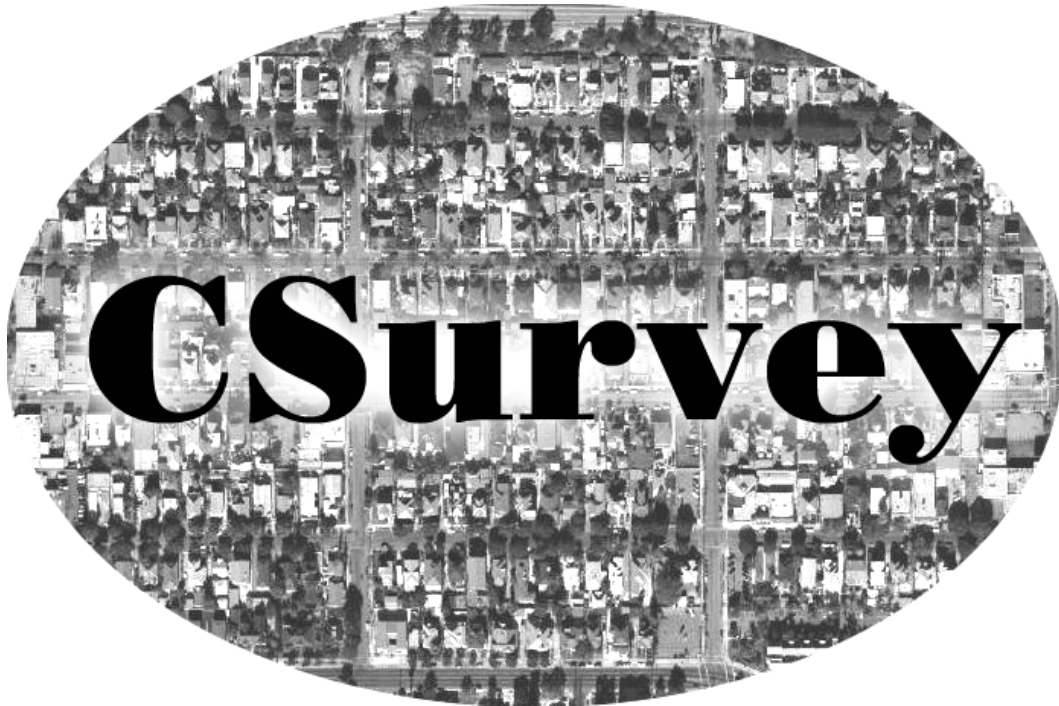


Software and Manual



Version 2.0

Muhammad N. Farid
Ralph R. Frerichs

Department of Epidemiology
University of California, Los Angeles (UCLA)
Los Angeles, CA 90095-1772 USA

June, 2007

The *CSurvey* program was originally programmed for DOS (IBM Compatible computers) by Iwan Ariawan of the University of Indonesia while doing graduate studies at UCLA in a program sponsored by the Fogarty International HIV/AIDS Training Program. *CSurvey* was based on a spreadsheet program that was created by Professor Ralph R. Frerichs and used for many years in his UCLA course, EPI 418 *Rapid Epidemiological Surveys in Developing Countries*. After attending the EPI 418 course, Muhammad N. Farid, also sponsored by the Fogarty International HIV/AIDS Training Program, designed and programmed Version 2 of *Csurvey* in Windows format. Following creation of Version 2, this manual was written by Professor Frerichs in collaboration with Muhammad Farid.

This manual and software program are in the public domain and may be copied and distributed without restriction. The manual and software program should not, however, be sold for financial compensation.

Table of Contents

Chapter 1: Introduction

What is <i>CSurvey</i> ?	1.1
Cluster selection	1.1
Sample size	1.1
Random number	1.2
How is this manual organized?	1.3

Chapter 2: Installation

Obtain from UCLA Epidemiology website	2.1
Install <i>CSurvey</i> on C:drive of computer	2.1
Removing <i>CSurvey</i> from computer	2.5

Chapter 3: Overview Example

Initial sample size	3.1
Parameter estimation	3.1
Hypothesis testing	3.4
Preparing for a rapid survey	3.6
Survey parameter	3.8
Cluster data	3.9
Sample size check	3.10
Conducting a rapid survey	3.12
PPS sample at first stage	3.12
PPS sample at first stage in multi-cluster communities	3.13
Other features	3.15
Spin dial	3.15
Random number	3.17

Chapter 4: Detailed Explanation

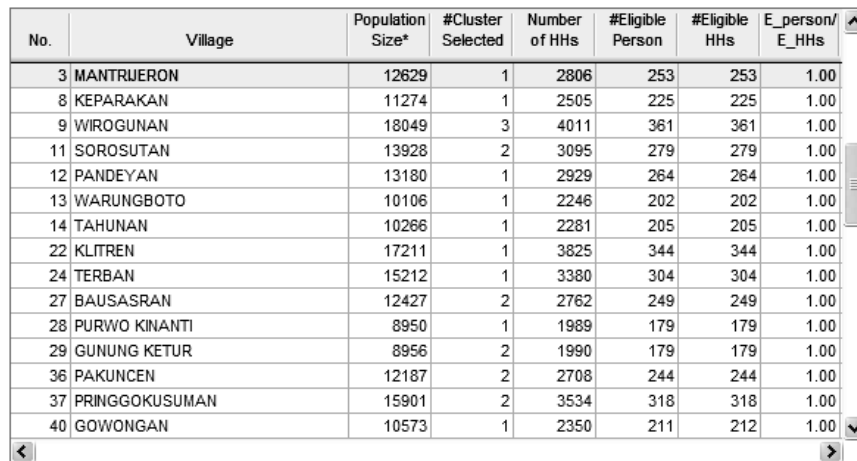
Sample size – parameter estimation	4.1
Sample size – hypothesis testing	4.5
PPS sample at first stage	4.9

Chapter 1: Introduction

What is CSurvey?

CSurvey is a Windows program that performs tasks necessary for conducting rapid surveys, otherwise termed two-stage cluster surveys with probability proportionate to size (PPS) sampling at the first stage and a constant number of households or persons at the second stage. Such surveys are typically small (i.e., about 300 households or subjects), although the methods can also be used for larger surveys. The *CSurvey 2.0* program is written for Windows-compatible microcomputers, following the earlier *CSurvey1.5* program written in DOS. The program helps select a cluster sample from a list of clusters, calculates the sample size for a cluster survey and creates a random number for selecting random start households or persons within households. There are three main modules in *CSurvey*.

Cluster selection. The first module selects a sample clusters from a list of all clusters using the probability proportionate to size (PPS) method. To sample clusters, users must create a source database consisting of the name and the size of each cluster in the population to be sampled. This database file can be created using *CSurvey*, or can be imported from other common spreadsheet or database programs. Figure 1.1 shows the selected clusters in a typical source database file.



No.	Village	Population Size*	#Cluster Selected	Number of HHs	#Eligible Person	#Eligible HHs	E_person/E_HHs
3	MANTRUJERON	12629	1	2806	253	253	1.00
8	KEPARAKAN	11274	1	2505	225	225	1.00
9	WIROGUNAN	18049	3	4011	361	361	1.00
11	SOROSUTAN	13928	2	3095	279	279	1.00
12	PANDEYAN	13180	1	2929	264	264	1.00
13	WARUNGBOTO	10106	1	2246	202	202	1.00
14	TAHUNAN	10266	1	2281	205	205	1.00
22	KLITREN	17211	1	3825	344	344	1.00
24	TERBAN	15212	1	3380	304	304	1.00
27	BAUSASRAN	12427	2	2762	249	249	1.00
28	PURWO KINANTI	8950	1	1989	179	179	1.00
29	GUNUNG KETUR	8956	2	1990	179	179	1.00
36	PAKUNCEN	12187	2	2708	244	244	1.00
37	PRINGGOKUSUMAN	15901	2	3534	318	318	1.00
40	GOWONGAN	10573	1	2350	211	212	1.00

* Persons in population as size unit

Figure 1.1 *CSurvey* cluster selection module.

Sample Size. The second module calculates the necessary sample size for a cluster survey to satisfy the needs of the investigator. Users can evaluate a proposed sample size, or calculate the minimum numbers of clusters or average persons per clusters that are needed for a specified confidence interval. Figure 1.2 shows the sample size calculation for a proposed cluster sample with prevalence estimate of 50 percent and a 95% confidence interval of 40 to 60 percent or less.

Parameter Estimation		Hypothesis Testing	
Calculation purpose <input checked="" type="radio"/> Test the proposed sample size <input type="radio"/> Calculate minimum number of clusters <input type="radio"/> Calculate average number in sample per cluster		Target standard error of proportion	0.0489
Estimated proportion with attribute	0.5000	Actual standard error of proportion	0.0408
One-half length of confidence interval	0.1000	Design effect (deff)	2.00
Desired level of confidence	95% ▼	Rate of homogeneity (roh)	0.1111
Homogeneity parameter	Design Effect ▼	Point estimate for proportion	0.5000
Level of homogeneity	Low ▼	Lower confidence limit	0.4165
Average number of eligible persons per HH	1.00	Upper confidence limit	0.5835
Number of clusters	30	Sample size for proposed cluster survey	300
Average number of selected HHs per cluster	10	Is sample size adequate for stated need?	YES
<input type="button" value="▲"/> <input type="button" value="Calculate"/> <input type="button" value="Print"/>		CI 0.50 90% 0.43 0.57 95% 0.42 0.58 99% 0.39 0.61 0 0.5 1	

Figure 1.2 CSurvey sample size module.

Random Number. The third module is used to create a printable random number table. This table is useful for selecting persons or households in the sampled clusters. Figure 1.3 shows a typical random number table for communities with less than 500 households.

Table																Spin Dial		
104	412	201	337	168	289	190	68	279	384	122	417	454	277	426				
121	445	8	153	192	362	52	194	198	259	309	323	441	444	142				
96	311	315	477	135	366	402	402	203	26	350	213	470	402	44				
121	338	104	145	202	290	59	39	141	73	122	469	20	474	464				
190	356	188	186	179	297	217	216	236	127	106	402	227	189	434				
122	115	401	476	94	460	328	447	406	152	149	397	251	307	316				
250	428	335	57	169	99	327	131	395	87	497	94	38	304	279				
366	308	202	330	233	177	253	42	304	110	433	332	57	109	180				
8	239	417	41	123	282	127	175	48	34	182	141	437	473	287				
350	305	478	283	68	418	397	10	494	274	248	427	398	200	375				
169	170	415	82	320	18	234	25	394	226	311	80	151	196	258				
445	76	200	284	470	160	235	335	111	39	157	392	193	252	472				
281	480	215	100	77	406	302	159	440	61	175	374	210	78	396				
445	373	200	287	279	320	376	376	297	156	431	34	456	486	368				
283	38	258	333	137	49	400	227	303	266	464	494	464	270	155				
394	282	68	134	279	411	352	300	22	234	459	280	278	91	368				
128	285	230	12	310	89	456	118	232	278	221	349	408	45	142				
367	287	67	155	55	109	279	149	320	314	299	2	85	220	84				
15	171	323	348	32	63	143	178	232	336	11	185	274	495	200				
The maximum number																500	<input type="button" value="Generate"/>	<input type="button" value="Print"/>

Figure 1.3 CSurvey random number module.

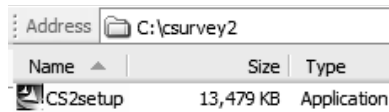
How is this manual organized?

Chapter 2 of the *CSurvey* manual describes how to install the program in a Windows-compatible computer with a C: hard drive that is using the Windows operating system. This is followed by **Chapter 3** which provides an overview example of a rapid survey which might be planned for the Yogyakarta region of Indonesia. Those familiar with the DOS version of *Csurvey* 1.5 will likely need no further information to use the new version. Finally, **Chapter 4** has the technical explanation of the various steps in the *CSurvey* program, including the mathematical formulas that are incorporated into the program.

Chapter 2: Installation

Obtain from UCLA Epidemiology Website

The *CSurvey* program needs first to be downloaded from the UCLA Epidemiology website, then installed on the C: drive of the destination computer. The program and instructions for this step are found at: <http://www.ph.ucla.edu/epi/rapidsurvey.html> in the software section. Once installed, the program should appear as in Figure 2.1.



Name	Size	Type
CS2setup	13,479 KB	Application

Figure 2.1 *CSurvey* program location in C: drive.

Install on C:drive of Computer

With your left mouse key, click on *CS2setup* (see Figure 2.1). The first screen of the installation process should appear as in Figure 2.2.

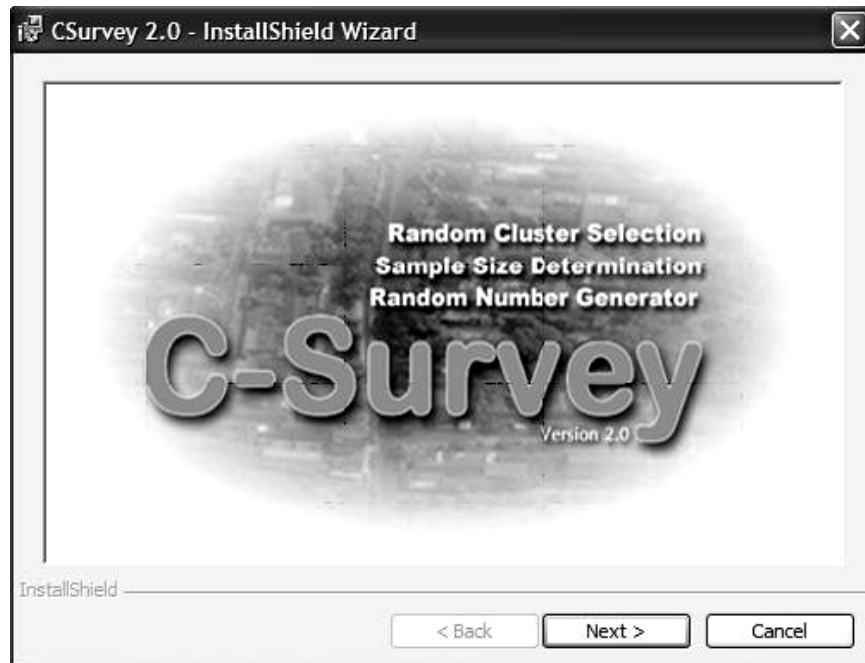


Figure 2.2 Opening screen of installation procedure.

Click *next* with the left mouse key and Figure 2.3 appears reminding the user that the material is copyrighted, not intended for commercial resale. Instead the program is available for free to all who want to do community-based surveys.

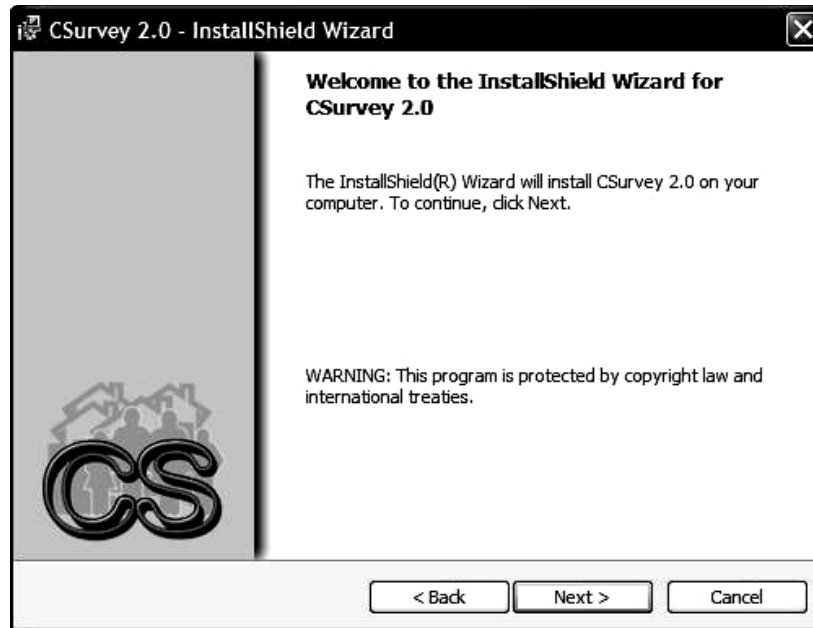


Figure 2.3 Welcome screen of installation process.

Click next again and the screen in Figure 2.4 should appear, showing where the program will be installed. If you want a different location, click with the left mouse key on *change* and enter the new directory and subdirectory.



Figure 2.4 Destination subdirectory for *CSurvey* program.

Note: in this instance the program is being installed as a subdirectory in C:\Program Files\CSurvey2. The sample files (with the extension *.csf) will also be installed in this

subdirectory, unless a new location is selected by clicking on *Change*. If the location is OK, click on *Next* and continue. Before the installation occurs, the program provides one last chance to view the destination subdirectory, as shown in Figure 2.5.



Figure 2.5 Review of destination subdirectory.

The appropriate files are copied by the installation program to the designated location. While this process is occurring, the screen shows the progress being made, as seen in Figure 2.6.

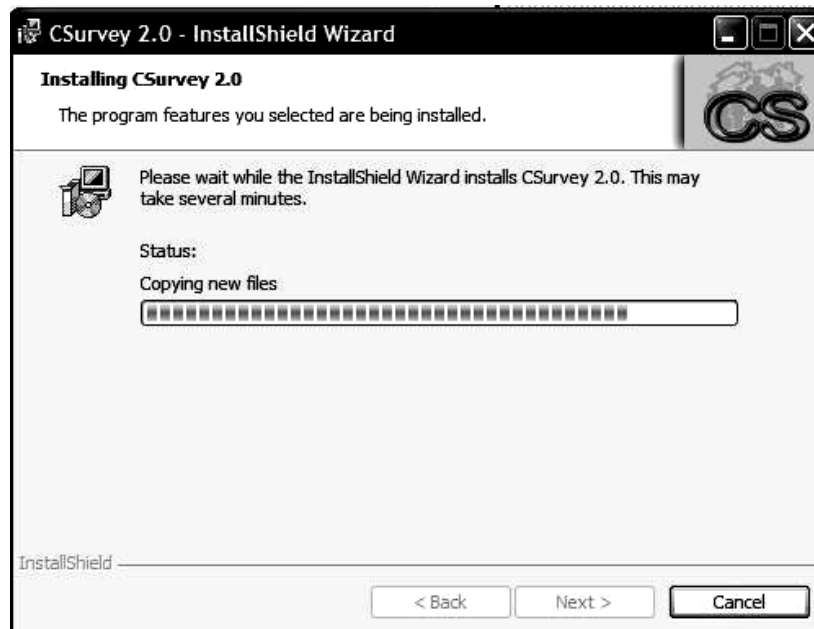


Figure 2.6 Installation of CSurvey files.

Then completed, the screen in Figure 2.7 appears, indicating that the program has been successfully installed.

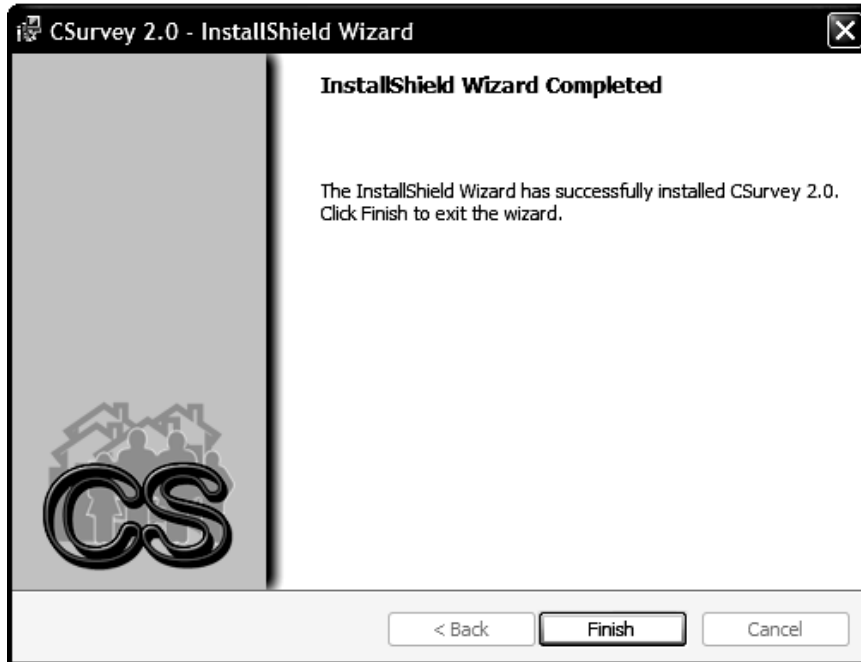


Figure 2.7 Successful installation of *CSurvey*.





Click *Finish* with the left mouse key.

Location of Files. If you look in the destination subdirectory in the C:drive, the files shown in Figure 2.8 should appear.


Name	Size	Type
FOXUSER.FPT	1 KB	FPT File
FOXUSER.DBF	1 KB	DBF File
csurvey	1,889 KB	Application
cluster	15 KB	Microsoft Excel Worksheet
vil9_yogya	6 KB	CSF File
yogya	13 KB	CSF File
csf	5 KB	Presentations 12 Master
msvcr71.dll	340 KB	Application Extension

Figure 2.8 Installed files in *CSurvey2* subdirectory.

The *CSurvey* program is now installed.

Start *CSurvey*. To start the program, go to the main Windows screen and click with the left mouse on . Then click on **All Programs** , followed by  **CSurvey 2.0**, and finally,  **CSurvey 2.0**.

Removing CSurvey from Computer

Uninstall CSurvey. If you want to uninstall *CSurvey*, the procedure is the same as starting *CSurvey*, but at the last step, click  Uninstall CSurvey 2.0. The program then asks if you are sure that you want to uninstall *CSurvey*, as seen in Figure 2.9.

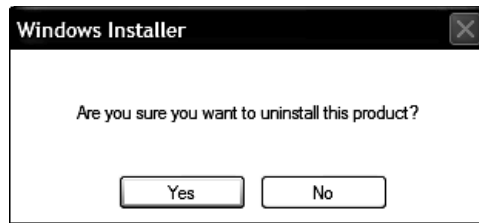


Figure 2.9 Uninstall *CSurvey*.

Click *yes* and the program begins the removal process, ridding the computer of *CSurvey*.

Chapter 3: Overview Example

Perhaps the easiest way to learn about *CSurvey* is to step through an example, using data from Indonesia that are included with the software program. The software is intended to assist with the various tasks of rapid surveys. More information on rapid surveys is found at: <http://www.ph.ucla.edu/epi/rapidsurvey.html>.

After starting the *CSurvey* program (as described at the end of Chapter 2), the screen in Figure 3.1 appears.

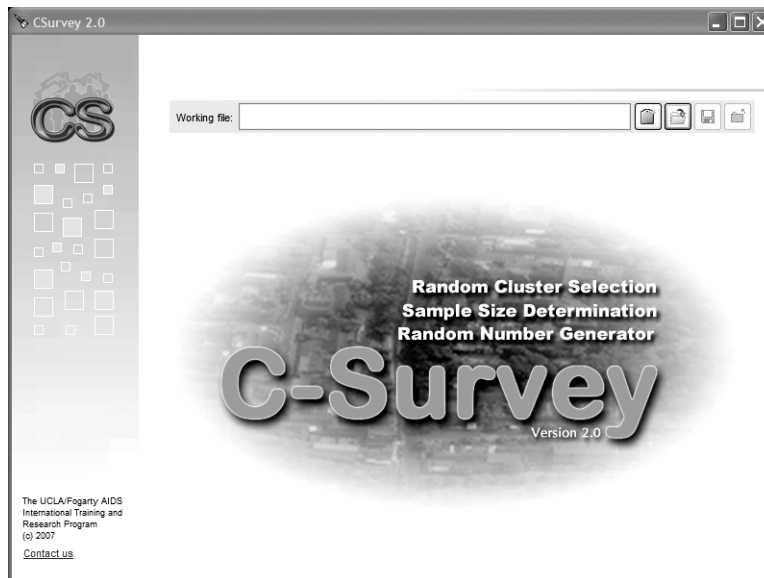




Figure 3.1 *CSurvey* opening screen.


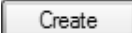
Assume that you are planning a rapid survey, but have not yet estimated the sample size that is



Figure 3.2 Creation of working file *samplesize.csf*.

needed to conduct the survey. To do so, consider the two boxes   at the top right of the screen.

Initial Sample Size

Parameter Estimation. Click with your left mouse key on  to create a temporary working file termed *samplesize.csf*, enter the text as shown in Figure 3.2. The click on  to create the working file. The screen shown in Figure 3.3 should appear.

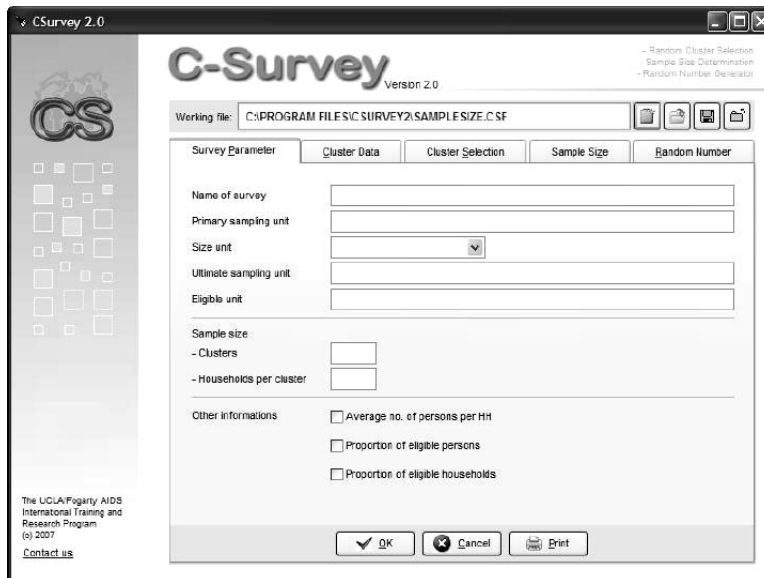


Figure 3.3 Opening screen (not used at this time).

The screen is divided into five sections, with tabs at the top showing the section names. Following opening, the first tab is highlighted, namely *Survey Parameter*. You will be using this at a later time once information is available for the specific survey to be done. The *Sample Size* section is divided into two parts, *Parameter Estimation* (to be presented first) and *Hypothesis Testing* (to be presented thereafter). For now to complete the planning process, click with your left mouse on and the screen shown in Figure 3.4 should appear.

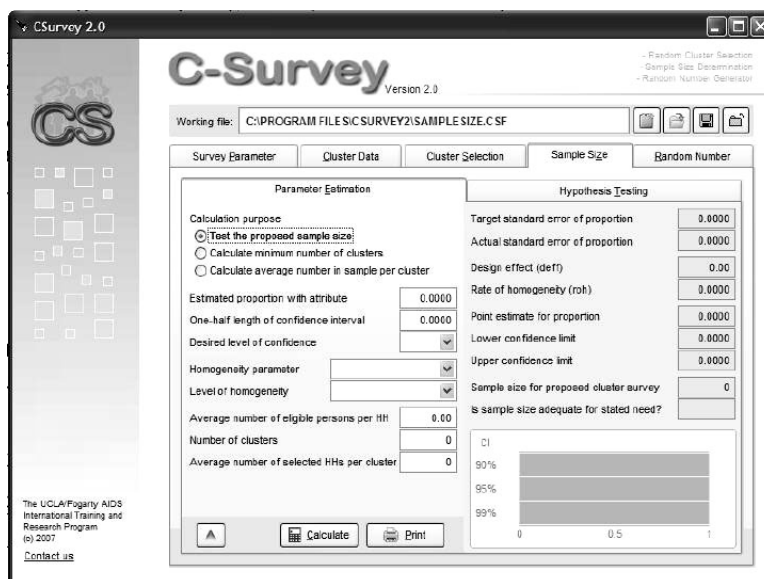
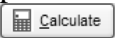


Figure 3.4 Sample size screen for estimating a proportion.

Since you will be considering various sample size estimates for a parameter of interest, click on *Test the proposed sample size*, as shown in Figure 3.4. To determine the sample size for a rapid

survey, you need four values: 1) your best estimate of the proportion with the attribute in the population to be sampled, 2) one-half the length of the maximum confidence interval that would be acceptable (i.e., the desired level of precision), 3) the desired level of confidence (either 90%, 95% – the usual level, or 99%), and 4) an estimate of the expected design effect or rate of homogeneity. The design effect is a measure of how much greater the variance is of a rapid survey (i.e., a two-stage cluster survey) than a similar-sized group with data collected as a simple random sample. For immunization surveys, the design effect for sample size estimation is often set to 2.0. The *rate of homogeneity* (or intraclass correlation coefficient) is often used by experienced surveyors with knowledge of the attribute based on past rapid surveys, while the *design effect* is more commonly used by those without such knowledge.

For this example, assume that about 20% of the sampled population will have the attribute of interest; hence, you need to enter 0.20 into the program. Further assume that the maximum acceptable 95% confidence interval is 5 percentage points (i.e., 0.05) or an upper limit of 25% and a lower limit of 15%. Next, assume that the design effect will be *low* (i.e. 2.0), there will 1.0 person per eligible household (a common assumption in immunization surveys of children 12-23 months of age), 30 clusters to selected at the first stage, and 10 households with one eligible person in each to be selected at the second stage. When through entering the data, click  and Figure 3.5 should appear.


Survey Parameter	Cluster Data	Cluster Selection	Sample Size	Random Number
Parameter Estimation Calculation purpose <input checked="" type="radio"/> Test the proposed sample size <input type="radio"/> Calculate minimum number of clusters <input type="radio"/> Calculate average number in sample per cluster Estimated proportion with attribute: 0.2000 One-half length of confidence interval: 0.0500 Desired level of confidence: 95% Homogeneity parameter: Design Effect Level of homogeneity: Low Average number of eligible persons per HH: 1.00 Number of clusters: 30 Average number of selected HHs per cluster: 10		Hypothesis Testing Target standard error of proportion: 0.0244 Actual standard error of proportion: 0.0327 Design effect (deff): 2.00 Rate of homogeneity (roh): 0.1111 Point estimate for proportion: 0.2000 Lower confidence limit: 0.1332 Upper confidence limit: 0.2668 Sample size for proposed cluster survey: 300 Is sample size adequate for stated need?: NO CI: 0.20 90%: 0.14 - 0.26 95%: 0.13 - 0.27 99%: 0.11 - 0.29		

Figure 3.5 Inadequate sample size for desired confidence limits.

Notice that confidence limits with the specified sample size would be 13.3% and 26.7%, wider than the 15% and 25% that was requested. To have the desired confidence limits, Figure 3.5 shows that the standard error for the estimated parameter should be no greater than 0.0244. With sample size selected, the actual standard error of the proportion is 0.0327, or too great for the “expectation.” Hence the program answers the question, *Is sample size adequate for stated need?* with a “No.” At this point, you can increase the acceptable confidence limits, increase the number of clusters, increase the number of selected households per cluster, or with additional knowledge about the sampling design, reduce the *intraclass correlation coefficient* towards 0

(the level of a simple random sample) . Assume for now that the size of the desired confidence limits remains fixed at plus or minus 5 percentage points, and that there are enough funds and time to sample a larger group, again with 30 clusters, but now set at 18 persons per cluster, as shown in Figure 3.6.

Figure 3.6 Adequate sample size for desired confidence limits.

Now the anticipated confidence limits are 15.0% and 25.0%, or at the level acceptable to the investigator. Rather than 300 persons being sampled, as in Figure 3.5 however, the sample size has now increased to 540 persons. Thus, increased precision has its price, and the cost is paid in increased time and labor used to sample an additional 240 persons. The small graph at the bottom of the panel shows the expected 90%, 95% and 99% confidence limits, useful for explaining the concept of confidence limits to persons not intimately familiar with statistical notions. If all parties involved with the planned survey deem these values to be acceptable, then click on  Print, sign and date the printed page, and leave it with the person or agency that is funding the planned survey.

Hypothesis Testing. Rather than determining the prevalence or incidence of an attribute in a population, you might be interested in comparing a change in an attribute over time, or in comparing the level of an attribute in one region versus another. Such studies are often done to evaluate changes, such as increases in vaccination levels, decreases in smoking behavior, increased use of condoms and the like. To conduct such an evaluation, the program provides information on two same-sized rapid surveys, and indicates if the sample size is sufficient to detect a difference in two proportions with an acceptable level of precision, as specified by the investigator.

In the *Sample Size* section, click on *Hypothesis Testing* in the right side of the panel. Notice that the left side of the panel changes, as shown in Figure 3.7.

Survey Parameter	Cluster Data	Cluster Selection	Sample Size	Random Number
Parameter Estimation Calculation purpose: <input type="radio"/> Test the proposed sample size <input type="radio"/> Calculate minimum number of clusters <input type="radio"/> Calculate average number in sample per cluster Estimated value of first proportion: 0.0000 Estimated value of second proportion: 0.0000 One-half length of confidence interval: 0.0000 Desired level of confidence: <input type="text"/> Homogeneity parameter: <input type="text"/> Level of homogeneity: <input type="text"/> Average number of eligible persons per HH: 0.00 Number of clusters: 0 Average number of selected HHs per cluster: 0		Hypothesis Testing Target standard error of different proportion: 0.0000 Actual standard error of different proportion: 0.0000 Design effect (deff): 0.00 Rate of homogeneity (roh): 0.0000 Point estimate for different proportion: 0.0000 Lower confidence limit: 0.0000 Upper confidence limit: 0.0000 Sample size for proposed cluster survey: 0 Is sample size adequate for stated need? <input type="text"/> CI 		
<input type="button" value="↑"/> <input type="button" value="Calculate"/> <input type="button" value="Print"/>				

Figure 3.7 Sample size screen for testing the difference between two proportions (i.e., hypothesis testing).




Assume that vaccination coverage is believed to be 20% in one region and 60% in another region, where a more active health care group is at work. Hence, the difference between the two regions is thought to be 40%. You want to conduct two rapid surveys to test the hypothesis that the two regions differ in vaccination coverage. While the investigator or funding agencies believes that the difference between the two regions is 40%, they are willing to accept with 95% confidence that the difference lies between 25% and 55%. That is, with a difference of 0.40 the 95% confidence interval should be no greater than ± 0.15 . As before the design effect is assumed to be *low*, the average number of eligible persons per household is assumed to be 1.0, the number of cluster is to be set at 30, and the number of households to be selected per cluster is set at various levels, but shown as 12.

The derived values that fulfill the precision requirements or the investigator or funding agency, are shown in Figure 3.8. As mentioned, the difference between the two proportions is estimated to be 0.40. Two surveys with 360 subjects in each would result in a 95% confidence interval for the difference between the two proportions of 0.30 to 0.50, acceptable to the criteria for precision set by the investigator. Once considered acceptable, the page should be printed, signed, dated and given to the agency or person funding the planned survey.

Survey Parameter	Cluster Data	Cluster Selection	Sample Size	Random Number									
Parameter Estimation Calculation purpose: <input checked="" type="radio"/> Test the proposed sample size <input type="radio"/> Calculate minimum number of clusters <input type="radio"/> Calculate average number in sample per cluster Estimated value of first proportion: 0.2000 Estimated value of second proportion: 0.6000 One-half length of confidence interval: 0.1000 Desired level of confidence: 95% Homogeneity parameter: Design Effect Level of homogeneity: Low Average number of eligible persons per HH: 1.00 Number of clusters: 30 Average number of selected HHs per cluster: 12		Hypothesis Testing Target standard error of different proportion: 0.0489 Actual standard error of different proportion: 0.0471 Design effect (deff): 2.00 Rate of homogeneity (roh): 0.0909 Point estimate for different proportion: 0.4000 Lower confidence limit: 0.3036 Upper confidence limit: 0.4964 Sample size for proposed cluster survey: 360 Is sample size adequate for stated need? YES CI: 0.40 <table border="1"> <tr> <td>90%</td> <td>0.32</td> <td>0.48</td> </tr> <tr> <td>95%</td> <td>0.30</td> <td>0.50</td> </tr> <tr> <td>99%</td> <td>0.27</td> <td>0.53</td> </tr> </table>			90%	0.32	0.48	95%	0.30	0.50	99%	0.27	0.53
90%	0.32	0.48											
95%	0.30	0.50											
99%	0.27	0.53											
<input type="button" value="↑"/> <input type="button" value="Calculate"/> <input type="button" value="Print"/>													

Figure 3.8 Adequate sample size for desired confidence limits.

Preparing for a Rapid Survey

The program assumes that the surveyor has demographic information available on the study population, but must decide on details of the two-stage cluster sampling design. At the top right are two boxes  . The one to left is used to create new program files with study population data while the one to the right is meant for the use of existing study population files. Since this section of Chapter 3 features the use of existing data, click with your left mouse on . The program should find two example files, *yogya.csf* and *vil9_yogya.csf* (and possible *SAMPLESIZE.csf* if you have used the *Initial Sample Size* section of this chapter) as seen in Figure 3.9.

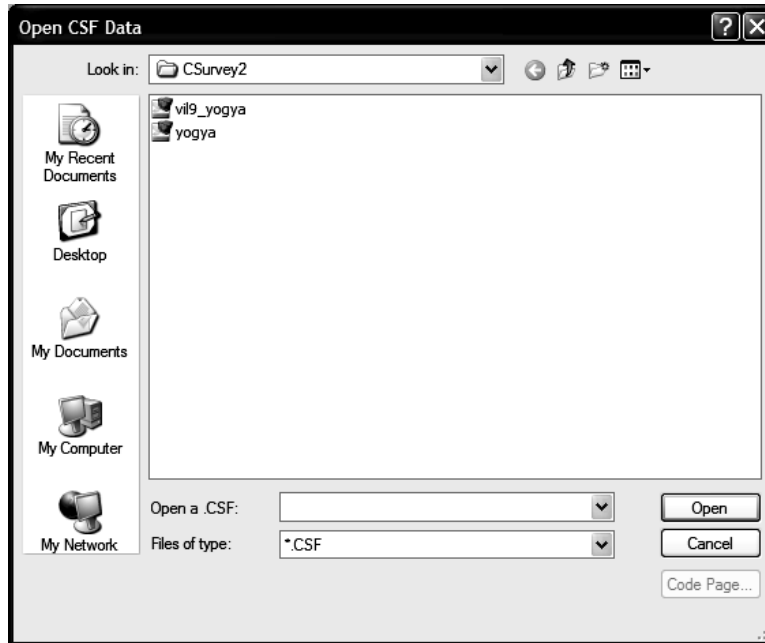


Figure 3.9 CSF files with *CSurvey* program.

Select *yogya* and click on *open* with the left mouse key, bringing forth the screen shown in Figure 3.10.

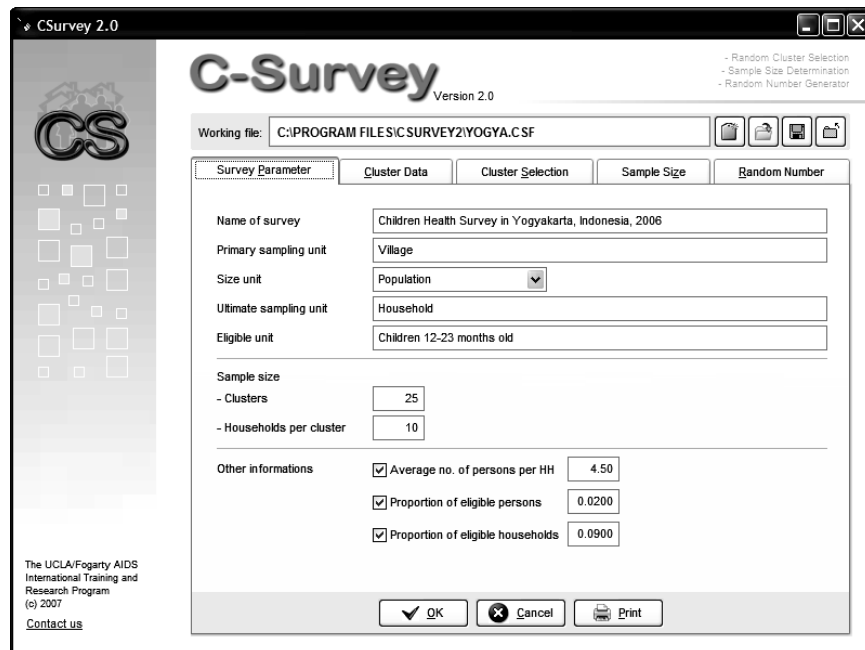


Figure 3.10 Opening *CSurvey* screen of *yogya.csf*, an example file.

The screen is divided into five sections, with tabs at the top showing the section names. Following opening, the first tab is highlighted, namely *Survey Parameter*.

Survey Parameter. Included in this screen is descriptive information on the intended survey and study population, to be filled in by the user. The information is used for the first stage of a rapid survey, namely the selection of clusters with probability proportionate to size (PPS). First is the *name of survey* followed by the *primary sampling unit* (i.e., PSU). PSUs are identified by size based on the number of people (i.e., *population*), or perhaps the *number of households*, or some *other* unit. In this example, the PSUs (or villages) are listed by *population* size. The *ultimate sampling unit* in the example is a household, while the *eligible unit* for inclusion in the survey is all children, aged 12-23 months. The number of clusters (or groups of individuals or households of a constant size) to be selected by the program at the first stage of sampling is 25, the minimum number for such surveys. The usual number of selected clusters is 30. Other information included by the user is the *population per household* (i.e., the average number of persons per household in the population to be sampled), the *proportion of eligible persons* (i.e., the proportion of the population that is 12-23 months of age), and the *proportion of eligible households* (i.e., the proportion of all households in the population to be sampled that have one or more children, aged 12-23 months). In general, the user will not know the exact values of the household variables, but should be able to provide reasonable estimates, appropriate for planning a survey. In the example shown in Figure 3.12, the average household in the Yogyakarta region has 4.5 persons per household. Children, aged 12-23 months (the eligible population), comprise 2 percent of the total population (i.e., 0.02). Finally, 9 percent of the households in the study population (i.e., 0.09) are estimated to have a child, aged 12-23 months.

Next click on *cluster data* to view the population information for the study population (see Figure 3.11).

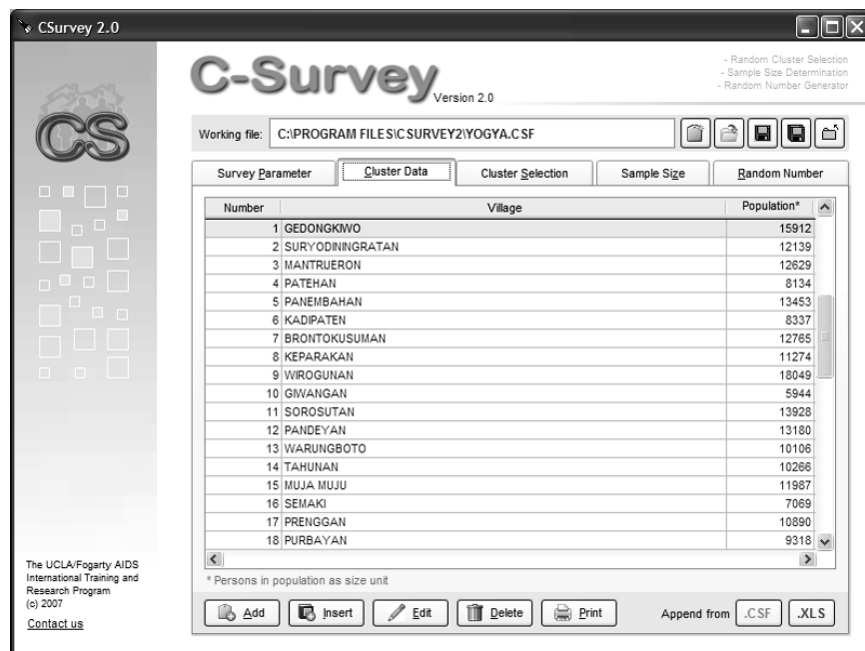


Figure 3.11 Population data by cluster.

For any rapid survey, such population data needs to be entered by the investigator for all communities in the study population. To do so, the person conducting the survey can create for the population to be sampled a new *.csf file, append an earlier-created *.csf (see bottom right of Figure 3.11), or created and append a *.xls file using the spreadsheet program, *MS Excel* (also see bottom right of Figure 3.11). If planning to use the *MS Excel* option, a screen appears that guides the data entry process, as seen in Figure 3.12.

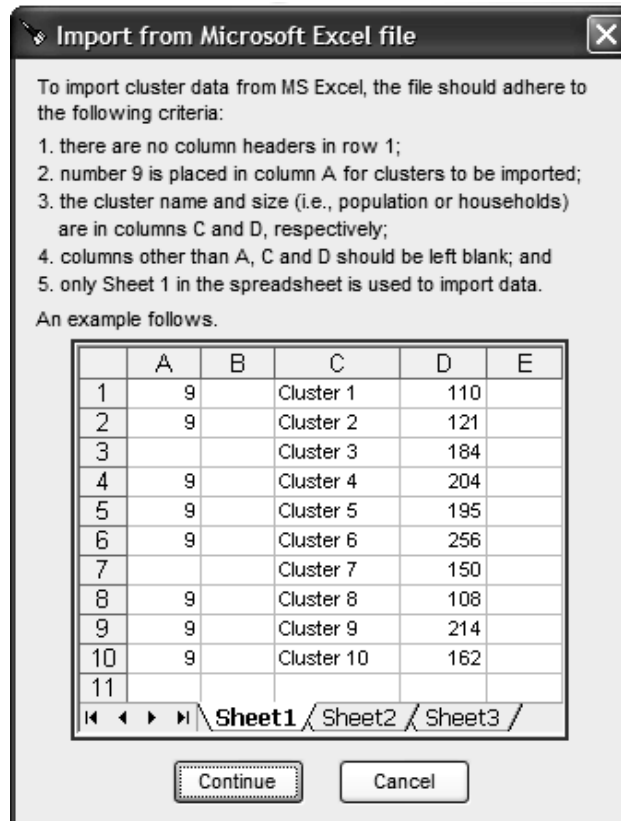


Figure 3.12 Format for importing data from MS Excel.

Cluster Data. The example data set shown in Figure 3.11 contains information on 45 villages, with the estimated population presented in the column at right. The data are easily edited or printed, using the keys at the bottom of the screen.

To make sure that the sample size specified in Figure 3.10 is adequate to fulfill the needs of the investigator, click on the tab *sample size*, as shown in Figure 3.13.

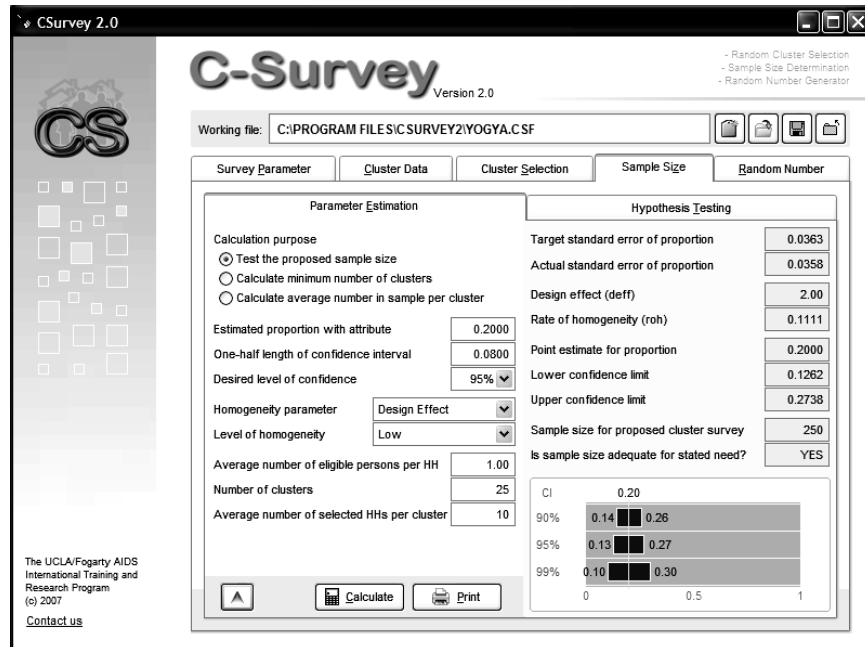

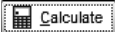



Figure 3.13 Check of specified sample size for two-stage cluster survey.

Sample Size Check. In this example, the attribute being surveyed has an estimated proportion value of 0.20 (or percentage value of 20%). The investigator is willing to accept 95% confidence limits of .12 to .28 (or 12 to 28%); that is, one half of the confidence interval size is 0.08. Being a cluster survey, the variance estimate will likely be larger than if done as a simple random sample survey. How much larger is estimated by either the *design effect* or the *rate of homogeneity*. In the example, the design effect is selected and the estimated value is set as *low*, which is equivalent to a design effect of 2.0. A small survey is specified, of 25 clusters and 10 children aged 12-23 months. For the Indonesia example, the 10 sampling units per cluster are 10 households with one or more 12-23 month olds. Is this sample size adequate? To make sure, click on  (which imports the appropriate information from *Survey Parameter*) followed by .

In the example in Figure 3.13, the sample size for the proposed survey would be 250 persons, or 25 clusters with 10 eligible households per cluster with 1.0 child, aged 12-23 months, in each eligible household. In this example, the target standard error of the proportion needed to be no more than 0.0363 to fulfil the criteria entered in the first column of *Sample Size* by the investigator. Based on the estimated sample size, the standard error of the proportion is 0.0358, less than the target maximum of 0.0363. Hence, the proposed sample size is adequate for the stated need, and the program responds *yes*. With a “low” level of homogeneity (as selected by the investigator), the program assumes a design effect of 2.0 (i.e., the variance of the cluster survey will be twice as great as the variance of a similar-sized survey done as a simple random sample) and have a *rate of homogeneity* of 0.1111. The mean and 95% confidence limits are estimated as a proportion as 0.2000 (0.1262, 0.2738), or as a percentage as 20% (12.6%, 27.4%).

Presented at the bottom of *Sample Size* is a small graph, as seen in Figure 3.14.

The information on sample size should be shared with the person or agency funding the survey to determine if the precision of the estimate is acceptable. That is, would it be acceptable to do a survey of an attribute that has a prevalence of 20% (i.e., 0.20) and estimated 95% confidence limits of 13 to 27%? In addition with this sample size, the investigator could be 90% confident that an interval of 14 to 26% would bracket the true value, or 99% confident that an interval of 10 to 30% would bracket the true value, assuming of course that there is no bias. If deemed acceptable, the investigator should click on  Print to print a copy of *Sample Size* and leave it with the funding agency to show what is expected. The page to be printed (including text information but no graph), appears as in Figure 3.15.

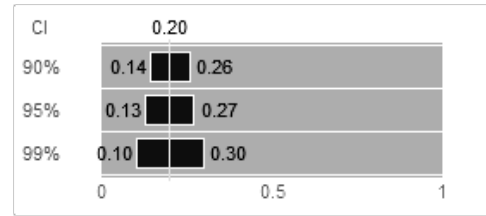


Figure 3.14 Graph of 90, 95 and 99% confidence limits for proposed survey parameter.



Sample Size: Parameter Estimation

Calculation purpose:	Test the proposed sample size
Estimated proportion with attribute:	0.2000
One-half length of confidence interval:	0.0800
Desired level of confidence:	95%
Homogeneity parameter:	Design Effect
Level of homogeneity:	Low
Number of clusters:	25
Average number of sample per cluster:	10
Target standard error of proportion:	0.0363
Actual standard error of proportion:	0.0358
Design effect (deff):	2.00
Rate of homogeneity (roh):	0.1111
Point estimation for proportion:	0.2000
Lower confidence limit:	0.1262
Upper confidence limit:	0.2738
Sample size for proposed cluster survey:	250
Is sample size adequate for stated need?	YES
90% confidence interval:	0.14 - 0.26
95% confidence interval:	0.13 - 0.27
99% confidence interval:	0.10 - 0.30

Figure 3.15 Printed information on sample size for parameter estimation.

Earlier in *Survey Parameter* (see Figure 3.10), the program was instructed to select 25 clusters with probability proportionate to size (PPS) from the population list of 45 villages or communities. Since the *Sample Size* module has shown that the number of clusters and households per cluster are acceptable, you are now set to proceed.

Conducting a Rapid Survey

PPS Sample at First Stage. Rapid surveys are conducted in a two-stage process. At the first stage, clusters are selected with probability proportionate to size (PPS) while at the second stage an equal number of households (or perhaps persons) are selected in each cluster selected at the first stage. Doing such sampling ensures that the survey data are self-weighted, and do not require special statistical weights in the analysis. Self-weighted surveys are easier to analyze than weighted surveys.

The population data in *yogya.csf* was presented in *Cluster Data*, as seen in Figure 3.11. To see the PPS sample that was drawn for the survey, click on *Sample Selection*, as presented in Figure 3.16


Cluster Selected by PPS-WR Method

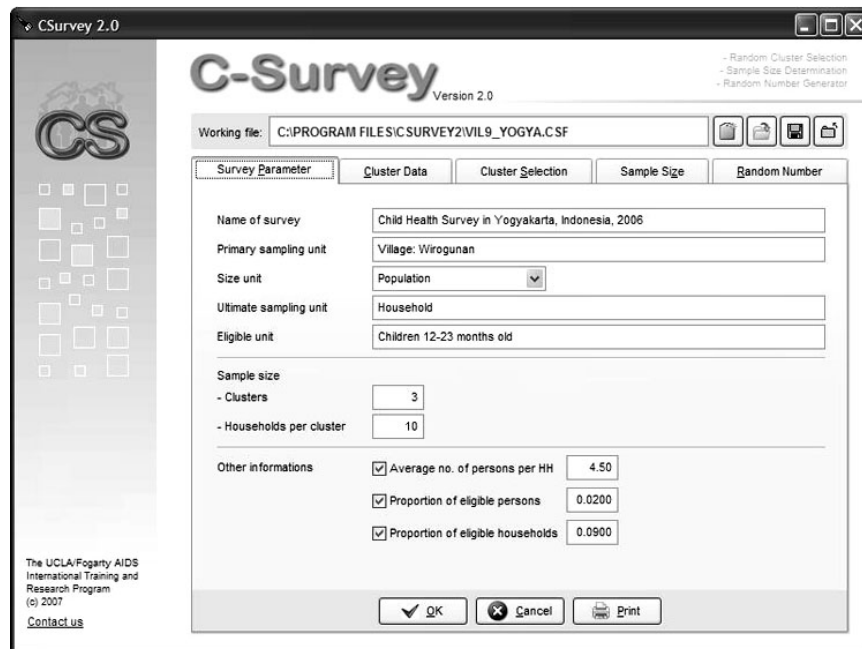
No.	Village	Population Size*	#Cluster Selected	Number of HHs	#Eligible Person	#Eligible HHs	E_person/E_HHs
3	MANTRUERON	12629	1	2806	253	253	1.00
8	KEPARAKAN	11274	1	2505	225	225	1.00
9	WIROGUNAN	18049	3	4011	361	361	1.00
11	SOROSUTAN	13928	2	3095	279	279	1.00
12	PANDEYAN	13180	1	2929	264	264	1.00
13	WARUNGBOTO	10106	1	2246	202	202	1.00
14	TAHUNAN	10266	1	2281	205	205	1.00
22	KLITREN	17211	1	3825	344	344	1.00
24	TERBAN	15212	1	3380	304	304	1.00
27	BAUSASRAN	12427	2	2762	249	249	1.00
28	PURWO KINANTI	8950	1	1989	179	179	1.00
29	GUNUNG KETUR	8956	2	1990	179	179	1.00
36	PAKUNCEN	12187	2	2708	244	244	1.00
37	PRINGGOKUSUMAN	15901	2	3534	318	318	1.00
40	GOWONGAN	10573	1	2350	211	212	1.00

* Persons in population as size unit

Figure 3.16 Sample of 25 clusters with probability proportionate to size (PPS).

Most of the selected villages have one cluster of 10 households to be selected at the second stage. Some of the villages, however, have more than one cluster of 10 households to be selected at the second stage. If the village is large, the investigator can repeat the selection process, but only for the number of cluster to be selected in that village.

PPS Sample at First Stage in Multi-Cluster Communities. An example follows for 9. *Wirogunan* which is shown in Figure 3.16 (line 3) as a village that has three cluster to be selected. To this end, the village of *Wirogunan* has been further divided, making the process easier on the field team. To see the data for the *Wirogunan* village, click on  followed by *vil9_yogya.csf* and ; Figure 3.17 should appear. Notice that the figure shows there are three clusters to be selected, not 25 as before, but still features 10 households per cluster. The remaining information about household size and the like is the same as in Figure 3.10.



The screenshot shows the C-Survey 2.0 software interface. The window title is "C-Survey 2.0". The main title is "C-Survey Version 2.0". The working file is "C:\PROGRAM FILES\C SURVEY2\WIL9_YOGYA.CSF". The interface has several tabs: "Survey Parameter", "Cluster Data", "Cluster Selection", "Sample Size", and "Random Number". The "Survey Parameter" tab is active, showing the following fields:

- Name of survey: Child Health Survey in Yogyakarta, Indonesia, 2006
- Primary sampling unit: Village: Wirogunan
- Size unit: Population (dropdown menu)
- Ultimate sampling unit: Household
- Eligible unit: Children 12-23 months old
- Sample size:
 - Clusters: 3
 - Households per cluster: 10
- Other informations:
 - Average no. of persons per HH: 4.50
 - Proportion of eligible persons: 0.0200
 - Proportion of eligible households: 0.0900

At the bottom, there are buttons for "OK", "Cancel", and "Print".

Figure 3.17 Sample of 3 clusters in village Wirogunan.

To see the data for the village of *Wirogunan*, click on *Cluster Data* at the top of the panel. Figure 3.18 should appear.

Other Features

There are two additional features in the *CSurvey* program that are useful for doing rapid surveys. These are a random-direction spin dial and the generation of a random number table.

In many regions of the world, households are not clearly identified or numbered. In such situations, the most common procedure for selecting a constant number of households (or eligible subjects) at the second stage is first to obtain a random start household, and then continue to the next nearest neighbor until the constant quota is met. The intention is for each household in the cluster to have an equal chance of becoming the random start household. The procedure has the surveyor start at the center of the selected village or sub-region. Next, he or she spins a dial to select a random direction to walk to the periphery of the village or sub-region (i.e., a randomly selected vector), and counts all the households passed along the directed vector (see Figure 3.20). The passed households are marked and numbered on a map form, sketched by hand in the field.

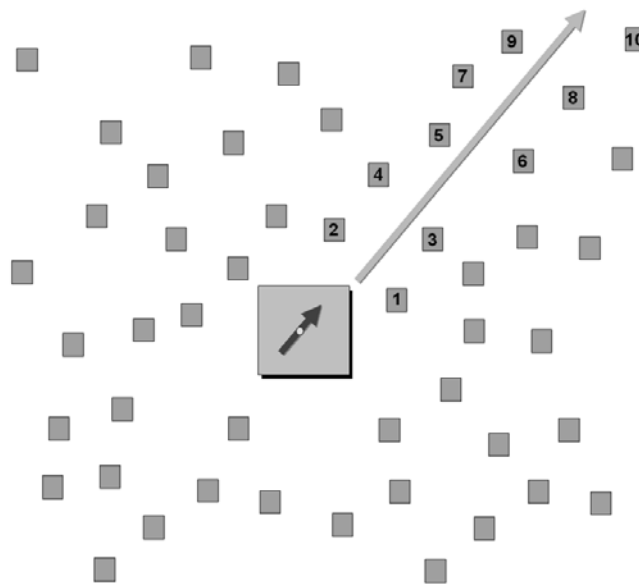


Figure 3.20 Count households along random vector to periphery.

Once all the households along the chosen vector are counted and marked on the map form, one of them is selected by sampling from a list of random numbers, specifically a number between 1 and the last house counted (i.e., #10 in the example). The selected household is deemed the *random start household* and is the starting point for obtaining the constant number of eligible households (or persons, if one eligible person per household) for the cluster.

Spin Dial. Click on *Random Number* at the top of the panels, followed by *Spin Dial* (the section to the right), as presented in Figure 3.20. Notice that the dial is divided into 8 numbered slices of a circular pie.

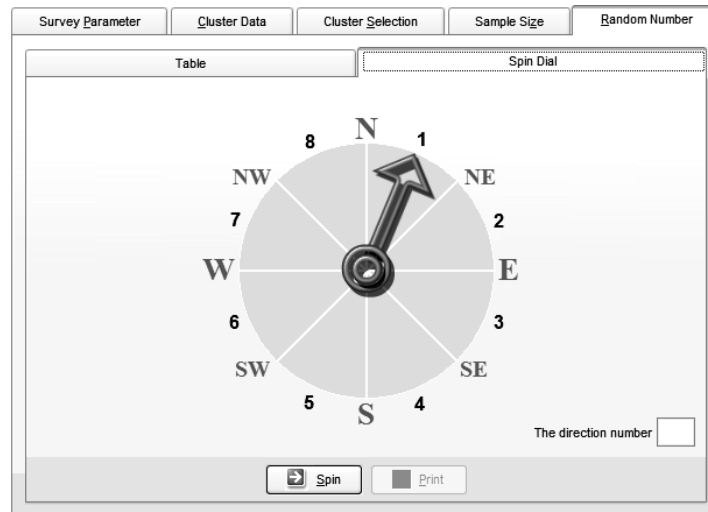
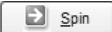
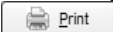
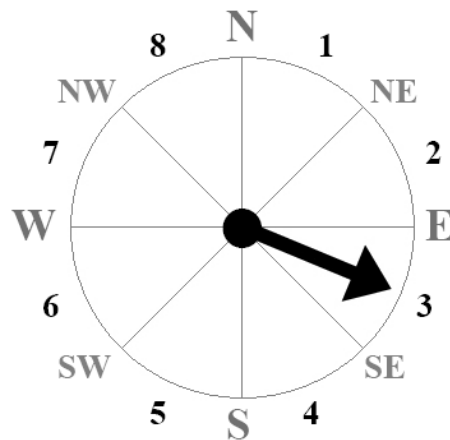


Figure 3.21 Spin dial for random direction for random-start household.

Click on  and the dial will proceed to spin and stop in a random direction, the number of which is shown at the point of the dial and in the small box at right. Once the spinning process is completed for the specific cluster (it should only be done once to conform to statistical theory), the page is printed. To do so, click on  and a random-direction spin dial image will print, similar to what is shown in Figure 3.22.



Spin Dial Direction



The direction number: 3

Figure 3.22 Printed spin dial form for cluster.

The process is repeated for all selected clusters, usually 30. The specific cluster number is

written at the top of the form and the page is given to the appropriate interviewer/examiner. With this page, the field persons needs only a small, inexpensive compass to determine the direction of the random vector. Using the compass, the field worker determines north, then walks along an imaginary line in the random direction shown on the spin dial (i.e., #3 in the example) towards the periphery of the village or sub-region. All households passed along the way are counted and listed on the map form, as previously shown in Figure 3.20.

Random Number. To select the random-start household, a table of random numbers is generated by *CSurvey* for each field team. If the villages or sub-regions are small, two-digit random numbers may be all that is needed. Conversely, if the villages or sub-regions are medium or large, then three-digit random numbers would be helpful. To generate a random number table, click on the *Random Number* screen followed by *Table*. Since the example shown in Figure 3.20 is very small, a list of two-digit random numbers from 1 to 50 would work well. To create such a list, enter 50 in the *Maximum Number* and click on to create the table of random numbers. An example of such a list is shown in Figure 3.23.

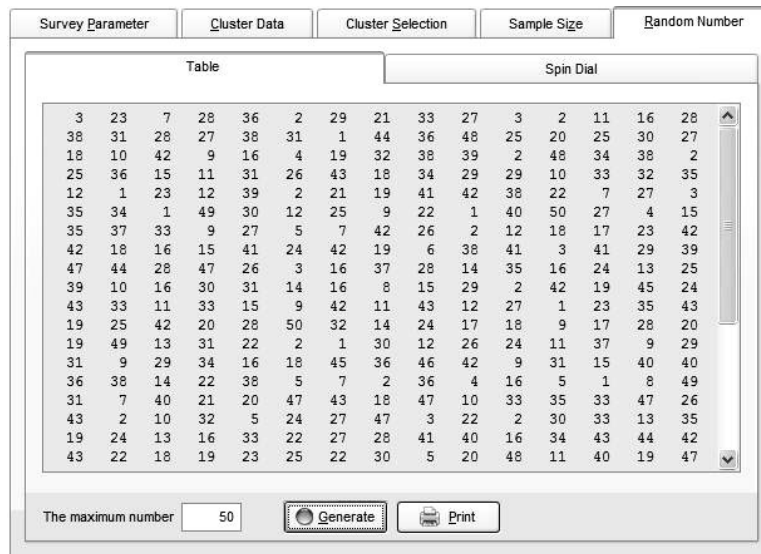


Figure 3.23 List of random numbers from 1 to 50.

This list can then be printed by clicking on . Starting at a random spot on the list, the surveyor goes down the column and over to the next until a random number between 1 and the last household along the vector to the periphery (i.e., 10) is located. For the example in Figure 3.20, a random number would be sought between 1 and 10. Assume that number was 8. If so, the household numbered as 8 on the map form would be identified as the random start household. The field team would then return to household number 8 and start searching for eligible subjects, going from one household to the next nearest neighbor, until the constant quota is reached.

For more details on rapid surveys, please visit: <http://www.ph.ucla.edu/epi/rapidsurvey.html>. This concludes the features of the *CSurvey* program.

Chapter 4: Detailed Explanation

This chapter provides a brief, but detailed, explanation of each procedure featured in *CSurvey*. For additional information on rapid surveys, please go to: <http://www.ph.ucla.edu/epi/rapidsurvey.html>.

Sample Size - Parameter Estimation

The sample size screen for parameter estimation was previously described in Chapter 3 and shown in Figure 3.6. The following descriptions use the values presented in Figure 3.6.

Values to be Entered by Investigator	
<p>Parameter Estimation</p> <p>Calculation purpose</p> <p><input checked="" type="radio"/> Test the proposed sample size</p> <p><input type="radio"/> Calculate minimum number of clusters</p> <p><input type="radio"/> Calculate average number in sample per cluster</p> <p>Estimated proportion with attribute <input type="text" value="0.2000"/></p>	<p>The investigator must estimate the proportion with the attribute in the sample population. This value is entered.</p>
<p>One-half length of confidence interval <input type="text" value="0.0500"/></p>	<p>The investigator enters the desired precision of the estimated proportion with the attribute (d). The precision is one-half the length of the confidence interval. It is equal to: $t \times se(p)$, where t is the Student's t corresponding to the desired level of confidence and $se(p)$ is the standard error of the proportion (also known as the standard deviation of the sample mean).</p>
<p>Desired level of confidence <input type="text" value="95%"/></p>	<p>The investigator sets one of three levels of confidence: 90%, 95% or 99%.</p>
<p>Homogeneity parameter <input type="text" value="Design Effect"/></p>	<p>The investigator sets the homogeneity parameter he or she intends to use. The choices are either the <i>design effect</i> which compares the variance of the cluster survey to the variance of a same-sized simple random sample, or the <i>rate of homogeneity</i> (ROH) which is a measure of the intraclass correlation coefficient.</p>

Level of homogeneity

The investigator sets the anticipated level of the homogeneity parameter. The choices are *same as a simple random sample* (i.e., either a design effect of 1.0 or the equivalent ROH), *low* (i.e., either a design effect of 2.0 or the equivalent ROH), *medium* (i.e., either a design effect of 4.0 or the equivalent ROH), *high* (i.e., either a design effect of 7.0 or the equivalent ROH), or *manual* (i.e., set by the investigator).

Average number of eligible persons per HH

The investigator either enters an estimate of the average number of eligible persons who reside in a household, or has the program provide this value based on information entered in the *Survey Parameter* screen (see Figure 3.10).

Number of clusters

The enters the number of cluster to be sampled at the first stage with probability proportionate to size (PPS) – shown here as the typical value of 30. This number should be 25 or greater to conform to statistical theory regarding an unbiased parameter estimate.

Average number of selected HHs per cluster

The investigator enters the constant number of households (or persons, if one person per eligible household) to be selected in each chosen cluster.

Once the investigator has entered the various values, the program calculates the corresponding sample values that go with the investigator's entries. As previously, the presentation is based on values shown earlier in Figure 3.6.

Values Derived by the Program

Target standard error of proportion

0.0244

Based on the entered values, the program determines the maximum standard error that would fulfil the wishes of the investigator. The value is the desired level of precision (d) divided by the value of the *Student's t* that corresponds to 1 minus the number of clusters. That is, $se(p) \leq \frac{d}{t}$.

Actual standard error of proportion

0.0243

The program calculates the standard error based on the entered values, the formula of which is:

$$se(p) = \sqrt{\frac{p q (roh(\bar{m} - 1) + 1)}{n \bar{m}}}$$

where p is the proportion with the attribute, q is $1-p$, roh is the rate of homogeneity (or intraclass correlation coefficient), \bar{m} is the mean number of persons per cluster, and n is the number of clusters.

Design effect (deff)

2.00

Based on entered value of the investigator the program calculates the design effect. If roh was entered, rather than the design effect ($deff$), the program calculates the design effect using the formula:
 $deff = roh(\bar{m} - 1) + 1$, where \bar{m} is as defined above.

Rate of homogeneity (roh)

0.0588

The *rate of homogeneity* (roh) is either the value entered by the investigator as a measure of the intraclass correlation coefficient, or is derived by: $roh = \frac{deff - 1}{\bar{m} - 1}$, where $deff$, and \bar{m} are as previously defined.

Point estimate for proportion

0.2000

The point estimate (p) was previously entered by the investigator and is again shown here.

Lower confidence limit	0.1502
Upper confidence limit	0.2498

The upper and lower confidence limits for the desired confidence interval (CI) are derived as: $p \pm [t \times se(p)]$, where p is the point estimate, t is the *Student's t* corresponding to 1 minus the number of clusters (i.e., the degrees of freedom for the analysis of a ratio estimator), and $se(p)$ is the standard error of the proportion.

Sample size for proposed cluster survey	540
---	-----

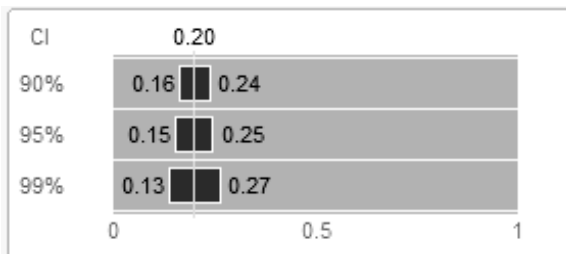
The sample size being proposed by the investigator is equal to: $n \bar{m}$, where n is the number of clusters and \bar{m} is the mean number of persons per cluster.

Is sample size adequate for stated need?	YES
--	-----

The program compares the derived $se(p)$ to the target $se(p)$ based on the wishes of the investigator and enters “yes” if $se(p) \leq \frac{d}{t}$ or

“no” if $se(p) > \frac{d}{t}$, where $se(p)$ is the

standard error of the proportion, d is one-half the length of the confidence interval and t is the value of the *Student's t* corresponding to 1 minus the number of clusters.



Finally, the program derives 90, 95 and 99% confidence intervals for the proposed sample. The formula for the confidence interval is $p \pm [t \times se(p)]$. For the example of 30 clusters (i.e., 29 degrees of freedom in the coming statistical analysis), the values of t are 1.699 for the 90% CI, 2.045 for the 95% CI and 2.756 for the 99% CI. The values of t used by the program are dependent on the number of clusters entered by the investigator. If the lower confidence limit is less than 0 or the upper confidence limit is greater than 1, the values are truncated to 0 and 1, respectively.

The program also calculates the *minimum number of clusters* that would be needed to fulfill the investigator’s wishes (assuming the average number of eligible persons per household and average number of households per cluster are included) or the *average number in sample per cluster* (assuming the average number of eligible persons per household and number of clusters are

included).

Sample Size - Hypothesis Testing

The sample size screen for hypothesis testing parameter estimation was previously described in Chapter 3 and shown in Figure 3.8. The following descriptions use the values presented in Figure 3.8.

Values to be Entered by Investigator	
Calculation purpose: <input checked="" type="radio"/> Test the proposed sample size <input type="radio"/> Calculate minimum number of clusters <input type="radio"/> Calculate average number in sample per cluster	The investigator must estimate the proportions with the attributed in the two sample populations to be compared. The program considers the absolute difference between the two proportions (i.e., $ p_2 - p_1 $), so order is not important.
Estimated value of first proportion	0.2000
Estimated value of second proportion	0.6000
One-half length of confidence interval	0.1000
Desired level of confidence	95% ▼
Homogeneity parameter	Design Effect ▼

Level of homogeneity ▼

The investigator sets the anticipated level of the homogeneity parameter for the difference between the two proportions. The choices are *same as a simple random sample* (i.e., either a design effect of 1.0 or the equivalent ROH), *low* (i.e., either a design effect of 2.0 or the equivalent ROH), *medium* (i.e., either a design effect of 4.0 or the equivalent ROH), *high* (i.e., either a design effect of 7.0 or the equivalent ROH), or *manual* (i.e., set by the investigator).

Average number of eligible persons per HH

The investigator either enters an estimate of the average number of eligible persons who reside in a household, or has the program provide this value based on information entered in the *Survey Parameter* screen (see Figure 3.10).

Number of clusters

The enters the number of cluster to be sampled at the first stage with probability proportionate to size (PPS) in the two surveys. In the example, each survey has 30 clusters selected, for a total of 60 clusters.

Average number of selected HHs per cluster

The investigator enters the constant number of households (or persons, if one person per eligible household) to be selected in each chosen cluster in the two surveys.

Once the investigator has entered the various values, the program calculates the corresponding sample values that go with the investigator's entries.

Values Derived by the Program

Target standard error of different proportion 0.0489

Based on the entered values, the program derives the maximum standard error of the difference between two proportions that would fulfil the wishes of the investigator. The value is the desired level of precision (d) divided by the value of the *Student's t* that corresponds to 1 minus the number of clusters in each survey. That is, $se(p_1 - p_2) \leq \frac{d}{t}$.

Actual standard error of different proportion 0.0471

The program calculates the standard error based on the entered values, the formula of which is:

$$se(p_2 - p_1) = \sqrt{\frac{p_1 q_1 + p_2 q_2}{n \bar{m}} \times deff},$$

where p_1 and p_2 are the two proportions with the attribute, q_1 and q_2 are $1 - p_1$ and $1 - p_2$, respectively, $deff$ is design effect, n is the number of clusters in each of the two surveys, and \bar{m} is the mean number of persons per cluster in each of the two surveys.

Design effect (deff) 2.00

Based on entered value of the investigator the program calculates the design effect. If roh was entered, rather than the design effect ($deff$), the program calculates the design effect using the formula:
 $deff = roh (\bar{m} - 1) + 1$, where \bar{m} is as defined above.

Rate of homogeneity (roh) 0.0909

The *rate of homogeneity* (roh) is either the value entered by the investigator as a measure of the intraclass correlation coefficient, or is derived by: $roh = \frac{deff - 1}{\bar{m} - 1}$, where $deff$, and \bar{m} are as previously defined.

Point estimate for different proportion 0.4000

The two point estimates (i.e., p_1 and p_2) were previously entered by the investigator and is shown here as $|p_2 - p_1|$ or $|0.60 - 0.20|$.

Lower confidence limit	0.3036
Upper confidence limit	0.4964

The upper and lower confidence limits for the desired confidence interval (CI) are derived as: $|p_2 - p_1| \pm [t \times se(p_2 - p_1)]$, where p_1 and p_2 are the two point estimates, t is the *Student's t* corresponding to 1 minus the number of clusters, and $se(p_2 - p_1)$ is the standard error of the difference between the two proportions.

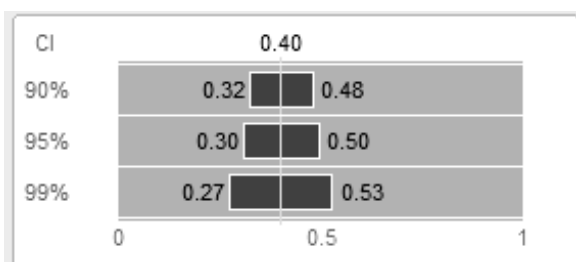
Sample size for proposed cluster survey	360
---	-----

The sample size being proposed by the investigator for each of the two cluster surveys is equal to: $n \bar{m}$, where n is the number of clusters and \bar{m} is the mean number of persons per cluster. The total in the example for the two surveys is 720.

Is sample size adequate for stated need?	YES
--	-----

The program compares the derived $se(p_2 - p_1)$ to the target $se(p_2 - p_1)$ based on the wishes of the investigator and enters “yes” if $se(p_2 - p_1) \leq \frac{d}{t}$ or “no” if

$se(p_2 - p_1) > \frac{d}{t}$, where $se(p_2 - p_1)$ is the standard error of the difference between the two proportions, d is one-half the length of the confidence interval and t is the value of the *Student's t* corresponding to 1 minus the number of clusters.



Finally, the program derives 90, 95 and 99% confidence intervals for the proposed sample. The formula for the confidence interval is $|p_2 - p_1| \pm [t \times se(p_2 - p_1)]$. For the example of 30 clusters (i.e., 29 degrees of freedom in the coming statistical analysis), the values of t are 1.699 for the 90% CI, 2.045 for the 95% CI and 2.756 for the 99% CI. The values of t used by the program are dependent on the number of clusters entered by the investigator.

The program also calculates the *minimum number of clusters* that would be needed to fulfill the

investigator's wishes (assuming the average number of eligible persons per household and average number of households per cluster are included) or the *average number in sample per cluster* (assuming the average number of eligible persons per household and number of clusters are included).

PPS sample at first stage

For rapid surveys (i.e., two stage cluster surveys), clusters (villages, communities, city blocks, etc) are selected at the first stage with probability proportionate to size. Once the population data is entered for each cluster, the program creates a cumulative list of the total sample population, and retains the location of each cluster in the cumulative list. A random number is then selected between 1 and the total sample population. The number is then assigned to the corresponding cluster in the cumulative list. The process is repeated for each of the clusters, typically 30. Hence the clusters are drawn randomly with probability proportionate to size (PPS), with replacement.

This ends Chapter 4 and the Csurvey Manual.