Notes on Cluster Sampling
for Statistics 36-303: Sampling, Surveys and Society

Stephen E. Fienberg
Department of Statistics
Carnegie Mellon University
March 18, 2003

I. Why Use Clustering?

After stratification the most natural extension to simple random sampling involves the use of clusters of the population of interest. In stratified sampling, we divide the population into distinct subpopulations called \textit{strata}, and within each stratum we select a separate sample. In cluster sampling, we divide the population up into clusters, and we select a sample of clusters and include all of the elements from these clusters in the sample. Figure 1, reproduced from Lohr (1999), illustrates diagrammatically the similarities and differences between these approaches.

There are two primary reasons for clustering:

1. It is often the case that a reliable list of elements of the population is unavailable and it is unreasonably expensive to try to compile such a list. We can, however, make a list of clusters and thus it is sensible to use them as the sampling units. This is often the case when we sample human populations and the clusters are households. This is because it is relatively easy to prepare and maintain a list of household locations, whereas it is virtually impossible to maintain a list of individuals in identifiable locations.

2. Even when a list of individual elements of the population is available, the cost incurred in actually locating one element of a cluster is far greater than that associated with getting the information from all other elements. For example the travel costs associated in going from one housing unit to another in a random sample of units may be substantial. Further, when the cluster consists of a household, one individual can provide information on all the other members.
Figure 1. Similarities and Differences Between Cluster Sampling and Stratified Sampling
Unlike stratification, which if used wisely increases the precision of our sample for a fixed overall sample size, clustering tends to decrease the precision of our sample for a fixed overall sample size. This is because individuals within a cluster tend to be more alike with respect to various characteristics than those between clusters. The basic ideas about clustering are relatively easy to understand and in these notes we explore the effects of clustering primarily by example, concentrating on the estimation of attributes. The basic ideas carry over to estimation using measurements and we indicate the nature of the impact of clustering in such situations. Later in the course, we will examine in detail the design actual surveys that combine both clustering and stratification to achieve the advantages in cost from the former while preserving some of the precision that stratification confers.

We limit our attention in these notes to clusters of equal size, primarily to keep the notation and the algebra simple, but the basic ideas of cluster sampling carry over to unequal size clusters.

II. Some Basic Elements of Cluster Sampling

Example 1: Suppose we have a population of size $2N$ composed of $N$ families of size 2, a husband and wife. We say that the families are clusters of size $M = 2$. Further suppose that the husband and wife in any given pair are exactly the same age. We are interested in the proportion of the population eligible for Medicare corresponding to those over the age of 65, and we take a sample of $n$ families. Since both members of each family have the same age we in effect have redundant information and instead of ending up with an overall sample of size $2n$ individuals, our effective sample size is only $n$.

Now suppose that husbands and wives don't have identical ages, but on average older husbands have older wives and younger husbands have younger wives. This positive association or correlation between the age of the husband and the age of a wife in a pair again reduces the "effective sample size" associated with our cluster sample of $n$ families. We get an estimate that is more accurate than a simple random sample of $n$ individuals from the population, but still less than a simple random sample of size $2n$.

This example illustrates the basic impact of clustering that we will tend to observe in the sampling of human populations. In general we consider a population of $NM$ elements subdivided into $N$ clusters of size $M$. We take a sample of $n$ of these clusters and incorporate into our sample information on all $M$ elements in each of the selected clusters. Thus we record information on $nM$ units. If the information from individuals within a cluster is positively related, then there is less variation among individuals within a cluster than for the same number of individuals drawn from different clusters. Thus we expect that our cluster sample will be less accurate than a simple random sample of the same size, $nM$.

The actual formulas for the variance of an estimate for a proportion or for a mean from a cluster sample involve measures of variation within and between clusters.
III. Cluster Sampling for Attributes

We now consider the estimation of the proportion \( p \) of the population of size \( NM \) possessing an attribute. Let \( p_i \) be the proportion of elements in the \( i \)th cluster possessing the attribute. Then the cluster sample estimate of \( p \), \( \bar{p}_c \), is just the average of the values of \( p_i \) for the sampled clusters,

\[
\bar{p}_c = \frac{1}{n} \sum_{i=1}^{n} p_i
\]

(1)

and, since the clusters are of equal size, this estimate is simply the overall sample proportion of individuals with the attribute of interest. Thus our cluster sample selection procedure assigns each individual in the sample the same chance of selection and the sample is self-weighting.

We get the variance of \( \bar{p}_c \) by treating the values of \( p_i \) as a sample of \( n \) measurements and looking at their variation when used to estimate the overall population proportion \( p \), i.e.,

\[
Var(\bar{p}_c) = \frac{(N-n)}{(N-1)} \frac{\left( \sum_{i=1}^{N} (p_i-p)^2 \right)}{n}.
\]

(2)

We typically want to compare the accuracy of our estimate, \( \bar{p}_c \), with that of a simple random sample of the same overall sample size, \( nM \), i.e.,

\[
Var(\bar{p}) = \frac{(NM-nM)}{(NM-1)} \frac{p(1-p)}{nM}
\]

(3)

One way to compare the variances involves the “correlation” between elements in the same cluster, \( \rho \). This quantity \( \rho \) is called the intrACLuster correlation coefficient. It turns out that

\[
\frac{Var(\bar{p}_c)}{Var(\bar{p})} = 1 + (M-1)\rho.
\]

(4)

Since variances cannot be negative, the quantity \( 1 + (M-1)\rho \) can’t be negative and the minimum possible value for \( \rho \) is \( \frac{-1}{(M-1)} \), which tends to 0 as the cluster size \( M \) gets large. Unlike a regular correlation coefficient which takes values between 1 and -1, the intrACLuster correlation coefficient runs between 1 and \( \frac{1}{(M-1)} \). If \( \rho > 0 \), the cluster provides less precision than a random sample of \( M \) individuals, whereas when \( \rho < 0 \), something which occasionally happens, the use of clusters is more precise.

**Example 2: Limited English Proficiency Students.** The U. S. Department of Education is interested in determining the number of elementary school children in public schools with limited proficiency in English (LEP). Suppose there are \( N = 20 \) schools in a given
district and that each school has $M = 100$ students. Investigators take a sample of $n = 5$ schools and gather information on the proportion of LEP students in each school. The total sample size is number of schools sampled times the number of children in each school, i.e., 500. Because of housing patterns we expect children in a given school to be more alike with respect to their proficiency in English than those in different schools. This is because immigrants to the district from a given country often live close to one another for economic or other reasons. Suppose we can determine that the intra-school correlation coefficient for the attribute LEP is 0.0383. Then the ratio of the variance of the cluster sample of 5 clusters to the variance we would have had if we had taken a simple random sample of students from a across the school district of size 500 is

$$1 + (M - 1)\rho = 1 + 99 \times 0.0383 = 4.79$$

i.e., the variance of the cluster sample is almost 5 times greater than that of a simple random sample of equivalent size. Put another way, we could have taken a simple random sample of 100 children from across the school district and achieved an estimate of the proportion of LEP students with equivalent accuracy.

Why then did investigators choose their sample in this way? The answer is cost. Suppose that the cost of going to a school and setting up a language test is $1000, whereas the cost of administering the test to the student one set up is $10. Taking a random sample of 100 students would have meant that they would have gone to at least, say, 10 schools, perhaps more. The cost of administering the test in 10 schools is

$$(10 \times $1000) + (100 \times $10) = $11,000.$$ 

Instead, the investigators went to only 5 schools and thus their cost was

$$(5 \times $1000) + (100 \times $10) = $6,000.$$ 

Thus, by taking a cluster sample the investigators incurred only $(6/11)$th of the cost associated with a simple random sample of equivalent precision.

When the population size is large and we can ignore the finite population correction, and we can rewrite the ratio of the variance for cluster sample versus that for a simple random sample directly as

$$\frac{Var(\bar{p}_c)}{Var(\bar{p})} \leq \frac{M \sum_{i=1}^N (p_i - p)^2}{Np(1 - p)} \quad (5)$$

What we have in these formulas for the variances is three different measures of variability: (i) the overall population variance, i.e., the variance of the observations for samples of size 1, $p(1 - p)$; (ii) the subpopulation for the $i$th cluster, i.e., variance of the observations for samples of size 1 within the $i$th cluster, $p_i(1 - p_i)$; and (iii) the variation of the cluster proportions about the overall population proportion, $\sum (p_i - p)^2$.

A very famous formula in statistics links these quantities as follows:
\[ NMp(1 - p) = M \sum_{i=1}^{N} (p_i - p)^2 + M \sum_{i=1}^{n} p_i(1 - p_i) \]

By substituting for the sum of squares between clusters the formula for the ratio of variances we get

\[ \frac{\text{Var}(\bar{p}_c)}{\text{Var}(\bar{p})} \approx \frac{M \sum_{i=1}^{N} (p_i - p)^2}{Np(1 - p)} = \frac{NMp(1 - p) - M \sum_{i=1}^{N} p_i(1 - p_i)}{Np(1 - p)} = M \left( 1 - \frac{\sum_{i=1}^{N} p_i(1 - p_i)}{Np(1 - p)} \right) \]  

The fraction in expression (7) in brackets tends is always less than or equal to 1. Thus the variance of cluster sampling is never more than \( M \) times that of an equivalently sized simple random sample. But often the fraction is non-negligible and we get little degradation from the clustering relative to simple random sampling. In the other extreme, when all of the clusters are identical and \( p_i = p \) for all clusters, the fraction is 1 and ratio of the variances is approximately equal to 0, i.e., cluster sampling has yielded a fantastic gain: in this instance, once you’ve seen one cluster you in effect have seen them all! Williams (1978, Chapter 11) gives a simple example of this phenomenon.

**Example 3: Numerical Illustrations.** Suppose that the proportion of individuals in a population with a specific attribute is \( p = 0.5 \). If the population consists of \( N = 20 \) clusters, where for 10 clusters \( p_i = 0.25 \) and for the other 10 \( p_i = 0.75 \), then

\[ \frac{\text{Var}(\bar{p}_c)}{\text{Var}(\bar{p})} \approx M \left( 1 - \frac{\sum_{i=1}^{N} p_i(1 - p_i)}{Np(1 - p)} \right) = M \left( 1 - \frac{3}{4} \right) = \frac{M}{4}. \]

Thus for clusters of size \( M = 2 \) and \( M = 3 \), there is less variability in a cluster sample than in a simple random sample! For \( M \geq 4 \), however, cluster sampling is less efficient, since the ratio is less than 1. As the cluster size \( M \) increases in this example, we have greater and greater degradation in precision associated with cluster sampling.

**Example 2 (Cont.):** Suppose the distribution of the number of LEP students in the 20 schools in our school district is as follows:
<table>
<thead>
<tr>
<th>Number of School</th>
<th>Proportion of LEP Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.1</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>10</td>
<td>0.4</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
</tr>
</tbody>
</table>

The overall proportion of LEP students is \( p = 0.326 \). Thus \( Np(1 - p) = 20 \times 0.326 \times 0.674 = 4.39 \) and from the information in the table we calculate

\[
\sum_{i=1}^{N} p_i(1 - p_i) = 4.19. \tag{9}
\]

Since \( M = 100 \) the ratio of the variances is

\[
\frac{Var(\bar{Y}_c)}{Var(\bar{Y})} \approx M \left( 1 - \frac{\sum_{i=1}^{N} p_i(1 - p_i)}{Np(1 - p)} \right) = 100 \left( 1 - \frac{4.19}{4.39} \right) = 4.56. \tag{10}
\]

Thus we get the precision of the cluster sample is equivalent to that of a simple random sample of about 20% the sample size. The difference between the ratio of 4.56 computed here and that of 4.79 computed in our earlier look at this example is due to ignoring the finite population correction here.

### IV. Cluster Sampling for Measurements

The same ideas carry over for cluster samples of measurements. We now consider the estimation of the mean \( \mu \) of the population of size \( NM \). Let \( y_i \) be the sum of measurements for the \( i \)th cluster. Then the cluster sample estimate of \( \mu \), \( \bar{y}_c \), is just the average of the all of measurements for all \( n \) sampled clusters

\[
\bar{y}_c = \frac{1}{nM} \sum_{i=1}^{n} y_i. \tag{11}
\]

Our cluster sample selection procedure assigns each individual in the sample the same chance of selection and the sample is self-weighting. We get the variance of \( \bar{y}_c \) by treating the values of \( y_i \) as a sample of \( n \) measurements and looking at their variation about their mean value for all \( N \) clusters. As we did with attributes, we compare the accuracy of our estimate, \( \bar{y}_c \), with that of a simple random sample \( \bar{y} \) of the same overall sample size, \( nM \), and the ratio turns out to be the same as it was for attributes:

\[
\frac{Var(\bar{y}_c)}{Var(\bar{y})} = 1 + (M - 1)\rho, \tag{12}
\]
where \( \rho \) is the intracluster correlation coefficient.

**Example 4:** Henry (1990, pp. 107-109) gives an example where and , in which the estimated variances are

\[
\begin{align*}
    s^2(\bar{y}_c) &= 30.17 \\
    s^2(\bar{y}) &= 20.23
\end{align*}
\]

Thus there is an increase in the variance of about 50% due to clustering. This means that a simple random sample of equivalent precision to the cluster sample would have required only a sample of size

\[
\left( \frac{30.17}{20.23} \right)^{15} = 10.05
\]

or 10. For this example, the estimated intracluster correlation coefficient is

\[
\bar{\rho} = \frac{1}{M - 1} \left( \frac{s^2(\bar{y}_c)}{s^2(\bar{y})} - 1 \right) = \frac{1}{4} \left( \frac{30.17}{20.23} - 1 \right) = 0.123. \tag{13}
\]

**V. Unequal Cluster Sizes**

The results for clustering described here work out in a similar way when the clusters are not of equal size although the actual formulas are somewhat more complicated. Choosing compact clusters of roughly equal size turns out to be the most efficient way to do cluster sampling, but in real life clusters often come in varying sizes and there is little we can do about this. For example, households come in sizes that typically vary from 1 (for single person households) to as many as 12, or even more. The ratio of the variance for a cluster sample to that of a simple random sample still takes approximately the same form,

\[
\frac{Var(\bar{y}_c)}{Var(\bar{y})} \approx 1 + (\bar{M} - 1)\rho, \tag{14}
\]

where \( \bar{M} \) is the average cluster size.
**Historical Note:** The formal ideas for the use of cluster sampling were introduced in the same the pioneering paper by Jerzy Neyman in 1934 in which he advocated the use of optimal allocation for stratified sampling, and it was his combination of stratification and clustering this had such a profound influence on subsequent developments in the field of sampling and in large-scale survey practice.

**Note on Proofs of Results:** Cochran (1977) provides formal derivations for some of the formulas presented here for cluster sampling as well as for multi-stage cluster sampling and cluster sampling when the clusters are of unequal size. Lohr (1999) also has an excellent presentation on the relevant formulas and their derivation.

**References:**


