36-303: Sampling Surveys and Society
Clustered Sampling Examples

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

March 16, 2010

Contents

1 Key Ideas

1.1 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 *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), indicates some similarities and differences between these approaches.

There are two primary reasons for clustering:

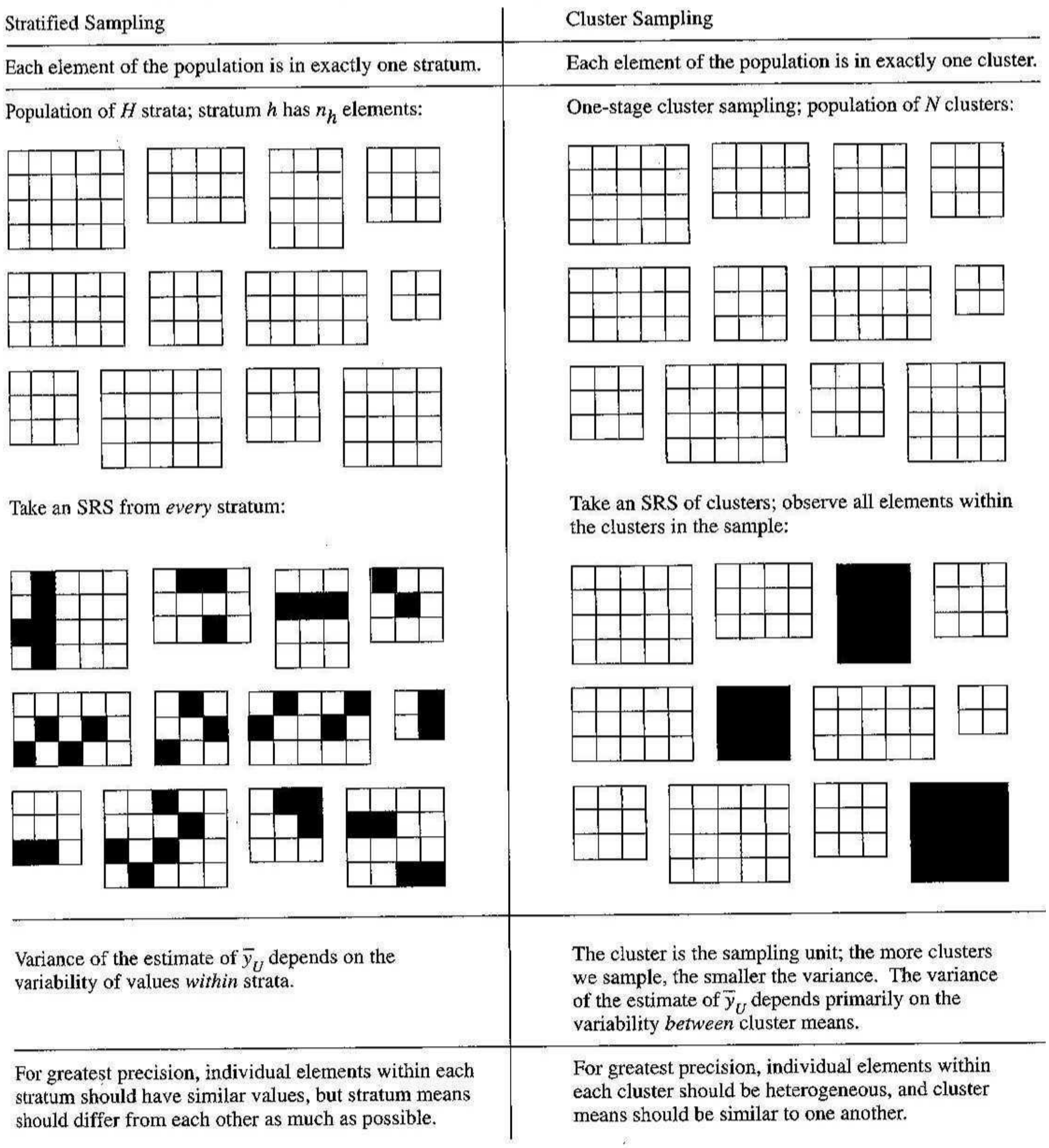
1. *A reliable list of elements of the population may be unavailable and it may be unreasonably expensive to try to compile such a list.* We can, however, make a list of clusters and thus it is sensible to use them as the sampling units. For example:

    - 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.
    - You could also imagine doing this on-campus. C-Book is a flawed frame for CMU undergraduates, but the Hub has an exhaustive list of classes and their locations, so you could take an SRS of classes from the Hub's list, rather than an SRS of students: the clusters are the classes.

2. *Even if a reliable list of population elements is available, it may be difficult, expensive or disruptive to take an SRS of individuals.* On the other hand, and SRS of clusters may be easier. For example:

    - The travel costs associated in going from one housing unit to another for a random sample of individuals may be substantial. Further, when the cluster consists of a household, one individual (e.g. "head of household") can provide information on all the other members.
    - In the National Assessment of Educational Progress, the survey form is an achievement test in math, science, or some other subject. While it would be possible to pull out individual students from a class and send them to a special room for testing, it is much easier and less disruptive to sample classrooms (so the classrooms are the clusters) and give the test to all eligible students in the class.

You might remember that stratification usually increases the precision of our sample (reduces standard errors), relative to an SRS. *Clustering has the opposite effect*: it tends to decrease the pre-



Stratified Sampling	Cluster Sampling
Each element of the population is in exactly one stratum.	Each element of the population is in exactly one cluster.
Population of H strata; stratum h has n_h elements:	One-stage cluster sampling; population of N clusters:
Take an SRS from *every* stratum:	Take an SRS of clusters; observe all elements within the clusters in the sample:
Variance of the estimate of $\bar{y}_U$ depends on the variability of values *within* strata.	The cluster is the sampling unit; the more clusters we sample, the smaller the variance. The variance of the estimate of $\bar{y}_U$ depends primarily on the variability *between* cluster means.
For greatest precision, individual elements within each stratum should have similar values, but stratum means should differ from each other as much as possible.	For greatest precision, individual elements within each cluster should be heterogeneous, and cluster means should be similar to one another.

Figure 1: Similarities and Differences Between Cluster Sampling and Stratified Sampling. From Lohr (1999).

cision of the sample (increase the standard errors). This is because individuals in the same cluster tend to be more alike than different, and this causes their responses to be positively correlated.

We will begin exploring the basic ideas of clustering by looking at examples of the influence of clustering on estimating proportions (p_{pop}'s). Then we will look at how clustering impacts estimationg numerical quantities ($\bar{y}_{pop}$'s).

One does not have to choose either clustering or stratification, and in large scale surveys the two methods are often combined. In coming lectures, we will examine some actual surveys that combine clustering and stratification to achieve the advantages in cost from the former while preserving some of the precision that stratification confers.

It is simpler to introduce the ideas when all the clusters have the same size (all families have exactly 4 members, all classes have exactly 30 students, etc.). The same ideas also work in the more realistic situation of clusters of differing sizes, and we will look briefly at that situation as well.

A key reference is Lohr's (1999) textbook [recommended text for the class].

1.2 Some Basic Elements of Cluster Sampling

Example 1: Using Family Clusters to Estimate the Proportion Eligible for Medicare

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 the 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.

After introducing some terminology and notation we will turn to the actual formulas for the variance of an estimate for a proportion or for a mean from a cluster sample, which involve measures of variation within and between clusters.

1.3 Terminology and Notation

In cluster sampling the notation is a little bit messier than in SRS or even stratified sampling, because of the need to keep track of which observations come from which clusters, and the need to keep track of posive correlations between elements from the same cluster.

- In a clustered sample, the clusters are sometimes called *primary sampling units* or *psu's*. The individuals within a cluster are called *secondary sampling units* or *ssu's*.
- In *one-stage cluster sampling*, we first take an SRS of psu's (clusters). Then *all* of the ssu's (individuals) within each cluster are included in the sample. This is the situation we will focus on in these notes.
- In *two-stage cluster sampling*, we first take an SRS of psu's. Then within each psu, we take an SRS of ssu's.

  For example in a survey of a school district we might take an SRS of schools (psu's, or clusters) and then take another SRS of students (ssu's, or individuals) because it is too expensive to go to every student in every sampled school.

I will try to stick with "clusters" and "individuals" or "clusters" and "units", but keep in mind that psu = cluster, and ssu = an individual or unit within a cluster.

In SRS, we talked of a population of N units. Now N will refer to the clusters or psu's. Within each cluster there are ssu's. The basic data we observe on each observation is

$$\begin{aligned} y_{ij} &= \text{measure for } j^{th} \text{ element of } i^{th} \text{ cluster} \\ &= \text{measure for } j^{th} \text{ ssu in the } i^{th} \text{ psu} \end{aligned}$$

So, y_{ij} might be 0 or 1 depending on which candidate the j^{th} member of household i supports, or y_{ij} might be the GPA of the j^{th} member of classroom i, etc.

- Some population quantities for psu's are:

$$\begin{aligned} N &= \text{number of psu's (clusters) in the population} \\ M_i &= \text{number of ssu's (individuals) in cluster } i \\ K &= \sum_{i=1}^{N} M_i = \text{total number of ssu's in the population} \end{aligned}$$

- Some population quantities for ssu's are:

$$\begin{aligned}
\bar{y}_{pop} &= \frac{1}{K}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = t/K = \text{ the population mean} \\
\bar{y}_{i,pop} &= \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij} = \text{ population mean in the } i^{th} \text{ psu} \\
S^2 &= \frac{1}{K-1}\sum_{i=1}^{N}\sum_{j=1}^{M_i}(y - \bar{y}_{pop})^2 = \text{ population variance of ssu's} \\
S_i^2 &= \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(y - \bar{y}_{i,pop})^2 = \text{ population variance of ssu's within a single psu}
\end{aligned}$$

- Some sample quantities of interest are:

$$\begin{aligned}
\mathcal{S} &= \text{ set of } i\text{'s (psu's; clusters) sampled} \\
\mathcal{S}_i &= \text{ set of } j\text{'s (ssu's; individuals) sampled in } i^{th} \text{ psu } = \text{ set of all ssu's in the sample} \\
n &= \text{ number of psu's in sample (size of } \mathcal{S}) \\
m_i &= \text{ number of ssu's in the sample from the } i^{th} \text{ psu (size of } \mathcal{S}_i) \\
\bar{y}_i &= \frac{1}{m_i}\sum_{j\in\mathcal{S}_i} y_{ij} = \text{ sample mean for the } i^{th} \text{ ssu}
\end{aligned}$$

1.4 Basic Ideas for Estimating Means

From the N clusters in the population we take an SRS without replacement of n of them. Let $\mathcal{S}$ be the set of clusters i sampled, so that $\mathcal{S}$ has n elements.

- The size of each cluster is M_i. For simplicity we assume *equal cluster sizes*: $M_i \equiv M\ \forall\ i$.
- We also will assume that every individual in the cluster is in our sample; this is *one-stage cluster sampling*.

For example, we could consider all two-person households ($M = 2$ for all clusters) in a survey to estimate mean income. y_{ij} is the income of the j^{th} person in the i^{th} household.

For each cluster i in $\mathcal{S}$, we can calculate the cluster mean

$$\bar{y}_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$$

In our example, $\overline{y}_i = \frac{1}{2}(y_{i1} + y_{i2})$, the average of the two person's incomes in the i^{th} household.

Since we have an SRS of clusters, we can apply our formulas for SRS's to estimate population quantities, but now we think of an SRS of *clusters*, where the measurement on each *cluster* is $\overline{y}_i$:

$$\overline{y}_{cl} = \frac{1}{n}\sum_{i\in\mathcal{S}}\overline{y}_i$$

The standard error (SE) needed for constructing confidence intervals is the *square root of*

$$\begin{aligned}
\text{Var}(\overline{y}_{cl}) &= (1-f)S^2_{cl,pop}/n \\
&= (1-f)\frac{1}{n}\left[\frac{1}{N-1}\sum_{i=1}^{N}(\overline{y}_i - \overline{y}_{pop})^2\right] \\
&\approx (1-f)\frac{1}{n}\left[\frac{1}{n-1}\sum_{i\in\mathcal{S}}(\overline{y}_i - \overline{y}_{cl})^2\right] \\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}\left[\frac{1}{n-1}\sum_{i\in\mathcal{S}}(\overline{y}_i - \overline{y}_{cl})^2\right] \\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}s^2_{\overline{y}_i}
\end{aligned}$$

Note that the cluster sample size is n but the individual sample size is $M \cdot n = 2n$. We might compare our SE here with the SE we would calculate if this were an SRS without replacement *of individuals*. So let

$$\overline{y}_{srs} = \frac{1}{Mn}\sum_{i,j}y_{ij} = \frac{1}{Mn}\sum_{j\in\mathcal{S}}\sum_{i=1}^{M}y_{ij}$$

(Note that

$$\begin{aligned}
\overline{y}_{srs} &= \frac{1}{n}\sum_{j\in\mathcal{S}}\frac{1}{M}\sum_{i=1}^{M}y_{ij} \\
&= \frac{1}{n}\sum_{i\in\mathcal{S}}\overline{y}_i \\
&= \overline{y}_{cl}
\end{aligned}$$

so that cluster sampling with equal cluster sizes is *self-weighting*: the complex estimator $\overline{y}_{cl}$ equals the simpler estimator $\overline{y}_{srs}$.)

The variance under SRS would be

$$\text{Var}(\overline{y}_{srs}) = (1-f)S^2_{pop}/(Mn)$$

$$\begin{aligned}
&= (1-f)\frac{1}{Mn}\left[\frac{1}{MN-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{pop})^2\right] \\
&\approx (1-f)\frac{1}{Mn}\left[\frac{1}{Mn-1}\sum_{i\in\mathcal{S}}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{srs})^2\right] \\
&= \left(1-\frac{Mn}{MN}\right)\frac{1}{Mn}s^2_{y_{ij}}
\end{aligned}$$

As with stratified sampling we can calculate a *design effect*

$$DEFF = \frac{\mathrm{Var}\,(\bar{y}_{cl})}{\mathrm{Var}\,(\bar{y}_{srs})} = \frac{\left(1-\frac{n}{N}\right)\frac{1}{n}\left[\frac{1}{n-1}\sum_{i\in\mathcal{S}}(\bar{y}_i-\bar{y}_{cl})^2\right]}{\left(1-\frac{Mn}{MN}\right)\frac{1}{Mn}\left[\frac{1}{Mn-1}\sum_{i\in\mathcal{S}}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{srs})^2\right]} = \frac{M\frac{1}{n-1}\sum_{i\in\mathcal{S}}(\bar{y}_i-\bar{y}_{cl})^2}{\frac{1}{Mn-1}\sum_{i\in\mathcal{S}}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{srs})^2} = \frac{Ms^2_{\bar{y}_i}}{s^2_{y_{ij}}}$$

to see what the effect on precision of clustering is. In stratified sampling, we also calculated a design effect $DEFF$ (it has a different formula).

- In stratified sampling we usually get $DEFF < 1$ if we design the strata carefully.
- In clustered sampling, we usually get $DEFF > 1$.

We will see more about the design effect below.

Example 2: Estimating Average GPA (Lohr, 1999)

A student wants to estimate the average GPA in his dormitory. There are $N = 100$ suites that hold $M = 4$ students each. There are three random sampling schemes he could use:

- *SRS without replacement:* From a list (frame) of all 400 students in the dorm, take an SRS without replacement of, say, 20 students.
- *Stratified sample:* From each of the 100 suites he could take an SRS without replacement of, say, 2 students (for a total sample size of $2 \times 100 = 200$). Here, *the suites are strata.*
- *Clustered sample:* He could take an SRS of, say, $n = 5$ suites from the list of suites, and then take all four students in each suite. Again the sample size is $Mn = 4 \times 5 = 20$. Here, *the suites are clusters.*

In this case, the least effort is probably the clustered sample, so that is what the student does. The results are contained in Table 1 (page 9).

For the clustered sample we have

$$\bar{y}_{cl} = \frac{1}{5}\sum_{i=1}^{5}y_i = \frac{1}{5}(3.04 + 2.84 + 2.24 + 3.24 + 2.77) = 2.826$$

Person	Suite (Cluster) (i)				
Number (j)	1	2	3	4	5
1	3.08	2.36	2.00	3.00	2.68
2	2.60	3.04	2.56	2.88	1.92
3	3.44	3.28	2.52	3.44	3.28
4	3.04	2.68	1.88	3.64	3.20
$\overline{y}_i$	3.04	2.84	2.24	3.24	2.77

Table 1: GPA data from a clustered random sample of dorm suites.

and the sample variance of the $\overline{y}_i$'s is

$$s^2_{\overline{y}_i} = \frac{1}{5-1}\left[(3.04-2.826)^2 + \cdots + (2.77-2.826)^2\right] = 0.14098$$

Therefore

$$\text{Var}\left(\overline{y}_{cl}\right) = \left(1-\frac{n}{N}\right)\frac{1}{n}s^2_{\overline{y}_i} = \left(1-\frac{5}{100}\right)\frac{1}{5}(0.14098) = 0.0268$$

So $SE = \sqrt{0.0268} = 0.164$, and a 95% CI for the mean GPA in the dorm would be

$$(2.826 - (1.96)\cdot(0.164),\ 2.826 + (1.96)\cdot(0.164))$$

which runs from about 2.51 to about 3.15.

If the data had been collected as an SRS of size $n = 20$ we would have gotten

$$\overline{y}_{srs} = 2.862$$

(same answer, since the cluster sample was self-weighted!), with sample variance[1] $s^2_{y_{ij}} = 2.648$, so that

$$\text{Var}\left(\overline{y}_{srs}\right) = \left(1-\frac{Mn}{MN}\right)\frac{1}{Mn}s^2_{y_{ij}} = (1-20/400)\frac{1}{20}(2.648) = 0.126$$

We can see that the design effect in this case is

$$DEFF = \frac{\text{Var}\left(\overline{y}_{cl}\right)}{\text{Var}\left(\overline{y}_{srs}\right)} = \frac{Ms^2_{\overline{y}_i}}{s^2_{y_{ij}}} = \frac{(4)(0.14098)}{(0.2648)} = 2.13$$

So we would need a little over twice[2] as many observations in a clustered sample, in this case, to get the same precision as an SRS.

[1] Lohr (1999, p. 142) points out that the simple sample variance 2.648 is an underestimate because the data was in fact collected as a clustered sample. She proposes a less biased estimator based on ANOVA sums of squares, but we will use the simpler estimate for our purposes.

[2] Lohr (1999) calculates $DEFF = 2.02$ based on her less-biased estimate of $s^2_{y_{ij}}$.

It turns out that

$$DEFF = \frac{\text{Var}\,(\overline{y}_{cl})}{\text{Var}\,(\overline{y}_{srs})} \approx 1 + (M-1)\rho$$

where ρ ("rho") is the *intracluster correlation (ICC)*. The ICC is the correlation between all pairs of observations within each cluster (remember that observations within a cluster tend to be positively correlated).

We can use this formula to see how correlated the observations are in this example, by solving for ρ:

$$\rho \approx (DEFF - 1)/(M-1) = (2.13-1)/3 = 0.38$$

That is to say, there is a correlation of abotu 0.38 between any two people's GPA in the same dorm suite, in this survey!

2 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 ith cluster possessing the attribute. Then the cluster sample estimate of p, $\overline{p}_c$, is just the average of the values of p_i for the sampled clusters,

$$\overline{p}_c = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{1}$$

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 $\overline{p}_c$ 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.,

$$Var(\overline{p}_c) = \frac{(N-n)}{(N-1)}\frac{\left(\frac{\sum_{i=1}^{N}(p_i-p)^2}{N}\right)}{n}. \tag{2}$$

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

$$Var(\overline{p}) = \frac{(NM-nM)}{(NM-1)}\frac{p(1-p)}{nM}. \tag{3}$$

One way to compare the variances involves the "correlation" between elements in the same cluster, ρ. This quantity ρ is called the *intracluster correlation coefficient.* It turns out that

$$\boxed{\frac{Var(\overline{p}_c)}{Var(\overline{p})} = 1 + (M-1)\rho.} \tag{4}$$

Since variances cannot be negative, the quantity $1 + (M-1)\rho$ can't be negative and the minimum possible value for ρ 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 3: 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

$$\frac{Var(\overline{p}_c)}{Var(\overline{p})} \cong \frac{M\sum_{i=1}^{N}(p_i - p)^2}{Np(1-p)} \tag{5}$$

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 ith cluster, i.e., variance of the observations for samples of size 1 within the ith 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:

$$\boxed{\begin{array}{c} NMp(1-p) = M\sum_{i=1}^{N}(p_i - p)^2 + M\sum_{i=1}^{n} p_i(1-p_i) \\ \text{total SS=SS between, clusters+SS within clusters} \end{array}} \tag{6}$$

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

$$\begin{aligned} \frac{Var(\overline{p}_c)}{Var(\overline{p})} &\cong \frac{M\sum_{i=1}^{N}(p_i - p)^2}{Np(1-p)} \\ &= \frac{NMp(1-p) - M\sum_{i=1}^{N} p_i(1-p_i)}{Np(1-p)} \\ &= M\left(1 - \frac{\sum_{i=1}^{N} p_i(1-p_i)}{Np(1-p)}\right) \end{aligned} \tag{7}$$

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 4: 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

$$\begin{aligned} \frac{Var(\overline{p}_c)}{Var(\overline{p})} &\cong M\left(1 - \frac{\sum_{i=1}^{N} p_i(1-p_i)}{Np(1-p)}\right) \\ &= M\left(1 - \frac{10(0.25 \times 0.75) + 10(0.75 \times 0.25)}{20(0.5 \times 0.5)}\right) \\ &= M\left(1 - \frac{3}{4}\right) = \frac{M}{4}. \end{aligned} \tag{8}$$

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 3 (Continued)

Suppose the distribution of the number of LEP students in the 20 schools in our school district is as follows:

Number of School	Proportion of LEP Students
2	0.1
3	0.2
3	0.3
10	0.4
2	0.5

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

$$\frac{Var(\overline{p}_c)}{Var(\overline{p})} \cong M\left(1 - \frac{\sum_{i=1}^{N} p_i(1-p_i)}{Np(1-p)}\right) = 100\left(1 - \frac{4.19}{4.39}\right) = 4.56. \tag{10}$$

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.

3 Cluster Sampling for Measurements

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

$$\overline{y}_c = \frac{1}{nM} \sum_{i=1}^{n} y_i. \tag{11}$$

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 $\overline{y}_c$ 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, $\overline{y}_c$, with that of a simple random sample $\overline{y}$ of the same overall sample size, nM, and the ratio turns out to be the same as it was for attributes:

$$\frac{Var(\overline{y}_c)}{Var(\overline{y})} = 1 + (M-1)\rho, \tag{12}$$

where ρ is the intracluster correlation coefficient.

Example 5

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

$$s^2(\overline{y}_c) = 30.17$$
$$s^2(\overline{y}) = 20.23$$

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

$$\left(\frac{20.23}{30.17}\right)15 = 10.05$$

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

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

4 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,

$$\frac{Var(\bar{y}_c)}{Var(\bar{y})} \cong 1 + (\overline{M} - 1)\rho, \tag{14}$$

where $\overline{M}$ 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.

5 References

Cochran, W. G. (1977). *Sampling Techniques.* 3rd. Edition. Wiley: New York.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection (with discussion). *Journal of the Royal Statistical Society*, **97**, 558–625.

Henry, G. T. (1990). *Practical Sampling.* Sage: Newbury Park, CA.

Lohr, S. (1999). *Sampling: Design and Analysis.* Duxbury: Pacific Grove, CA.

Williams, B. (1978). *A Sampler on Sampling.* Wiley: New York.