CHAPTER 4

AN ALTERNATIVE ESTIMATOR UNDER GMS DESIGNS

As a sequel to the previous chapter, we discuss an alternative estimator under the Generalized Midzuno-Sen (GMS) designs. A mixture estimation strategy of developing separate estimators from component samples is proposed.

4.1 Introduction and Summary

A HTE-type estimator was used under a GMS design in the discussion of the previous chapter. We now investigate the use of a different type of estimator in the same context. We assume equal probability sampling for the second design. A mixture estimation strategy of constructing separate estimators from the component samples is discussed. A simple NNUVE is proposed in addition to a componentwise variance estimator. It is well known that in UPS increasing the sample size does not ensure increase in the precision of estimation though, unfortunately, the computational complexity goes up. Thus, researchers have proposed use of replicated samples or additional samples from the depleted population rather than a single large sample to make UPS efficient. (For example, see Mukhopadhyay (1974), Singh (1978)). A few others will be mentioned in the more general context of extended designs in the next chapter.

The notations and the set-up are as in the previous chapter. However, the size of the second sample is now denoted by \( m \) to indicate that it can be determined separately without reference to the overall sample size \( n \). The discussion needs the following additional notations, terminology and definition. The first design used is UPSWOR and the second is SRSWOR which are now respectively denoted by design \( p \) and design \( q \) and just \( p \) and \( q \) when there is no cause for confusion. The design \( q \) is applied to the depleted population \((U-s^p)\), where \( s^p \) denotes the sample based on \( p \). Thus, the GMS design used is \( p'=(p,q) \). The estimators based on \( p \), \( q \) and \( p' \) are respectively denoted by \( e_p, e_q, e_{p'} \).
The pair \((p,e_p)\) denotes a sampling strategy, where \(e_p = e_p(s_p,y)\) is a sample function depending on \(y = (y_1, y_2, ..., y_n)\) only through \(s_p\). The design \(p'\) is called an extension of \(p\) by \(q\). The sampling variance \(V_p(e_p,y)\) is denoted by \(V_p\). The notation \(V_{p'}\) has a similar meaning.

4.2 The Estimator

As an estimator of \(Y\) based on \(s_p\) we take the traditional \(HTE\) given by

\[
e_p = \sum_{i \in s_p} y_i / \pi_i,
\]

(4.2.1)

which is \(p\)-unbiased for \(Y\) in the sense that \(E_{p'}(e_p) = Y\) and \(\pi_i\) indicates \(\pi(k)\), suppressing \(k\) for brevity. A conditionally unbiased estimator of the total of the depleted population \((U-s_p)\) is

\[
e_q = (N-k) \sum_{i \in s_q} y_i / m_q,
\]

(4.2.2)

since the design \(q\) is SRSWOR. It may be noted that the estimator \(e_q\) is \(q\)-unbiased for \((Y - \sum_{i \in s_p} y_i)\). It easily follows that

\[
e_{p'} = \sum_{i \in s_p} y_i + e_q,
\]

(4.2.3)

is \(p'\)-unbiased for \(Y\). It is useful and interesting to note that

\[
E_{p'}(e_{p'}) = (\sum_{i \in s_p} y_i) + (Y - \sum_{i \in s_p} y_i) = Y.
\]

Thus, \(e_{p'}\) is in fact \(q\)-unbiased for \(Y\) and hence \(p'\)-unbiased. Accordingly, \((p',e_{p'})\) can be considered as an extended strategy for unbiasedly estimating \(Y\). As a final estimator of \(Y\) we may take a convex combination of \(e_p\) and \(e_{p'}\), given by

\[
e_w = w e_p + (1-w)e_{p'}, \quad 0 < w < 1,
\]

(4.2.4)

which is clearly unbiased for \(Y\). It is interesting to note that the estimators \(e_p\) and \(e_{p'}\) are uncorrelated. This can easily be proved using the fact that \(e_{p'}\) is \(q\)-unbiased since

\[
E_{p'}(e_p,e_{p'}) = E_p E_q (e_p e_{p'})
\]

\[
= E_p \left[ e_p E_q (e_{p'}) \right]
\]
\[ E_p(e_p, e_{p'}) = E_p \left[ e_p Y \right] = Y^2 \]

As a consequence, we have,
\[ V(e_w) = w^2 V(e_p) + (1-w)^2 V(e_{p'}). \] (4.2.5)

The variance of \( e_w \) is minimum when the weight,
\[ w = \frac{V(e_{p'})}{V(e_p) + V(e_{p'})} = w_{opt} \text{ (say)} \]

This choice of \( w \) makes the extended strategy \( (p', e_w) \) optimal in the sense of minimum variance.

4.3 Sampling Variance

We have several choices for \( p \). The estimator \( e_p \) and its variance depend on the choice made. In particular, if \( e_p \) is HTE, we have
\[ V(e_p) = \sum A_{ij} \left( \frac{y_i}{\pi_i} - \bar{y} / \pi_j \right)^2. \] (4.3.1)

When \( e_p \) is not HTE, \( V(e_p) \) may have different forms. Some other examples of UPS designs not using HTE are in Brewer and Hanif (1983, Chapter 4).

Next, we derive the sampling variance of \( e_p \), using the familiar conditional argument,
\[ V_p(e_p) = E_p V_q(e_p) + V_p E_q(e_p). \] (4.3.2)

Since \( E_q(e_p) = Y \), the second component on the right side of (4.3.2) vanishes so that we need evaluate only the first component which is done as follows. Since \( q \) is SRSWOR and, in (4.2.3) the first component on the right side is conditionally fixed, we have,
\[ V_q(e_p) = V_q(e_p) \]
\[ = V_q \left[ \frac{(N-k)}{\sum I \in s_q} \bar{y}_I / m \right] \]
\[ = \frac{(N-k) (N-k-m)}{m} S^2_1. \] (4.3.3)

where
\[ S^2_1 = \sum \frac{(y_i - Y)^2}{(N-k-1)}, \] (4.3.4)

and \( \bar{Y} \) is the mean of the character \( y \) in the depleted population given
by \( \sum_{1 \in (U-x)} y_1/(N-k) \). To evaluate \( p \)-expectation of \( V_p \), we essentially need such an expectation for \( S^2 \). As a preliminary step consider,

\[
\sum_{1 \in (U-x)} (y_1 - \bar{y})^2 = \sum_{1 \in (U-x)} y_1^2 - (\bar{y} - \sum_{1 \in (U-x)} y_1)^2/(N-k). \tag{4.3.5}
\]

Adding and subtracting \( \sum_{1 \in (U-x)} y_1^2 \) and expanding the square in the second term on the right side of (4.3.5) and then evaluating the expectation leads to,

\[
E_p \sum_{1 \in (U-x)} (y_1 - \bar{y})^2 = \sum_{1=1}^N (1-\pi_1) y_1^2 - \left[ \sum_{1=1}^N y_1 \pi_1 + \sum_{1=1}^N y_1^2 \pi_1 \right. + \left. \sum_{1 \neq j} \sum_{1=1}^N y_1 \ y_j \ \pi_{1j} \right]/(N-k). \tag{4.3.6}
\]

To simplify the second term on the right side, it is handy to use an identity given by the following lemma.

**Lemma 4.1** In UPSWOR with \( k \) draws, the first and second-order inclusion probabilities denoted by \( \pi_i \) and \( \pi_{1j} \), respectively,

\[
\sum_{1=1}^N y_1^2 \pi_1 + \sum_{1 \neq j} \sum_{1=1}^N y_1 \ y_j \ \pi_{1j} = (\sum_{1=1}^N y_1 \ \pi_1)^2 + \sum_{1=1}^N (y_1 - y_j)^2 \Delta_{1j},
\]

where \( \Delta_{1j} = \pi_i \pi_{1j} - \pi_{1j} \)

**Proof**

First consider,

\[
\sum_{1 \neq j} \sum_{1=1}^N (y_1 - y_j)^2 \Delta_{1j} = \sum_{i \neq j} (y_i^2 + y_j^2 - 2y_i y_j) \ (\pi_i \pi_j - \pi_{1j})
\]

\[
= \sum_{i \neq j} \sum_{1=1}^N \left[ (y_i^2 + y_j^2) \pi_i \pi_j - (y_i^2 + y_j^2) \pi_{1j}
- 2 y_i \ y_j \ \pi_i \pi_j + 2 y_i \ y_j \ \pi_{1j} \right].
\]

Evaluating the double sum and simplifying, we get,

\[
\sum_{i \neq j} \sum_{1=1}^N (y_1 - y_j)^2 \Delta_{1j} = 2 \left[ \sum_{1=1}^N y_1^2 \pi_1 - (\sum_{1=1}^N y_1 \pi_1)^2 + \sum_{1 \neq j} \sum_{1=1}^N y_1 \ y_j \ \pi_{1j} \right],
\]

so that
\[\sum_{i=1}^{N} y_i^2 \pi_i + \sum_{i \neq j} y_i y_j \pi_{ij} = \sum_{i=1}^{N} (y_i - \bar{y})^2 \pi_{ij} + (\sum_{i=1}^{N} y_i \pi_i)^2.\]

This completes the proof. We use Lemma 1 for evaluating sampling variance of \(e_p\). This lemma is again used in a later chapter for a similar purpose in the context of use of ranks in PPS sampling.

Returning to the derivation of \(E_p \sum_{i=1}^{N} (y_i - \bar{y}')^2\) and using Lemma 4.1 to replace the last two terms on the right side of (4.3.6) leads to

\[E_p \sum_{i=1}^{N} (y_i - \bar{y}')^2 = \frac{\sum_{i=1}^{N} y_i^2 (1-\pi_i) - \left[ (Y-\sum_{i=1}^{N} y_i \pi_i)^2 + \sum_{i=1}^{N} (y_i - \bar{y})^2 \pi_{ij} \right] / (N-k)}{\sum_{i=1}^{N} (1-\pi_i)}.\]

Writing \(N-k\) as \(\sum_{i=1}^{N} (1-\pi_i)\) and \(Y\) as \(\sum_{i=1}^{N} y_i\) in the second term on the right side gives,

\[E_p \sum_{i=1}^{N} (y_i - \bar{y}')^2 = \frac{\sum_{i=1}^{N} y_i^2 (1-\pi_i) - \left[ \sum_{i=1}^{N} y_i (1-\pi_i) \right]^2 / \sum_{i=1}^{N} (1-\pi_i) - \sum_{i=1}^{N} (y_i - y_j)^2 \pi_{ij} / (N-k)}{\sum_{i=1}^{N} (1-\pi_i)}.\]

The first two terms on the right side together reduce to \(\sum_{i=1}^{N} (y_i - \bar{y})^2 (1-\pi_i)\) with \((1-\pi_i)\) as weights and with \(\bar{y} = \frac{\sum_{i=1}^{N} y_i (1-\pi_i)}{\sum_{i=1}^{N} (1-\pi_i)}\).

Consolidation of these results gives \(E_p (S'^2)\) as,

\[E_p (S'^2) = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2 (1-\pi_i) - \sum_{i=1}^{N} (y_i - y_j)^2 \pi_{ij} / (N-k)}{\sum_{i=1}^{N} (1-\pi_i)} / (N-k-1). \quad (4.3.7)\]

Finally, we evaluate the p-expected value of \(V_q (e_p')\) in (4.3.3) using (4.3.7).

\[E_p V_q (e_p') = (N-k-m) \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2 (1-\pi_i) - \sum_{i=1}^{N} (y_i - y_j)^2 \pi_{ij} / (N-k-1)m. \quad (4.3.8)\]

Thus, in view of (4.2.5), we get the sampling variance of the final estimator \(e_w\) as
\[ V_p'(e_w) = w^2 V_p(e_p) + (1-w)^2(N-k-m) \left[ (N-k) \sum_{i=1}^{N} (y_i - \bar{y})^2 (1-\pi_i) - \sum_j (y_j - \bar{y})^2 \Delta_{ij} \right] / (N-k-1)m. \tag{4.3.9} \]

The expression for the first term on the right side will depend on the choice of the design \( p \) for the first \( k \) draws. A few remarks regarding use of a weighted combination of \( e_p \) and \( e_p' \), are in order even though \( (p', e_p') \) is itself unbiased, and is based on both the samples.

1. Given the extended design \( p' \), the estimator \( e_w \) can be superior to the component estimators \( e_p \) and \( e_p' \), with a suitable choice of \( w \) which is not very difficult to accomplish as described in the next section.

2. Since \( e_p \) and \( e_p' \), are uncorrelated, their covariance vanishes and the final sampling variance consists of just the two component variances.

3. The variance estimation essentially concerns estimation of the component variances as a consequence of (ii).

We next consider the practical question of determining suitable weights for constructing \( e_w \).

### 4.4 Choice of \( w \)

In order to construct the estimator \( e_w \), a choice of the weight \( w \) has to be made. This is to be done keeping in view the precision of the estimator, feasibility of assessing optimum weight and simplicity of variance estimation. From (4.2.5) it readily follows that the variance-minimising weight is

\[ w_{opt} = V_p'(e_p') / \left[ V_p(e_p) + V_p'(e_p') \right] \tag{4.4.1} \]

and the corresponding minimized variance is

\[ V_{min}(e_w) = V_p(e_p) w_{opt} = V_p'(e_p') (1-w_{opt}) \tag{4.4.2} \]

However, \( w_{opt} \) can rarely be computed as it is a parametric function. As a practical choice, one may replace \( V_p(e_p) \) and \( V_p'(e_p') \) in (4.4.1) by their estimates (Cochran and Carroll, 1953). However, this renders the
weights random which introduces bias in $e_w$ as an estimator of $Y$. It also invalidates (4.2.5) for $V_p'(e_w)$. A few other choices for $w$ are the following.

(i) $w = 1/2$. This allows a very simple variance estimator as in (4.5.1).
(ii) $w = k/(k+m)$, where the weight is proportional to the support of the design.
(iii) $w = p(s_p)/(p(s_p) + p(s_q))$, where $p(s_p)$ stands for probability of the sample $s_p$, and similarly $p(s_q)$.

The suggestions (i) and (ii) are simple for implementation and the former is well suited for easy variance estimation. Suggestion (iii) may be rather difficult for implementation.

4.5 Variance Estimation

We now consider estimation of sampling variance. This essentially concerns estimation of the component variances $V_p(e_p)$ and $V_p'(e_p')$ in (4.2.5). The form of estimator of $V_p(e_p)$ depends on the choice of $p$, $e_p$, and $k$. For example, if $e_p$ is HTE, the SYC variance estimator may be used particularly if $k = 2$ (Vijayan 1975, Rao and Vijayan 1977). If $k > 2$, there are other choices (Rao 1979). If $e_p$ is not HTE, the choices for variance estimator have been discussed in Brewer and Hanif (1983, Chapter 4) for the following cases.

(i) Das' estimator with Yates-Grundy draw-by-draw procedure.
(ii) Raj's and Murthy's estimators with Yates-Grundy draw-by-draw procedure.
(iii) The Rao-Hartley-Cochran (RHC) estimator with RHC procedure.
(iv) Unbiased and ratio estimators with Poisson sampling procedure.
(v) Unbiased and ratio estimators with modified Poisson sampling.
(vi) Unbiased and ratio estimators with collocated sampling.
(vii) Lahiri's estimator with Lahiri's procedure or Ikeda-Midzuno procedure.

In the extreme case of one of the two samples $s_p$ or $s_q$ having only one unit, componentwise variance estimation is not feasible. The choice $w = 1/2$ allows a simple variance estimator in view of $e_p$ and $e_p'$, being
uncorrelated. In this case, \( e_w = (e_p + e_p')/2 \) and the simple unbiased variance estimator is

\[
\nu(e_w) = (e_p - e_p')^2/4. \tag{4.5.1}
\]

This choice of \( w \) is also handy when the size of one of the samples is unity and hence componentwise variance estimation is not feasible. The choice \( w = 1/2 \) and the associated simple variance estimator (4.5.1) may be particularly preferred for computational ease when there are several study characters (Wolter 1985, pp. 2-5).

Coming to the question of estimating \( V_p(e_p) \), we note the following. For this estimator, we essentially need an estimator of \( V_q(e_p) \) from the sample \( s_q \) since \( E V_p(e_p) = V_p(e_p) \). For instance, if \( \bar{y}' \) denotes the sample mean under the second design SRSWOR, then

\[
(N-k)(N-k-m) \sum_{i \in s_q} (y_i - \bar{y}')^2 / m(m-1) \tag{4.5.2}
\]

is \( p \)-unbiased for \( V_p(e_p) \).

4.6 Conclusion

We have illustrated use of UPSWOR design in combination with SRSWOR from the depleted population constructing a non-HTE type estimator as a convex combination of two estimators. The procedure bears a similarity with the Raj (1956) procedure for ordered estimators; but there is a difference in that more than one sampling design is employed here. Also, the estimators correspond to each design rather than to each draw. However, these estimators are noted to be uncorrelated.

The sampling variance can be estimated componentwise. Alternatively, uncorrelated estimators from the component samples allow the facility of a very simple NNUVE when equal weights are used for combining them. This is particularly attractive when a large number of characters are of interest.

A few practicable suggestions have been made for choosing the weight \( w \) needed for constructing a convex combination of the two initial estimators from the two designs. As a general result, the suggested
estimator $e_w$ is more precise than the component estimators and $e_p$ and $e'_p$, as long as $0 < w < 2w_{opt}$. This points to the flexibility of choosing $w$ particularly when $w_{opt}$ is near half or higher. A formal discussion of this aspect is in Chapters 5 and 6.

It is pertinent to note that the computation of the estimator $e_w$ needs only a suitable $w$ given the estimators $e_p$ and $e'_p$. The estimator $e_p$ requires inclusion probabilities with respect to the first $k$ draws only which is not computationally complex for a small $k$. The other estimator $e'_p$, is a simple estimator. Thus, as compared to the estimator suggested in the previous chapter, there is a slight computational advantage. However, attempting an overall IPPS design is a point in favour of the strategy proposed in the previous chapter. A fair numerical comparison of the two strategies is not easy to carry out in view of their differing requirement with respect to sample sizes, sampling designs and weights etc.

The overall precision of the estimator $e_w$ may be ensured by means of (i) choosing a right size measure, (ii) choosing the first design to be IPPS and (iii) using an appropriate weight $w$.

The contents of the present chapter are comparable with those of the previous one in that both assume a CMS design. They differ with respect to the estimator employed. We recall that in the set-up of the previous chapter, the design can be made IPPS provided that $k$ exceeds a certain number, which implies a condition on $k$. The same effect is accomplished in the extended strategy of the present chapter without a condition on $k$ through the considerations stated in the previous paragraph.

In the next chapter, we put the concept of using a second design in the same survey on a firmer footing of extended designs and examine their role in resolving several of the traditional problems in UPS.