Kevin M. Sullivan, PhD, MPH, MHA, Associate Professor
Rollins School of Public Health at Emory University

Sampling for Epidemiologists

This document describes how to calculate proportions with confidence intervals assuming simple random sampling (SRS), one-stage cluster surveys (1sc), probability proportional to size (PPS) cluster sampling, and stratified cluster sampling. Sample size calculations are also presented.

SIMPLE RANDOM SAMPLING (SRS)

Point and Variance Estimation for a Proportion Assuming SRS

Simple random sampling (SRS), when applied to population surveys, is when every eligible individual in the population has the same chance or probability of being selected. This usually means the availability of a list of all eligible individuals and, using a random selection scheme, a sample of individuals is selected to be surveyed. This type of sampling could be used in a situation where a listing of the population is available, such as a university, where a listing of all enrolled students could be obtained, or a survey of voters could use a voter registration list. However, in some survey situations no such listing of individuals of interest in a population is available. For example, for a national survey of the immunization level of children 12-23.9 months of age, rarely would there be a listing of all children in this age group at a national level.

When analyzing data, most statistical software assumes that SRS was used. The formula for calculating a proportion, variance, and confidence interval are presented next.

Point estimate for simple random sampling

$\hat{p}_{srs} = a / n$

where

$\hat{p}_{srs}$ = the estimated proportion assuming SRS

a = the number of individuals with the attribute of interest

n = the number of individuals sampled

Variance estimate for simple random sampling with the fpc

$$\hat{\text{var}}(\hat{p}_{srs}) = \frac{\hat{p}_{srs}\hat{q}_{srs}}{n-1}\left(\frac{N-n}{N}\right)$$

where

$\hat{q}_{srs} = 1 - \hat{p}_{srs}$

N = population size

The $(N\text{-}n)/N$ term is called the "finite population correction" or *fpc*. If the size of the population N is large relative to the number sampled n, then this term will have little effect on the variance estimate. For example, say the population size is 1,000,000 and 900 individuals are sampled, the *fpc* would be:

fpc = (1,000,000-900)/1,000,000 = .9991 ≈ 1

In this example the *fpc* will have little influence on the variance estimate. When the proportion of the population sampled is relatively high, the use of the *fpc* will decrease the size of the variance estimate, which will in turn reduce the width of the confidence interval. Many textbooks ignore the *fpc* in their presentation on how to calculate the variance or sample size for a proportion. The formula for a two-sided confidence interval is presented below. The t-value is shown in the formula with n-1 degrees of freedom. If the number sampled is large, around 500 or more, then the Z-value could be used (i.e., 1.96 for two-sided Z-value). Most statistics textbooks present the use the Z-value under the assumption that the sample size is large.

Two-sided confidence interval for the point estimate (SRS)

$$\hat{p}_{srs} \pm t_{1-\alpha/2,n-1}\sqrt{\text{var}(\hat{p}_{srs})}$$

Example

Assume that an SRS survey was performed in a specific area for immunizations and it was found that 48 children out of 70 surveyed were properly immunized. Assume that the population from which the 70 children were selected was large, say >100,000, and therefore the *fpc* would have little impact on the variance. The SRS point estimate, variance estimate, and 95% confidence interval are calculated as:

$$\hat{p}_{srs} = 48/70 = .686 \text{ or } 68.6\%$$

The variance will be calculated assuming that *fpc* will be close to 1 and therefore ignored.

$$\hat{\text{var}}(\hat{p}_{srs}) = \frac{.686 \times .314}{70-1} = .00312$$

For confidence interval, a t-value with 69 degrees of freedom is use (70 surveyed – 1) which is 1.995:

$$.686 \pm 1.995\sqrt{.00312} = .686 \pm .111 = (.575, .797)$$

The interpretation would be that, assuming SRS, the immunization level in the area sampled is estimated to be 68.6%; we would be 95% confident that the true immunization level is captured between 57.5% and 79.7%. (Note that **if** the *fpc* had been used, to narrow the confidence interval in this example by 0.1% or more, the population size *N* would need to be 2,500 or less).

Sample Size Calculation for Simple Random Sampling

The formula for calculating a sample size with simple random sampling (SRS) using the "specified absolute precision" approach is presented below. This formula assumes that the investigator desires to have a 95% confidence interval (the 1.96 value in the formula). The Z-value of 1.96 is used under the assumption that a relatively large sample size will be selected. In addition, while it might be more correct to use a t-value, the t-value requires the degrees of freedom which is based on the sample size. The formula also incorporates the *fpc*.

Sample size formula for simple random sampling (SRS) ***with*** *the finite population correction factor (fpc)*

$$n_{srs} = \frac{N\hat{p}_{srs}\hat{q}_{srs}}{\frac{d^2}{1.96^2}(N-1) + \hat{p}_{srs}\hat{q}_{srs}}$$

where

n_{srs} = sample size

N = population size

$\hat{p}_{srs}$ = the estimated proportion

$\hat{q}_{srs} = 1 - \hat{p}_{srs}$

d = desired absolute precision

If a very small proportion of the population is to be sampled, then the *fpc* could be dropped from the formula as shown below:

*Sample size formula for simple random sampling (SRS) with****out*** *the finite population correction factor (fpc)*

$$n_{srs} = \frac{1.96^2\,\hat{p}_{srs}\hat{q}_{srs}}{d^2}$$

For both of the above sample size formulae (with or without the *fpc*), the investigator must come up with an estimate or educated guess for the proportion *p* of the population that will have the factor under investigation and the desired level of absolute precision *d*. If the investigator is unsure of the proportion, usually a value of .5 or 50% is used. The reason for selecting .5 is that, for a given level of precision, a *p* of .5 has the largest sample size. To see this, in the numerator of the sample size formula is *pq*. The larger the value of *pq*, the larger will be the sample size. When *p*=.5 and *q*=.5, then pq = .25. When *p*=.6, *pq* = .24. Finally, as one more example, when *p*=.9, *pq* = .09.

The other value the investigator must provide is the level of desired absolute precision *d*. The level of precision is how far (in absolute terms) the lower and upper bound of the confidence limits should be from the point estimate. For "common" events (10% to 90%), the *d* value is usually set at .05. For example, say the investigator has decided that the proportion *p* is 50% and the level of precision *d* is 5%. If the investigator's estimate of *p* was correct, then the 95% confidence limits would be from 45% to 55% (i.e., $\pm$5%).

Example

Investigators want to determine the proportion of health providers in a large metropolitan hospital who have received three doses of Hepatitis B vaccine. They go to the personnel office and are told that there are 1,536 health providers employed. The investigators believe that 80% of health providers have received 3 doses of Hepatitis B vaccine, but to be sure they want to perform a survey. They decide that they want their estimate to be $\pm$3% (i.e., the *d* value) with 95% confidence and use the formula with the *fpc*.

$$n_{srs} = \frac{1536(.8)(.2)}{\frac{.03^2}{1.96^2}(1536-1) + (.8)(.2)} = \frac{245.76}{.5196} = 472.98 = 473$$

Therefore, 473 health care providers need to be surveyed. If the sample size formula with*out* the *fpc* had been used, the sample size would be 683 providers.

CLUSTER SAMPLING

Another approach to sampling is cluster sampling. With cluster sampling, the population is divided into exclusive and exhaustive groups, sometimes referred to as primary sampling units (PSUs). With one-stage cluster sampling, a sample of PSUs is selected and, within all selected PSUs or clusters, all elements assessed. With two-stage cluster sampling, at the first stage a sample of PSUs is selected, and at the second stage, a sample of elements within each of the selected clusters are assessed. With population-based surveys, the PSUs are usually geographic areas, such as enumeration units, census tracts, and communities. The elements are frequently households or individuals within households.

Frequently cluster sampling is used because a listing of all eligible individuals is **not** available and therefore simple random sampling cannot be performed. Another reason for using cluster sampling rather than SRS is that from a logistical viewpoint, cluster surveys tend to be more efficient in terms of transportation and time in large geographic areas.

In the following two sections we will describe two approaches to the analysis of cluster surveys. In both instances it is assumed that the selection at the first stage is by probability proportional to size (PPS; described in Appendix 1). There are other ways to sample PSUs, such as through simple random or systematic sampling, but here we concentrate on the selection by PPS. The first method is called the One-Stage Cluster Approach, and the other the PPS Equal Weighting Cluster Approach.

One-Stage Cluster Approach

The formulae for estimating a proportion, its variance, and 95% confidence interval assuming a one-stage cluster (*1sc*) approach are presented next.

Point estimate for 1sc

The point estimate for *1sc* is the same as *srs*:

$$\hat{p}_{1sc} = a/n$$

where

$\hat{p}_{1sc}$ = the estimated proportion assuming a one-stage cluster survey

a = the number of individuals with the attribute of interest

n = the number of individuals sampled

Approximate variance estimate for a proportion assuming 1sc

$$\hat{\text{var}}(\hat{p}_{1sc}) = \frac{\sum_{i=1}^{m} (a_i - \hat{p}_{1sc} n_i)^2}{\bar{n}^2 m(m-1)}$$

a_i = number with attribute in ith cluster

n_i = number assessed in ith cluster

$\bar{n}$ = average number sampled per cluster

m = number of clusters

Approximate two-sided confidence interval for a proportion assuming 1sc

$$\hat{p}_{1sc} \pm t_{1-\alpha/2,m-1} \sqrt{\hat{\text{var}}(\hat{p}_{1sc})}$$

Note that when using complex sample commands in SAS, SPSS, and Epi Info, the point estimate and variance are calculated assuming a 1-stage cluster design when a cluster variable is specified and no sample weight variable is provided. Note also the use of the *t*-statistic for the confidence interval where the degrees of freedom are based, in part, on *m*-1, i.e., the number of clusters minus one (assuming a single stratum/no stratification). For example, in a 30 cluster survey, a Z-value for a two-sided 95% confidence interval assuming *srs* is 1.96 whereas for a one-stage cluster survey the equivalent *t*-value would be 2.0452, a 4.3% larger multiplier of the standard error.

Example

An example of a 10-cluster immunization survey is presented in Table S.1. Usually around 30 clusters are selected, but for purposes of performing the calculations by hand, only 10 clusters are presented. Within each cluster, seven children were selected and it was determined if they were completely immunized ("VAC"=1) or not completely immunized ("VAC"=2). The *1sc* point estimate, variance estimate, and 95% confidence interval are calculated as:

$$\hat{p}_{1sc} = 48/70 = .686 \text{ or } 68.6\%$$

$$\hat{\text{var}}(\hat{p}_{1sc}) = \frac{(3-.686*7)^2 + (7-.686*7)^2 + ... + (6-.686*7)^2 + (1-.686*7)^2}{(70/10)^2 * 10(10-1)} =$$

$$\frac{3.24 + .4.84 + ... + 1.44 + 14.44}{49 * 10(9)} = \frac{31.600}{4410} = .00717$$

The 95% two-sided *t*-value with 10-1 degrees of freedom is 2.2621.

$$.686 \pm 2.2621\sqrt{.00717} = .686 \pm .192 = (.494, .878)$$

The interpretation would be that the immunization level of children in the area sampled is estimated to be 68.6%; we would be 95% confident that the true immunization level is captured between 49.4% and 87.8%. Note that the confidence interval assuming *1sc* is wider than when SRS is assumed (95% CI for SRS, 57.5%, 79.7%; see Figure S.1). In general, confidence intervals calculated from a *1sc* survey will be wider than those calculated assuming the data were collected using SRS. The wider confidence interval for *1sc* surveys is attributed the cluster design.

Table S.1. Example Data

	VAC		
CLUSTER	1	2	Total
1	3 (42.9%)	4	7
2	7 (100.0%)	0	7
3	4 (57.1%)	3	7
4	5 (71.4%)	2	7
5	5 (71.4%)	2	7
6	7 (100.0%)	0	7
7	4 (57.1%)	3	7
8	6 (85.7%)	1	7
9	6 (85.7%)	1	7
10	1 (14.3%)	6	7
Total	48 (68.6%)	22	70

PPS Equal Weighting Cluster Approach

The formulae for estimating a proportion, its variance, and 95% confidence interval for the PPS equal weighting cluster approach (PPS) are presented next.

Point estimate for PPS

$$\hat{p}_{pps} = \frac{\sum_{i=1}^{m} \hat{p}_i}{m}$$

$\hat{p}_i$ = proportion estimate in the *i*th cluster

m = the number of clusters

Note that when analyzing data, the point estimate for SRS will be the same as the point estimate for PPS **when** the number of individuals sampled in each cluster is the same. The variance estimate and confidence interval formula are presented below. As with 1sc, for the *t*-value, the degrees of freedom is the number of clusters – 1 (i.e., *m*-1).

Variance estimate for PPS sampling

$$\hat{\text{var}}(\hat{p}_{pps}) = \frac{\sum_{i=1}^{m} (\hat{p}_i - \hat{p}_{pps})^2}{m(m-1)}$$

Two-sided confidence interval (PPS)

$$\hat{p}_{pps} \pm t_{1-\alpha/2,m-1}\sqrt{\hat{\text{var}}(\hat{p}_{pps})}$$

Example

Using the example data presented in Table S.1, the PPS point estimate, variance estimate, and 95% confidence interval are calculated as:

$$\hat{p}_{pps} = \frac{.429 + 1.0 + .571 + .714 + .714 + 1.0 + .571 + .857 + .857 + .143}{10} = \frac{6.856}{10} = .686 \text{ or } 68.6\%$$

$$\hat{\text{var}}(\hat{p}_{pps}) = \frac{(.429 - .686)^2 + (1.0 - .686)^2 + ... + (.857 - .686)^2 + (.143 - .686)^2}{10(10-1)} =$$

$$\frac{.066049 + .098596 + ... + 029241 + .294849}{10(9)} = \frac{.64459}{90} = .007162$$

The 95% two-sided t-value with 10-1 degrees of freedom is 2.2621.

$$.686 \pm 2.2621\sqrt{.007162} = .686 \pm .191 = (.495, .877)$$

The interpretation would be that the immunization level of children in the area sampled is estimated to be 68.6%; we would be 95% confident that the true immunization level is captured between 49.5% and 87.7%. Note that the point estimates assuming SRS calculated earlier for 48 out of 70 children and PPS and 1sc are identical because the number of children sampled per cluster was the same. Also note that the confidence interval assuming PPS and 1sc are wider than when SRS is assumed (95% CI for SRS, 57.5%, 79.7%; see Figure S.1). In general, confidence intervals calculated from a PPS survey will be wider than

those calculated assuming the data were collected using SRS. Also note that when there are the same number of individuals sampled per cluster, the 1sc and PPS point and variance estimates will be the same. When analyzing cluster survey data in many programs, the variance is estimated using the 1sc rather than the PPS method.

Design Effect (DEFF) and Intra-cluster Correlation Coefficient (ICC)

The design effect (*deff*) is a measure of the variability between clusters and is calculated as the ratio of the variance calculated assuming a complex sample design divided by the variance calculated assuming SRS:

Formula for calculating the design effect (deff)

$$\hat{deff} = \frac{\hat{\text{var}}(\hat{p}_{cluster})}{\hat{\text{var}}(\hat{p}_{srs})}$$

In most circumstances, the *deff* will be greater than 1, indicating that the variance estimated assuming cluster sampling is larger that the variance assuming SRS. However, sometimes the DEFF can be less than one. From the example data in Table S.1, the *deff* based on PPS is:

$$\hat{deff} = \frac{.007162}{.00312} = 2.3$$

The interpretation would be that the variance assuming PPS is 2.3 times larger than the variance assuming SRS. What effect does the *deff* have in planning a study? An estimate of the *deff* is frequently used for sample size calculations for a cluster survey. Note that while the *deff* estimated above was 2.3, the confidence interval width is increased by a smaller amount. In the example in Table S.1, to derive the confidence interval limits, for the SRS they are 68.6%$\pm$11.1%; compare this to the PPS which is 68.6%$\pm$19.1%; therefore, the PPS interval is approximately 1.7 times wider than the SRS interval in this example.

The three most important factors that affect the size of the *deff*:

1. The inherent variability of the proportion of the factor between clusters; the more the clusters differ in the proportion with the attribute, the larger the *deff*.
2. The number of individuals sampled in each cluster; the more individuals sampled per cluster, the larger the *deff*.
3. Estimates near 50% tend to have larger *deff*s than estimates near the extremes (given equal sample sizes)

The investigator has little or no control over 1 and 3 above, but in designing a survey, can determine the number of individuals to sample per cluster. Sample size issues are discussed further in the next section.

ICC is the *intra-cluster correlation coefficient* and is a measure of the relatedness of observations within each cluster. The ICC is also sometimes referred to as the *rate of homogeneity* or ROH. For a given DEFF and average number of individuals sampled per cluster ($\bar{n}$), the ICC can be calculated as:

$$\text{ICC}=(\text{DEFF}-1)/(\bar{n}-1)$$

An important feature of the ICC is that it is not affected by the average number of observations per cluster, whereas the DEFF is strongly affected. As an example of calculating the ICC, using the above example data where DEFF = 2.3 and $\bar{n}$ =7:

ICC = (2.3-1)/(7-1) = (1.3)/(6)= 0.2167

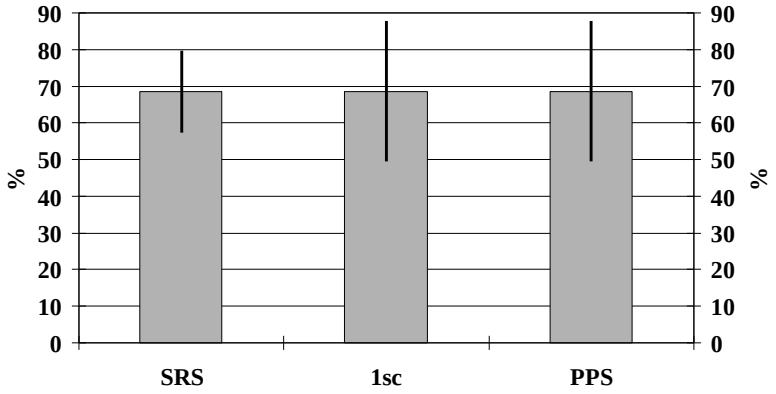
The DEFF can also be calculated from a known ICC and $\bar{n}$:

$$\hat{DEFF} \cong 1 + (\bar{n} - 1) \times ICC$$

As an example where the ICC=0.2167 and $\bar{n}$ =7:

DEFF=1+(7-1)x0.2167 = 1 + 6 * 0.2167 = 1 + 1.3 = 2.3

Figure S.1. Comparison of 95% Confidence Intervals between Simple Random Sampling (SRS), One-Stage Cluster (1sc), and Proportionate to Population Size (PPS) Sampling



Comparison of 1sc and PPS

As previously mentioned, if there is the same number of elements sampled in each cluster, e.g., 30, then the point and variance estimates for the 1sc and PPS methods will be the same as shown in the previous example. If the number of elements differs by cluster, then the 1sc and PPS methods *may* provide different point and variance estimates, although frequently the difference is not that great. Using the 2004 Afghanistan survey for the prevalence of anemia among children 6-59 months of age, 32 clusters were assessed. In these 32 clusters, the number of children 6-59 months of age who participated in the survey and had anemia status determined ranged from 8 to 47 children per cluster, with an average of 27.2 children per cluster. The analyses of these data for the three methods (SRS, 1sc, and PPS) are shown in Table S.2 and Figure S.2.

Figure S.2. Comparison of 95% Confidence Intervals between Simple Random Sampling (SRS), One-Stage Cluster (1sc), and Proportionate to Population Size (PPS) Sampling, Afghanistan 2004 survey, anemia in children 6-59 months of age

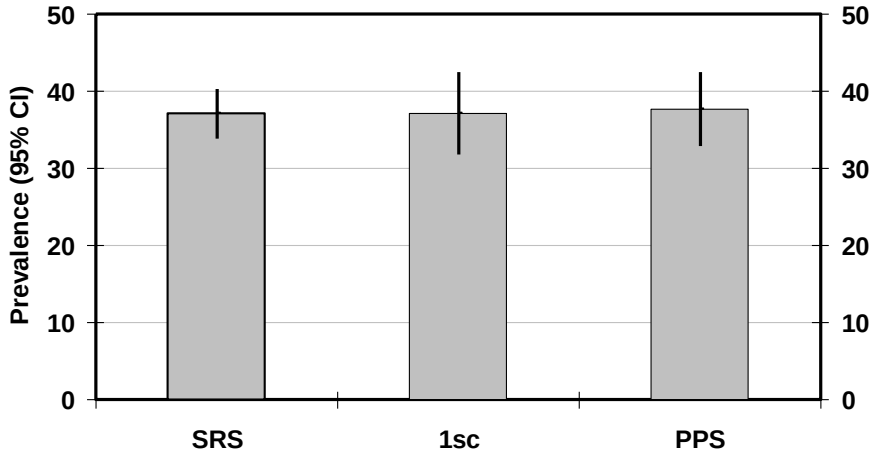


Table S.2. Comparison of estimates assuming SRS, 1sc, and PPS, Afghanistan 2004 survey*, anemia in children 6-59 months of age.

Method	Point estimate	95% CI	DEFF	ICC
SRS	37.1%	(33.9, 40.3)	-	-
1sc	37.1%	(31.8, 42.5)	2.57	.0600
PPS	37.7%	(32.9, 42.4)	2.04	.0397

*A 32 cluster survey with a total of 870 children with anemia information

Sample Size Calculation for Cluster Sampling

The sample size formula for a cluster survey, 1sc or PPS, is the *deff* times sample size estimate assuming SRS. Formula with and without the *fpc* are shown below. Again, 1.96 could be used for sample size estimation in place of *t*, although if one knows how many clusters are going to be sampled, the *t*-value should be used where *m*-1 is the number of clusters – 1. For a 30 cluster survey, the degrees of freedom would be 29, and the 95% CI two-sided t-value would be 2.0452

Sample size formula for probability proportional to size (PPS) sampling with the fpc

$$n_{pps} = deff \times \frac{N\hat{p}_{srs}\hat{q}_{srs}}{\frac{d^2}{t_{1-\alpha/2,m-1}^{\ 2}}(N-1) + \hat{p}_{srs}\hat{q}_{srs}} = deff \times n_{srs}$$

*Sample size formula for probability proportional to size (PPS) sampling with**out** fpc*

$$n_{pps} = deff \times \frac{t_{1-\alpha/2,m-1}^{\ 2}\hat{p}\hat{q}}{d^2}$$

The investigator needs to have an estimate of the *deff*. This estimate is usually from surveys of the same size performed previously in the area or based on the experience in other areas. For surveys on immunization and anthropometry, usually a value of 2 is used for the *deff*. For water and sanitation-related factors, usually a larger estimate of the *deff* is used, generally in the range of 5 to 9. Once the total

sample size has been calculated, the next step is to determine the number of individuals to be sampled in each cluster. This would be:

Formula for calculating the number of individuals to sample per cluster in a PPS survey

$$\text{number to sample per cluster} = \frac{n_{pps}}{m}$$

where

m = the number of clusters

In many surveys there are 30 clusters. Always round up on the number of individuals to survey per cluster, which will slightly increase the total sample size.

Example

The Ministry of Health is interested in determining the proportion of households using iodized salt. It has been decided to conduct a 30-cluster *pps* survey. They are unsure of the iodized salt coverage so an estimate of 50% is used and they want the precision to be ±5% with 95% confidence. The *deff* is estimated to be 2 and they will ignore the *fpc* in the sample size calculation. They are also planning on a 30 cluster survey so they use the *t*-value of 2.0452 rather than 1.96. What is the total sample size and how many per cluster? Ignoring the *fpc*, in this example:

$$n_{pps} = 2 \times \frac{2.0452^2(.5)(.5)}{.05^2} = 836.57 = 837$$

The total sample size is 837. How many would need to be sampled in each cluster? Assuming a 30-cluster survey, the number to sample per cluster would be 837/30=27.9. This would be rounded to 28 households per cluster; therefore the total sample size would be 30 x 28 = 840. Note that if the value of 1.96 been used in the above formula rather than 2.0452, the sample size is 769 overall compared to 837. One other issue is the need to deal with nonresponse. If it is estimated that 90% of households would participate in the survey, in this example the number of households initially selected would need to be increased to assure that 840 households would be assessed. To do this, take the sample size and divide the by response proportion. In the above example, 840/.9 = 933.3 or 934. That is, if 934 households are invited to participate in the survey and 90% agree, then there will be around 840 households participating.

The Expanded Program on Immunization (EPI) sample size is based on a *p*=.5, *d*=.1, and *deff*=2 with 95% confidence and ignoring the *fpc*. They also used the *Z* value of 1.96 rather than the *t* value. What is the sample size?

$$n_{pps} = 2 \times \frac{1.96^2(.5)(.5)}{.10^2} = 192.08 = 193$$

The EPI is a 30-cluster survey, so the number to sample in each cluster is 193/30=6.4, which is rounded to 7. Therefore, the sample size is 30 x 7 = 210 children.

STRATIFIED CLUSTER SAMPLING

For national surveys, sometimes the country is divided into two or more areas and a separate survey carried out in each area. The separate areas could be provinces or states, by urban/rural status, by topography (mountainous area vs. coastal region), or other political/geographic designations. For example, a country may want to assess the immunization status of children and has a number of choices concerning the sampling. One choice could be to perform one survey nationwide that would provide a national estimate. Another option would be to perform a separate survey in each province which would be used to calculate provincial estimates, and then combine all of the provincial surveys to derive a

national estimate. This latter method is referred to as *stratified sampling*. With stratified sampling, the geographic area of interest is divided into mutually exclusive and exhaustive strata. Mutually exclusive means that there is no overlap between the strata/geographic areas, and exhaustive means that all areas of interest in the geographic area must fall into one of the strata. For this section we will assume that PPS sampling was performed in each stratum/subnational area. The point and variance estimate and confidence interval for a single national estimate based on the stratum-specific values are shown below. These estimates are "weighted", that is, they take into account the differences in the population size of each stratum.

Point Estimate for Stratified PPS Sampling

$$\hat{p}_{.} = \frac{\sum_{j=1}^{s} N_j \hat{p}_j}{\sum_{j=1}^{s} N_j}$$

where

$\hat{p}_{.}$ = the combined (national) estimate

s = the number of strata

N_j = total population in the jth stratum

$\hat{p}_j$ = the proportion with the factor of interest in the jth stratum

Variance Estimate for Stratified PPS Sampling

$$\hat{\text{var}}(\hat{p}_{.}) = \frac{\sum_{j=1}^{s} N_j^2 \, \hat{\text{var}}(\hat{p}_j)}{\left(\sum_{j=1}^{s} N_j \right)^2}$$

where

$\hat{\text{var}}(\hat{p}_j)$ = the variance in the jth stratum

Two-sided confidence interval for Stratified PPS sampling

$$\hat{p}_{.} \pm t_{1-\alpha/2,m-s} \sqrt{\hat{\text{var}}(\hat{p}_{.})}$$

Note that the *t*-value has m-s degrees of freedom, that is, the total number of clusters surveyed minus the number of strata

Example

A stratified PPS survey to assess immunization levels in children 12 months up to 24 months of age is performed in a country. The country was divided into 3 strata, and an EPI (Expanded Program on Immunizations) survey performed in each stratum (30 clusters, 7 children in each cluster; note that exactly 210 children were not surveyed in each cluster in Table S.3; in some clusters, 8 children were surveyed). Table S.2 has the information relevant to the survey. In the first column are the strata numbered as 1, 2, and 3. In columns 2 and 3 is information on the estimated population size and percent distribution by stratum for children in the selected age group. The estimated population size is from the most recent national census. The number and percent of children surveyed in each stratum are shown in columns 4 and 5. The number surveyed by stratum differs slightly, but approximately one-third of the

surveyed children are from each stratum. Columns 6 through 8 are the results of the survey in each stratum. Generally one would like to calculate the national immunization coverage based on the stratum estimates. A naïve approach would be to add the number of children immunized, in this example 369, and divide this by the total surveyed: 369/656=.563 or 56.3% (95% confidence interval assuming *srs*: 52.4%, 60.0%). Sometimes this is referred to as an "unweighted" estimate. This unweighted estimate ignores the fact that population size in stratum 2 is around three times larger than stratum 1 and two times larger than stratum 3. To derive an unbiased or corrected estimate of the proportion of children immunized in the nation, there is a need to take into account the differences in the population size of each stratum. This is where the statistical weighting is important, which in this situation where the weight is the number of individuals in the population in each stratum.

Table S.3. Example stratified PPS data

Strata	Population distribution		Survey distribution		Survey results		
	N	%	n	%	p_j	Var(p_j)	$deff_j$
1	9,870	17.14	225	34.30	.8133	.0008779	1.301
2	33,599	58.33	219	33.38	.5479	.0015555	1.379
3	14,130	24.53	212	32.32	.3113	.0038950	3.851
Sum	57,599	100.00	656	100.00	-	-	-

The point estimate would be

$$\hat{p}_{.} = \frac{(9870 \times .8133) + (33599 \times .5479) + (14130 \times .3113)}{9870 + 33599 + 1430} = \frac{30834.8321}{57599} = .535 \text{ or } 53.5\%$$

The weighted estimate of the immunization coverage would be 53.5%; note that this estimate is has a lower value (but more valid) than the ***un***weighted estimate of 56.3%. The variance for the weighted estimate is:

$$\hat{\text{var}}(\hat{p}_{.}) = \frac{(9870^2 \times .0008779) + (33599^2 \times .0015555) + (14130^2 \times .0038950)}{(9870 + 33599 + 1430)^2} = \frac{2619178.674}{3317644801} = .0007895$$

The 95% confidence interval for the weighted estimate would be calculated as follows. Note that the t-value for a 95% two-sided confidence interval, in this example, has 90-3 degrees of freedom, a t-value of 1.9876.

$$.5353 \pm 1.9876\sqrt{.0007895}$$

$$.5353 \pm .0558$$

The weighted point estimate and 95% confidence interval would be 53.5% (48.0%, 59.1%). A comparison of the naïve estimate with 95% confidence interval assuming *srs* and the weighted estimate can be seen in Figure S.3. The naïve approach in this example has a biased estimate and too narrow a confidence interval compared to the more valid weighted stratified *pps* estimate and its confidence interval which takes into account the stratified PPS survey design. Table S.4 presents the confidence interval width for each stratum and the national estimate as well as the design effects. In this example, the weighted national estimate is more precise than the stratum-specific estimates. Figure S.4 presents a bar graph with 95% confidence intervals for the national estimate and each stratum.

Figure S.3. Comparison of Point Estimates and 95% Confidence Intervals between Simple Random Sampling (SRS) and Stratified Proportionate to Population Size (PPS) Sampling

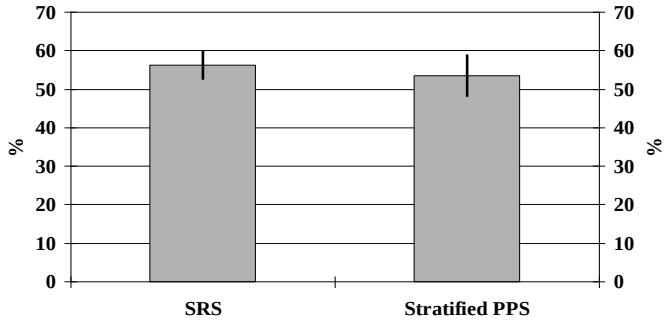
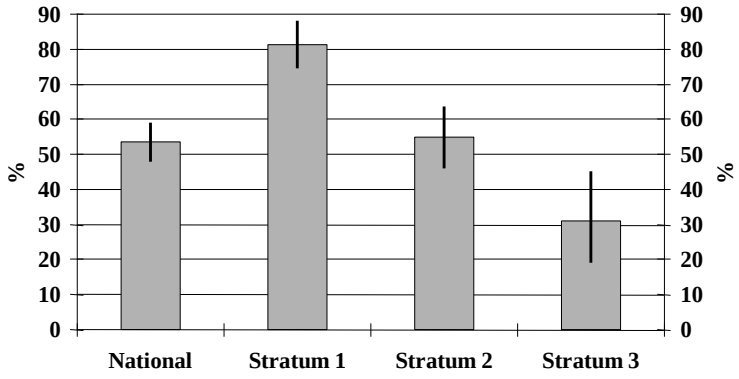


Table S.4. Example stratified PPS data

Strata	Survey results			
	Percent	Var(p_i)	± for CI (%)	DEFF
1	81.3%	.0008779	±6.7%	1.301
2	54.8%	.0015555	±8.9%	1.379
3	31.1%	.0038950	±14.1%	3.851
Weighted National estimate	53.5%	.0007895	±5.6%	2.077

Figure S.4. Comparison of Point Estimates and 95% Confidence Intervals for the national and stratum-specific estimates.



Adding Weights to a Computer File

The above approach for a weighted point estimate and variance work when directly weighting the proportions and variances from summary stratum data as presented. Usually one would be analyzing data

in a statistical computer program. To add a statistical weight variable to a data set requires a slightly different approach for weighted analyses. Two weighting methods are provided. Method 1 is to take the percent of the population in each stratum (*PD*) divided by the percent of those in the survey in each stratum (*SD*) and is sometimes referred to as a *normalized* weight:

Method 1: PD_i / SD_i

For example, in stratum 1 in Table S.4, the weight would be 17.14%/34.30%=0.4997. Similar calculations would be applied to strata 2 and 3. Another approach, Method 2, to develop a weight is to divide the number of individuals in the population by the number surveyed at each stratum level:

Method 2: N_i / n_i

For example, in stratum 1 in Table S.5, the weight would be 9870 / 225 = 43.87. This could be thought of as for every child in the survey in stratum 1, they represented 43.87 children. Similar calculations would be applied to strata 2 and 3.

Table S.5. Example stratified PPS data – weights for computer program

Strata	Population distribution		Survey distribution		Weight	
	N	% (*PD*)	*n*	% (*SD*)	Method *1* (w_i)	Method 2 (w_i)
1	9,870	17.14	225	34.30	.4997	43.87
2	33,599	58.33	219	33.38	1.7475	153.42
3	14,130	24.53	212	32.32	.7590	66.65
Sum	57,599	100.00	656	100.00	-	-

Once the weights are calculated, they need to be accounted for in the analysis, with one approach to use IF statements. For example, assuming in the data file the name for the stratification variable is **strata** and the name for the weight variable is **popwt**, in SAS:

```
IF strata = 1 THEN popwt = 0.4997;
IF strata = 2 THEN popwt = 1.7475;
IF strata = 3 THEN popwt = 0.7590;
```

The analyses should take into account the stratification by using the appropriate software and commands. In SAS one could use PROC SURVEYFREQ or the other survey procedures with the weight option; in Epi Info (Windows version) use COMPLEX SAMPLE FREQ or other complex survey commands, again using the weight option; or use of SUDAAN or the complex sample module for SPSS (when using SPSS use the Method 2 weighting scheme). Note that the only advantage to using Method 1 of weighting is that, when not accounting for the complex survey design but accounting for the weights in the analyses, Method 1 does not inflate the sample size.

Sample size for Stratified PPS surveys

To calculate the sample size for stratified PPS surveys, generally one would determine the desired level of precision for each stratum using the sample size formula described in the section on PPS surveys. Note that the nationally stratified PPS estimate will generally be more precise than the estimates for each stratum as shown in Table S3.

ADDITIONAL DISCUSSION ON THE NUMBER OF CLUSTERS AND NUMBER OF INDIVIDUALS TO SAMPLE PER CLUSTER

The number of clusters to select in a cluster survey

In general, it has been found that collecting information on around 30 clusters will provide good estimates of the true population with an acceptable level of precision (Binkin et al., 1992) when: 1) the

percentage with the outcome is between 10% to 90%; 2) the desired level of precision is around 5%; and 3) the DEFF is around 2. For a fixed number of individuals selected per cluster (e.g., 10 individuals per cluster or 30 individuals per cluster), collecting information on more than 30 clusters can improve precision, however, beyond around 60 clusters the improvement in precision is minimal. Some surveys, such as the UNICEF Multiple Indicator Cluster Survey (MICS) (UNICEF, 2000) and Demographic Health Surveys (DHS), many more clusters are recommended, up to 300 or more. Some of the reasons for a large number of clusters include:

- The survey is almost always stratified to provide region- or province-specific estimates.
- Some of the indicators occur infrequently with some clusters having few if any eligible individuals. Things that occur infrequently (in some populations) include: the number of children within any one-year age interval (such as for immunizations or anthropometry); the number of children 0-4 months by breastfeeding status; and the number of women who have given birth in the previous year. In some populations there may only be one or two eligible individuals within each cluster.
- Some of the factors studied have very large design effects (*deff*), such as factors relating to access to potable water and adequate sanitation.
- The factors under study are presented by sex, by urban/rural status, age groups, and other factors, therefore requiring larger sample sizes to assure precise estimates for subgroups analyses.

For relatively frequent events, such as the prevalence of stunting or anemia, around 30 clusters should be sufficient for a geographic area. For rare events and events with a large design effect, selection of more than 30 clusters may be necessary. If the investigators are willing to accept less precision, fewer than 30 clusters could be sampled.

The number of individuals to sample in a 30-cluster survey

Based on the analysis of many 30-cluster surveys, it is recommended that the minimum number of samples to be collected in each cluster is 10 and the maximum 40. Collecting information on fewer than 10 samples per cluster can lead to unstable variance estimates. Collecting information on more than 40 per cluster results in little improvement in precision. An example of a survey where the number of individuals sampled per cluster was varied from 6 to 48 is shown in Figure S.5. In this figure, on average 48 children were selected in each cluster. From the data file, every other child was selected and the analysis performed again on 24 children per cluster. This procedure was repeated selecting every third child, every fourth child, etc. Note that the point estimates and confidence interval width vary little from around 16 sampled per cluster and above. Below 10 sampled per cluster the point estimates become unstable and the confidence intervals tend to get wider. If the collection of information is costly, such as collecting blood specimens, then the fewest samples per cluster with adequate precision should be collected. If the cost of collecting the information is minimal, such as palpating children for goiter, then doing more than 40 would be acceptable.

A reason to increase the number of samples per cluster would be to compare two or more subgroups. For example, say the investigator wants to determine if the prevalence of anemia in females is different than the prevalence in males. If this comparison is important, then the sample size information mentioned in the previous paragraph might need to be increased to assure adequate precision for each subgroup.

While it would seem that collecting more samples per cluster would lead to improved precision in PPS surveys, the improvement in precision in minimal beyond 40 samples. The reason for this is that as more samples are collected per cluster, the DEFF increases (see Figure S.6). In this figure, as more individuals were sampled per cluster, the distance between the variance calculated assuming SRS and variance calculated assuming PPS gets wider. This results in the DEFF becoming larger as the number sampled per cluster increases. Also note that as the number sampled per cluster increases, there is little reduction in the variance, assuming PPS, with the larger numbers sampled per cluster. Also note the instability of the variance estimates assuming PPS when the sample size is less than 10. This instability in the variance estimate assuming PPS results in instability in the DEFF estimate. Figure S.7 presents an

example of the effect on precision as the number sampled per cluster increases for various prevalence levels.

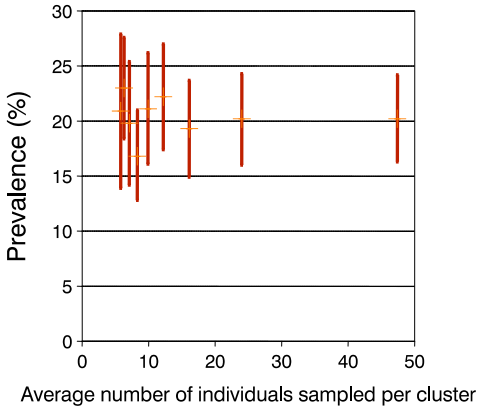


FIGURE S.5 Prevalence of high thyroid stimulating hormone (TSH >5mU/L whole blood) and 95% confidence intervals by average number of individuals sampled per cluster. Based on a 30 cluster probability proportional to size (PPS) survey

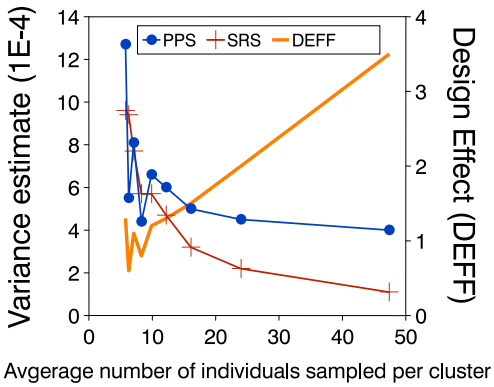


FIGURE S.6 Comparison of variance estimates (PPS SRS) and design effect (DEFF) for high levels of thyroid stimulating hormone (TSH >5mU/L whole blood) by average number of individuals sampled per cluster. Based on a 30 cluster probability proportional to size (PPS) survey; SRS=simple random sampling

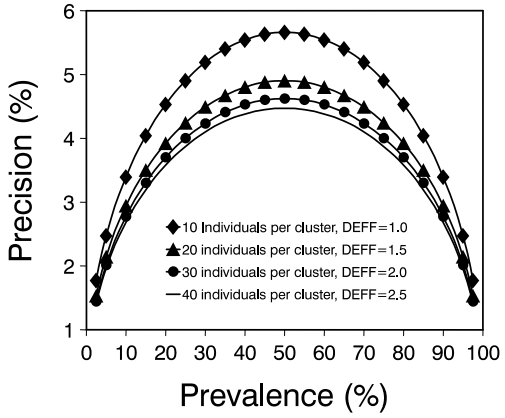


*FIGURE S.7 Precision for various levels of prevalence number of individuals sampled per cluster assuming a 30 cluster probability proportional to size (PPS) survey. Precision is defined as the 1.96*standard error*

Some common misperceptions and errors in sampling

There are a number of misperceptions in sampling populations. One misperception is that the larger the target population, the sample size should be larger. While this may be true with small populations where use of the *fpc* can be used to reduce the sample size for small populations, for large populations, whether the population is 100,000 or 100,000,000, the sample size for a survey would be the same. In some situations where there is a large population, there may be a decision to perform stratified *pps* surveys, which would increase the overall sample size.

Another misperception is that rather than sampling 30 children in each of 30 clusters, why not sample 60 children in 15 clusters? This would result in the same overall sample size and be less costly because fewer clusters would need to be sampled. This is an issue of precision in that the 30x30 design would be more precise than a 15x60 design. The loss in precision would be that the latter would have a larger design effect and would use a *t*-value with 14 degrees of freedom (95% CI two-sided *t*-value with 14 df = 2.1148) rather than 29 degrees of freedom 95% CI two-sided *t*-value with 29 df = 2.0452). For example, if a 30x30 cluster had a DEFF=2, then the ICC would be 0.0345. Assuming a prevalence of 50% and 900 assessed, the point estimate and 95% confidence interval would be:

50% (95% CI: 45.2, 54.8); $\pm$4.8%; 30x30 design with a DEFF=2

A 15x60 cluster survey would be expected to have a DEFF = 3.04. Assuming a prevalence of 50% and 900 assessed, the point estimate and 95% confidence interval would be:

50% (95% CI: 43.8, 56.2); $\pm$6.2%; 15x60 design with a DEFF=3.04

So the trade off in doing fewer clusters but more individuals per cluster is a loss of precision.

Another common error is that if a 30-cluster survey were performed in an area, that several clusters in a subarea could be grouped together as an estimate for that area. For example, if a national survey was performed and five clusters were located in a specific area of the country, the results of these five clusters could be summed together as a sub-national estimate for that area. This problem is similar to the one discussed in the previous paragraph, where when substantially fewer than 30 clusters are used to represent a geographic area, there is a chance that the estimate could be quite different from the truth, which is an issue of precision. The confidence interval for the five clusters would be relatively wide. If a point estimate with a desired level of precision is needed for a specific area, then a stratified *pps* survey should be performed.

In some situations, a *pps* sampling approach does not work well. For example, say there are three refugee camps and there is a desire to estimate the proportion of children less than 5 years of age who are malnourished. Should the investigator divide 30 clusters among the three camps? Or should the investigator perform a 30-cluster survey in each camp? While there are many possible ways to approach this problem, here are two suggested approaches. If only one estimate is needed for all three camps (i.e., camp-specific estimates are not a priority), then one approach would be to calculate a sample size assuming simple random sampling, and then divide the sample proportionally among the camps. For example, assume it is estimated that the prevalence of malnutrition is 50% and the *d* value is .05. The sample size (ignoring the *fpc*) would be $[1.96^2 (.5)(.5)]/.05^2=385$. Next, based on population size estimates in each camp, determine the proportion of refugees in each camp. For example, if exactly one-third of the refugees were in each camp, then one would sample (.333)(385)=128.2 or 129 children in each camp. Ideally, within each camp, the children to be surveyed would be randomly selected. This would result in a total sample size of 3 x 129 = 387.

In the refugee situation with three camps, if an estimate is needed for *each* camp, then a stratified simple random sampling approach could be used. Assuming the *p* and *d* values in the previous paragraph and a large target population size, one would sample 385 children in each camp and calculate camp-specific point estimates with confidence intervals. Then, for an overall estimate, a stratified *srs* approach similar to the stratified *pps* method described earlier would be used for a weighted point estimate and confidence interval.

Frequently individuals will state that one cannot use information from a single cluster - only the combined 30-cluster information can be analyzed and presented. While this is correct in terms of presenting overall estimates, the results of individual clusters can be useful in identifying problem areas. For example, if an EPI survey on immunizations is performed, if 29 clusters had immunization coverage levels of 90% or better and one cluster had a coverage of 14%, the cluster with a low coverage should be investigated further to determine the cause of the low coverage. Is the problem only in the PSU selected

or is the coverage also low in the surrounding communities? Why is the coverage low? Is it due to an inadequate supply of vaccines? Was there an error by the survey team or in the data analysis? Therefore, individual cluster information can be used to investigate potentially problematic areas.

Summary

This chapter presents the formulae and examples on how to analyze data from simple random sampling (*srs*), one-stage cluster survey (1sc), probability proportional to size (*pps*) sampling, and stratified cluster surveys. Sample size formulae were presented and a number of issues discussed in the application of *pps* surveys in populations.

Notation

srs = simple random sampling

pps = proportional to population size

$deff$ = design effect

$\hat{p}_{srs}$ = estmated proportion assuming simple random sampling

a = the number of individual s with the attribute of interest

n = the number of individual s sampled

$\hat{\text{var}}(\hat{p}_{srs})$ = variance assuming simple random sampling

$\hat{q}_{srs} = 1 - \hat{p}_{srs}$

N = population size

N_i = population size in stratum i

n_i = number in sample in stratum i

d = desired absolute precision

$\hat{p}_{pps}$ = estimated proportion assuming pps sampling

$\hat{p}_i$ = proportion estimate in the ith cluster

m = the number of clusters

$\hat{p}_.$ = the combined (national) estimate

s = the number of strata

w_j = a weighting factor for the jth stratum

$\hat{p}_j$ = the proportion with the factor of interest in the jth stratum

ICC = intra class correlation coefficient

Exercises

1. Which study design *usually* has a larger variance, and therefore a wider confidence interval?
   A. *srs*
   B. *pps*

2. What is the formula for the *deff*?
   A. Estimates of var(p_{srs})/ var(p_{pps})
   B. Estimates of var(p_{pps})/var(p_{srs})
   C. Estimates of p_{srs}/p_{pps}
   D. Estimates of p_{pps}/p_{srs}

3. The inclusion of the *fpc* into the *srs* sample size formula may have the following effect:
   A. Can reduce the sample size
   B. Can increase the sample size
   C. Has no effect on the sample size

4. A 30-cluster survey on the prevalence of goiter in school children was performed. Results from 8 of the clusters are shown in Table S.4 (only 8 clusters are presented to simplify hand calculations). Calculate the following from the data in Table S.4:

a. Prevalence of goiter, variance, and 95% confidence limits assuming *Simple Random Sampling*:

   Prevalence (%) = Variance =

   95% confidence interval (%) =

b. Same as question 4.a. except perform the calculations assuming *Proportional to Population Size* sampling:

   Prevalence (%) = Variance =

   95% confidence interval (%) =

c. Calculate the design effect:

   DEFF =

Table S.4. Results from 8 clusters on the prevalence of goiter in school children.

Cluster	Goiter	Total	PERCENT
1	27	40	67.5
2	37	40	92.5
3	34	40	85.0
4	36	40	90.0
5	34	40	85.0
6	40	40	100.0
7	37	40	92.5
8	34	40	85.0
Total	279	320	

5. The results of a stratified PPS survey are presented in Table S.5. Fill in the blank cells in the table. Then, calculate the following:

a. Weighted prevalence (%)

b. 95% confidence interval around the weighted prevalence (%)

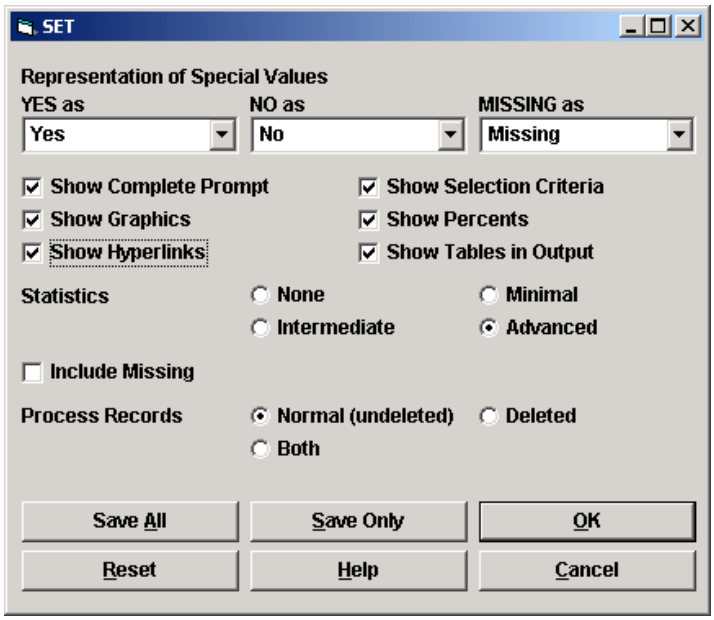
Table S.5. Stratified PPS data

Strata	Population distribution		Sample distribution		Survey results		
	N	%	n	%	p_i	Var(p_i)	$deff_i$
1	45,993		1,200		.836	.0006017	5.3
2	21,023		1,200		.571	.0007975	3.9
Sum		100.00		100.00	-	-	-

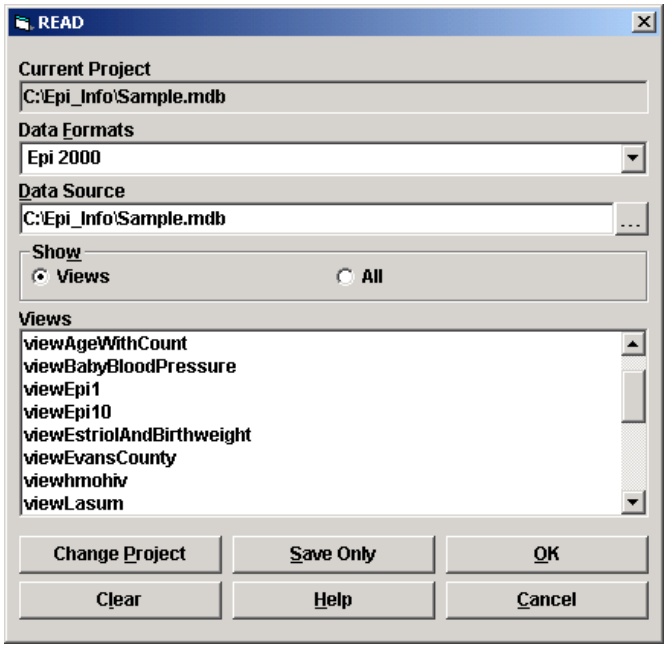
6. Use Epi Info to calculate the immunization coverage with 95% confidence interval and the DEFF. You can use either the DOS version or Windows version of Epi Info, instructions for both are below:

For the ***DOS*** version of Epi Info, use the Csample program in Epi Info (DOS version) to calculate the immunization coverage with 95% confidence interval and DEFF. This data file contains the results of a 30-cluster survey where the immunization status was determined on 7 children in each cluster. Open Epi Info, under "Programs" select "CSAMPLE." In the first screen of CSAMPLE, select the file "EPI1.REC." On the second screen, under "Main" put VAC (the variable for vaccinated; 1= yes and 2= no); under "PSU" put "cluster." Then click on the "Tables" button." Write the vaccine coverage with 95% confidence interval below; also write down the DEFF.

For the ***Windows*** version of Epi Info, from the main screen, click on the Analyze Data button. Before reading data, make sure the SET command (the very last command in the command window) has the Statistics option set to Advanced (see below). This will assure you will be provided with 95% confidence intervals in the output.



On the next screen, click on the Read(Import) command in the left window. It is important that the Data Source and Current Project are as shown on the next page:



Select the file called veiwEpi1 in the Views window. Use the Complex Sample Frequencies command and in the dialog box for the Frequency of select the variable VAC and under PSU select CLUSTER. Write the vaccine coverage with 95% confidence interval below; also write down the DEFF.

Coverage _____95% CI (_____, _____) DEFF _____

7. Using either the DOS or Windows version of Epi Info, answer the following question for a stratified cluster survey.

For the **DOS** version, use the CSAMPLE program to perform the analysis of a stratified cluster survey on immunization levels. There are 10 strata in this survey. To analyze the data correctly, a weighted approach is needed. On the first screen of CSAMPLE, select the file "EPI10.REC." On the second screen, under "Main" put VAC (the variable for vaccinated; 1= yes and 2= no); under "Strata" put "Location"; under "PSU" put "cluster"; under "Weight" put "Popw." Then click on the "Tables" button." Write the vaccine coverage with 95% confidence interval below; also write down the DEFF.

For the ***Windows*** version of Epi Info, using the Sample.MDB again, select the file viewEpi10. Similar to the previous question, use the Complex Sample Frequencies command and in the dialog box for the Frequency of select the variable VAC and under PSU select CLUSTER; also, for Weight select POPW. Write the vaccine coverage with 95% confidence interval below; also write down the DEFF.

Coverage _____95% CI (_____, _____) DEFF _____

8. Table S.6 lists 100 enumeration units in an area of Nepal (in Nepal referred to as "wards"). A 30-cluster survey is to be performed in this area. Perform the following:

a. Calculate the cumulative population in Table S.6
b. Calculate the sampling interval (the total population divided by the number of clusters)
c. Assume the random starting point is 356; select the clusters in Table S.6

Appendix 1 presents the details for selecting clusters using the PPS methodology.

Table S.6. Population size in 100 communities/wards

Ward #	Pop.	Cum.	Cluster #	Ward #	Pop.	Cum.	Cluster #
1	259			51	283		
2	207			52	327		
3	664			53	319		
4	450			54	395		
5	483			55	542		
6	302			56	590		
7	398			57	564		
8	148			58	331		
9	281			59	490		
10	696			60	521		
11	518			61	364		
12	565			62	379		
13	450			63	917		
14	790			64	423		
15	684			65	172		
16	984			66	232		
17	563			67	286		
18	440			68	256		
19	267			69	174		
20	273			70	245		
21	324			71	278		
22	346			72	372		
23	380			73	208		
24	506			74	481		
25	643			75	245		
26	376			76	306		
27	367			77	292		
28	536			78	328		
29	382			79	257		
30	401			80	212		
31	891			81	598		
32	303			82	257		
33	1149			83	297		
34	482			84	267		
35	454			85	262		
36	1251			86	340		
37	324			87	344		
38	554			88	370		
39	511			89	380		
40	463			90	247		
41	435			91	403		
42	841			92	224		
43	943			93	163		
44	1186			94	262		
45	923			95	143		
46	448			96	233		
47	475			97	543		
48	292			98	298		
49	189			99	539		
50	353			100	329		

References

Binkin N, Sullivan K, Staehling N, Nieburg P. Rapid nutrition surveys: how many clusters are enough? *Disasters*, 16(2):97-103, 1992.

Lemeshow S, Stroh G Jr. Sampling Techniques for Evaluating Health Parameters in Developing Countries. National Academy Press, Washington D.C. 1988.

Schaeffer RL, Mendenhall W, Ott L. Elementary Survey Sampling, Fourth Edition. Duxbury Press, Belmont, California 1990.

Sullivan K. The effect of sample size on validity and precision in probability proportionate to size (PPS) cluster surveys (abstract). 28th Annual Meeting of the Society for Epidemiologic Research, Snowbird, Utah, June 21-24, 1995; *American Journal of Epidemiology* 141(11), S47, 1995.

UNICEF. End-Decade Multiple Indicator Survey Manual: Monitoring Progress Toward the Goals of the 1990 World Summit for Children. Division of Evaluation, Policy and Planning Programme Division, UNICEF, New York, 2000.

Appendix 1

PPS selection of clusters

As mentioned, clusters should ideally be selected using a technique called "probability proportionate to size" or PPS sampling. Using the PPS method, the likelihood of a PSU being selected is proportional to its population size, i.e., larger PSUs are more likely to be selected than smaller ones. The first step is to obtain the "best available" census data for all the PSUs in the geographic area to be surveyed (e.g., a country). This information is usually available from the government agency that performs the census for the country, such as a national bureau of statistics.

Countries with very organized census information will frequently have PSUs or enumeration units that are relatively small geographic areas with a population size between 100-1,000 or 20-200 households. It may be necessary to designate a minimum PSU population size to assure that enough potential respondents are available to meet the sample size per cluster; consequently, there may be situations where two or more contiguous enumeration units will need to be combined to form a single PSU.

If the enumeration unit information is either not readily accessible or is very inaccurate, the most recent estimates of population size should be obtained by village, towns, and cities that would serve as the PSUs. It is essential to include all areas, including those which may be remote and/or rural.

With the PSU information, make a list with four columns (see Table 3.4). The first column lists the name of each PSU; the second column contains the population of each PSU; the third column contains the cumulative population that is obtained by adding the population of each PSU to the cumulative population of PSUs preceding it on the list. As a general rule, it is best for the list to be in geographic order by districts or provinces. A sampling interval (*k*) is obtained by dividing the total population size by the number of clusters to be surveyed. A random number between 1 and the sampling interval (*k*) is chosen (see Appendix 6 for a table of random numbers) as the starting point and the sampling interval is added cumulatively until thirty clusters are chosen; the selected clusters are shown in the 4th column of Table 3.4.

3.7.1.i. Example of Selecting PSUs for a Cluster Survey

In the fictitious area of El Saba, there are fifty PSUs (Table 3.4). In practice there are usually many more than fifty PSUs in a survey area. With a large number of PSUs, the selection process is usually performed using a computer. For SAS users, there is PROC SURVEYSELECT which has an option to select data using PPS. With SPSS, the optional Complex Samples module has a “Select Sample…” option. Use of spreadsheets is another method for performing the selection.

Table 3.4. Selecting communities for a cluster survey in El Saba using the PPS method

PSU	Pop.	Cum.	Cluster	PSU	Pop.	Cum.	Cluster
Utural	600	600		BanVinai	400	10,880	13
Mina	700	1,300	1	Puratna	220	11,100	
Bolama	350	1,650	2	Kegalni	140	11,240	
Taluma	680	2,380	3	Hamali-Ura	80	11,320	
War-Yali	430	2,810		Kameni	410	11,730	14
Galey	220	3,030		Kiroya	280	12,010	
Tarum	40	3,070		Yanwela	330	12,340	
Hamtato	150	3,220	4	Bagvi	440	12,780	15
Nayjaff	90	3,320		Atota	320	13,100	
Nuviya	300	3,610		Kogouva	120	13,220	16
Cattical	430	4,040	5	Ahekpa	60	13,280	
Paralai	150	4,190		Yondot	320	13,600	
Egala-Kuru	380	4,570		Nozop	1,780	15,380	17,18
Uwanarpol	310	4,880	6	Mapazko	390	15,770	19
Hilandia	2,000	6,880	7,8	Lotohah	1,500	17,270	20
Assosa	750	7,630	9	Voattigan	960	18,230	21,22
Dimma	250	7,880		Plitok	420	18,650	
Aisha	420	8,300	10	Dopoltan	270	18,900	
Nam Yao	180	8,480		Cococopa	3,500	22,400	23,24,25,26,27
Mai Jarim	300	8,780		Famegzi	400	22,820	
Pua	100	8,880		Jigpelay	210	22,840	
Gambela	710	9,590	11	Mewoah	50	22,890	
Fugnido	190	9,880	12	Odigla	350	23,240	28
Degeh Bur	150	10,030		Sanbati	1,440	24,680	29
Mezan	450	10,480		Andidwa	260	24,940	30

Follow the four steps below to select clusters to be included in the survey:

Step 1: Calculate the sampling interval by dividing the total population by the number of clusters to be surveyed. In this example, 24,940 / 30 = 831.
Step 2: Choose a random starting point between 1 and the sampling interval (*k*, in this example, 831) by using the random number table in Appendix 6. For this example, the number 710 is randomly selected.
Step 3: The first cluster will be where the 710th individual is found based on the cumulative population column, in this example, Mina since it includes the population from 601 to 1,300.
Step 4: Continue to assign clusters by adding 831 cumulatively. For example, the second cluster will be in the PSU where the value 1,541 is located (710 + 831 = 1541), which is Bolama. The third cluster is where the value 2,372 is located (1541 + 831 = 2372), and so on. In PSUs with large populations, more than one cluster could be selected. Note that if two clusters were selected in one PSU, when the survey is performed, the survey team would divide the area into two sections of approximately equal population size and treat each area as independent clusters. Similarly, if three or more clusters were in a PSU, the PSU would be divided into three or more sections of approximately equal population size, as is the case with Cococopa in Table 3.4 (described in more detail later).

3.7.2 Random and systematic selection of clusters

When a list of PSUs is available but the population size for each PSU is not known or very inaccurate, simple random sampling or systematic selection can be used. Systematic sampling tends to be easier to implement by hand and is described next, although simple random sampling (see Appendix 4) could also be performed. With the availability of computer programs that can sample records from a file, the preference would be to use simple random sampling. The steps for systematic sampling, should it be more convenient to implement, are as follows:

Step 1: Obtain the list of the PSUs and number them from 1 to N (the total number of PSUs)
Step 2: The number of PSUs to sample (*n*) should have already been determined.
Step 3: Calculate the "sampling interval" (*k*) by N/n (always round down to the nearest whole integer).
Step 4: Using the random number table (Appendix 6), select a number between 1 and *k*. Whichever number is randomly selected, go to the PSU list and include that PSU in the survey.

Step 5: Select every *k*th PSU after the first selected PSU.

For illustrative purposes, Table 3.5 lists fifty PSUs and below demonstrates how to select 8 PSUs.
Step 1: There are fifty PSUs, therefore N=50.
Step 2: The number of PSUs to sample is eight, therefore n=8.
Step 3: The sampling interval is 50/8 = 6.25; round down to the nearest whole integer which is 6; therefore, *k*=6.
Step 4: Using a random number table, select a number from 1 to (and including) 6. In this example, let's say the number selected was 3. Therefore, the first PSU to be selected is the third PSU on the list, which in this example is Bolama.
Step 5: Select every 6th PSU thereafter. In this example, the selected PSUs would be the 3rd, 9th, 15th, 21st, 27th, 33rd, 39th, and 45th PSUs on the list.

In some circumstances you might actually end up selecting more than the number of clusters needed. In the above example, had the random number chosen in Step 4 been 1 or 2, nine PSUs would have been selected rather than eight. To remove one cluster so that only eight are selected, again go to the random number table, and pick a number and the cluster that corresponds to the random number is removed from the survey. To properly analyze the data collected using systematic sampling, an estimate of the population size in each cluster should be collected when the survey team arrives on site. (Note that usually more clusters are selected; the 8 selected in this example was for illustrative purposes only).

Table 3.5 Selection of PSUs using the systematic selection method

PSU		Selected?	PSU		Selected?
1	Utural		26	BanVinai	
2	Mina		27	Puratna	Y
3	Bolama	Y	28	Kegalni	
4	Taluma		29	Hamali-Ura	
5	War-Yali		30	Kameni	
6	Galey		31	Kiroya	
7	Tarum		32	Yanwela	
8	Hamtato		33	Bagvi	Y
9	Nayjaff	Y	34	Atota	
10	Nuviya		35	Kogouva	
11	Cattical		36	Ahekpa	
12	Paralai		37	Yondot	
13	Egala-Kuru		38	Nozop	
14	Uwanarpol		39	Mapazko	Y
15	Hilandia	Y	40	Lotohah	
16	Assosa		41	Voattigan	
17	Dimma		42	Plitok	
18	Aisha		43	Dopoltan	
19	Nam Yao		44	Cococopa	
20	Mai Jarim		45	Famegzi	Y
21	Pua	Y	46	Jigpelay	
22	Gambela		47	Mewoah	
23	Fugnido		48	Odigla	
24	Degeh Bur		49	Sanbati	
25	Mezan		50	Andidwa	