AN ESTIMATOR OF CHANGE FROM AN UPDATED FRAME

Chapman F. Gleason, U.S. Department of Agriculture

Introduction

Estimation of totals and change is a basic procedure from periodic surveys. Current text books (Cochran [1], Raj [6], Hansen, Hurwitz, Madow [3]) treat only the case of successive sampling from finite populations when the sampling frame remains fixed between occasions. Lowerre [5] treats the birth/death estimation problem by using regression estimators for estimates of totals on the second occasion. Population sizes on both occasions are unknown. The case where population sizes are known (in all strata) on both occasions and where the number of births and deaths to the "old" sampling frame to construct the "new" sampling frame are also known will be treated here. A combined ratio estimate will be used to estimate the total and the percent change on the current occasion.

As an example consider a frame of business, units can:
1) merge with another unit,
2) spin off a subsidiary,
3) go out of business, or
4) a new unit can be created and go into business.

In general only two types of changes can happen to a sampling frame; a frame can:
1) add units (called birth or adds), or
2) loose units (called deaths or deletes).

Any type of merger or subsidiary spinoff of a new company or business enterprise can be treated as a death to the "old" sampling frame and an add to the "new" sampling frame. A unit which remains the same on each sampling frame is said to be "static." Changes in stratum membership are treated as deletes to one stratum and adds to a stratum on the new sampling frame. The mathematical and statistical formulation of the problem will be given in the next section.

Sampling Design and Estimators

Let F be a stratified sampling frame at a previous point in time (call it the "old" frame), and let F* be the sampling frame F updated with adds FA and deletes FD. Suppose the number of strata and the stratum definitions do not change between updates. However, the i th unit can change stratum membership in the update process. So in set notation,

\[ F* = (F - FD) \cup FA. \]

Conceptually the following figure depicts the situation, \( F_S = (F \cap F*) \) is the "static" part of the frame. Figure 1

Let S be a sample of size n selected from frame F on the previous occasion and S* be an independently drawn sample of size n* from frame F* on the current occasion, further let S' be a subsample of size n' selected from S. Let SA be a sample of size nA for FA.

Enumerate the samples S', S* and S A on the current occasion. The previous sample S has already been collected. The subscript h will denote stratum membership. Let \( \bar{y}_{ih} \) be the value of the i th unit in the h th stratum on the current occasion selected from the S* sample, let \( \bar{y}_{ih} \) and \( \bar{y}_{ih} \) be the value of a variable on the i th unit in the S' subsample on the current occasion and the previous occasion respectively. The m subscript denotes matched.

Further let the number of units in the h th stratum on the sampling frame F be Nh, similarly define the number on the frame F* as Nh*, and let NhA be the population size of the add frame F for the h th stratum. The design can be summarized as:

\[
\begin{array}{c|c|c|c}
\text{Frame} & F & F* & FA \\
\hline
\text{Population size} & N_h & N_{h*} & N_{hA} \\
\hline
\text{Samples selected} & S, S* & S', S* & S_A \\
\text{from frame} & & & \\
\hline
\text{Sample size} & & & \\
\text{in} & & & \\
\text{h-th stratum} & & & \\
\hline
\end{array}
\]

Mathematically, a more complex description of the estimators is given below. The simple unbiased estimator of the total for the frame F* on the current occasion using only the S* sample is

\[ Y_{F*} = \sum_{h=1}^{L} N_h \bar{y}_{ih}, \]

where,

\[ \bar{y}_{ih} = \frac{\sum_{i=1}^{n_h} y_{ih}}{n_h} \text{ i.e. the h-th stratum mean.} \]

Another estimator of the frame F* total based on the sample of adds (SA) and the subsample S' is the combined ratio estimate:

\[ Y_{RF*} = \sum_{h=1}^{L} N_h \bar{y}_{ih} + \bar{y}_{FA}, \]

where,

\[ \bar{y}_{FA} = \sum_{h=1}^{L} N_{hA} \bar{y}_{hA}. \]
\( \hat{y}_{hA} = \frac{n_{hA}}{1-f_{hA}} \hat{y}_{hA1} / n_{hA} \) (add sample mean for the \( h \)th stratum) and

\[
(3) \quad \hat{y}_{F_S} = \hat{Y}_F = \frac{1}{N_{h}} \sum_{h=1}^{L} \hat{y}_{h} \hat{y}_{hm} \left( \sum_{h=1}^{L} N_{h}^{-1} \hat{y}_{h} \right).
\]

So the estimator \( \hat{Y}_{F_S} \) is the sum of a combined ratio estimate of the portion of the frame which is static from the previous occasion to the current occasion \( (F_S) \) and a simple unbiased estimator of the add population total.

Other estimators could also be used for the situation just described. Consider the frames \( F_D, F_S \), and \( F_A \) as the domains \( a, ab \) (ba) and \( b \) in the multiple frame situation described by Hartley [2]. Hartley's estimator could also be used to estimate for the frame \( F^* \) total.

Variances

In this section we show the variances of the estimators, \( (1) \) and \( (2) \). Now the variance of \( (1) \) is the usual variance of a simple linear unbiased estimator of the total, namely:

\[
\text{Var}(\hat{Y}_{F_S}) = \sum_{h=1}^{L} \frac{N_{h}^{2}(1-f_{h})}{N_{h}^{2} n_{h}^{2}} \text{Var}(\hat{y}_{h})
\]

where \( f_{h} = n_{h} / N_{h} \) is the finite population correction factor for the updated frame \( F^* \). We now derive the variance of the estimator \( (2) \).

\[
\text{Var}(\hat{Y}_{R,F^*}) = \text{Var}(\hat{Y}_{F_A} + \hat{Y}_{F_S}) = \text{Var}(\hat{y}_{F_A}) + \text{Var}(\hat{y}_{F_S}),
\]

since the add sample \( S_A \) was selected independent of \( S' \).

Now,

\[
\text{Var}(\hat{y}_{F_A}) = \sum_{h=1}^{L} \frac{N_{h}^{2}(1-f_{h})}{N_{h}^{2} n_{h}^{2}} \text{Var}(\hat{y}_{h})
\]

where \( f_{h} = n_{h} / N_{h} \h_{h} \) the sampling fraction for the add frame. Now consider

\[
\text{Var}(\hat{y}_{F_S}) = \text{Var}(\hat{Y}_F),
\]

where, \( R \) is defined in \( (3) \) and \( \hat{Y}_F = \sum_{h=1}^{L} \hat{y}_{h} \).

\( \hat{Y}_F \) is a double sampling combined ratio estimator of the total for frame \( F \) on the 2nd occasion. But noting that units that were deleted from the frame \( F \) are 0 on the current occasion, and since \( F = F_D U F_S \), any estimator of \( F \) on the current occasion is 0, therefore \( \hat{Y}_F \) estimates \( F_S \) for the current occasion.

Since \( S' \) was a subsample of \( S \) and estimates the total \( Y \) for the frame \( F \) on the 2nd occasion, generalizing the variance formula which appears in Knoijn [4, page 126] to a combined ratio estimator for a stratified random sample we get

\[\begin{align*}
(5) \quad \text{Var}(\hat{Y}_F) &= \sum_{h=1}^{L} \frac{N_{h}^{2}(1-f_{h})}{N_{h}^{2} n_{h}^{2}} \text{Var}(\hat{y}_{h}) \\
&= \sum_{h=1}^{L} \frac{N_{h}^{2}(1-f_{h})}{N_{h}^{2} n_{h}^{2}} \text{Var}(\hat{y}_{h}) \\
&= \left( \frac{n_{h} - n_{h}^{'}}{n_{h}^{''}} \right) \text{Var}(\hat{y}_{h}) \\
&= \left( \frac{n_{h} - n_{h}^{'}}{n_{h}^{''}} \right) \left( \sum_{h=1}^{L} \frac{N_{h}^{2}(1-f_{h})}{N_{h}^{2} n_{h}^{2}} \text{Var}(\hat{y}_{h}) \right) \\
&= \left( \frac{n_{h} - n_{h}^{'}}{n_{h}^{''}} \right) \left( \sum_{h=1}^{L} \frac{N_{h}^{2}(1-f_{h})}{N_{h}^{2} n_{h}^{2}} \text{Var}(\hat{y}_{h}) \right) \\
&= \left( \frac{n_{h} - n_{h}^{'}}{n_{h}^{''}} \right) \left( \sum_{h=1}^{L} \frac{N_{h}^{2}(1-f_{h})}{N_{h}^{2} n_{h}^{2}} \text{Var}(\hat{y}_{h}) \right) \\
&= \left( \frac{n_{h} - n_{h}^{'}}{n_{h}^{''}} \right) \left( \sum_{h=1}^{L} \frac{N_{h}^{2}(1-f_{h})}{N_{h}^{2} n_{h}^{2}} \text{Var}(\hat{y}_{h}) \right) \\
&= \left( \frac{n_{h} - n_{h}^{'}}{n_{h}^{''}} \right) \left( \sum_{h=1}^{L} \frac{N_{h}^{2}(1-f_{h})}{N_{h}^{2} n_{h}^{2}} \text{Var}(\hat{y}_{h}) \right) \\
&= \left( \frac{n_{h} - n_{h}^{'}}{n_{h}^{''}} \right) \left( \sum_{h=1}^{L} \frac{N_{h}^{2}(1-f_{h})}{N_{h}^{2} n_{h}^{2}} \text{Var}(\hat{y}_{h}) \right) \\
&= \left( \frac{n_{h} - n_{h}^{'}}{n_{h}^{''}} \right) \left( \sum_{h=1}^{L} \frac{N_{h}^{2}(1-f_{h})}{N_{h}^{2} n_{h}^{2}} \text{Var}(\hat{y}_{h}) \right)
\end{align*}\]

We note that if \( n_{h} = n_{h}^{'} \) (i.e. reinterview the entire sample \( S \)), then \( (3) \) becomes

\[
\hat{Y}_{F_S} = \sum_{h=1}^{L} \frac{N_{h}^{2}(1-f_{h})}{N_{h}^{2} n_{h}^{2}} \text{Var}(\hat{y}_{h})
\]

and \( (5) \) becomes the variance of the appropriate linear unbiased estimator. Now if \( n_{h} = N_{h} \)

(i.e. the previous survey was a census) the estimator \( (3) \) is the ordinary combined ratio estimator with a known total, \( (3) \) reduces to the appropriate formula given in Cochran [1, page 176 formula (6.25)].

Application

The Economics, Statistics, and Cooperative Service conducts a quarterly survey in December, March, June and September to estimate hog inventories. A sample is selected from a stratified list frame, this sample is used four quarters with simple linear unbiased estimates \( \hat{Y}_F \) made of totals for each item each quarter. The list frame is updated with births, deaths, and stratum changes after the September survey and a completely new sample is drawn from the updated frame for the next sequence of quarterly surveys. In the theory section developed earlier the old frame is \( F \) and the updated frame is \( F^* \).

A State Statistical office agreed to collect data from a subsample of their September 1979 sample \( S' \) in the theory developed earlier) in addition to the required sample for December 1979 \( S^* \). Data was already available for the entire September 1979 sample \( S \). Further, adds to the sampling frame \( F \) were identified and a Sample \( S_A \) selected from this frame \( F_A \) in our theory section. We now have the necessary data to compute the estimators \( (1) \) and \( (2) \) and compare their precision using \( (4) \) and \( (5) \). Table I gives the estimates and their precision.

<table>
<thead>
<tr>
<th>Occasion</th>
<th>Estimator</th>
<th>Estimate</th>
<th>S.E.</th>
<th>C.V.</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous</td>
<td>( \hat{Y}_F )</td>
<td>3,265</td>
<td>149</td>
<td>4.6</td>
<td>1801</td>
</tr>
<tr>
<td>Current</td>
<td>( \hat{Y}_{F*} )</td>
<td>3,650</td>
<td>113.9</td>
<td>3.1</td>
<td>1800</td>
</tr>
<tr>
<td>Current</td>
<td>( \hat{Y}_{R,F*} )</td>
<td>3,549</td>
<td>493</td>
<td>13.9</td>
<td>507</td>
</tr>
</tbody>
</table>

Note that the sample size on the estimator \( \hat{Y}_{R,F*} \) is about 3.5 times smaller than the sample size for estimator \( \hat{Y}_{F*} \). For a fixed size sample, estimator \( \hat{Y}_{R,F*} \) is more precise than \( \hat{Y}_{F*} \), since
CV(Previous occasion) > 1/2 CV(Current occasion) in all strata [1, p. 165]. This is shown in Table 2.

Table 2. Correlation Coefficients between September and December 'total hogs and pigs' based on subsample S' of September sample S.

<table>
<thead>
<tr>
<th>Stratum Definition</th>
<th>Correlation Coefficient</th>
<th>CV(September)</th>
<th>CV(December)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No livestock</td>
<td>1.00</td>
<td>.75</td>
<td></td>
</tr>
<tr>
<td>No hogs</td>
<td>.83</td>
<td>.54</td>
<td></td>
</tr>
<tr>
<td>1-99</td>
<td>.79</td>
<td>.50</td>
<td></td>
</tr>
<tr>
<td>100-199</td>
<td>.63</td>
<td>.58</td>
<td></td>
</tr>
<tr>
<td>200-299</td>
<td>.92</td>
<td>.56</td>
<td></td>
</tr>
<tr>
<td>300-399</td>
<td>.69</td>
<td>.49</td>
<td></td>
</tr>
<tr>
<td>600-1249</td>
<td>.97</td>
<td>.53</td>
<td></td>
</tr>
</tbody>
</table>

Conclusion

This paper attempts to apply existing theory to an important problem in estimation from periodic surveys when the sampling frame changes between survey periods. More research is needed in the following areas:

1. More efficient estimation of both change and totals on the current occasion.
2. Optimum allocation of the sample (given the estimators of course).

Bibliography