_{(k)}}{\hat{d}_k}\right). \qquad (35)$$

These results are related to the literature concerning testing interviewer differences and other

non-sampling or measurement errors. Mahalanobis (1946) used interpenetrating subsamples to test to test interviewer differences. The assumption of equal workload and a linear random component model leads under simple random sampling to an F-test of no interviewer effects as well as an estimate of the total variance (Cochran, 1977, section 13.15). Hartley and Rao (1978) provided a general theory, using linear mixed models, to estimate the overall variance for stratified multistage sampling designs in which the last stage units are drawn with simple random sampling. If in the experimental designs considered in this paper the subsamples are assigned to the different interviewers, then the Wald statistic (34) can be interpreted as a weighted sum of squares of interviewer means.

7 Discussion

Two variance estimation procedures for the analysis of CRD's embedded in complex sampling schemes are proposed. The first procedure directly estimates the covariance matrix of the population treatment means. An imputation procedure is applied for the unobservable paired observations. The second procedure makes strong assumptions about the purpose of the analysis of the experiments, by deriving a variance estimator of the contrasts between the finite population treatment means.

The estimator for the covariance matrix of the K population treatment means, is less restrictive since we do not assume a particular hypothesis of no treatment effects in the variance estimation procedure in advance. Generally, the approximation of the covariance term is biased. In an attempt to reduce this bias, the imputation method is refined by using auxiliary information to construct homogeneous imputation groups. It follows that the bias in the covariance approximation can be reduced with the covariance between these imputation groups. As a result, the quality of this covariance approximation is determined by the extend in which we can construct homogeneous imputation groups.

The estimator for the covariance matrix of the $K-1$ contrasts has the structure as if the K subsamples are drawn independently from each other by means of simple random sampling with replacement and with selection probabilities π_i/n. This result is obtained due to the randomization mechanism of a CRD, which comes down to simple random sampling without replacement. Under this randomization mechanism it follows that if the variance of a contrast between two subsample means is derived, then the finite population corrections in the variances of the two subsample means cancels out against the covariance between these two subsample means. See Van den Brakel and Renssen (2000) for more details. As a result, no second order inclusion probabilities are required, which simplifies the variance estimation considerably.

References

[1] Cochran, W.G. (1977). Sampling Techniques. Wiley, New York.

[2] Fienberg, S.E. and Tanur, J.M. (1988). From the inside out and the outside in: Combining experimental and sampling structures. The Canadian Journal of Statistics, Vol. 16, No. 2, pp. 135-151.

[3] Hartley, H.O. and Rao, J.N.K. (1978). Estimation of nonsampling variance components in sample surveys. In N.K. Namboodiri (ed.), Survey Sampling and Measurement, Academic Press, New York, pp. 35-43.

[4] Lehmann, E.L., (1975). Nonparametrics: Statistical Methods Based on Ranks. McGraw-Hill, New-York.

[5] Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. Journal of the Royal Statistical Society, Vol. 109, pp. 325-370.

[6] Van den Brakel, J.A. and Binder, D.A. (2000). Variance estimation of treatment effects for experiments embedded in complex sampling schemes. Research paper, Statistics Netherlands.

[7] Van den Brakel, J.A. and Renssen, R.H. (1996). The analysis of completely randomized designs embedded in complex sampling designs. Research paper, BPA no. 8090-96-RSM. Statistics Netherlands.

[8] Van den Brakel, J.A. and Renssen, R.H. (1998). Design and analysis of experiments embedded in sample surveys. Journal of Official Statistics, Vol. 14, no. 3, pp. 277-295.

[9] Van den Brakel, J.A. and Renssen, R.H. (2000). Analysis of experiments embedded in complex sampling designs. Contributed paper of the International Conference on analysis of survey Data, University of Southampton, August 1999.