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.

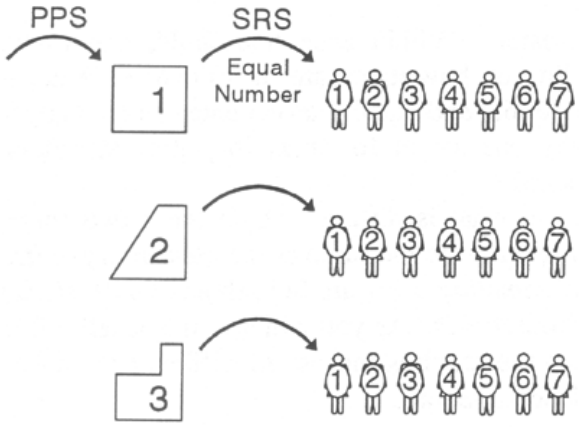

Figure 5-1. PPS sampling at first stage and SR sampling of equal number of persons at second stage.

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.

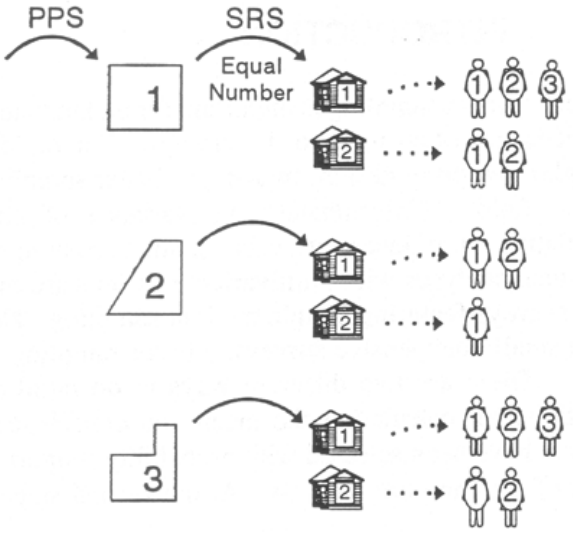

Figure 5-2. PPS sampling at first stage and SR sampling of equal number of households at second stage.

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...

$$p = \frac{a}{m} \tag{5.1}$$

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...

$$a = \sum_{i=1}^{n} a_i \tag{5.2}$$

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...

$$m = n\bar{m} \tag{5.3}$$

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

$$p = \frac{\sum_{i=1}^{n} a_i}{n\bar{m}} \tag{5.4}$$

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.

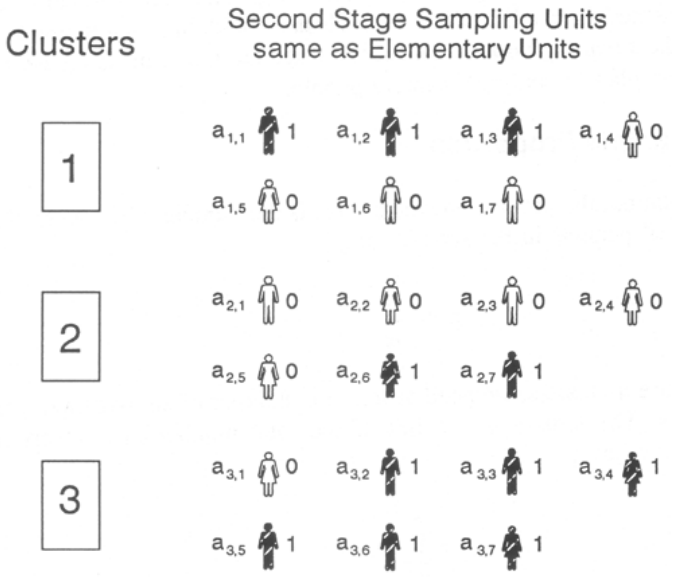

Figure 5-3. Smoking status with persons serving as both sampling units and elementary units.

$$a = \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \sum_{j=1}^{\bar{m}} a_{i,j} \tag{5.5}$$

Since there are $\bar{m}$ persons being sampled in each cluster, the identifying subscript *j* counts each person from 1 to $\bar{m}$. 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...

$$p_i = \frac{\sum_{j=1}^{\bar{m}} a_{i,j}}{\bar{m}} \tag{5.6}$$

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

$$p_1 = \frac{1+1+1+0+0+0+0}{7} = 0.43$$

$$p_2 = \frac{0+0+0+0+0+1+1}{7} = 0.29$$

$$p_3 = \frac{0+1+1+1+1+1+1}{7} = 0.86$$

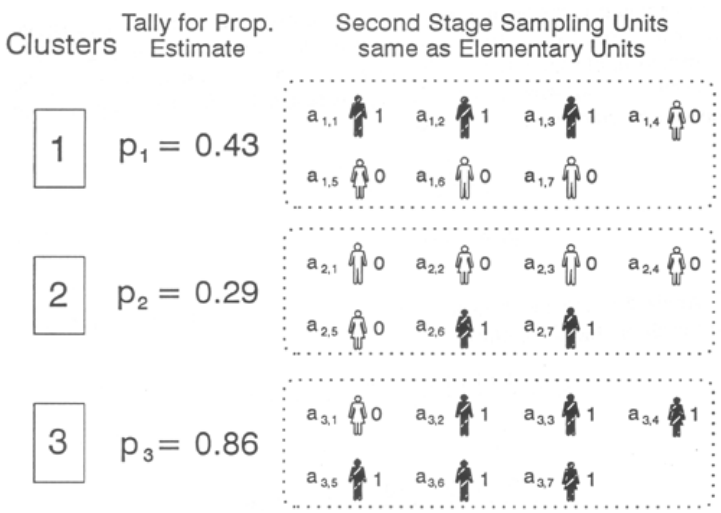

Figure 5-4. Proportion of current smokers per cluster) persons serve as both sampling units and elementary units.

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...

$$p = \frac{3+2+6}{3\,x\,7} = 0.52$$

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...

$$p = \frac{\sum_{i=1}^{n} p_i}{n} \qquad (5.7)$$

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.

$$v(p) = \frac{\sum_{i=1}^{n} (p_i - p)^2}{n(n-1)} \tag{5.8}$$

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

$$se(p) = \sqrt{v(p)} = \sqrt{\frac{\sum_{i=1}^{n} (p_i - p)^2}{n(n-1)}} \tag{5.9}$$

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...

$$CI_{90\%}(p) = p \pm 1.64\ se(p) \tag{5.10}$$

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

$$CI_{95\%}(p) = p \pm 1.96\ se(p) \tag{5.11}$$

and lastly, the 99 percent confidence interval...

$$CI_{99\%}(p) = p \pm 2.58\ se(p) \tag{5.12}$$

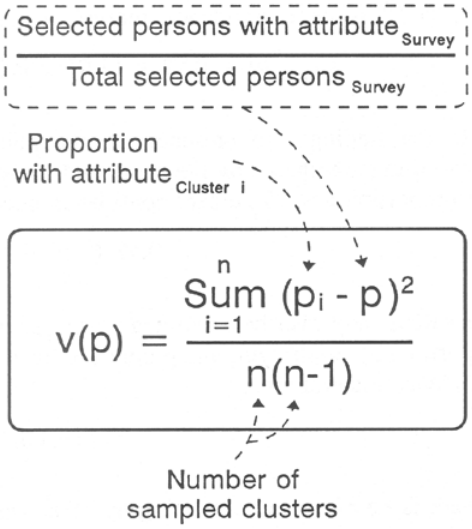

Figure 5-5. Variance formula of a proportion for surveys where persons are both sampling units and elementary units.

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

$$v(p) = \frac{(0.43 - 0.52)^2 + (0.29 - 0.52)^2 + (0.86 - 0.52)^2}{3\,(2)} = 0.029$$

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...

$$CI_{95\%}(p) = 0.52 \pm (1.96 \; x \; 0.17) = 0.52 \pm 0.34$$

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...

$$0.52 \; (0.19, 0.86)$$

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...

$$52 \; (19, 86)$$

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...

$$\bar{y} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}}{n \bar{m}} \tag{5.13}$$

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...

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

$$\bar{y} = \frac{(3.5)+(2.5)+(7.0)}{21} = \frac{13}{21} = 0.62$$

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...

$$\bar{y} = \frac{\sum_{i=1}^{n} \bar{y}_i}{n} \qquad (5.14)$$

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

$$\bar{y}_i = \frac{\sum_{j=1}^{\bar{m}} y_{i,j}}{\bar{m}} \qquad (5.15)$$

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....

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

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

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

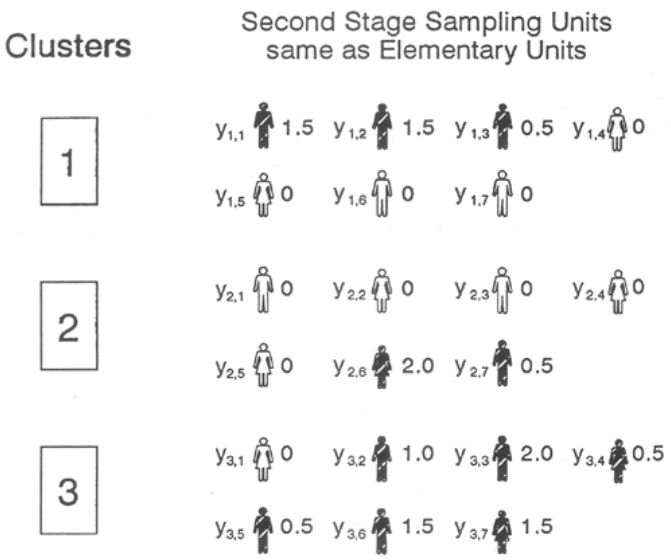

Figure 5-6. Packs smoked per day among persons serving as both sampling units and elementary units.

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 $y_{i,j}$ 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.

$$v(\bar{y}) = \frac{\sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2}{n(n-1)} \qquad (5.16)$$

The standard error of the mean is...

$$se(\bar{y}) = \sqrt{v(\bar{y})} = \sqrt{\frac{\sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2}{n(n-1)}} \qquad (5.17)$$

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

$$CI_{90\%}(\bar{y}) = \bar{y} \pm 1.64\ se(\bar{y}) \qquad (5.18)$$

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

$$CI_{95\%}(\bar{y}) = \bar{y} \pm 1.96\ se(\bar{y}) \qquad (5.19)$$

and the 99 percent confidence interval is...

$$CI_{99\%}(\bar{y}) = \bar{y} \pm 2.58\ se(\bar{y}) \qquad (5.20)$$

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...

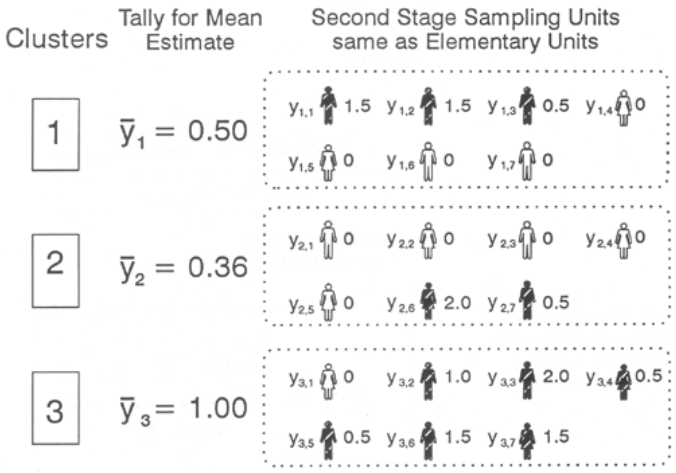

Figure 5-7. Mean packs smoked per day per cluster) persons serve as both sampling units and elementary units.

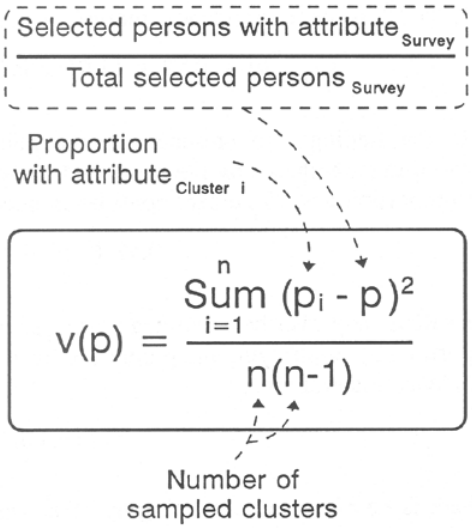

Figure 5-8. Variance formula of a mean for surveys where persons are both sampling units and elementary units.

$$v(\bar{y}) = \frac{(0.50-0.62)^2 + (0.36-0.62)^2 + (1.00-0.62)^2}{3\,(2)} = 0.038$$

while the standard error is...

$$se(\bar{y}) = \sqrt{0.038} = 0.19$$

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...

$$CI_{95\%}(\bar{y}) = 0.62 \pm (1.96 \; x \; 0.19) = 0.62 \pm 0.38$$

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. 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*.

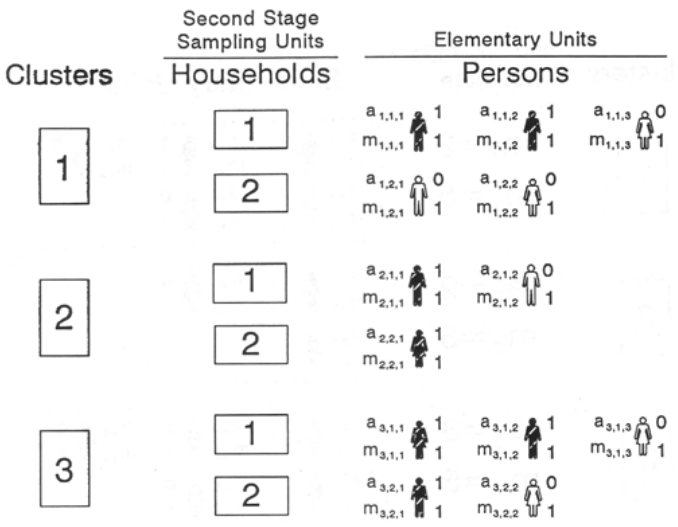

Figure 5-9. Smoking status with households serving as sampling units and persons as elementary units.

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

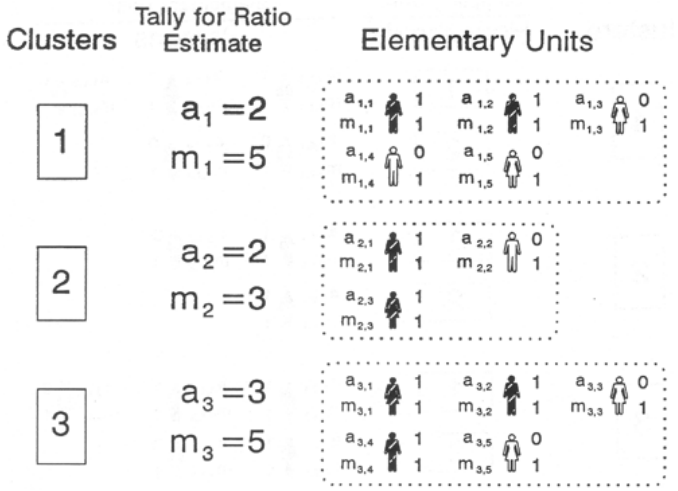

Figure 5-10. Tally of smokers per cluster) households serve as sampling units and persons as elementary units..

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....

$$a = \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \sum_{j=1}^{m_i} a_{i,j} \tag{5.21}$$

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

$$m = \sum_{i=1}^{n} m_i = \sum_{i=1}^{n} \sum_{j=1}^{m_i} m_{i,j} \tag{5.22}$$

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 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...

$$p = \frac{\sum_{i=1}^{n} \sum_{j=i}^{m_i} a_{i,j}}{\sum_{i=1}^{n} \sum_{j=i}^{m_i} m_{i,j}} \tag{5.23}$$

where n is the number of clusters, m_i is a count of the number of sampled persons in cluster i, $a_{i,j}$ is the smoking status of person i,j (0 or 1) and $m_{i,j}$ 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...

$$r = \frac{\sum_{i=1}^{n} \sum_{j=i}^{m_i} a_{i,j}}{\sum_{i=1}^{n} \sum_{j=i}^{m_i} m_{i,j}} \tag{5.24}$$

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...

$$p = \frac{(1+1+0+0+0)+(1+0+1)+(1+1+0+1+0)}{(1+1+1+1+1)+(1+1+1)+(1+1+1+1+1)}$$

$$p = \frac{(2)+(2)+(3)}{(5)+(3)+(5)} = \frac{7}{13} = 0.54$$

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

a description of the terms, and in Formula 5.25.

$$v(p) = \frac{\sum_{i=1}^{n} (a_i - p\, m_i)^2}{n(n-1)\bar{m}^2} \qquad (5.25)$$

The equation calculates the deviations of the observed number with the attribute in each cluster from the expected number with the attribute based on the number of persons in the cluster. The standard error and confidence intervals for proportions in household surveys are derived the same as in surveys of persons. Specifically, the standard error is...

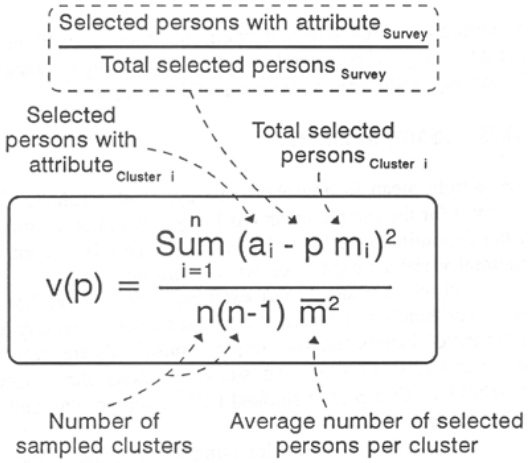

Figure 5-11. Variance formula of a proportion for surveys where households are sampling units and persons are elementary units.

$$se(p) = \sqrt{v(p)} = \sqrt{\frac{\sum_{i=1}^{n} (a_i - p\, m_i)^2}{n(n-1)\bar{m}^2}} \qquad (5.26)$$

and the confidence intervals are: 90 percent confidence...

$$CI_{90\%}(p) = p \pm 1.64\ se(p) \qquad (5.27)$$

95 percent confidence....

$$CI_{95\%}(p) = p \pm 1.96\ se(p) \qquad (5.28)$$

and 99 percent confidence...

$$CI_{99\%}(p) = p \pm 2.58\ se(p) \qquad (5.29)$$

Using the date in Figure 5-10, the calculation of the variance of the proportion is...

$$v(p) = \frac{[2-(0.54\,x\,5)]^2 + [2-(0.54\,x\,3)]^2 + [3-(0.54\,x\,5)]^2}{3\,(2)\,4.3^2} = 0.0064$$

Notice that this formula does **not** require that the same number of persons is sampled per cluster, as does Formula 5.8. You do, however, need to sample the same number of households per cluster. The standard error is..

$$se(p) = \sqrt{0.0064} = 0.08$$

The proportion who smoke in the sample households is 0.54. Therefore using Formula 5.28, the 95 percent confidence interval for the proportion is...

$$CI_{95\%}(p) = 0.54 \pm (1.96\ x\ 0.08) = 0.54 \pm 0.16$$

and presented as the proportion followed in parentheses by the upper and lower limits of the confidence interval, or...

$$0.54\ (0.38, 0.70)$$

Since percentages are easier to discuss than proportions, I multiply the values by 100 as....

$$54\ (38, 70)$$

Assuming the absence of bias or confounding, I am 95 percent confident that the true percentage who currently smoke in the survey households lies between 38 and 70 percent, with the best estimate of 54 percent.

5.3.3 Sample Mean

The sample mean in household surveys is also calculated as a *ratio estimator*. The ratio for the sample mean has in the numerator an equal interval variable and in the denominator, a binomial variable. This is in contrast to the ratio of two binomial variables used to derive a proportion

So let us return to our example survey to see how to derive the sample mean. The number of packs of cigarettes smoked per day and number of persons in the three clusters and six sampled households are shown in Figure 5-12.

For example, there were two smokers among three persons in cluster 1 and household 1. One person smoked half a pack per day and the other smoked two packs per day.

As was done when calculating the proportion, the persons per set of two households are combined in each cluster (see Figure 5-13). Keep in mind that this should only be done if there is an equal number of households sampled in each cluster, with each household having the same probability of being selected. The variable *y* is use to identify the number of packs 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, while for the cluster, *y* has only one subscript, *i*.

The formula for the mean is...

$$\bar{y} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} y_{i,j}}{\sum_{i=1}^{n} \sum_{j=1}^{m_i} m_{i,j}} \qquad (5.30)$$

where n is the number of clusters, m_i is a count in cluster i of the number of sampled persons, $y_{i,j}$ is the packs of cigarettes smoked daily by person i,j (a range of values), and $m_{i,j}$ is the existence status of person i,j (0 or 1).

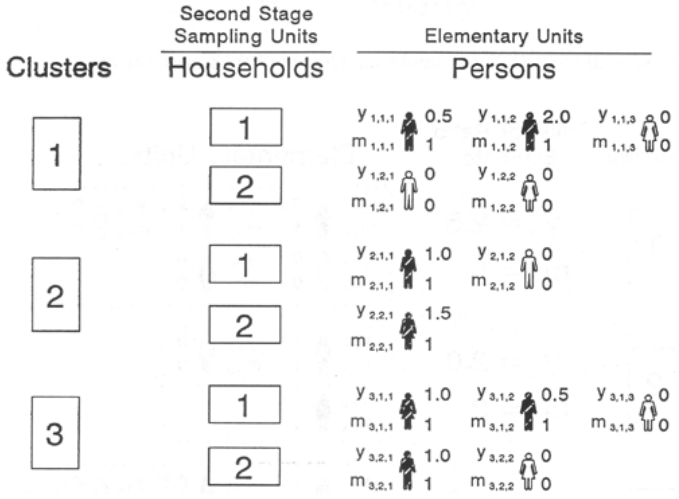

Figure 5-12. Packs smoked per day with households serving as sampling units and persons as elementary units.

Since Formula 5.30 is a *ratio estimator*, it should be written as...

$$r = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{m_i} m_{i,j}} \qquad (5.31)$$

where *r* is the mean ratio estimator. Yet as mentioned earlier in discussing the *ratio estimator* as a proportion, to conform to common usage when describing a sample mean, I will use Formula 5.30 rather than Formula 5.31.

Using the data in Figure 5-13 and Formula 5.30, the mean packs smoked per day is calculated as...

$$\bar{y} = \frac{(0.5 + 2.0 + 0 + 0 + 0) + (1.0 + 0 + 1.0) + (1.0 + 0.5 + 0 + 1.5 + 0)}{(1 + 1 + 1 + 1 + 1) + (1 + 1 + 1) + (1 + 1 + 1 + 1 + 1)}$$

which reduces to...

$$\bar{y} = \frac{(2.5) + (2.0) + (3.0)}{(5) + (3) + (5)} = \frac{7.5}{13} = 0.58$$

Thus persons in the population smoke an average of 0.58 pack per day.

While this finding is interesting, it might be useful to learn how many packs are consumed per day by smokers. For such an analysis, we would limit the ratio estimation to those who are current smokers (see Figure 5-14). Rather than having 13 persons in three clusters, we would restrict the analysis to 7 smokers, also in three clusters. The numerator of the ratio remains the same, but denominator is reduced from the former value of 13 persons to the new value of 7 persons. Using the data in Figure 5-14, the mean number packs consumed per day among smokers is...

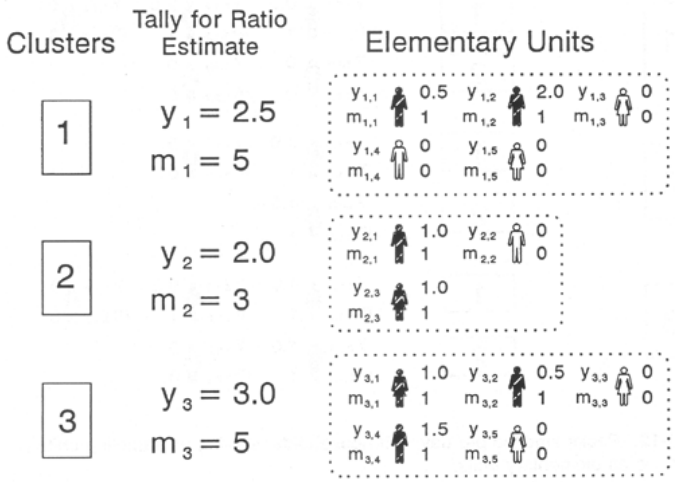

Figure 5-13. Tally of packs smoked per day – households serve as sampling units and persons as elementary units.

$$\bar{y} = \frac{(0.5 + 2.0) + (1.0 + 1.0) + (1.0 + 0.5 + 1.5)}{(1 + 1) + (1 + 1) + (1 + 1 + 1)}$$

which reduces to...

$$\bar{y} = \frac{(2.5) + (2.0) + (3.0)}{(2) + (2) + (3)} = \frac{7.5}{7} = 1.07$$

Smokers in the sampled population smoke slightly more than one pack per day.

5.3.4 Confidence Interval of Mean

For the confidence interval, we need the mean and variance of the mean. The variance of the sample mean is self-weighted, like the mean, as long as each household has the same probability of being selected. The equation for variance of the mean, accompanied by descriptions of the terms, is shown in Figure 5-15 and in Formula 5.32.

$$v(\bar{y}) = \frac{\sum_{i=1}^{n} (y_i - \bar{y} m_i)^2}{n(n-1)\bar{m}^2} \qquad (5.32)$$

The formula calculates the deviations of the sum of values in the individual cluster from the expected sum of value for that sized cluster in the total sampled. The standard error of the mean is...

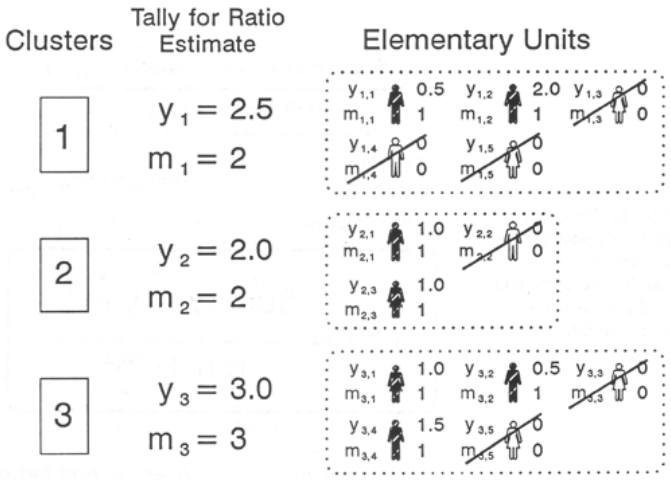

Figure 5-14. Tally of packs smoked per day by smokers – households serve as sampling units and persons as elementary units.

$$se(\bar{y}) = \sqrt{v(\bar{y})} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y} m_i)^2}{n(n-1)\bar{m}^2}} \qquad (5.33)$$

The confidence interval is calculated the same as before using information on both the mean and standard error. The formula for the 90 percent confidence interval of a mean is...

$$CI_{90\%}(\bar{y}) = \bar{y} \pm 1.64\ se(\bar{y}) \qquad (5.34)$$

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

$$CI_{95\%}(\bar{y}) = \bar{y} \pm 1.96\ se(\bar{y}) \qquad (5.35)$$

and the 99 percent confidence interval is...

$$CI_{99\%}(\bar{y}) = \bar{y} \pm 2.58\ se(\bar{y}) \qquad (5.36)$$

Using the data for all members of the household shown in Figure 5-13 and Formula 5.32, the variance of the mean number of packs smoked per day is...

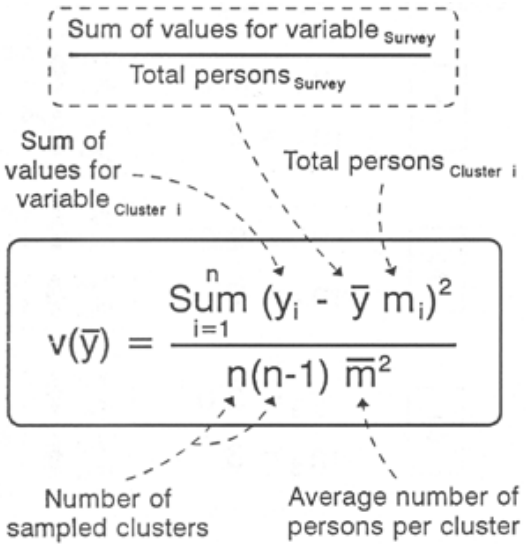

Figure 5-15. Variance formula of a mean for surveys where households are sampling units and persons are elementary units.

$$v(\bar{y}) = \frac{[2.5-(0.58\,x\,5)]^2 + [2.0-(0.58\,x\,3)])^2 + (3.0-(0.58\,x\,5)]^2}{3\,(2)\,4.33^2} = 0.002$$

and using Formula 5.33 the standard error is...

$$se(\bar{y}) = \sqrt{0.002} = 0.05$$

The mean number of packs smoked per day was previously calculated as 0.58. Therefore using Formula 5.35, the 95 percent confidence interval for the mean is...

$$CI_{95\%}(\bar{y}) = 0.58 \pm (1.96\ x\ 0.05) = 0.58 \pm 0.09$$

The 95 percent confidence interval is derived as the mean followed in parentheses by the upper and lower limits of the confidence interval, or...

$$0.58\ \ (0.49, 0.67)$$

The interpretation of the findings is that 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 0.49 and 0.67, with my best estimate being 0.58.

For the variance of the mean number of packs smoked per day by smokers, we also use Formula 5.32 but limit the number of persons to those who smoked. As noted earlier, the mean number of packs smoked per day by smokers is 1.07. Using the data for smokers only from Figure 5-14, the mean estimate of 1.07 and Formula 5.32, the variance of the mean number of packs smoked per day by smokers is...

$$v(\bar{y}) = \frac{[2.5-(1.07\,x\,2)]^2 + [2.0-(1.07\,x\,2)])^2 + (3.0-(1.07\,x\,3)]^2}{3\,(2)\,2.33^2} = 0.006$$

and using Formula 5.33 the standard error is...

$$se(\bar{y}) = \sqrt{0.006} = 0.08$$

The mean number of packs smoked per day was previously calculated as 0.58. Therefore using Formula 5.35, the 95 percent confidence interval for the mean is...

$$CI_{95\%}(\bar{y}) = 1.07 \pm (1.96\ x\ 0.08) = 1.07 \pm 0.15$$

Observe that the numbers are not quite as shown due to rounding during intermediate steps. The 95 percent confidence interval is derived as the mean followed in parentheses by the upper and lower limits of the confidence interval, or...

$$1.07\ \ (.92, 1.22)$$

The interpretation of the findings is that if there is no bias or confounding, I am 95 percent confident that the mean number of packs smoked per day by smokers in the sampled population lies between 0.92 and 1.22, with my best estimate being 1.07.

Notice that different from the first survey in which persons are *sampling units*, in the second survey we did not need to have a constant number of persons per cluster. In fact the number per cluster varied from 3 to 5 when all persons were considered and 2 to 3 when we focused only on smokers. Instead the second survey required only a constant number of households per cluster. By analyzing the data as a ratio estimator, we were able to derive a mean and variance for *elementary units* (i.e., either persons or smokers) that were different from *sampling units* (i.e., households).

Keep in mind that the mean and variance formulas for household surveys are only valid in certain circumstances. First, households in the population must be selected with equal probability so that all households have the same chance of appearing in the study. That is, the sample must be self-weighted. Second, an equal number of households 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 households is sampled from each selected cluster. If you are doing a household survey and the number of sampled households varies from one cluster to the next, you will need to use a more complicated set of formulas for the mean and variance estimates. These formulas are beyond the scope of this book and require the active assistance of a sampling statistician. A list of advanced sampling texts is presented in the Appendix.

5.4 SAMPLING OF PERSONS/HOUSEHOLDS

In some surveys it is not clear if persons or households are being sampled. For example, the Expanded Program of Immunization (EPI) of the World Health Organization suggests doing small surveys for vaccination coverage involving 30 clusters and seven children per clusters. The 30 clusters are selected with probability proportionate to size, and thereafter an equal number of children are selected in each cluster. Usually the children are 12 through 23 months of age, a time when all should have already been vaccinated. The surveys report the prevalence of vaccination coverage for BCG (tuberculosis), DPT (diphtheria, pertussis and tetanus), OPV (poliomyelitis) and measles. Because of the limited age range, the children most likely come from different households. Thus the second stage of the EPI style surveys could be viewed as samples of seven young children per cluster or seven households per cluster, but limited to households with at least one 12-23 month old child.

If we viewed the EPI surveys as a sample of children, the *sampling unit*s would be deemed the same as the *elementary unit*s and as such, we should use of the equations in Section 5.2. Yet if we viewed the survey as a sample of households, each with one eligible young child, the *sampling unit*s would he households, different from the *elementary units* of young children. This view would leading us to use the formulas in Section 5.3. So which approach is correct for the EPI surveys? The answer is that it does not matter because both sets of formulas will give you the same results

For example, let us return to the first survey shown in Figure 5-1. Here we sampled seven persons per cluster, with the *sampling unit* being people. The data were presented for smoking status in Figure 5-3. Now assume that we are doing a survey of households, but limited to those with eligible people. Also assume that we are only interested in persons aged 18 years. With this restriction there will rarely be more than one eligible person per household.

5.4.1 Proportion and Variance of Proportion

If the same data for persons in Figure 5-1 came from a household survey, they could be analyzed as a ratio estimator. The count of smokers, *a*, would be in the numerator while the count of persons who exist, *m*, would be in the denominator (see Figure 5-16). We would tally the number of sampled smokers per cluster and the number of sampled people per cluster, recognizing they come from an equal number of households per cluster. The proportion who smoke is derived using Formula 5.23 in Section 5.3 rather than Formula 5.4 in Section 5.2. Using the data from Figure

5-16, the proportion is...

$$p = \frac{(1+1+1+0+0+0+0)+(0+0+0+0+0+1+1)+(0+1+1+1+1+1+1)}{(1+1+1+1+1+1+1)+(1+1+1+1+1+1+1)+(1+1+1+1+1+1+1)}$$

$$p = \frac{(3)+(2)+(6)}{(7)+(7)+(7)} = \frac{11}{21} = 0.52$$

or the same as derived earlier in Section 5.21 using Formula 5.6 for a proportion. When the number of persons per cluster is the same in all clusters, the proportion and ratio estimator are the same. But what about the variance? Instead of using Formula 5.8, we will analyze the data in Figure 5-17 using Formula 5.25.

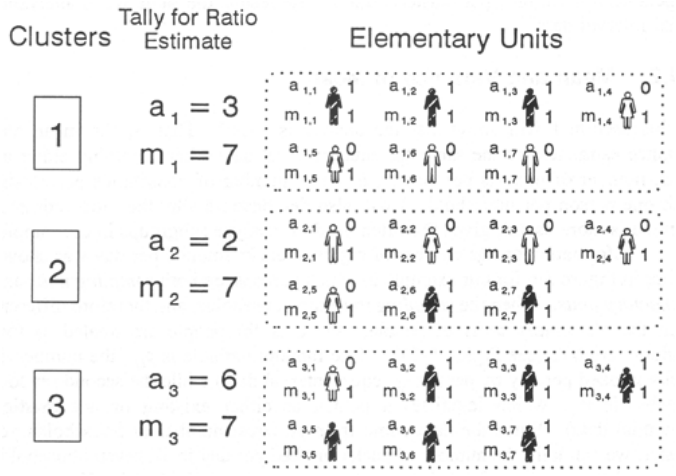

Figure 5-16. Tally of smokers – households serve as sampling units and persons as elementary units, with constant number of persons per cluster.

$$v(p) = \frac{[3-(7\,x\,0.52)]^2 + [2-(7\,x\,0.52)]^2 + [6-(7\,x\,0.52)]^2}{3\,(2)\,7^2} = 0.029$$

The variance is exactly the same as we calculated in Section 5.2 using Formula 5.8. The reason is that the two formulas are identical when a constant number of persons are sampled per cluster. But are the results the same when analyzing equal interval data?

5.4.2 Mean and Variance of Mean

In this section I will show that the answer is "yes." That is, the mean and variance estimates are the same for equal interval data when sampling either an equal number of persons per cluster, or equal number of households per cluster with one person per household. I will also demonstrate why the ratio estimator formula is more useful, given the often need to analyze subgroups in the sample.

Information on the number of cigarette packs smoked per day was shown earlier in Figure 5-6 for our example in which persons are both *sampling units* and *elementary units*. When the *sampling unit*s are households, and therefore different from the *elementary units* of people, the data for people are treated as two random variables (see Figure 5-17). One random variable is $y_{i,j}$, the number of packs smoked per day by person i,j (equal interval data) while the second random variable is $m_{i,j}$, which identifies a person as either existing or not existing (binomial data). Since the

data come from an equal number of households per cluster, we can tally the number of packs smoked per day in all seven households as y_i and the number of existing people as m_i. , where i is the identifier for the three clusters.

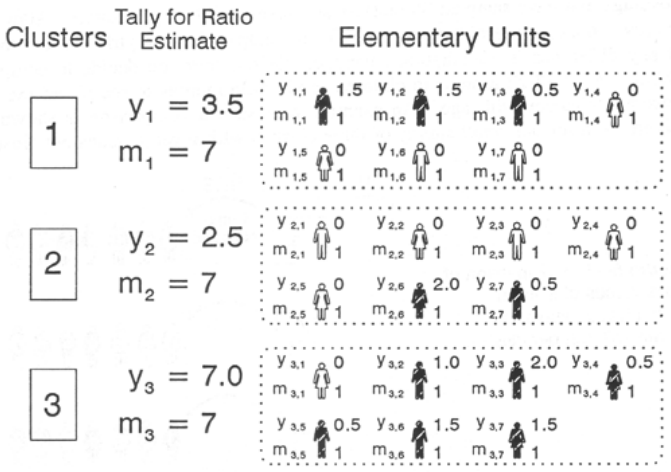

Figure 5-17. Tally of packs smoked per day – households serve as sampling units and persons as elementary units, with constant number of persons per cluster.

$$\bar{y} = \frac{(1.5+1.5+0.5+0+0+0+0)+(0+0+0+0+0+2.0+0.5)+(0+1.0+2.0+0.5+0.5+1.5+1.5)}{(1+1+1+1+1+1+1)+(1+1+1+1+1+1+1)+(1+1+1+1+1+1+1)}$$

$$\bar{y} = \frac{(3.5)+(2.5)+(7.0)}{7+7+7} = \frac{13}{21} = 0.62$$

The mean number of packs smoked per day is the same as analyzed in Section 5.2.3, namely 0.62.

For calculating the variance, we use Formula 5.29 for the variance of a ratio estimator rather than Formula 5.16. The findings are...

$$v(\bar{y}) = \frac{[3.5-(0.62\,x\,7)]^2+[2.5-(0.62\,x\,7)]^2+[7.0-(0.62x7)]^2}{3\,(2)\,7^2} = 0.038$$

the same value as previously calculated in Section 5.2.4. Thus with both a proportion and a mean, the two sets of formulas give identical results. Keep in mind, however, that we assumed that there was one eligible person per household and that we were actually doing a household survey.

5.4.3 Analysis of Subgroups

Based on what has been presented in the prior two sections, its seems that I have made things more complicated, but for no apparent reason. After all, why insist on doing a household survey when it is much easier to sample persons. By selecting an equal number of persons per cluster, there is no need to understand the concept of a ratio estimator or use more complicated equations. Yet things are not always as easy as they seem.

In many instances, you will want information on more than one narrowly defined group. For example, assume that you are doing a survey of immunization coverage and have sampled 30 clusters and seven children per cluster. After the survey is done, you decide you would like to compare boys to girls to see if there is any difference in vaccination coverage. Or possibly you decide to compare children aged 12-23 months to children aged 24-36 months to see if vaccination

coverage changes with age. An example of a such a comparison is shown in Figure 5-18 for our small survey of three clusters with seven persons per cluster. Here I have separated the persons into males (black number on white background) and females (white number on black background). For such an analysis, the clusters no longer have an equal number of persons. Cluster 1 in Figure 5-18 has five males, cluster 2 has three males and cluster 3 has four males. Thus without constant numbers per cluster, we cannot use the simpler formulas to derive our basic statistics for males. We have violated the basic assumption of equal number of sampled units per cluster. The same situation holds true with females.

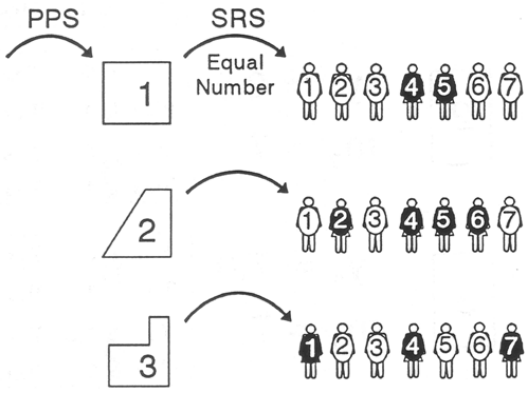

Figure 5-18. Comparison of boys versus girls in an immunization survey with seven children per cluster.

Another example is shown in Figure 5-19 for packs of cigarettes smoked per day. In Section 5.4.2 we found that an average of 0.62 packs were smoked per day in the sampled population. Some of the people, however, do not smoke. Thus in our analysis we want to determine the number of packs smoked per day by smokers. To do this, we eliminate non-smokers from the analysis as shown in Figure 5-19. We can still derive the mean number of packs smoked in each cluster, but the mean for cluster 1 of 1.17 is based on three persons, the mean for cluster 2 of 1.25 is based on 2 persons and the mean for cluster 3 of 1.17 is based on 6 persons. Clearly the three cluster-specific means cannot be averaged together as they were with Formula 5.14 to obtain the mean for the sample. Nor can we calculate the variance of the mean by using Formula 5.16 to tally the deviations of each mean from the group mean. Instead, we are faced with a set of three clusters with varying numbers of persons per cluster (see Figure 5-20), similar to what occurs with a household survey in which the *sampling unit* is households rather than people. Thus our only option for analysis of the smoking subgroup is to use the ratio estimator formulas.

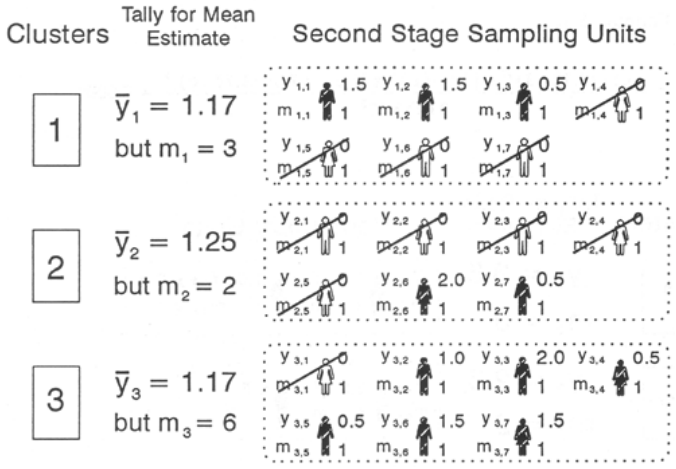

Figure 5-19. Tally of packs smoked per day by smokers (eliminate non-smokers) – households serve as sampling units and persons as elementary units.

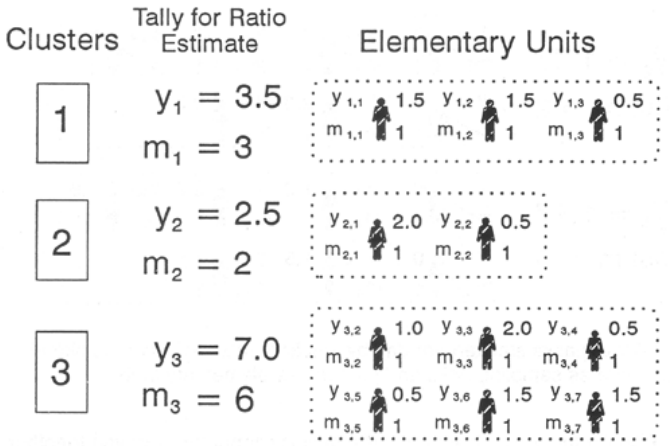

Figure 5-20. Tally of packs smoked per day by smokers – households serve as sampling units and persons as elementary units.

Using Formula 5.30 the mean number of packs smoked per day by smokers is...

$$\bar{y} = \frac{(1.5 + 1.5 + 0.5) + (2.0 + 0.5) + (1.0 + 2.0 + 0.5 + 0.5 + 1.5 + 1.5)}{(1 + 1 + 1) + (1 + 1) + (1 + 1 + 1 + 1 + 1 + 1)}$$

which reduces to...

$$\bar{y} = \frac{(3.5) + (2.5) + (7.0)}{(3) + (2) + (6)} = \frac{13.0}{11} = 1.18$$

For the variance of the mean number of packs smoked per day, we use the ratio estimator Formula 5.32

$$v(\bar{y}) = \frac{[3.5 - (1.18 \times 3)]^2 + [2.5 - (1.18 \times 2)])^2 + (7.0 - (1.18 \times 6)]^2}{3\,(2)\,3.67^2} = 0.0004$$

and using Formula 5.33 the standard error is...

$$se(\bar{y}) = \sqrt{0.0004} = 0.019$$

Finally using Formula 5.35, the 95 percent confidence interval for the mean is...

$$CI_{95\%}(\bar{y}) = 1.18 \pm (1.96 \times 0.018) = 1.18 \pm 0.037$$

Observe that the numbers are not quite as shown due to rounding during intermediate steps. The 95 percent confidence interval is derived as the mean followed in parentheses by the upper and lower limits of the confidence interval, or...

$$1.18\ (1.14, 1.22)$$

which indicates that if there is no bias or confounding, we can be 95 percent confident that the mean

number of packs smoked per day by smokers in the sampled population lies between 1.14 and 1.22, with the best estimate being 1.18.

5.5 CONCLUSION

Two approaches have been presented for doing two-stage cluster surveys. The first selects clusters with probability proportionate to the number of persons residing in the cluster and then selects a constant number of persons per cluster. People are both the *sampling unit*s and *elementary units*. This approach is easiest to analyze with hand calculations of the mean and variance and is favored by the Expanded Program on Immunization of the World Health Organization. Its is limited, however, to an analysis of all sampled persons combined and cannot be used to examine subgroups within the survey.

An alternative approach is to select clusters with probability proportionate to the number of households in the cluster (equivalent to PPS sampling based on population counts), and thereafter select a constant number of households per cluster. Households are the *sampling units* while people in the households are the *elementary units.* The analysis is done as a ratio estimator and can be done both for the total sample as well as subgroups.

When a constant number of persons are sampled per cluster, the two sets of formulas give the same results. Thus to avoid confusion, I recommend using the ratio estimator formulas when doing two-stage cluster surveys, with either a constant number of persons per cluster or constant number of households per cluster. In most instances, I recommend sampling a constant number of households per cluster, with households limited to those with members of the eligible population.

5.5.1 Eligible Persons and Households

Most surveys specify who is eligible to be included in the sample. For some surveys it may be young children, ages 5 years or less, with the mothers being the respondent. For others it may be adolescents and young adults, who are currently enrolled in high school. For still others it may be married adults, age 20-44 years. If persons are both the s*ampling units* and *elementary units*, then enough households must be visited in a random manner to obtain interviews or examinations of a constant number of eligible persons per cluster. If households are the *sampling units* and persons within households are the *elementary units*, then the surveyor must sample at least one person per household in an equal number of households per cluster.

An eligible person is easy to define, following the specifications of the survey. Eligible households are those with one or more eligible person. Thus to be more exact, *sampling units* are either eligible persons or eligible households, not just persons or households. Similarly, *elementary units* are eligible persons in the households, and not all persons who reside in the household.