# Simple Analytic Procedures for Rapid Microcomputer-Assisted Cluster Surveys in Developing Countries

RALPH R. FRERICHS, DVM, DrPH

Dr. Frerichs is Chairman and Professor of Epidemiology at the School of Public Health, University of California at Los Angeles (UCLA). This work was partially supported by the Western Consortium for Public Health, Berkeley, CA, and funded by the U.S. Agency for International Development (USAID) Project 482-004, Rangoon, Burma. It was also partially funded under Contract No. DPE-5920-A-00-5056-00 between the Center for Human Studies, Chevy Chase, MD, and USAID.

Tearsheet requests to Dr. Ralph R. Frerichs, Professor of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90024-1772.

**Synopsis**

*Surveys are often deemed necessary in developing countries when routine sources of data are not considered adequate to answer important policy-related questions. Although field work often goes smoothly, many surveys become bogged down in the analysis stage. With the availability of microcomputers and contemporary software, investigators in developing countries can use rapid survey methodology (RSM) to process, analyze, and report survey findings more quickly than ever before.*

*Presented in this paper are three simple analytic procedures for planning and doing two-stage, rapid cluster surveys. All were successfully used in three rapid surveys in rural regions of Burma and Thailand. By use of a spreadsheet and graphics software package, the three procedures (a) derive the first-stage selection of 30 cluster sites with probability proportionate to size, (b) calculate variance estimates and confidence limits for the parameters of interest and graphically present the findings as 90, 95, and 99 percent confidence intervals, and (c) estimate the necessary sample size for planning two-stage, rapid cluster surveys. The procedures can be used both in the field and in teaching workshops or courses on survey methods. Examples are given from three rapid surveys conducted in Hlegu Township, Burma, and Sisaket Province, Thailand. In both countries, local health professionals were first taught the methods in a 1-week workshop before they used the procedure for conducting the rapid computer-assisted surveys.*

LACK OF POPULATION-based information has traditionally been one of the key drawbacks to formulating timely, responsive health policies in much of the developing world. In the usual situation, the administrator or policymaker requests data from an information or evaluation unit which in turn presents either an analysis of existing data or conducts a field survey. Often the administrator's questions are not complex. They may focus on the prevalence of diseases such as acute respiratory infection, diarrhea, or malnutrition, the level of immunization coverage, the use of condoms, or the availability of community health workers. When a survey is deemed necessary, most likely the survey workers are effective at data collection but then find themselves hopelessly bogged down in the complexities of data analysis. By the time the information is made available, 1 or 2 years may have gone by, and the administrator is focusing on new problems and no longer is interested in answers to the original questions. Furthermore, if statistical analyses are done, the findings are usually presented as point estimates with standard errors and levels of statistical significance (that is, *P*-values), measures of little relevance to statistically less sophisticated administrators.

Rapid survey methodology (RSM) has been developed to provide administrators with quick information on problems facing persons at the community level. As described in the accompanying article (*1*), RSM is a procedure that uses data forms with clearly defined questions, portable computers and printers, contemporary software, and a two-stage cluster sampling method favored by the World Health Organization (WHO), to determine in a few short weeks measures of health status of the population. Findings are presented in simple graphs and tables that can be generated in the field using portable, battery-powered printers. For each variable of interest, a set of three confidence intervals is graphically displayed. By showing the administrator a graph with 90, 95, and 99 percent confidence intervals, the surveyors are able to emphasize that (*a*) survey findings are from a sample rather than the total count of the population, (*b*) estimates based on samples can only have certain levels of precision, and (*c*) given the sample size, the level of precision is dependent on how confident the administrator or surveyor wants to be that the interval brackets the true value in the sampled population. From my personal experience in the field and a recent review of studies reported in the Weekly Epidemiological Record of WHO, I have observed that confidence intervals are rarely included in material presented to administrators or policymakers. Furthermore, because the formula for variance estimates from cluster samples may seem too complicated, correct analytic procedures may be ignored in favor of a less complex analysis which treats data as if they were drawn from a simple random sample of the population. In following this procedure, the derived variance estimates are usually too small: they give the impression that the survey findings are more precise than they really are. RSM tries to address this problem by having survey workers in less developed countries quickly calculate proper variance estimates for two-stage cluster samples and present their findings as confidence intervals.

RSM was used for the first time in Burma, in May 1987, to conduct a health survey of all births during the prior 3 years. This survey is described in an accompanying article (*1*); the major findings were published elsewhere (*2,3*). More recently, two additional rapid surveys were done during December 1987, in Sisaket Province, Thailand. The first focused on antenatal care among women who had a pregnancy outcome during the prior 24 months. The second dealt with family planning practices of married women, ages 15 through 44 years.

In this paper I will present three analytic procedures used in doing rapid surveys, including one which is useful for planning future rapid surveys. The three procedures are (*a*) the selection with probability proportionate to size (PPS) of 30 clusters, (*b*) the rapid derivation of a variance estimate and three levels of confidence intervals, and (*c*) a procedure for estimating the necessary sample size for rapid surveys. All three procedures can be done in the field using a spreadsheet and graphics software package and a battery-powered, portable microcomputer.

## Description of Analytic Procedures

**Two-stage cluster sampling.** The two-stage cluster sampling procedure included in RSM has been used throughout the world in immunization surveys, both to assess the current level of vaccination coverage and to verify coverage estimates provided by routine reporting systems (*4*). Typically, 30 clusters are selected in a region of interest with probability proportionate to the size of the resident population. Within each of the 30 clusters, an initial household is randomly selected. Seven persons are selected per cluster, starting with an eligible person in the initial household and proceeding to the nearest neighboring households. Thus, the total sample comprises 30 × 7 or 210 eligible people. This form of survey has been characterized as a two-stage PPS cluster sample without random selection at the second stage (*5*). As advocated by WHO, the parameter estimated by this method of sampling is to have a 95 percent chance of being within 10 percentage points of the true value in the sampled population (*6*). Computer simulation studies of Lemeshow and Robinson and of Henderson and Sundaresan have provided recent statistical verification of the method, at least for the precision levels set by WHO (*5,7*). The same cluster sampling method, with modifications in cluster size and number of clusters, has also been used to assess the occurrence of uncommon diseases such as poliomyelitis and tetanus (*8*), the occurrence and treatment patterns of diarrhea (*9*), and the use of health services (*10*).

**Theory for sample selection at first stage.** The two-stage cluster surveys use equal probability of selection method sampling. The population is initially divided into a series of clusters. At the first stage, 30 (or possibly more) clusters are selected with probability proportionate to the number of persons in the clusters. Thus, large clusters are more likely to be selected than small clusters. Since at the second stage the same number of persons are selected from each cluster, the fraction of persons selected in large clusters will be less than the selected fraction in small clusters. PPS sampling at the first stage, coupled with a constant number per cluster at the second stage, results in a self-weighted sample in which all persons in the population have the same probability of being selected.

For the PPS sample, the population is divided into geographically defined clusters of known size. The clusters are listed, with the population size included in a cumulative tally. After a random start, a systematic sample of 30 clusters is drawn from the cumulative population list. Since larger clusters contribute more to the cumulative population list than smaller clusters, the probability of being included in the sample is proportionate to the

*'The two-stage cluster random sampling procedure included in RSM has been used throughout the world in immunization surveys, both to assess the current level of vaccination coverage and to verify coverage estimates provided by routine reporting systems.'*

size of the cluster. The exact procedure for PPS sampling with examples is shown in most sampling textbooks (*11–15*).

**Theory for variance analysis.** The variance analysis procedure of RSM assumes that we are estimating a proportion with some attribute in the population of interest. Examples are parameters such as the prevalence of disease, the proportion who have been immunized, or the proportion who have received antenatal care. This proportion can be viewed as a typical ratio estimate with the numerator being the number with the attribute of interest and the denominator being the total number of observed persons. Using the terminology of Cochran (*11*), the approximate variance for the proportion derived in the cluster sample is

$$v(p) = (1-(n\,\overline{m} \div N)) \times \sum_{i=1}^{n} (a_i - p\,m_i)^2 \div (n\,(n-1)\,\overline{m}^2) \qquad (1.1)$$

where $n$ is the number of clusters, $\overline{m}$ is the average cluster size, $N$ is the size of the total population from which the sample is drawn, $a_i$ is the number of persons with the attribute in cluster $i$, $p$ is the proportion with the attribute in the total sample, and $m_i$ is the number of persons in cluster $i$. Note that $(a_i - p\,m_i)$ is the observed number of persons with the attribute in a given cluster minus the expected number based on the proportion with the attribute in the total sampled population. This number is calculated for each cluster; the sum for all clusters is divided by $(n\,(n-1)\,\overline{m}^{-2})$. The term at the beginning of the formula, $(1-(n\,\overline{m} \div N))$, is the finite population correction which is included in all sample variance estimates. The standard error of the proportion, ($SE(p)$), is the square root of the above variance. This same formula was reported by Rothenberg and colleagues when using the two-stage cluster survey method to estimate disease incidence (*8*).

Approximate confidence limits for the estimated proportions are derived by taking the estimated proportion plus (upper limit) or minus (lower limit) $z$ times $SE(p)$, where $z$ is the standardized normal deviate with values for the 90, 95, and 99 percent confidence limits of 1.64, 1.96 and 2.58 respectively (*11*).

## Determining Sample Size for Rapid Surveys

**Design effect and intraclass correlation coefficient.** Two interrelated concepts are important for determining the sample size for two-stage cluster samples; the design effect (*deff*) and the intraclass correlation coefficient (*roh*). The design effect is derived by dividing the variance of the estimated proportion obtained from the cluster sample (that is, $v(p)$) by the variance if the same data had been analyzed as a simple random sample (*12*). *Deff* was used to establish that 210 children (that is, 30 clusters with 7 children per cluster) are necessary for immunization coverage surveys. The investigators first stated that they wanted coverage estimates that 95 times out of a 100 were within 10 percentage points of the true value in the underlying population. They next determined what the sample size would be if they were able to draw a simple random sample using the standard formula,

$$n = (z^2\,p\,q) \div d^2 \qquad (1.2)$$

where $n$ is the sample size, $z$ is the standardized normal deviate, $p$ is the proportion immunized, $q$ is the proportion not immunized, and $d$ is the precision or one-half the length of the desired confidence interval. Setting values for $z$, $p$, and $d$ of 1.96, 0.5, and 0.1, respectively, the investigators determined that if data were collected in a simple random sample, 96 children would be sufficient to estimate with 95 percent confidence that the proportion immunized in the sampled population lies within an interval 0.2 in length (that is, with $d$ = 0.1 and $p$ = 0.5, the confidence interval is 0.5 ± 0.1 or 0.4 to 0.6). Based on the design effect of prior two-stage cluster samples, they further estimated that twice as many children would have to be sampled to obtain an estimate with the same confidence limits. Thus, the sample was increased from 96 to 210 children (*7*).

The design effect is also related to another measure which is frequently used by statisticians to help assess sample size, the intraclass correlation coefficient. Kish uses the label *roh* for "rate of homogeneity" to identify the intraclass correlation coefficient (*12*). It provides a measurement of the portion of the total variance due to group membership in a cluster. The formula for estimating *roh* follows:

$$roh = (deff - 1) \div (\overline{m} - 1) \qquad (1.3)$$

where $\overline{m}$ and *deff* are as defined previously. The values of *roh* lie between $+1$ and $-1/(\overline{m}-1)$ and can be interpreted as the degree to which persons within clusters resemble one another with respect to the attribute of interest (that is, their degree of homogeneity).

While the design effect provides a rough guide for determining necessary sample size, the process cannot be easily generalized to samples involving either more clusters or more persons per cluster. For example, assume that we intend to measure the proportion of children, ages 12 through 35 months who have received the third dose of DPT (diphtheria, pertussis, and tetanus) vaccine in a rural region of a developing country. Children in some regions are more likely to have received the third dose than others, depending on the quality of the local immunization program staff. Thus, the likelihood of having received the third dose of DPT would be more homogeneous within clusters than between clusters. Also, assume that some other immunization survey of DPT had reported a design effect for the third dose coverage of 4. We might assume that the sample size in our proposed rapid survey should be 4 times as large as it would be if done as a simple random sample. What is not clear, however, is how the increased number of sampled children should be allocated within and between clusters. Should we have four times as many clusters with the same number of children per cluster or the same number of clusters but four times as many children per cluster? The sample size program included as part of RSM and described subsequently helps the investigator to answer this question.

**Determining the sample size.** Four steps are necessary for determining the sample size for rapid surveys. First, the investigator must estimate the size of the "true" proportion to be measured in the sampled population. Second, the desired precision of the estimate must be stated. If the survey was done over and over again, the precision of a parameter estimate refers to the size of deviations for individual surveys from the average value of the parameter derived from all the surveys (*8*). Precision is also defined as one-half the total length of the confidence interval. Third, the investigator needs to identify the acceptable level (that is, 90, 95, or 99 percent) of the confidence interval. By doing these three steps, the investigator has established the maximum size of the variance which can occur in the sample if it is to fulfill the specified criteria. Rearranging equation 1.2 and changing the terms to reflect the number of people sampled in a cluster survey, we see that the square of the precision divided by the square of the standardized normal deviate is equivalent to the variance of the proportion if drawn by a simple random sample. That is . . .

$$d^2 \div z^2 = p\,q \div n\,\overline{m}$$

where $d$, $z$, $p$, and $q$ are as previously defined. The number of persons in the sample is $(n\,\overline{m})$, since $n$ is the number of clusters and $\overline{m}$ is the average number of persons per cluster. If a rapid survey is done using two-stage cluster sampling, then the variance, $v(p)$, defined previously in equation 1.1, can be no larger than $(d^2 \div z^2)$ if the sample is to fulfill the criteria for precision specified at the onset.

Continuing, we rearrange equation 1.3 to derive the design effect, *deff*.

$$deff = roh\,(\overline{m} - 1) + 1 \qquad (1.4)$$

Since we know that the design effect is defined as the variance of the cluster sample divided by the variance if the same number of persons had been drawn as a simple random sample, we can derive the variance of the proposed cluster sample, $v(p)$, by rewriting equation 1.4 as follows . . .

$$v(p) = \frac{pq}{n\,\overline{m}} \times roh\,(\overline{m} - 1) + 1 \qquad (1.5)$$

If our estimate of $v(p)$ computed in equation 1.5 is less than $(d^2 \div z^2)$, the size of the sample should be adequate to provide a confidence interval of acceptable precision.

## Software to Simplify the Process

While the formulas just presented are not overly complex for someone with statistical training, they do create a barrier for many investigators in developing countries who might otherwise benefit from doing rapid community-based surveys. The

Table 1. Example of selection of 30 clusters in the first state of PPS cluster

| | | | |
|---|---|---|---|
| POPULATION PER HOUSEHOLD | 5.25 | NUMBER OF CLUSTERS | = 30 |
| PROP. IN GROUP OF INTEREST | .088 | RANDOM START NUMBER | = 1,945 |
| | | SAMPLING INTERVAL | = 4,134 |

| Community Name (A) | Estimated Population in Community (B) | Cumulative Population (C) | Range of Cumulative Population Count in Community Low (D) | High (E) | Sequence No. of Cluster (F) | Cumulation of Sampling Interval from Random Start (G) | Total No. of Independent Clusters to be Selected in Community (H) | Est. No. of Households in Selected Community (I) | Est. No. of Pop. of interest in Selected Community (J) |
|---|---|---|---|---|---|---|---|---|---|
| Town A | 5,000 | 5,000 | 1 | 5,000 | 1 | 1,945 | 1 | 952 | 440 |
| Village B | 1,250 | 6,250 | 5,001 | 6,250 | 2 | 6,079 | 1 | 238 | 110 |
| Village C | 1,750 | 8,000 | 6,251 | 8,000 | 3 | 10,213 | | | |
| Village D | 1,000 | 9,000 | 8,001 | 9,000 | 4 | 14,347 | | | |
| Town E | 7,500 | 16,500 | 9,001 | 16,500 | 5 | 18,481 | 2 | 1,428 | 660 |
| Village F | 500 | 17,000 | 16,501 | 17,000 | 6 | 22,615 | | | |
| Village G | 2,500 | 19,500 | 17,001 | 19,500 | 7 | 26,749 | 1 | 476 | 220 |
| Village H | 2,000 | 21,500 | 19,501 | 21,500 | 8 | 30,883 | | | |
| Village I | 3,750 | 25,250 | 21,501 | 25,250 | 9 | 35,017 | 1 | 714 | 330 |
| Village J | 2,500 | 27,750 | 25,251 | 27,750 | 10 | 39,151 | 1 | 476 | 220 |
| Village K | 2,750 | 30,500 | 27,751 | 30,500 | 11 | 43,285 | | | |
| Village L | 1,500 | 32,000 | 30,501 | 32,000 | 12 | 47,419 | 1 | 285 | 132 |
| Village M | 1,000 | 33,000 | 32,001 | 33,000 | 13 | 51,553 | | | |
| Town N | 4,500 | 37,500 | 33,001 | 37,500 | 14 | 55,687 | 1 | 857 | 396 |
| Village O | 1,750 | 39,250 | 37,501 | 39,250 | 15 | 59,821 | 1 | 333 | 154 |
| Village P | 2,000 | 41,250 | 39,251 | 41,250 | 16 | 63,955 | | | |
| Town Q | 4,500 | 45,750 | 41,251 | 45,750 | 17 | 68,089 | 1 | 857 | 396 |
| Town R | 4,750 | 50,500 | 45,751 | 50,500 | 18 | 72,223 | 1 | 904 | 418 |
| Village S | 1,000 | 51,500 | 50,501 | 51,500 | 19 | 76,357 | | | |
| Village T | 1,500 | 53,000 | 51,501 | 53,000 | 20 | 80,491 | 1 | 285 | 132 |
| Village U | 1,296 | 54,296 | 53,001 | 54,296 | 21 | 84,625 | | | |
| City V | 8,345 | 62,641 | 54,297 | 62,641 | 22 | 88,759 | 2 | 1,589 | 734 |
| Village W | 1,056 | 63,697 | 62,642 | 63,697 | 23 | 92,893 | | | |
| Village X | 3,789 | 67,486 | 63,698 | 67,486 | 24 | 97,027 | 1 | 721 | 333 |
| City Y | 7,903 | 75,389 | 67,487 | 75,389 | 25 | 101,161 | 2 | 1,505 | 695 |
| City Z | 11,256 | 86,645 | 75,390 | 86,645 | 26 | 105,295 | 3 | 2,144 | 990 |
| Village AA | 158 | 86,803 | 86,646 | 86,803 | 27 | 109,429 | | | |
| Village BB | 2,575 | 89,378 | 86,804 | 89,378 | 28 | 113,563 | 1 | 490 | 226 |
| City CC | 12,678 | 102,056 | 89,379 | 102,056 | 29 | 117,697 | 3 | 2,414 | 1,115 |
| Village DD | 2,365 | 104,421 | 102,057 | 104,421 | 30 | 121,831 | | | |
| Village EE | 965 | 105,386 | 104,422 | 105,386 | | | 1 | 183 | 84 |
| Village FF | 3,672 | 109,058 | 105,387 | 109,058 | | | | | |
| Town GG | 4,593 | 113,651 | 109,059 | 113,651 | | | 2 | 874 | 404 |
| Village HH | 1,768 | 115,419 | 113,652 | 115,419 | | | | | |
| City II | 8,592 | 124,011 | 115,420 | 124,011 | | | 2 | 1,636 | 756 |
| TOTAL | 124,011 | | | | | | 30 | 19,361 | 8,945 |

Instructions for table 1

SuperCalc program entries for column H (independent clusters to be selected in community) in table 1.

```
For "1st Community"    If((G14<=E14),1+INT((E14-G14)/H5),0)
For "2nd Community"    IF(E15-((SUM($H$14.H14)*$H$5)+$H$4)>0,INT((E15-((SUM($H$14.H14)*$H$5)+$H$4))/$H$5)+1,0)
For "all others"       IF(E16-((SUM($H$14.H15)*$H$5)+$H$4)>0,INT((E16-((SUM($H$14.H15)*$H$5)+$H$4))/$H$5)+1,0)
                         .    .    .    .    .    .    .    .    .    .
                         .    .    .    .    .    .    .    .    .    .

                       IF(E52-((SUM($H$14.H51)*$H$5)+$H$4)>0,INT((E52-((SUM($H$14.H51)*$H$5)+$H$4))/$H$5)+1,0)
```

- E16 and H15 are "adjusted" as part of the copying procedure while the other variables with the $ sign are not.
- The "if" statement instructs the computer to write "0" if the cluster is not selected. SuperCalc can then be instructed to delete the "O"s from the output (that is, turn them into blanks).

analysis procedure becomes much easier when the formulas are included in spreadsheet programs and run on microcomputers. In this section, I will first mention the hardware (that is, computers and printers) and software we used to do rapid surveys in Burma and Thailand, second describe the specific software programs, and third give examples of the application. The format and output of the spreadsheet programs are included in the tables and graph. Most of the spreadsheet formula entries are self-explanatory and can be easily programmed by spreadsheet users. The exceptions are explained in notes accompanying the tables.

**Hardware and software requirements.** The procedure for doing the rapid analysis requires a microcomputer and SuperCalc (*A*), Lotus 1-2-3 (*B*) or other similar spreadsheet software. Although a portable, battery-powered computer is not essential, this type of computer works very well in the harsh physical and electrical environments often found in developing countries. Using a battery-powered computer, all the calculations to be presented in this paper can be done in the field by a rapid survey team.

In Burma, we used both the Toshiba T1100 (*C*) and Hewlett-Packard Portable Plus (*D*) laptop computers. "Laptop" is the computer industry designation for battery-powered computers weighing 6–14 pounds which can fit on the person's lap. For printers, we used the Diconix 150 (*E*) and Hewlett-Packard ink-jet models; both of them are battery-powered. In Thailand, we used the Toshiba T1000, T1100+ and T1200 laptop computers and the Diconix 150 printer. For software, we used Lotus 1-2-3 with the Hewlett-Packard computers in Burma and SuperCalc with the Toshiba computers in both Burma and Thailand. The software programs for the three tables and one graph presented in this paper were developed using SuperCalc and then converted to Lotus 1-2-3 using the SuperCalc conversion program.

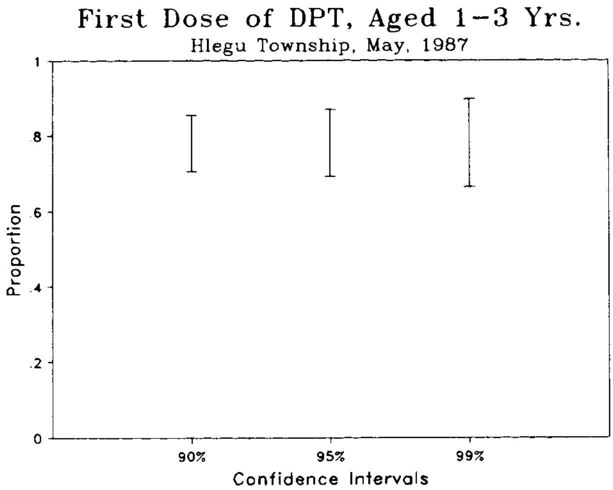
**Software for sampling of clusters.** Once identified, the population to be sampled needs to be divided into geographically defined "clusters." These clusters may be all persons in a given city, town, village, or sections of an aerial map. Next, some estimate must be made of the population in each cluster. Such estimates can often be provided by the national census bureau or by local governmental officials. Table 1 shows the spreadsheet for drawing the PPS sample of 30 clusters. The study investigator must enter four items or sets of items into the table: (*a*) community name, (*b*) estimated community population, (*c*) population per household, and (*d*) proportion of the population in the group of interest. The name of the community and the most recent population estimates are entered into columns A and B of table 1. Using estimates from other surveys or from a recent local census, the population per household and the proportion of the population in the group of interest are entered at the top of the table. These two values do not have to be exact, since they are only used to provide the survey team with a crude guide of the number of eligible persons to be found in the selected communities.

The program derives the cumulative population count in column C, and the low and high range of that count for each community in columns D and E. The NUMBER OF CLUSTERS has been set at 30. The SAMPLING INTERVAL is determined by dividing the total population by 30, the number of clusters. Using a random number function, the RANDOM START NUMBER is derived by multiplying a random number between 0 and 1 times the SAMPLING INTERVAL. The sequence number of the 30 clusters to be sampled is shown in column F. Adding multiples of the sampling interval (4,134) to the random start number (1,945), a cumulative list is made in column G for the selection of the 30 clusters. The numbers in column G are linked with the range of the cumulative population count in columns D and E to determine if each community is to be included in the sample. For example, the third number in column G is 10,213, a value that lies between 9,001 and 16,500 for Town E. Thus Town E is one of the selected clusters. The number of selected clusters is shown in column H (see notes with table 1 for the cell formulas).

A large cluster may be selected more than once. If so, the interview team travels twice to this cluster community and independently selects the second set of households to be visited (*6*). Once all clusters are selected, the program lists the estimated number of households and estimated number of eligible people in the specified communities. This provides the survey team organizer with information for planning the second stage of the survey.

**Software for calculating confidence intervals.** Once the survey has been completed, the values are tabulated by cluster for doing the confidence interval analysis. We did this tabulation by hand on tally sheets in Burma and then, more recently in Thailand, using Survey Mate (*F*), a data entry, editing, and analysis program which can be run on

Confidence intervals for the estimated proportion of persons with the attribute of interest analyzed in table 2




portable computers (*1*). Using either method, the tallies for the binomial variables of interest are entered individually into the spreadsheet program for analysis in table and chart form (table 2).

For a rapid analysis, three items are entered into the spreadsheet shown in table 2: (*a*) the number of persons sampled in each cluster (column B), (*b*) the number observed to have the attribute in each cluster (column C), and (*c*) the estimated total number of eligible persons in the population from which the sample was selected (bottom left side of table). The program then derives (*a*) the proportion with the attribute in each cluster, (*b*) the proportion with the attribute in the total sample, (*c*) the average number of persons per cluster, (*d*) the average number of clusters, (*e*) the standard error of the estimated sample proportion, (*f*) the standard error of the estimated sample proportion if it had been a simple random sample, (*g*) the design effect relating the variance of the cluster sample to the variance if it had been a simple random sample, (*h*) the estimated intraclass correlation coefficient showing the degree of homogeneity within clusters, and (*i*) confidence intervals at the 90, 95, and 99 percent levels. Besides producing the table, the SuperCalc program also produces the graph of the confidence intervals for presentation to the administrator (see chart). This graph is the most important aspect of the program output since it clearly shows the interval estimates that were derived from the sample and does not rely on complicated statistical concepts such as standard errors or *P* values to be understood.

Examples of intraclass correlation coefficients and design effects for selected variables measured in Burma and Thailand rapid surveys are shown in table 3. In both countries, we selected 30 clusters with the indicated number of persons per cluster. Two of the less common variables, infant deaths and severely malnourished children, were distributed homogeneously throughout the 30 clusters, as noted by the coefficients of zero and the design effects of slightly less than 1.0. Conversely one variable, third dose of DPT, had a coefficient of 0.47 and a design effect of 4.12, indicating that children within clusters were more likely to have the same immunization pattern than children in other clusters. In a preferred immunization program, children would be vaccinated in an equal manner throughout the area of interest, thus providing maximum protection through herd immunity. If children are equally as likely to be vaccinated in one region as another, the rate of homogeneity in clusters would be low. That is, the children would be as similar to one another within clusters as they are to other children in surrounding clusters. When surveying such an area, the intraclass correlation coefficient would be approximately zero and the design effect would be 1.0. In contrast, a high intraclass correlation coefficient (and design effect) would suggest to a program administrator that the quality of the vaccination delivery system varies considerably throughout the study area since children within clusters are more alike than children in neighboring clusters.

**Software for determining sample size.** Prior to doing a rapid survey, the investigator should determine from the administrator the desired level of precision for the estimated parameter. Since administrators are usually not experts in either sampling methodology or statistics, most likely they will not be able to articulate the desired level of precision in statistical terms. The concept of a confidence interval is relatively easy to understand, especially if presented in a graph. It can be pointed out that because we are not going to count all members of a population, we will not know for sure the "true" value of the parameter we are interested in. A sample survey is less expensive than a total count of the population and can be done more quickly. Yet, we cannot be completely certain that the parameter value we will find in our sample is close to the "true" value. Instead, we can only be certain that if we select the sample in an unbiased manner, we can construct an interval that brackets the "true" value. If we want to be 99 percent certain that the interval brackets the "true" value, the interval will have to be larger. If we only want

Table 2. Example of analysis for estimating proportion of the sampled population with the attribute of interest

| Sequence no. of cluster (A) | No. of sampled persons (B) | Number of persons with the attribute: Observed (C) | Number of persons with the attribute: Expected (D) | Observed minus expected squared (E) | Proportion with attribute in each cluster (F) |
|---|---|---|---|---|---|
| 1 | 12 | 9 | 9.37 | .14 | .750 |
| 2 | 12 | 12 | 9.37 | 6.91 | 1.000 |
| 3 | 11 | 11 | 8.59 | 5.81 | 1.000 |
| 4 | 12 | 2 | 9.37 | 54.33 | .167 |
| 5 | 12 | 12 | 9.37 | 6.91 | 1.000 |
| 6 | 12 | 10 | 9.37 | .40 | .833 |
| 7 | 12 | 11 | 9.37 | 2.65 | .917 |
| 8 | 12 | 9 | 9.37 | .14 | .750 |
| 9 | 11 | 10 | 8.59 | 1.99 | .909 |
| 10 | 12 | 11 | 9.37 | 2.65 | .917 |
| 11 | 12 | 9 | 9.37 | .14 | .750 |
| 12 | 12 | 11 | 9.37 | 2.65 | .917 |
| 13 | 12 | 12 | 9.37 | 6.91 | 1.000 |
| 14 | 12 | 4 | 9.37 | 28.85 | .333 |
| 15 | 12 | 11 | 9.37 | 2.65 | .917 |
| 16 | 11 | 11 | 8.59 | 5.81 | 1.000 |
| 17 | 12 | 5 | 9.37 | 19.10 | .417 |
| 18 | 12 | 12 | 9.37 | 6.91 | 1.000 |
| 19 | 12 | 9 | 9.37 | .14 | .750 |
| 20 | 13 | 11 | 10.15 | .72 | .846 |
| 21 | 12 | 9 | 9.37 | .14 | .750 |
| 22 | 12 | 12 | 9.37 | 6.91 | 1.000 |
| 23 | 12 | 10 | 9.37 | .40 | .833 |
| 24 | 11 | 5 | 8.59 | 12.89 | .455 |
| 25 | 12 | 12 | 9.37 | 6.91 | 1.000 |
| 26 | 12 | 6 | 9.37 | 11.36 | .500 |
| 27 | 12 | 12 | 9.37 | 6.91 | 1.000 |
| 28 | 12 | 4 | 9.37 | 28.85 | .333 |
| 29 | 11 | 9 | 8.59 | .17 | .818 |
| 30 | 12 | 7 | 9.37 | 5.62 | .583 |
| Total | 356 | 278 | 278.00 | 235.97 | .781 |

EST. TOTAL PERSONS IN POPULATION = 200,000

AVERAGE NUMBER OF PERSONS PER CLUSTER = 11.87

AVERAGE NUMBER OF CLUSTERS = 30

| | Confidence Intervals: 90% | 95% | 99% |
|---|---|---|---|
| Upper | .853 | .867 | .894 |
| Lower | .709 | .695 | .668 |

STANDARD ERROR OF EST. SAMPLE PROP.

| | |
|---|---|
| For Cluster Sample | .044 |
| If Simple Random Sample | .022 |
| DESIGN EFFECT | 4.00 |
| ESTIMATED INTRACLASS CORRELATION COEFFICIENT | .28 |

Instructions for table 2

SuperCalc program entries for the upper portion of table 2.

| Number of persons with the attribute: Observed (C) | Expected (D) | Observed minus expected squared (E) | Proportion with attribute in each cluster (F) |
|---|---|---|---|
| 9 | $D$39/$C$39*C8 | (D8-E8)^2 | IF(D8<1,0,D8/C8) |
| 12 | $D$39/$C$39*C9 | (D9-E9)^2 | IF(D9<1,0,D9/C9) |
| 11 | $D$39/$C$39*C10 | (D10-E10)^2 | IF(D10<1,0,D10/C10) |
| . | . | . | . |
| . | . | . | . |
| 7 | $D$39/$C$39*C37 | (D37-E37)^2 | IF(D37<1,0,D37/C37) |
| SUM(D8.D37) | SUM(E8.E37) | SUM(F8:F37) | D39/C39 |

• The observed entry is in table column C (column D in the spreadsheet since column A is left blank). If the number in table column C is zero (that is, less than 1, the program enters a O in table column F. Otherwise, the formula in table column F derives the proportion with the attribute in the specific cluster.
• The sum of all 30 entries in each column are entered at the bottom of the table.

SuperCalc program entries for the bottom right portion of table 2.

| STANDARD ERROR OF EST. SAMPLE PROP. | |
|---|---|
| For Cluster Sample | SQRT((1-((E46*E44)/E42))*(F39/(E46*(E46-1)*E44^2))) |
| If Simple Random Sample | SQRT(G39*(1-G39)/(C39-1)) |
| DESIGN EFFECT | (I44^2)/(I46^2) |
| ESTIMATED INTRACLASS CORRELATION COEFFICIENT | (I48-1)/(E44-1) |

• The mathematical formulas calculated in the spreadsheet are based on the formulas included in the text.

SuperCalc program entries for the bottom left portion of table 2.

| | Confidence Intervals: 90% | 95% | 99% |
|---|---|---|---|
| Upper | (D39/C39)+1.64*I44 | (D39/C39)+1.96*I44 | (D39/C39)+2.58*I44 |
| Lower | IF(B59<0,0,B59) | IF(C59<0,0,C59) | IF(D59<0,0,D59) |
| | (D39/C39)-1.64*I44 | (D39/C39)-1.96*I44 | (D39/C39)-2.58*I44 |

• The mathematical formula calculated in the spreadsheet for the upper limit of the confidence intervals is described in the text.
• The lower limit of the confidence interval is set to O if the confidence interval extends below 0.
• The spreadsheet calculation formula for the lower limit of the confidence interval is included in the spreadsheet several rows below the table.

to be 90 percent certain it brackets the "true" value, the interval will be smaller. To reduce the size of the confidence interval, the administrator must either accept a lower level of certainty that the interval brackets the "true" value or be willing to pay more and increase the sample size.

The interactive program for determining sample size for rapid surveys is shown in table 4. The program also produces the graph shown on page 30 to illustrate what the study findings will look like, given the input characteristics. The specific SuperCalc program steps are listed with table 4. When considering a rapid survey of a population, the investigator enters (*a*) the estimated proportion with attribute of interest, (*b*) one-half the length of the desired confidence interval (that is, the level of precision or *d*), (*c*) the desired level of confidence, (*d*) the number of clusters to be sampled (25 clusters is the minimum for methodological reasons; see Cochran (*11*)), (*e*) the average number of persons per cluster, and (*f*) the intraclass correlation coefficient indicating the level of homogeneity for the attribute of interest. The program then calculates, among other items, the sample size for the cluster sample and indicates if it is large enough to fulfill the criteria specified in the top portion of the table. The investigator next generates a graph with the stated specifications to visually determine if the sample size is acceptable.

Since the program is interactive, the investigator can experiment with different combinations of the various input criteria to see the effect that they have on the sample size. For example, increasing the number of clusters has a greater effect on the variance of the proposed cluster sample than increasing the average number per cluster. Yet the cost of traveling to more clusters may indicate that more persons per cluster are desirable. As noted in table 3, the intraclass correlation coefficients for many variables are quite high. The investigator might want to consider ways of reducing the intraclass correlation coefficients to increase the heterogeneity within clusters. For example, Lwanga and Abiprojo greatly reduced the variance of cluster immunization surveys in Indonesia by randomly sampling children in each cluster rather than selecting neighboring children after a random start, as is the usual practice (*16*). Yet they reported that the additional cost did not justify the change in the conventional procedure. Perhaps for other variables with very high intraclass correlation coefficients, the cost would be justified.

In summary, the sample size program focuses the attention of the investigator and the administrator

Table 3. Intraclass correlation coefficient and design effect of selected variables in three rapid surveys conducted in rural regions of Burma and Thailand

| Measured attribute | Mean number of children or women per cluster | Proportion with attribute in total sample | Intraclass correlation coefficient (roh) | Design effect (deff) |
|---|---|---|---|---|
| *Births during prior 3 years* [1] | | | | |
| Infant deaths | 13.9 | 0.04 | 0.00 | 0.95 |
| Trained birth attendant | 13.9 | 0.91 | 0.27 | 4.49 |
| *Children, ages 0–35 months* [1] | | | | |
| Severely malnourished [2] | 13.2 | 0.02 | −0.02 | 0.80 |
| *Children, ages 12–35 months* [1] | | | | |
| First dose of DPT | 7.9 | 0.78 | 0.26 | 2.81 |
| Second dose of DPT | 7.7 | 0.58 | 0.31 | 3.12 |
| Third dose of DPT | 7.7 | 0.41 | 0.47 | 4.12 |
| Single dose of BCG | 8.0 | 0.78 | 0.28 | 2.92 |
| Presence of BCG scar | 8.2 | 0.73 | 0.27 | 2.93 |
| *Married women, ages 15–44 years* [3] | | | | |
| Using a family planning method | 7.0 | 0.54 | 0.11 | 1.67 |
| *Women with pregnancy outcome in last 24 months* [3] | | | | |
| Received antenatal care | 7.0 | 0.68 | 0.37 | 3.21 |
| Received tetanus toxoid | 7.0 | 0.69 | 0.23 | 2.36 |

[1] Hlegu Township, Burma, May 1987. [2] "Red" zone in Burmese weight-for-age growth chart. [3] Sisaket Province, Thailand, December 1987.

on the outcome of the rapid survey so that both can decide in advance if the new knowledge is worth the cost. In addition, the program allows the investigator to experiment with different options to see their effects on the variance of the proposed cluster sample. By highlighting the intraclass correlation coefficient, the program emphasizes the importance of the rate of homogeneity on the variance estimates and stimulates thinking about ways to increase the degree of intra-cluster heterogeneity with modifications in the second stage of the sampling procedure.

## Discussion

Microcomputers are becoming increasingly common in developing countries (*17,18*). Based on our earlier experiences in Bangladesh, we have shown that it is possible to train quickly the educated, but computer-illiterate, health professionals in the use and operation of computers (*19,20*). The most recent application of computer usage has been presented in this paper. The three spreadsheet programs featured in this article can be used both to do rapid health surveys and as a learning tool for understanding the two-stage cluster sampling method. One-week workshops on rapid survey methodology were held in both Burma and Thailand before going into the field. During the afternoon sessions of both workshops, the participants experimented with the various spreadsheet programs and learned first hand about components of the statistical theory. The material is also currently being taught to students interested in health problems of developing countries at the University of California at Los Angeles. By altering parameters in the spreadsheets and seeing the effect on the variance or the confidence intervals, workshop participants and students are able to understand more clearly the interrelationships between the various equations. In addition, students can experiment with other procedural modifications such as changing the number of clusters from 30 to 40 and the number per cluster from 7 to 20 or 25 as was done in a study of neonatal mortality (*8*).

The design effect and intraclass correlation coefficient have long been used by sampling statisticians to estimate the desired sample size for complex surveys. Since the two-stage cluster sampling method favored for immunization coverage is now being used to measure many different attributes, knowing the design effect and intraclass correlation coefficient will help investigators determine if the rapid survey findings can be presented with a desired level of precision. By seeing the three levels of confidence and noting the increasing size associated with the 90, 95, and 99 percent confidence limits, the person requesting the data can decide if the information will be of value, given the cost of the survey. The graph (see chart) is much easier to understand than a statistical discussion using abstract ideas and complex formulas, and it brings a certain level of common sense into the conversation.

Table 4. Program for determining sample size for interval estimation in rapid two-stage cluster surveys

| | |
|---|---|
| **TO BE COMPLETED BY THE INVESTIGATOR** | |
| Estimated proportion with the attribute | .300 |
| One-half length of confidence interval | .100 |
| Desired level of confidence (90% = 1.64; 95% = 1.96; 99% = 2.58) | 1.96 |
| Number of clusters (should be > 25) | 38 |
| Average number per cluster | 10.0 |
| Intraclass correlation coefficient (ROH) | .400 |
| **DERIVED BY THE PROGRAM** | |
| Necessary variance of sample proportion | .002603 |
| Sample size if SIMPLE RANDOM SAMPLE | 81 |
| Variance of proposed cluster sample | .002542 |
| Sample size for proposed CLUSTER SAMPLE | 380 |
| Est. DESIGN EFFECT for cluster sample | 4.60 |
| Sample size specifications are OK | Yes |

Instructions for table 4.

SuperCalc program entries for columns A (text) and B (formulas) in the bottom portion of table 4.

| | |
|---|---|
| **DERIVED BY THE PROGRAM** | |
| Necessary variance of sample proportion | E7^2/E10^2 |
| Sample size if SIMPLE RANDOM SAMPLE | (E5*(1-E5))/E22 |
| Variance of proposed cluster sample | (E5*(1-E5))*(E16*(E14-1)+1)/(E12*E14) |
| Sample size for proposed CLUSTER SAMPLE | E12*E14 |
| Est. DESIGN EFFECT for cluster sample | E26/((E5*(1-E5))/E28) |
| Sample size specifications are OK | IF(E26<E22,'Yes','No') |

• The mathematical formulas calculated in the spreadsheet are included in the text.
• If the variance of the proposed cluster sample is less than the necessary variance of the sample proportion based on the values set by the operator, the program will print "yes" in cell B32. Otherwise the program will print "No."

Because of many problems with routine data collection systems, administrators in technologically less developed countries find that health surveys are often the only method of either monitoring or evaluating the impact of preventive or curative efforts. Through the use of portable computers and the improved software described in this article and reference *1*, community-based surveys can now be done more rapidly than ever before.

## References


1. Frerichs, R. R., and Tar Tar, K.: Computer-assisted rapid surveys in developing countries. Public Health Rep 104: 14–23, January-February 1989.
2. Frerichs, R. R., and Tar Tar, K.: Use of rapid survey methodology to determine immunization coverage in rural Burma. J Trop Pediatr 34: 125–130 (1988).
3. Frerichs, R. R., and Tar Tar, K.: Breastfeeding, dietary intake and weight-for-age of children in rural Burma. Asian-Pacific J Public Health 2: 16–21 (1988).
4. World Health Organization: Immunization coverage survey. Weekly Epidemiological Record 62: 183–185 (1987).
5. Lemeshow, S., and Robinson, D.: Surveys to measure program coverage and impact: a review of the methodology used by the Expanded Program on Immunization. World Health Stat Q 38: 65–75 (1965).
6. Expanded programme on immunization: Evaluation and monitoring of national immunization programmes. EPI/GEN/86/4 REV 1. World Health Organization, Geneva, 1986.
7. Henderson, R. H., and Sundaresan, T.: Cluster sampling to assess immunization coverage: a review of experience with a simplified method. Bull WHO 60: 253–260 (1982).
8. Rothenberg, R. B., Lobanov, A., Singh, K. B., and Stroh, Jr, G.: Observations on the application of EPI cluster survey methods for estimating disease incidence. Bull WHO 63: 93–99 (1985).
9. World Health Organization: Diarrheal diseases control program. Weekly Epidemiological Record 62: 361–363 (1987).
10. Expanded Programme on Immunization: Program review. Weekly Epidemiological Record 61: 21–23 (1986).
11. Cochran, W. G.: Sampling techniques. Ed 3, John Wiley and Sons, New York, 1977.
12. Kish, L.: Survey sampling. John Wiley and Sons, New York, 1965.
13. Levy, P. S., and Lemeshow, S.: Sampling for health professionals. Lifetime Learning Publications, Belmont CA, 1980.
14. Hansen, M. H., Hurwitz, W. N., and Madow, W. G.: Sample survey methods and theory. Vol. 1. Methods and applications. John Wiley and Sons, New York, 1953.
15. Serfling, R. E., and Sherman, I. L.: Attribute sampling methods for local health departments. Centers for Disease Control, Atlanta, GA, 1965.
16. Lwanga, S. K., and Abiprojo, N.: Immunization coverage survey methodology studies in Indonesia. Bull WHO 65: 847–853 (1987).
17. Berge, N., Ingle, M. D., and Hamilton, M.: Microcomput-

ers in development: a manager's guide. Kumarian Press, West Hartford, CT, 1986.

18. Bertrand, W. E.: Microcomputer applications in health population surveys: experience and potential in developing countries. World Health Stat W 38: 91–100 (1985).
19. Frerichs, R. R., and Miller, R. A.: Introduction of a microcomputer for health research in a developing country—the Bangladesh experience. Public Health Rep 100: 638–647, November–December 1985.
20. Gould, J. B., and Frerichs, R. R.: Training faculty in Bangladesh to use a microcomputer for public health: followup report. Public Health Rep 101: 616–624, November–December 1986.

**Software and Equipment**

*A.* SuperCalc, Computer Associates International, Inc., 2195 Fortune Dr., San Jose, CA 95131.
*B.* Lotus 1-2-3, Lotus Development Corporation, Cambridge, MA.
*C.* Toshiba America Inc., 2441 Michelle Dr., Tustin, CA 92680.
*D.* Hewlett-Packard Personal Computer Group, 10520 Ridgeview Ct., Cupertino, CA 95014.
*E.* Diconix Inc., 3100 Research Blvd., Dayton, Ohio 45420.
*F.* Survey Mate, Henry Elkins and Associates, Inc., 15 Willow Circle, Bronxville, NY 10708.