5

Cluster Sampling


5.1 INTRODUCTION

Simple random sampling is important for understanding the principles of sampling. Yet it is not often used to do surveys. For rapid surveys we will use a more complex sampling design, two-stage cluster sampling, that is much easier to use in the field. Unfortunately the variance of cluster surveys, necessary for calculating confidence intervals, is not as easy to derive. In addition, advanced statistical analyses with multivariate relations are more complex to calculate than with surveys featuring simple random sampling. Nevertheless, for those wanting to do small, inexpensive surveys, cluster sampling is often the method of choice.

There are two different ways to do rapid surveys, each having its own equations for calculating the mean and confidence interval. Both assume that clusters have been selected with probability proportionate to size (PPS) at the first stage of the sampling process. At the second stage within clusters, one method assumes the selection of an equal number of persons while the second method assumes the sampling of an equal number of households.

As an example, we will be studying two small surveys of smoking behavior. Specifically, we will be measuring the prevalence of current smoking and the average number of packs smoked per day. For the first survey a sample of three clusters is selected with probability proportionate to size (PPS), followed by a simple random sample of seven person per cluster (see Figure 5-1). The survey is limited to three clusters only to simplify the example. For actual surveys you should not sample fewer than 25 clusters, or else the findings might be biased. If the example had been a survey in which 30 clusters were selected rather than three, it would have followed the design of the Expanded Program on Immunization (EPI) of the World Health Organization. The second survey is shown Figure 5-2. Here three clusters are also selected with PPS sampling, but thereafter two households rather than seven people are randomly selected from each cluster. Within each household, one to three persons are interviewed, depending on how many are in residence. With these two surveys, I will show how binomial and equal interval data can be analyzed using two sets of formulas, and why one set of formulas for ratio estimators...
works quite well with both approaches.

So what is different about these two surveys? In the first survey, the sampling units are the same as the elementary units, that is, people. In the second survey, sampling units are households while elementary units are persons living in the households. As you will see, this small difference in approach between the two surveys requires the use of different variance formulas necessary to derive confidence intervals.

5.2 SAMPLING OF PERSONS

For our first survey, I will start with the analysis of smoking, coded as 0 for current non-smokers and 1 for current smokers. As such, smoking is a binomial variable, the mean of which is the proportion or percentage who smoke in the population. For terminology, I will use \( n \) to denote the number of clusters, \( a \) to represent the number of persons with the attribute of interest (in this case smoking), \( m \) to signify the number of persons and \( p \) as the proportion with the attribute. The subscripts \( i \) and \( j \) are used to designate variables at the two levels of the sampling process. The example will make this clearer.

The first survey, as shown in Figure 5-1, follows a two-stage sampling process with the three clusters selected with probability proportionate to size at the first stage. At the second stage, the sampling units (i.e., persons) are listed only for those clusters that were selected at the first stage. Thereafter the sample is selected from the list by simple random sampling. The same number of sampling units are selected from a list within each cluster. With this first method of cluster sampling, the sampling units at the second stage are the same as elementary units (the units we plan to analyze), namely people.

5.2.1 Sample Proportion

In our population, the proportion who smoke is the number of smokers divided by the number of persons in the sample, or...

\[ p = \frac{a}{m} \]  

(5.1)

Since there are \( n \) clusters, we need to tally the number of smokers and persons in each clusters. The term \( a \) is a count of the total number of smokers in the \( n \) clusters, defined as...

\[ a = \sum_{i=1}^{n} a_i \]  

(5.2)

Notice that \( a_i \) is a random variable that varies from one cluster to the next depending on the number of smokers that appear in the sample. The number of sampled persons per cluster is not a random variable since it is a constant number set by the surveyor. The total number of persons included in
the survey is the average number per cluster (equal for all clusters) times the total number of clusters, or...

\[ m = n \bar{m} \]  

(5.3)

With these changes, the proportion of smokers in the total sample is defined as...

\[ p = \frac{\sum a_i}{n \bar{m}} \]  

(5.4)

As noted in Figure 5-1, seven persons were sampled from each of three clusters. Since the clusters were selected with probability proportionate to size and an equal number of persons were selected per cluster, each person in the population had the same probability of being selected. The values of variables in our survey, therefore, represent single persons and do not need to be weighted in our analysis. Having self-weighted data makes the analysis much easier. The findings of our example survey are shown in Figure 5-3.

The symbols \( a \) in Figure 5.3 has been expanded one more level to represent counts for individuals rather than clusters. This expansion is done using additional subscripts as shown in Formula 5.5 and for smokers, in Figure 5.3.

\[ a = \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \]  

(5.5)

Since there are 16 persons being sampled in each cluster, the identifying subscript \( j \) counts each person from 1 to 16. In our example survey shown in Figure 5-3, the first person in cluster 1 is a smoker and therefore \( a_{1,1} \) is counted as 1. The fourth persons is a non-smoker and thus \( a_{1,4} \) is counted as 0. The proportion in each cluster who smoked is defined as...

\[ p_i = \frac{\sum_{j=1}^{n} a_{ij}}{m} \]  

(5.6)
Using Formula 5.6 and the data in Figure 5-3, we derive the proportion who smoke for each of the three clusters.

\[ p_1 = \frac{1 + 1 + 0 + 0 + 0 + 0}{7} = 0.43 \]

\[ p_2 = \frac{0 + 0 + 0 + 0 + 0 + 1 + 1}{7} = 0.29 \]

\[ p_3 = \frac{0 + 1 + 1 + 1 + 1 + 1 + 1}{7} = 0.86 \]

The proportion who are current smokers is now designated for each cluster. These proportions can be used to describe the smoking experience of the cluster without mentioning the people in the cluster (see Figure 5-4).

For the total sample, the proportion who smoke can be calculated two ways. For the first method, I use Formula 5.4 as...

\[ p = \frac{3 + 2 + 6}{3 \times 7} = 0.52 \]

which shows that 52 percent of the surveyed population currently smokes. Since the sample is self-weighted and all the clusters are the same size, we could also have obtained the proportion (or percentage) who smoke by calculating the average of the three cluster proportions. The general formula for this calculation is...

\[ p = \frac{\sum_{i=1}^{n} p_i}{n} \] (5.7)

while the specific calculation for our example is...

\[ p = \frac{0.43 + 0.29 + 0.86}{3} = 0.52 \]

Notice that each the cluster-specific proportions in the equation must represent the same number of people or else the average of the three proportions using Formula 5.7 would not be the same as the total number of smokers divided by the total number of sampled persons as calculated with Formula 5.4.

### 5.2.2 Confidence Interval of Proportion

If the sample is self-weighted and there is an equal number of selected persons per cluster, the
confidence interval of the sample proportion is easy to derive. We first calculate the variance, then the standard error, and finally, the confidence interval. The variance formula for a proportion is shown in Figure 5-5, with descriptions of the various terms, and in Formula 5.8. The equation calculates the deviations of the proportions in the individual clusters from the proportion for the sample as a whole.

\[ \nu(p) = \frac{\sum_{i=1}^{n} (p_i - p)^2}{n(n-1)} \]  
(5.8)

The standard error of the proportion is the square root of the variance or...

\[ se(p) = \sqrt{\nu(p)} = \sqrt{\frac{\sum_{i=1}^{n} (p_i - p)^2}{n(n-1)}} \]  
(5.9)

Finally, we use the proportion and standard error to derive the confidence interval of the proportion. Intervals with 95 percent confidence limits are the most common used by surveyors. Yet you also might want to derive 90 percent or 99 percent confidence intervals to show the relationship between the level of confidence and the size of the interval. Thus equations for three confidence intervals are presented. First is the 90 percent confidence interval...

\[ CI_{90\%}(p) = p \pm 1.64 \, se(p) \]  
(5.10)

followed by the more common 95 percent confidence interval....

\[ CI_{95\%}(p) = p \pm 1.96 \, se(p) \]  
(5.11)

and lastly, the 99 percent confidence interval...

\[ CI_{99\%}(p) = p \pm 2.58 \, se(p) \]  
(5.12)

Returning to our example in Figure 5-4, the variance of the proportion is...

\[ \nu(p) = \frac{(0.43 - 0.52)^2 + (0.29 - 0.52)^2 + (0.86 - 0.52)^2}{3 \, (2)} = 0.029 \]

and the standard error is...

\[ se(p) = \sqrt{0.029} = 0.17 \]

Earlier we calculated the proportion who smoke as 0.52. Therefore using Formula 5.11, the 95 percent confidence interval for the proportion is...
Often the findings are presented as the point estimate (here the proportion) followed in parentheses by the upper and lower limits of the confidence interval. The proportion and 95 percent confidence interval are...

\[
CI_{95\%}(\hat{p}) = 0.52 \pm (1.96 \times 0.17) = 0.52 \pm 0.34
\]

The values may also be presented as the percentage who smoke, rather than a proportion, by multiplying the proportion by 100. The percentage and 95 percent confidence interval are...

\[
52 \ (0.19, 0.86)
\]

If there is no bias or confounding by other variables, I am 95 percent confident that the true percentage who smoke in the survey population lies between 19 and 86 percent. My best estimate is that 52 percent of the survey population is currently smoking.

Keep in mind that Formula 5.8 can only be used to derive the variance if the sample size is the same in each cluster. The equation will not work as intended if some clusters have more sampled persons than others, as may occur if there are missing data or if only certain subgroups are to be analyzed. Fortunately, there is another formula available that is more flexible, as described in 5.3 Sampling of Households.

### 5.2.3 Sample Mean

Equal interval data are analyzed in a similar manner to binomial data. Here, however, we will calculate the sample mean, rather than proportion. The persons in our survey smoked a certain number of packs per day ranging from 0 to 2. Most of the smokers consumed 1.5 packs per day. The variable \( y \) is use to identify the number of pack smoked per day. For the individual, \( y \) has two subscripts, \( i \) and \( j \), showing the identify of the cluster and the person in the cluster. The formula for the mean is...

\[
\bar{y} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} y_{ij}}{nm}
\]

where \( n \) is the number of clusters and \( \bar{m} \) is the average number of persons per cluster. The data for the number of packs smoked per day are shown in Figure 5-6. Using Formula 5.13, the mean for the population is...

\[
\bar{y} = \frac{(1.5 + 1.5 + 0.5 + 0 + 0 + 0)}{3(7)} + \frac{(0 + 0 + 0 + 0 + 0 + 2.0 + 0.5) + (0 + 1.0 + 2.0 + 0.5 + 0.5 + 1.5 + 1.5)}{3(7)}
\]

\[
\bar{y} = \frac{(3.5) + (2.5) + (7.0)}{21} = \frac{13}{21} = 0.62
\]
Because the number of sampled persons per cluster is equal in all three clusters, we could have derived the average for the survey by tallying the means for each cluster and dividing by the number of clusters, or...

\[
\bar{y} = \frac{\sum_{i=1}^{n} \bar{y}_i}{n} \tag{5.14}
\]

First, however, we need calculate the mean per cluster with the following equation:

\[
\bar{y}_i = \frac{\sum_{j=1}^{m} y_{ij}}{m} \tag{5.15}
\]

Using Formula 5.15 and the data in Figure 5-6, the mean number of packs smoked per day for the persons in the three clusters are....

\[
\bar{y}_1 = \frac{1.5 + 1.5 + 0.5 + 0 + 0 + 0 + 0}{7} = 0.50
\]

\[
\bar{y}_2 = \frac{0 + 0 + 0 + 0 + 2.0 + 0 + 0.5}{7} = 0.36
\]

\[
\bar{y}_3 = \frac{0 + 1.0 + 2.0 + 0.5 + 0.5 + 1.5 + 1.5}{7} = 1.00
\]

Since these means are based on the same number of persons, the mean for each cluster, \( \bar{y}_i \), can be viewed as a comparable unit to describe the sampled clusters, the same way that \( y_{ij} \) is used as a comparable unit to describe the sampled persons. This point is illustrated in Figure 5-7 where the three clusters are now represented by their means, rather than by individual values.

Because the three cluster means have an equivalent base, we can use Formula 5.14 to derive the mean for the sample, as...

\[
\bar{y} = \frac{0.50 + 0.36 + 1.00}{3} = 0.62
\]

Notice 0.62 packs per day is the same value as shown earlier for the mean based on individual observations.
5.2.4 Confidence Interval of Mean

The ingredients to calculate the confidence interval of the mean are the same as for a proportion. You need both the mean and the standard error of the mean, calculated as the square root of the variance of the sample mean. The sampled persons being measured are self-weighted and there is an equal number of selected persons per cluster. As a result, the variance of the sample mean is easy to derive. The formula for variance of the mean in the cluster survey is shown in Figure 5-8 with descriptions of the various terms and as Formula 5.16.

\[ \nu(\bar{y}) = \frac{\sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2}{n(n-1)} \]

The standard error of the mean is...

\[ se(\bar{y}) = \sqrt{\nu(\bar{y})} = \sqrt{\frac{\sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2}{n(n-1)}} \]

The confidence interval is derived the same as for a proportion. The formula for the 90 percent confidence interval of a mean is...

\[ CI_{90\%}(\bar{y}) = \bar{y} \pm 1.64 \cdot se(\bar{y}) \]

while the 95 percent confidence interval of the mean is...

\[ CI_{95\%}(\bar{y}) = \bar{y} \pm 1.96 \cdot se(\bar{y}) \]

and the 99 percent confidence interval is...

\[ CI_{99\%}(\bar{y}) = \bar{y} \pm 2.58 \cdot se(\bar{y}) \]

As with the proportions, the variance equation (Formula 5.16) calculates the deviations of the sample means in the individual clusters from the mean for the sample as a whole. Based on the data in Figure 5-7, the variance of the mean number of packs smoked per day is...
The mean number of packs smoked per day was previously calculated as 0.62. Therefore using Formula 5.19, the 95 percent confidence interval for the mean is...

\[
C_{95\%}(\bar{y}) = 0.62 \pm (1.96 \times 0.19) = 0.62 \pm 0.38
\]

As with proportions, I present the confidence interval as the mean followed in parentheses by the upper and lower limits of the confidence interval, or...

\[0.62 \ (0.24, 1.00)\]

If there is no bias or confounding, I am 95 percent confident that the mean number of packs smoked per day in the sampled population lies between 0.24 and 1.00, with my best estimate being 0.62.

While the formulas for variance, standard error and confidence intervals of the proportion and mean are easy to derive, they are only valid in certain circumstances. First, persons in the population must be selected with equal probability so that the values for one person can be directly compared to those of another. That is, the sample must be self-weighted. Second, an equal number of persons must be selected in each cluster so that the cluster proportions or means can also be directly compared with one another. These conditions are met when the cluster is selected with probability proportionate to the size of the population, and a constant number of persons is sampled from each selected cluster. If for one reason or another the number of people sampled within each cluster is not the same, you will need to use a different equation, as described in the following section.

### 5.3 SAMPLING OF HOUSEHOLDS

Instead of sampling persons at the second stage, assume that we have sampled households. Our *sampling unit* then becomes households rather than persons. Yet only occasionally are we interested in household characteristics. For example, we might want to know if the household as a whole subscribes to a certain newspaper or magazine, or uses a certain brand of detergent. Or we might be interested in the style of construction, or the total household income. Most of the time, however, we will want to analyze data on people. Therefore, the *elementary unit*, but possibly not the *sampling unit*, is a person.

When the units being sampled (households) are not the same as the elementary units (people), we must use a different set of formulas to derive the average values and confidence intervals for variables describing people. While in the last section each person had the same probability of being selected, in this section each household has the same chance of selection. In both types of surveys, clusters are selected with probability proportionate to size at the first stage. Thereafter, however, an equal number of households, rather than persons, is selected at the second stage.

\[
\nu(\bar{y}) = \frac{(0.50 - 0.62)^2 + (0.36 - 0.62)^2 + (1.00 - 0.62)^2}{3 \times 2} = 0.038
\]

while the standard error is...

\[
se(\bar{y}) = \sqrt{0.038} = 0.19
\]
stage. Following this sampling procedure, each household (but not each person) will have the same chance of being selected.

The selection stages of our example household survey were previously shown in Figure 5-2. At the first stage, three clusters are selected with probability proportionate to size, followed by a random selection of two households in each cluster. The households vary in size with the smallest having one person and the largest having three people. There are a total of six households in our example survey, containing 13 people. The smoking status of the 13 persons is shown, by household, in Figure 5-9. Notice that each person is now characterized by two binomial variables, \( a \) or their smoking status and \( m \), or their existence. If a person smokes, the variable \( a \) has a value of 1. If the person does not smoke, \( a \) is coded 0. Since all persons included in the analysis exist (I refuse to sample nonexisting people), all are given a value of 1 for the binomial variable, \( m \).

Technically, the two variables \( a \) and \( m \) have three levels of subscripts used to describe values at the person level. The binomial variable smoking status takes the form \( a_{i,j,k} \) while the binomial variable existence is described as \( m_{i,j,k} \). For example, the first person in the first household in cluster 1 is a smoker. Therefore \( a_{1,1,1} \) has the value 1. The last person in the second household in cluster 3 is not a smoker. The value of \( a_{3,2,2} \) is thus 0. Also notice that since all the people exist, the different values of \( m_{i,j,k} \) are all coded 1.

While our sampling unit is households, we do not need to keep track of households in our analysis. Because the number of sampled households is equal in each cluster, we can combine the households and treat all persons in the cluster as coming from an equal number of households. We then use \( a \) and \( m \) to count the number of persons in the combined set of households. The example will make this clearer. As shown in Figure 5-10, the elementary units (that is, people) in the two households are combined in each cluster. Since each cluster contains exactly two sampled households, knowing the identity of the cluster means knowing the set of two households. That is, the cluster designator gives us the same information as the household designator. Instead of having three subscripts identifying each person, we will combine the household and cluster designators and use only two. For example

### Figure 5-9. Smoking status with households serving as sampling units and persons as elementary units.

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Second Stage Sampling Units</th>
<th>Elementary Units</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Households</td>
<td>Persons</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a_{1,1,1} 1</td>
<td>a_{1,1,1} 1</td>
</tr>
<tr>
<td></td>
<td>m_{1,1,1} 1</td>
<td>m_{1,1,1} 1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a_{2,1,1} 0</td>
<td>a_{2,1,1} 0</td>
</tr>
<tr>
<td></td>
<td>m_{2,1,1} 1</td>
<td>m_{2,1,1} 1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a_{3,2,2} 0</td>
<td>a_{3,2,2} 0</td>
</tr>
<tr>
<td></td>
<td>m_{3,2,2} 1</td>
<td>m_{3,2,2} 1</td>
</tr>
</tbody>
</table>

### Figure 5-10. Tally of smokers per cluster)

households serve as sampling units and persons as elementary units.
as seen in Figure 5-9, the second person in cluster 1, household 2, is coded as $a_{1,2,2}$ (with value 0) and $m_{1,2,2}$. The same individual is recoded as the fifth person in cluster 1, but this time using two subscripts, $a_{1,5}$ and $m_{1,5}$.

Earlier in Formula 5.5, I showed that the random variable, $a$, is actually a count of the number of smokers in the total sample, in each cluster and in each person (at this level, limited to 0 or 1). The various counting levels are described using the subscripts $i$ and $j$. The only change in our second survey is in the number of persons sampled per cluster. The variable, $m_i$, varies from cluster to cluster and can no longer be described with $\bar{m}$, the mean value for the entire survey. Therefore Formula 5.5 must be changed slightly to remove $\bar{m}$ (the average number per cluster) and insert $m_{i}$ (the number of persons in cluster $i$). That is....

$$a = \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \sum_{j=1}^{m_i} a_{ij}$$

(5.21)

Similarly, our second random variable, $m$, is described at three levels as...

$$m = \sum_{i=1}^{n} m_i = \sum_{i=1}^{n} \sum_{j=1}^{m_i} m_{ij}$$

(5.22)

These two random variables, $a_i$ and $m_i$, will be used in the coming sections to derive both the proportion and the 95 percent confidence interval for the proportion.

### 5.3.1 Sample Proportion

Earlier in Formula 5.1, I described a proportion as $a$, the number of smokers, divided by $m$, the number of persons in the sample. That is...

$$p = \frac{a}{m}$$

In its expanded form, the proportion was shown in Formula 5.4 as....

$$p = \frac{\sum_{i=1}^{n} a_i}{n \bar{m}}$$

where the numerator is a random variable, $a_i$, and the denominator is a constant, $n \bar{m}$, set by the investigator. It is in this denominator that things change somewhat in our households survey. Since the sampling units are households, not people, the denominator of the proportion can no longer be set by the investigator. The sampled households may have a small or large number of people. Thus the number of persons should correctly be viewed as a random variable.

The correct form of the proportion equation for a household survey is...
where \( n \) is the number of clusters, \( m_i \) is a count of the number of sampled persons in cluster \( i \), \( a_{ij} \) is the smoking status of person \( i,j \) (0 or 1) and \( m_{ij} \) is the existence status of person \( i,j \) (also 0 or 1). Since this is a ratio of two random variables, this proportion is termed a ratio estimator. In a pure sense, Formula 5.23 should be written as...

\[
p = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} a_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{m_i} m_{ij}} \tag{5.23}
\]

where \( r \) is the symbol for the ratio estimator. Yet because in common usage the ratio estimator is viewed as a proportion, I will continue to use \( p \) (as in Formula 5.23) rather than the more correct \( r \) (as in Formula 5.24).

Using the data in Figure 5-9 and 5-10 and Formula 5.23, the proportion who are currently smoking in our survey is calculated as...

\[
p = \frac{(1 + 1 + 0 + 0 + 0) + (1 + 0 + 1) + (1 + 1 + 0 + 1 + 0)}{(1 + 1 + 1 + 1) + (1 + 1 + 1) + (1 + 1 + 1 + 1)}
\]

\[
p = \frac{7}{13} = 0.54
\]

Hence, 0.54 or 54\% of the sampled population is estimated to be current smokers.

The ratio estimator, as you have learned earlier in Chapter 3, provides a good estimate of the proportion or mean in the sample so long as \( m_i \), the number of persons in the sampled clusters, does not vary too much. If at least 25-30 clusters are sampled with no fewer than 6-8 households per cluster, the variation in \( m_i \) will be minimal. Many rapid surveys focus only on certain categories of people, such as males or females, or on age categories such as children less than 5 years or senior persons 65 years and older. Only households that contain one or more eligible persons are sampled. Thus clusters in these surveys will have minimal variation in the size of \( m_i \), resulting in almost no bias in the ratio estimator.

### 5.3.2 Confidence Interval of Proportion

The confidence interval for proportions in household data use the proportion and square root of the variance of the proportion, the same as in other surveys. What is different, however, is the variance equation. Because the proportion in household surveys is a ratio estimator, the variance formula for the proportion must also be specific for a ratio estimator. The equation is shown in Figure 5-11 with...