### CHAPTER 10

## DESIGN–AND MODEL BASED

## SAMPLING INFERENCES-1

**10.1 Introduction**

Survey sampling (also known as finite population sampling) is unique within statistics in that there are two distinct probability distributions which can serve as source of inference.

One, peculiar to survey sampling, describes that way in which the sample is selected. The distribution is exactly known, because it is imposed on the population by the designer. The other, shared with all other areas of statistics, is a structure innate to the population itself, unknown, but capable to being modeled.

In design-based inference, bias, variance and mean square error (mse) are defined in terms of the *expectation* over all samples permissible under the sample design. The design-based approach uses sample selection probabilities to provide the basis for its inference. For most part it is only the first and second order inclusion probabilities of the individual population units that are relevant. In general the first order probability varies from unit to unit and the second order probabilities from pair to pair. However, the special case where the population is stratified, and simple random sampling without replacement is used within each stratum, is both simpler to consider and is extensively used in practices.

The properties of a design-based estimator are defined in terms of its behavior over repeated sampling (i.e. in expectation over the set of all possible samples permitted by the sample design). In its pure form, design-based inference rests on what may be termed as the representation principle. Imagine that a population is divided into three stratum on the basis of size. From each of these stratum a simple random sample of fixed size is drawn with equal probabilities without replacement. The probabilities of inclusion in sample are unity for each (large) unit in the first stratum, 1/10 for each (medium-sized) unit, in the second stratum and 1/100, for each (small) population unit in the third.( A unit in the first stratum clearly needs only to represents itself, and the representative principle accordingly specifies that it has have might unity.) But for every unit included in the sample in the second stratum there are nine others in the population that are not in the sample, so each sample unit has weight ten. Similarly, each sample unit in the third stratum represents itself and 99 other population units in the sample, S it has weight 100.

More generally, for any simple random sample selections without replacement, the representative principle requires that the weight given to each unit included in the sample should be the reciprocal of its probability of inclusion. The resulting estimator of the total Y for the survey variable yi is the sum over all stratum of the straightforward expansion estimators (which are N/n times the sample total of y for any given stratum). This is a special case of the Horvitz and Thompson estimator, which reflects and embodies the representative principal in full generality.

For many purposes, however Horvitz and Thompson Estimator is clearly inefficient. Whenever there is a variable *X* for which total population is known, and it is sufficiently closely related to the survey variable *Y ,*is any departure of the Horvitz and Thompson Estimator of *X* from its known value to be likely to approximate the departure of the Horvitz and Thompson Estimator of *Y* from it known value. The classical (conventional) ratio estimator [the ratio estimator of the two Horvitz and Thompson estimators time the known value *X*] will be more acceptable than the simple Horvitz and Thompson estimator of *Y* in this sort of situation.

All this, presupposes that the sample is probability sample. For samples that are not probability samples, a radically different type of inference is necessary. Such inferences were regularly made from purposively selected samples before the arguments for the excessive use of probability samples because generally accepted, but there inferences are generally adhoc in nature and these short comings were *incisively* exposed by Neyman (1934). The serious advocacy for the use of an alternative source, called model-based, of inference commence with Royall (1970).

Royall (1970) argued that survey sampling was out of step unit statistics as a whole. Statisticians working in other fields used their data to build models and analyzed their data in those terms (using model-based inference) but survey statisticians had allowed themselves to be reduced into using an entirely undesired source of probability structure, not related to the data themselves but only to the manner in which the data had been collected. He suggested that in many instances a suitable model for inferential purposes was one for which the classical ratio estimator would be optimal. This is one where the survey variable y is a linear homogeneous function in an explanatory variable x and the variable function for y is also linear homogeneous in x.

The classical ratio estimator can therefore be arrived at from two quite different inferential standpoints. If the sample is a probability sample, the classical ratio estimator can be seen as an ad hoc but highly plausible refinement of Horvitz and Thompson Estimator, but whether it is a probability sample or not, the classical ratio estimator is the optimal (Best linear unbiased under generalized least square} estimator under the linear model.

In this example, the estimators happen to coincide, but generally this does not happen. A choice must then be made between design-based and model based –based inference.

In model-based and prediction approaches, bias, variance and mean square error relate strictly to the particular sample selected and the relevant expectations are computed over all possible realizations of a stochastic model.

 , (6.8.2)

where *β* and *σ*2 are constant, E denotes the expectation overall possible hypothetical populations.(In accordance with the accepted practice, the designed-based approach will be referenced to as p-bias, p-variance and p-mse and the model-based ones as ξ-bias, ξ-variance and ξ-MSE.)

The analyst who assumes a super-population model may be led to a different approach, given the observed sample. He may adopt a simple and commonly used model (Cochran – 1953) and conclude from prior experience, or from examining a scatter plot of the individual sample observations, or both, that the relationship between variables of interest could be represented approximately by a straight line through the origin, and that the variability of y around the regression line increases as x increases.

On the one hand, many design-based statisticians now acknowledge that the explicit use of modeling can help them to make more useful inferences than they previously supposed. In fact, all competent design-based sampling statisticians have; unconsciously and implicitly; been using population models from the beginning. For evidence of this, see Basu’s entertaining elephant fable and the lively discussions concerning it (Basu 1971, Hanif 1994).

On the other hand, there had long been recognition in the model-based inference that randomization can play a useful role. Royall (1976) likened this to the use of a coin toss at the start of a ball game.

The earliest modern finite population sample, which can be considered as starting point with the work of Kiaer (1897), were characteristically purposive selections of sample units which were held to be ‘representative’ of the population. Since no explicit models were formulated, however, this early approach lacked the theoretical *underpinning* which might have led to the construction of more acceptable survey estimates.

Quite early, however Bowley (1913), survey designers started to use randomized sample designs. This was particularly the case when there was a lack of prior information about the population. Neyman (1934) described how Gini and Galvani settled for a purposive sample in designing sample of Italian population census areas.

Following Neyman’s paper only one discussant was prepared to argue the case for purposive selection. The randomization-based (or design-based) approach became within a few years the only respectable basis for survey sampling. This was particularly the case after the publication of the influential sampling textbooks by Yates (1949), Deming (1950), Hansen, Hurwitz and Madow (1952) and Cochran (1953). Its universal acceptance by the statistical community and the fruitfulness of the innovations which is inspired, such as unequal probability sampling (Hansen and Hurwitz 1943), are the hallmarks of the successful scientific paradigm described by Kuhn (1970).

The model (6.8.2) has been extensively used in this book, in a sense this has been inevitable in a book on unequal probability sampling. To use unequal selection probabilities more than any other kind of probability sampling, is to admit that the manner in which the sample is selected is relevant to the inferences to be made form it, to design-based inference.

But according to Basu’s elephant fable, design-based inference must not be used blindly. Statisticians who use it in unequal probability sampling have always required that there be some rough kind of relationship between the size measures and to estimand variable. One might say that they carry an informal model in their head. And it is rare now to find a design-based statistician who refuses to allow population modeling some role in the choice of a sample design (Sarndal, 1985).

There are, however, statisticians who believe that model-based inference can supply all the answers and either deny that random selection is a useful tool or relegate it to a very subordinate role. They argue that is wrongheaded to base inferences on the set of all samples that might have been selected, when only one – i.e. the sample in fact selected –can be observed. When this argument is driven to its logical conclusion it leads in the first instance to a class of sampling designs which are super-efficient as long as the model used describes the data accurately, but which can be totally misleading once it breaks down.

Typically such designs involve the purposive inclusion of the n largest population units in sample, these being the ones which in terms of model (6.8.2) contain most information. This may be seen as follows. The best linear unbiased (BLU) predictor of  based on any given sample is:

 , (10.1.1)

where  (10.1.2)

The predictor  has model-based variance

  (10.1.3)

It may be seen that provided the sample size n is fixed and  does not increase faster than , i.e. provided  in (6.8.2) then  is indeed smallest when the sample consists of the largest n units.

If the population is better described by the slightly more general model:

  (10.1.4)

Then the of (10.1.1) is no longer the BLU or even an unbiased predictor of Y but has bias:

  (10.1.5)

If the sample values of  are much larger than the non-sample values, this bias can approach ; in which case the bias in the predictor of the mean would approach . The simplest design-unbiased sampling strategy will clearly be preferable to this, provided only that the sample is large enough for the variances of its estimators of total and mean to be smaller than the squares of these respective biases.

**10.2 Horvitz-Thompson Ratio Estimator – **

The Horvitz and Thompson estimator suggested by Brewer (1963b) and also proposed by Hajek (1971) is

 . (10.2.1)

The derivation of mean square error of  can be easily made by using the concept given in Chapters 6 and 8 and is

  (10.2.2)

* 1. **A more General Asymptotically Optimal Sampling Strategy**

Cassel et al (1976) introduced a design-unbiased General Difference Estimator (GDE) which was optimal under the following model.

  (10.3.1)

(Note There is no loss of generality here if u is set equal to zero since the  are sufficiently flexible to be able to include a term such as . However the original form of the CSW model will be adhered to here, although the notation has been changed to confirm with that used elsewhere in this book).

The GDE proposed by Cassel et al. (1976) has general form

 (10.3.2)

Where  were initially arbitrary. Latter they showed that . The variance of  is given in the following theorem.

**Theorem 10.1**

The general difference estimator proposed by Cassel et al.(1976) has variance is:

  (10.3.3)

**Proof**

For the time being the value of  in (10.3.1) will be set to zero and the modified (10.3.1\*) will be used. Under (10.3.1\*) the expected variance of , is

 

 

  ◊ (10.3.3)

(10.3.3) is at the Godambe–Joshi lower bound. The optimal values of  are those which minimize this expression, i.e.:

  (10.3.4)

in which case

  (10.3.5)

Turning now to the question as to how the  might be guessed, Cassel et al (1976) considered a special case of (6.2.14) where  where were known,  was unknown. This model is very close to that of (6.8.2.) The minimum expected variance strategy if  is known is then  where

 

with  (10.3.5)

* + 1. **The Generalized Regression Estimator – **

Cassel, et al. (1976) has proposed an other estimator which has the form

  (10.3.6)

where . (10.3.7)

The expected variance of (10.3.6) under the super population model (6.8.2), asymptotically achieves Godambe-Joshi lower bound. If  then , transform to Horvitz and Thompson (1952) estimator like Brewer’s ratio estimator, but for  does not become conventional ratio estimator, it turns out to be a slightly unconventional form of ratio estimator. A further alternative (or class of alternative) is provided by the Generalized Regression Estimator of Cassel, Sarndal and Wretman (1976). In its most general form, it is written as

  (10.3.8)

where  is essentially an estimator of the β of the model (6.8.2)

The use of the BLUE of β minimize the variance of generalized regression estimator.

* 1. **Sample Strategies with both Design–Based and Model–Based Properties**

It is now increasingly common for sample strategies to be put forward with both design-based and model-based justification. As indicated in the previous section, this is little more than a formation of what competent design-oriented statisticians have been doing for decades. The formation itself has, however, considerable value. If the strategy is not close to be optimal under a plausible model, there is every reason to look for an improved version. Moreover a strategy that has both design-based and model-based properties carries a double insurance policy. If the model breaks down but the sample is reasonably typical (given the selection procedure) one can rely on the design-based properties; while if the sample is a typical but the model is a good approximation, there are the model-based properties to fall back on.

The search for a completely optimal sampling strategy under a simple model such as (6.8.2) or (10.1.4) has nearly but not quite attained its goal. The absence of the intercept term from (6.8.2) makes the search comparatively easy, so this will be described first.

* + 1. **A Ratio Estimator Proposed by Brewer – **

Brewer (1979) devised a sampling strategy for use in official surveys of businesses, farms and the like, where the population units were of very different sizes and both the population and sample were large. He assumed that the dependence of the estimated values  on the auxiliary variable values  was approximately given by model (6.8.2). Assuming linear and homogeneous dependences he used the following class of estimators.

  (10.4.1)

where  (10.4.2)

The  being arbitrary. Equation (10.4.1) ensured that the estimation process was limited to the non-sample values. Since  was model-unbiased under (6.8.2) it, therefore, has the ratio-estimator property. Brewer (1979) then sought to optimize simultaneously the weights  and the inclusion probabilities . He first required that his estimator be asymptotically unbiased as n and N jointly approached infinity. Since the design expectation of a ratio is then asymptotically equal to the ratio of the design expectations, he obtained the approximate result.

  (10.4.3)

Since the scale of the weights  is arbitrary, it can be chosen to make

  (10.4.4)

Substituting (10.4.4) in (10.4.3), setting  and comparing the coefficients of, it followed that

  (10.4.5)

Thus the condition of asymptotic unbiasedness gave a relationship between  and  which enabled them to be optimized simultaneously as follows:

**Theorem 10.2**

The expected variance of

  (10.4.6)

is  (10.4.7)

#### Proof:

By Schwartz’s inequality,



  (10.4.8)

Hence  (10.4.9)

The right hand side is achieved and expression (10.4.7) is minimized when

  (10.4.10)

and hence from (10.4.5) when

  (10.4.11)

which is analogous to Neyman allocation.

The expected variance of when N is large may be written as

 =. (8.9.1 )

This is equivalent to minimum variance of any design based estimator as given by Godambe and Joshi (1965). Since the  values have also been optimized, the strategy consisting of  is asymptotically efficient. Finally since its relative variance tends to zero as N, and n tend to infinity in a constant ratio,may be described as design consistent.

The Brewer’s (1979) results has been commented by Robinson and Tsui (1979) and by T.J. Rao (1984). Robinson and Tsui pointed out that Brewer’s assumptions were so restrictive that, given a large enough sample, an alternative sampling strategy could have estimated the total entirely without error. They derived an identical estimator using a less restrictive asymptotic approach of their own. T.J. Rao also derived the same estimator, starting from a wider initial class and using a slightly modified proof. An alternative approach to the same problem was presented by Isaki and Fuller (1982), Sarandal (1980a) for his asymptotic analysis adopted a formulation used by Brewer (1979) Tsui 1983) showed that Brewer’s estimator was admissible for any fixed population size].

# 10.5 A Design- and Model-Based Estimator

A general class of estimators was introduced in Chapter-6 as

 

In the following we will give some properties of the estimator . We start by proving the design and model unbiasedness of . The variance of  is obtained in Section (10.5.1)

 (a) For the design-unbiased ness we can write

 . (10.5.1)

 Taking design expectation, we have

  

 

 . (10.5.2)

 So  will be design-unbiased if

  i = 1, 2, 3, ……, N (10.5.3)

 and i)  (10.5.4)

 ii)  (10.5.5)

 iii)  (10.5.6)

(b) For model–unbiasedness we use model (6.8.2) in (10.5.1) then we have

  

 .

 Applying the model-expectation, we have

  (10.5.7)

or

  (10.5.8)

Now  or  if

  for all S (10.5.9)

Conditions (10.5.3) and (10.5.9) ply useful role in deriving the variance of 

* + 1. **Anticipated Variance of Design- and Model-unbiased Estimator **

In this section we have derived the anticipated variance of . Since  is both a design- and model-unbiased estimator, therefore the anticipated variance [Brewer (1963), Isaki and Fuller (1982)] is given as

  

 

 . (10.5.10)

Using (6.8.2) in (10.5.10) we have

 

  (10.5.11)

Since , so (10.5.11), can be written

 

or  (10.5.12)

Squaring the right hand side of (10.5.12) and using the model expectation we get

  

 

 

 

  (10.5.13)

For a given sampling scheme, the best  is obtained by minimizing  or in fact  subject to (10.5.3) and (10.5.9) unconditionally. The function will take the following form

 , (10.5.14)

where  and  are Lagrange multipliers. Differentiating (10.5.14) with respect to CIS and equating to zero, we have

 

or . (10.5.15)

Substituting (10.5.15) in (10.5.9) we have

 . (10.5.16)

Substituting the value of  from (10.5.16) in (10.5.15), we get

  (10.5.17)

where  are to be solved iteratively by using (10.5.3)

Now substituting (10.5.17) in (10.5.3), we get

 . (10.5.18)

Since , (10.5.16) therefore (10.5.18) will be

  (10.5.19)

Now substituting the value of CIS from (10.5.17) in (10.5.1), we have

  (10.5.20)

Since  is a lease-square estimator of , therefore

 . (10.1.2)

Therefore (10.5.20) takes the form

  (10.5.21)

Taking the design expectation of (10.5.21)

  (10.5.22)

If n and N are both large, then following the arguments of Brewer (1979), (10.5.22) becomes

 

 . (10.5.24)

Equating the coefficients of YI on both sides of (10.5.23), we get

 . (10.5.24)

If we put , then (10.5.24) is satisfied, therefore the asymptotic form of  is

  (10.5.25)

or  (10.3.9)

It is indeed extremely interesting to note that the asymptotic form of the estimator under investigation is the Generalized Regression estimator of Cassel et .al (1976), and it appears with the BLUE of β rather than any other. Equation (10.5.25) provides the essential part of the proof, since it has been obtained by minimizing the variance of the general estimator given in (10.5.1) – of which the Generalized Regression estimator is a special case – subject to design- and model-unbiasedness, and then using asymptotic result. If Generalized Regression estimator with smaller variance existed, then it would have emerged as (10.5.25), or else the optimum model-unbiased and asymptotically design-unbiased estimator would not be Generalized Regression estimator of any kind.

It is interesting to note that the asymptotic form (10.5.25) is not the one suggested by Brewer (1979).

Some special cases of (10.5.25) can be derived for different values of  as under

If we put  in (10.5.25), then the estimator becomes

  (10.5.26)

which is Brewer’s estimator (1979).

Further if  then  is:

  (10.5.27)

which is the Generalized Regression estimator of Cassel, Sarndal and Wretman (1976).

It seems fairly obviously that if optimum selection probabilities are  in all three cases, the three estimators become

 = unchanged for 

 = for 

and

=  for the new estimator

The new estimator (10.5.25) is equal to  unconditionally if  or if  it requires the Zi all be equal.

**10.5.2 Predictive form of Estimator**

We now turn to give a predictive form of  in the following. Since

  (10.5.28)

If we put  (which is close to  [to be discussed in Section 10.7], then (10.5.28) turns out to be

  (10.5.29)

Equation (10.5.29) can be written as

 

or  (10.5.30)

Equation (10.5.30) is a predictive form of one suggested by Basu (1971).

The first part of (10.5.30) is the predictive estimate and the second part is some kind of correction to the predictive estimate of Basu (1971). However, it does not stop here. For the sampled unit, we can look at . These terms provide some idea about the validity of the model. If the model is appropriate, the best estimate of error  for  is zero. If the model is not appropriate (for example, if ’s reveal some trend with ZI), then the estimate of  is . Therefore, the last part of (10.5.30) provides correction to Basu’s (1971) estimator.

It is not reasonable to think that the non-BLUE suggested by the Brewer (1979) was the result of deliberately sacrificing a small amount of efficiency in order to ensure that the Generalized Regression estimates could be written in the predictive form. Samiuddin et al,(1992) used alternative approach which strictly minimizes the anticipated variance, is left with an additional term over and above the predictive form, namely, the last term in (10.5.23). If Samiuddin et al. (1992) had used the Brewer (1979) estimator of β - instead of the BLUE the last term in (10.5.30) would be zero. Even if we look minutely with the BLU estimator of β, this term will generally be small – or asymptotically zero – whether (6.8.2) is an appropriate model or not. It is, however, not merely a simple term but a summation of individual terms, each of which is proportional to an estimate of the random error term . Regardless of what estimator is used for β (within a wide range of reasonable choices), the estimates of the individual  provide a useful diagnosis as to the appropriateness or otherwise of the model (6.8.2).

**10.6 A New Design Based Estimator**

In this section, an attempt has been made to derive a design-based estimator for unequal probability sampling using a strictly designed based approach.

Brewer et al (1982) derived a designed based estimator for comparison purpose using an exclusively design-based approach to the problem. For this consider a population of N units, Yi is estimand variable and Zi is benchmark variable. A fixed size sample of n units is selected with unequal probabilities without replacement. The probability of selection of sample s is denoted by P(s), the set of all possible samples by s, and the probability of inclusion of the i-th unit by . Consider the estimator

 , ((10.6.1)

where the coefficients Cis are subject to the following conditions and to make the design variance as small as possible.

**Condition-1**.  is to be design-unbiased. This requires that

  i = 1, 2, ……, N (10.5.3)

**Condition-2**.  is to have the ratio estimator property. This implies that

   (10.5.10)

**Condition-3**. The weighted variance of the CiS – or equivalently the weighted sum of the  – is to be a small as possible, i.e.,

  (10.6.2)

is to be minimized with respect to Cis, where Wi are arbitrary weights.

While conditions (1) and (2) are familiar, condition (3) requires an explanation. The need for a third condition arises because the first two admit infinite of possible solutions. The third condition would ensure minimum variance under repeated sampling.

Without resourse to a model, the best that can be done is to impose adhoc conditions on the quintities . An attempt to keep within bounds is clearly called for, since if they are allowed to take extreme values,  becomes unstable. It is especially important to ensure this for the more frequently occurring samples; hence the weights P(s). Finally, since condition (2) implies that Yi are proportional to Zi then Y has the ratio estimator property provided CiS are chosen such that  for all . It is relatively more important to keep bounds on the Cis for those units with large Zi values; hence the weights WI, which, although still arbitrary at this stage in the analysis, may be expected to increase with increasing Zi.

In order to minimize (10.6.2) subject (10.5.3) and (10.5.10), we use the Lagrange multipliers i.e.

 , (10.6.3)

where  and  are Lagrangian multipliers. Differentiating (10.6.2) with respect to CIS and equating to zero, we get

 

or  (10.6.4)

Substituting Cis in (10.5.10) we get

 

 

or   (10.6.5)

similarly we have

  I = 1, 2, 3, ……, N. (10.6.6)

Now summing (10.6.5) over  and substituting in (10.6.6), we get

 =. (10.6.7)

Since (10.5.3) and (10.5.10) combined contain one equation which is a linear combination of the reminder, the N equations (10.6.7) admit infinity of solutions. But if  and ; I = 1, 2, …, N are any two solution sets for (10.6.7), it can be shown that  must be proportional to Zi. Further, substitution in (10.6.5) and (10.6.6) of solution sets which differ only by multipliers of the Zi results in identical sets of the coefficients CIS. Consequently, all solution sets are essentially equal. In an empirical study of Section (10.6.3) equation (10.6.7) is solved using as starting values

 . (10.6.8)

The equation set (10.6.7) is, however, unwieldy, particularly for n > 2. An approximate solution, asymptotically correct as N, n → ∞, may be obtained using the  of (10.6.8) without iteration. Using (10.6.8) and (10.6.5) in (10.6.6), we get

  (10.6.9)

and putting it in (10.6.1), we get

 . (10.6.10)

All the model-based estimators considered in this chapter, other than the exact solution to (10.6.7), are special cases of (10.6.10) with appropriate choices of  and Wi.

10.6.2 Unbiasedness, variance and expected variance of the design-based estimator

In this section, we will prove the unbiased ness of and derive the formulae for its variance and expected variance. For unbiasedness we intend to derive the expression in which  is not involved. This can be achieved by substituting (10.2.7) in (10.2.6)

 , (10.6.10)

where  are to be solved iteratively.

Using (10.3.1) in (10.2.1) we get

 

or  

 

Using the concept of Brewer (1979) for large N and n, we have

 . (10.6.11)

The variance of  may be easily derived using the basic concept

 , (10.6.12)

where . (10.6.13)

A unique non-negative unbiased variance estimator for the sample size 2 consisting of population units I and J may be written in a simply way

  (10.6.14)

where  is the value of  associated with the sample , [Rao(1979)].

The corresponding unique non-negative unbiased variance estimator for n = 2, given by Sen (1953) and Yates and Grundy (1953) is

  (3.3.6)

The expected variance of  may be derived under the model (6.8.2). Substituting CiS from (10.6.3) and  for WI, the expected variance may be written

  (10.6.15)

where ED denote design-expectation and EM denotes model-expectation.

Expanding the right hand side and substituting the value of  from (10.6.5) and on simplifying we get

 (10.6.16)

The first term in (10.6.16) is the Godambe-Joshi lower bound (6.8.2). The second term, which is non-negative, will be small when  is nearly proportional to P(s) for all . It may be seen from (10.6.4) that P(s) enters multiplicatively into the formula for , but the remaining terms and factors ensure that in general the proportionality is not exact.

The remaining three model-based estimators are closely related to  and consequently their estimators of mean square error are closely related to the Sen-Yates-Grundy variance expression. Considering only the leading term in the Taylor expansion, the mean squared error of all the three can be estimated by

  , (10.3.9)

where for 

 

**10.6.3 Empirical and Semi-Empirical**

For Semi empirical studies 25 populations used in this study are the artificial population 1-5 and the natural population 1-20 employed by Rao and Bayless (1969).  is derived under the condition to be design-unbiased, therefore, it cannot be expressed in an explicit form. It requires iteration to solve a system of equations. Therefore, the study is limited to n = 2. Two simplifications in the calculation of  are brought about by restricting the sample to two units.

1. In equation (10.2.8) the summation  is replaced by double summation over i and j (j > i) and two units summed over by  are the units i and j.
2. A simple formula suggested by Brewer (1963), Durbin (1967), Rao (1965) and Sampford (1967) is available for n = 2 and for P(s) satisfying the conditions that P(s) > 0 for all s and that . For the sample s consisting of units
i and j this is

  (8.6.12)

In summary, the three model-based estimators are very similar in their small sample behavior. In fairly extreme cases they may be subject to a bias of 4 – 5%, but there is no clear tendency for the bias to be in either direction. In a few cases, they are appreciably less efficient than the ideal estimator , but in many cases they are by that same standard super-efficient. Their Taylor expansion estimators of mean squared are generally only moderately biased (downwards) and tend to be appreciably more stable than the unique non-negative unbiased estimator of the variance of . On all counts other than bias, these three estimators are distinctly preferable to .

The three model-based estimators differ among themselves to some extent. There are slight but consistent tendencies for  to be less biased than  and for  to be less biased than . The study nevertheless suggests that for most practical purposes, it would be acceptable to use any of those three estimators in small-sample situations. The simplicity of  compared with other two estimators may well be the deciding factor between them.

**10.7 Alternative to Godambe-Joshi Lower Bound**

Godambe–Joshi lower bound to the variance of Horvitz–Thompson estimator was derived under the assumption that . This relates to estimators which are design-unbiased. Unless , or  this bound cannot be achieved. So the Horvitz and Thompson estimator achieves the lower bound for. However, the minimum of (8.9.1) for given  will depend on the choice of . Let us now try to investigate that for what value ’s, (6.8.2) achieve minimum (8.9.1). Note that for fixed sample size ; so we minimize  with the condition that . Then we have

  (10.7.1)

Unconditionally, where  is the Lagrangian multiplier. Differentiating (11.2.1) with respect to , we get

  (10.7.2)

or .

If , then from (10.7.2),  i.e. . We have, therefore, shown that a sampling scheme with  together with the Horvitz-Thompson (1952) estimator constitutes the optimum strategy when .

In fact, the lower bound (8.9.1) is rarely if ever achieved. This is because the class of estimators considered is very wide. A natural course of action is to limit the class, and then look for the optimum estimator and the lower bound. In this chapter, we propose to limit the class of estimators which are both design- and model-unbiased. We shall see that within the class we can find the optimum estimator and that the lower bound is then achieved. The choice of  which minimizes (8.9.1) is clearly . Many of the commonly used estimators satisfy both conditions i.e. design- and model-unbiased ness. Some estimators are asymptotically design-unbiased [see Cassel, Sarndal and Wretman (1976), and Brewer (1979)]. The fact is that we can find a better bound than (8.9.1) for some restricted class of estimators which are both design- and model-unbiased.

It follows from (6\_\_\_\_\_) that

 

Since, by construction,  is both design- and model-unbiased for Y, therefore its anticipated variance is

  .

  (10.7.3)

Using (6.8.2) in (10.7.3), we have

  

 

 

The first term of R.H.S. of the above expression is zero because of (3.3.2). Therefore,

 

 

 . (10.7.4)

From (10.7.2), we have

  (10.7.5)

Comparing the coefficient of Yi of both sides of (10.7.5), we have

  for all I (10.7.6)

Using (10.7.6) in (10.7.4), we have

   (10.7.7)

Now the design-expectation over the set of samples  is

 ,

and the corresponding design-based variance is

  (10.7.8)

This leads to  and consequently using (10.7.8), we get

  (10.7.9)

Equality in (10.7.9) holds iff  or  for all I, where  is a constant over all . Putting the value of  in (10.7.\_\_\_) we get , for all s. This must be true regardless of which of the N population units make up the n in the sample. So it must be true for all I separately. Since there are n sample units in s, therefore  and . Also since  which leads to . Since  is minimized when  for any sample s, then the lower bound will involve . We, therefore, consider estimators which are both design- and model-unbiased and rework the lower bound.

It is possible to choose  in such a way that (10.7.6) and (8.9.1) are satisfied and  at (10.7.7) is minimized. To do this we in fact minimize

 ,

Unconditionally, where  and  are Lagrangian multipliers. Differentiating the above expression with respect to CiS and equating to zero we get,

 

or , (10.7.10)

Multiplying (10.7.10) by Zi and summing over  gives

 .

This leads to  (10.7.11)

Using (10.7.11) in (10.7.10) gives

 , where , (10.7.12)

For a given , the  can be solved iteratively using (10.7.10). This yields the following implicit equation for the 

 

or . (10.7.13)

For large sample sizes  is necessarily close to , since the  is the Horvitz-Thompson estimator for Z and approaches it asymptotically. Further, for numerous small sample cases that we have solved numerically  is indeed very close to . At any rate this provides a good first approximation to . The second approximation to  is given by

 

If we write  as the r-th approximation to  (when ), we get

 . (10.7.14)

Usually it look 3 to 4 iterations to get to the solutions in the examples that we worked out.

To simplify the expression for  in (10.7.7) we proceed as follows. Multiplying (10.7.12) by  and summing over  we get

  

 

Multiplying the above by  and summing over all I

 

  (10.7.15)

Now  and  for all . Putting the value of Cis we have

 

Taking the design-expectation we have

 .

Putting  in the above expression

 

  (10.7.16)

Using (10.7.16) in (10.7.15) we get

  (10.7.17)

So (10.7.8) simplifies to

  (10.7.18)

  (10.7.19)

provided  and  are chosen in an optimum way. Consequently, (10.7.18) provides the lower bound (achievable) for a given  to. Expression (10.7.18) provides an exact result so that V2 provides the lower bound for a given .

It would be useful to work out a case where the solution for  is explicit. In the Ikeda-Lahiri-Midzuno scheme, we set. We then have, where A is any constant. It does not really matter what value of A is chose since y’S is always independent of A. Setting A = 0 we can obtain  and  works out to

 . (10.7.20)

Using the value of  in (10.7.18)

  (10.7.21)

Substituting  from (10.7.20) and on simplification  works out to be

 , (10.7.22)

which is the exact lower bound to  for a given P(s) and is achievable.

**10.7.1 An Approximation to the Lower Bound**

 Equation (10.7.18) gives lower bound to  for estimators which are simultaneously design- and model-unbiased for a given . This however requires the solution for *Ci*’s. Except in special cases this is not available in explicit form and has to be solved iteratively. We now indicate an approximation to the lower bound at (10.7.18).

Put

   

and . From (10.7.13) *CI* is given by

  (10.7.23)

where 

and

  (10.7.24)

Writing  the second term on R.H.S. of (10.7.24) becomes

.

The product term vanishes by writing  in the form of  because

 

 

 = 0,

since  by definition.

Otherwise this is a standard result  where  is the weighted average , the product term vanishes.

So

 

  (10.7.25)

Multiplying both sides of (10.7.23) by , summing over  and rearranging terms we get

 

and 

so that 

This finally leads to

 

From (9.3.3), we get

  (10.7.26)

Equation holds iff . Using Brewer’s (1979) type approximation

 

Consequently we may except for n, from (7.3.4)

  (say) (10.7.27)

It is interesting to note that one (first) part of the bound is minimized if  and the other part is minimized if  (when this part becomes zero). Optimum choice of  would thus seem to lie between the two extreme cases. The numerator of the second part is the variance of the Horvitz and Thompson estimator of Zi values which for large n will be small. This suggested by Brewer (1979), namely . This is borne out by the numerical study in the next section.

**10.7.2 Numerical Comparison**

The calculation of V1, (8.9.1), V2 (10.7.18) and V3 (10.7.27), given earlier, depends on the values of *Zi*,  and *Ci*,  and  [which in turn depends on ]. We consider the populations given in Bayless and Rao (1970) and others from different textbooks. This was done so that *Zi* values correspond to quite a number of real populations. We have set  and 4 different values of γ are selected. These are γ = 0.50, 0.625, 0.750 and 0.875. We have dropped γ = 1 because in this case the optimum strategy is completely known.

For the purpose of conducting empirical studies, two selection procedures have been selected form 100 selection procedures. These are

1. Rao-Sampford (1967) procedure-11
2. Samiuddin-Asad (1981) procedure-51

The main reason for choosing these selection procedures is that they give  and  in explicit form quite easily for any n. We used firstly (i) Rao-Sampford procedure-11 for which

  (10.7.28)

where in all these calculations, we have set , when γ = 0.500, 0.625, 0.750, 0.875 and . For the calculation of V2 (exact bound) first P(s) is used in (10.7.14), i.e.

 . (10.7.14)

Three to four iterations were used when the last two values were close to 4 or 5 decimal places and then the values of *Ci* were used to calculate V2. The value of Ci was taken equal to  as first approximation and this provides us a good starting point for calculations. The following populations, with their Zi’s values given, were used.

Very interesting conclusions emerge from these calculations. However, we must bear in mind that under both systems we are using sampling schemes which would be considered good.

Again, the first important conclusion in that Godambe-Joshi lower bound provides a rather tight bound even for n as low as 2. One need only to compare V1 (Godambe-Joshi bound) with V2 (the exact achievable lower bound) for each pair of . The second important conclusion is that choice of  reflected by the choice of δ has practically no effect on , given the selection procedure in Samiuddin–Asad (1981). To compare the values of V1 and V2 in the same row, notice that V1 is always less than V2 as it should be. A third point of some interest is that the approximate lower bound V3 is always between V1 and V2. Note that the lowest value of V2 does not occur when γ = δ in case n = 2 but the lowest value of V2 does occur when γ = δ as Brewer (1979) suggests in case n = 4. In fact for n = 2, it matches the closing remarks in Section 3. Finally, the conclusions for all other cases remain almost exactly as is the case with Rao-Sampford procedure.

It must be recalled that using this approximation we remarked that optimum choice is likely to lie between  and . This is borne out by the value of  given under Exact Bound (V2) for each population. One has to look to the minimum in each row in V2. It will be noticed that generally the lowest value does not lie along the diagonal but is placed slightly towards the case  which corresponds to δ = 1. The most important finding that emerges from these calculations are rather surprising.

For each given γ, the values of the new bound for different values of δ differ very little. This implies that the choice  is of little consequence when one is using optimum estimator. This matches with the comment of Smith (1976) that a good estimator can overcome a bad design, but a good design can be ruined by the use of a bad estimator.

Another possibility of course is that Godambe’s criterion is not very sharp in discriminating between different strategies. It is known, for example, that when , then the Horvitz-Thompson estimator with  and the ratio estimator with Lahiri’s scheme leads to the same value of . A choice between the two strategies cannot, therefore, be made on the basis of the criterion. We now see that the criterion is rather insensitive to the choice of  when optimum estimator (i.e. in the class of simultaneously design- and model-unbiased estimators) are used. However, we have not pursued the point further.

**10.8 The Class of Generalization Murhty Estimator**

Consider a draw – by – draw procedure (Yates and Grundy procedure) in which the first unit (say *i*) of the population is drawn with probability proportional to sizes i.e. , where  is the measure of sizes and . After this the order units of the sample are drawn with whatsoever probabilities we wish. Lets *P* (s|*I*) denote the probability of drawing sample *s* given the first unit drawn is *I*. We have then

  (10.8.1)

A general class of estimator (Generalized Murthy’s estimator)  therefore is

  (10.8.2)

  (10.8.3)

It can be easily proved that is design and model – unbiased. For design – unbiasedness, the design – expectation of (10.8.2) gives

 

  (10.8.4)

Using (10.8.1) in (10.8.4) we get

 

Therefore, the generalized Murthy’s estimator is design – unbiased.

For model – unbiased ness,  using the model (6.8.2)

 

 

Taking model – expectation

 

 

therefore is model – unbiased as well.

 It can be easily proved that under the Ikeda – Midzuno – Lahiri selection procedure, the *generalized Murthy’s class of estimator* reduces to classical ratio estimator. For this, the first unit is selected with probability proportional to size and the remaining *n* – 1 units are selected with equal probability and without replacement.

Then  (10.8.5)

and  (10.8.6)

Using (10.8.5) and (10.8.6) in (10.8.2) we get

 

which is a classical ratio estimator.

* + 1. **The New Selection Procedure**

The new suggested selection goes like Raj’s (1956) [Yates and Grundy – procedure] up to the second unit selection and like Ikeda – Lahiri – Midzuno, for the rest of the units to be in the sample, i.e.

**Step 1:** *Select the unit (say i) with probability proportional to size, i.e., , select the second unit (say J) with probability proportional to the remaining (N – 1) units, i.e., .*

**Step 2:** *Select the remaining (n – 2) units from N – 2 population units with equal probability and without replacement, i.e. .*

Under this selection procedure, the *P(S|i)* will be

 

 

  (10.8.7)

where .

Using (10.8.7) in (10.8.2), we get

  (10.8.8)

This is *like a classical ratio estimator* where the larger units are given relatively less weight than the smaller units in the computation of ratio.

 This is quite a fascinating result. Its most appealing feature which combines model – unbiasedness under (6.8.2) with exact design – unbiasedness is shared with Raj (1956) and Murthy (1957) estimators. Because of this, a name has be suggested as Generalized Murthy’s Estimator. A more appropriate name would be Generalized Ikeda – Lahiri – Midzuno estimator, since it was Lahiri who first suggested the sampling scheme in 1951 to make the classical ratio estimator design – unbiased and Midzuno’s (1952) contribution was merely to show how that sampling could be achieved most simply. The new scheme (selection procedure) stands in the same relation to the Horvitz – Thompson ratio estimator as the unbiased Ikeda – Lahiri – Midzuno ratio estimator for a sample selected with  does to the classical ratio estimator for a sample selected with simple random sampling without replacement (of course, the unbiased Ikeda – Lahiri – Midzuno ratio estimator is identical in form to the classical ratio estimator.

Note that this class of estimators will be helpful in devising a sample selection procedure which commences with a single draw for the selection probabilities are proportional to  and for which the eventual probabilities , are proportional to .

The possibility of -balancing the sample (to use Cumberland and Royall’s terminology) is not peculiar to this new class of estimators. It could equally apply to the Horvitz – Thompson ratio estimators, to Brewer’s (1979) estimators, to the generalized regression estimator (Cassel – Sarndal and Wretman – 1976) and to several others.

Corresponding to a specified *P(s|i)* we have from (10.8.2)

 

Using the model (6.8.2) we will have

 

  (10.8.9)

where . Now

 

Since  (10.9.1), we therefore have

 

and

 

  (10.8.10)

Then

 

 

  (9.\_\_\_\_\_\_)

which gives the Godambe – Joshi (1965) lower bound. The equality sign holds only if  for *s* containing *I*.

Now

 

which leads to

 

where . This class of Murthy’s estimator is interesting because of the possibility of setting  [Brewer – 1979] and at the same time of achieving

 

at least in some cases. This contradiction of  suggest some kind of balancing of samples [Royall and Herson = 1973; Cumberland and Royall – 1981] as mentioned before.

An empirical studies of the performance of this new estimator for samples of sizes 2, 3 and 4 using Rao-Samford (1965, 1967) selection procedure. The new estimator is compared with the

1. Horvitz and Thompson estimator (8.1.1). Ratio estimator using Lahir (1951) and Midzuno (1951) selection procedures.
2. Murthy’s estimator

 . (6.2.3)

1. The new or revised ratio estimator is (10.8.8)

For this purpose, 19 real and artificial population have been chosen which have been either used by Rao and Bayless (1969) or Bayless and Rao (1970), or are available in different books on sampling techniques.

We now consider the results of an empirical investigation in terms of the variances of the four estimators considered in Section 4. the prime interest of the investigation is to see if  improves upon  in a general sense and how  performs in relation to other estimators. We consider samples of size *n* = 2, 3 and 4. Table 6.3 gives the variances of the four estimators for *n* = 2, 3 and 4 for each of the 19 populations.

The following conclusion can be made from this study.

1. For *n* = 2,  = .
2. Generally, the  lies between  and 
3. Whenever the performance is very good for ratio estimator,  (see populations 1, 2, 5, 6, 8, 16), the performance of  is also very good and generally does not lose much on . Whenever the performance of  is bad (see populations 3, 9, 11, 12, 13, 14, 15, 17, 18 and 19), then the performance of  definitely improves over . We take it to imply that the performance of  is more stable than that of .
4. The overall performance of  is as good as that of  and 