Software Training Manual
(Windows)

Ralph R. Frerichs, D.V.M., Dr.P.H.
Professor
Department of Epidemiology
University of California, Los Angeles (UCLA)

Rapid Survey Course
UCLA, November, 2008

TABLE OF CONTENTS

Chapter One: Epi Info and Stata

Chapter Two: Form-making

Chapter 1

EPI INFO and STATA

This training manual was last updated for the Spring Quarter 2008 UCLA course, EPI 418 *Rapid Epidemiological Surveys in Developing Countries*. It has been slightly modified for the Rapid Survey Course offered on the web. The main software programs for rapid surveys to be presented in this course is *Epi Info*. It is a shareware program (free to copy) produced by the United States Centers for Disease Control and Prevention (CDC) and distributed in collaboration with the World Health Organization (WHO). The program has been used by thousands of epidemiologists around the world, including most developing countries. The authors of the *Epi Info* program have included helpful tutorials with their program, along with an electronic version of an instruction manual.

OBTAINING SOFTWARE PROGRAM

The programs for this course can be obtained on the Internet or from a friend.

■ **Internet**. I assume you are using the *Microsoft Internet Explorer*. Once you have logged on to the world wide web, enter: *http://www.ph.ucla.edu/epi/* and the screen shown in Figure 1.1 should

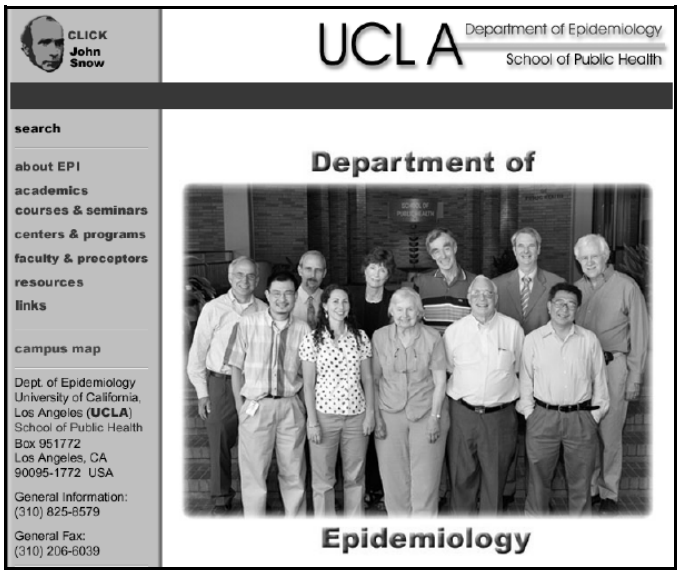


Figure 1.1 Screen for Epidemiology Department web page

appear.

Click with your mouse on *resources* in the column at left, then when the new page appears scroll down to *software* and click on it. When you do, the screen presented in Figure 1.2 should appear, showing a list of software programs that can be down loaded from the Department of Epidemiology website. You should be at http://www.ph.ucla.edu/epi/software.html. Only a few of the programs are actually stored at UCLA. The web page has instructions, however, that link you along the electronic highway to another computer where the software is stored. Such a computer is termed a "file server" or simply, a "server." The first software to be obtained is *Epi Info*. To do so, left click with the mouse *EpiInfo (Windows)*, then click on *Downloads*, and the screen in Figure 1.3

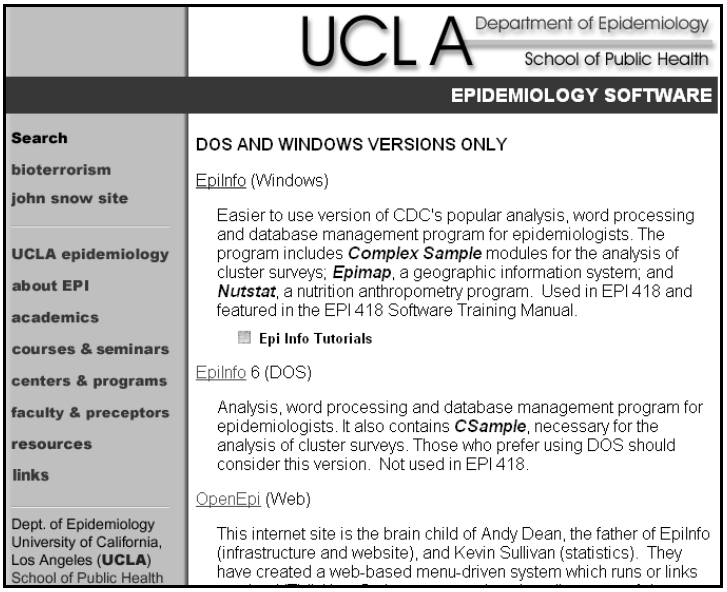


Figure 1.2 Screen for public domain software

should appear. Notice that the program is now at the CDC address.

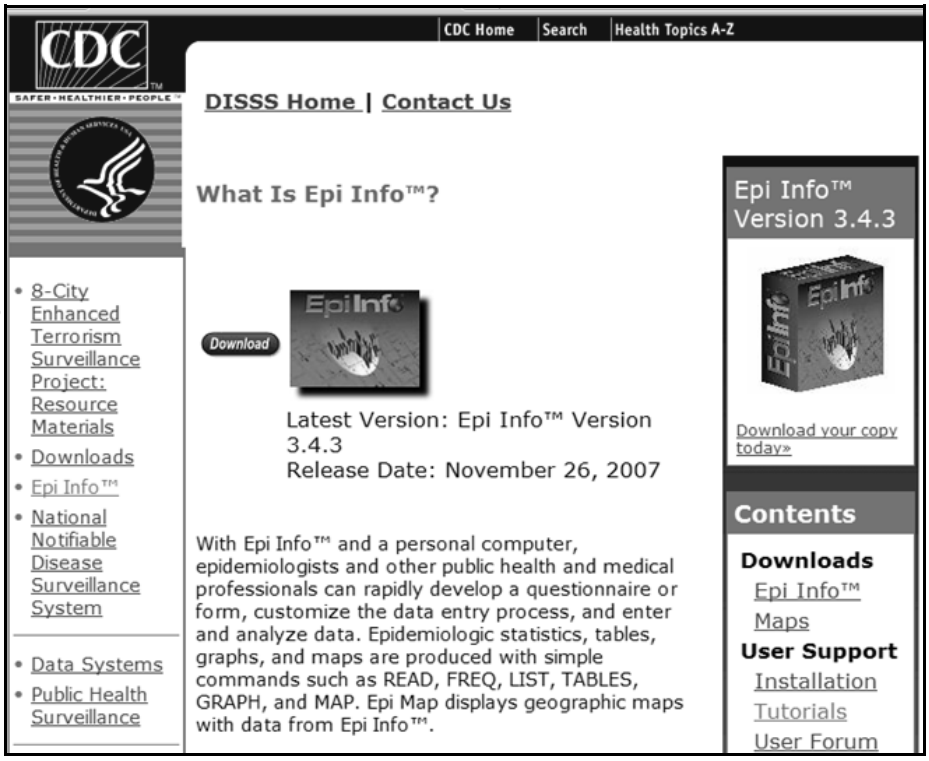


Figure 1.3 Location of Epi Info software

Left click with your mouse on *Download*, then *Download* again, and then either *Web Install* or *Download Setup* to transfer the program through your modem to your computer. When completed, the Epi Info icon  should appear on your main computer screen. Later, you will click on the icon to start the Epi Info program.

Csurvey. In addition to *Epi Info*, you should obtain the *Csurvey 2.0* program. This *Windows* program automates several steps necessary for doing rapid surveys. In collaboration with Professor

Frerichs, the program was written by Muhammad N. Farid, a graduate student in the Department of Epidemiology, sponsored by the Fogarty International HIV/AIDS Training Program. An earlier DOS version, also in collaboration with Dr. Frerichs, was written by Iwan Ariawan, M.D., M.P.H., a former graduate student in Epidemiology, now on the faculty of the University of Indonesia. When through with getting *EpiInfo*, return to the Epidemiology Department software web site by left clicking with your mouse on **<-Back** at the top of your screen. Move down the screen to *Csurvey 2.0 Windows* and with your mouse, left click on **Csurvey**. The next screen will appear, as shown in Figure 1.4. More your cursor down to the Windows Version section at the bottom. Down load the program, as before, by left clicking with the mouse. Save the file on your C: drive in a subdirectory named "download." Use zip program if necessary. Note that these are DOS programs (rather than Windows), having been written some while ago. To install the program on your computer, change directories to C:\download\ and enter **install**. The program will automatically create a directory C:\CSURVEY on your computer and copy the necessary files.

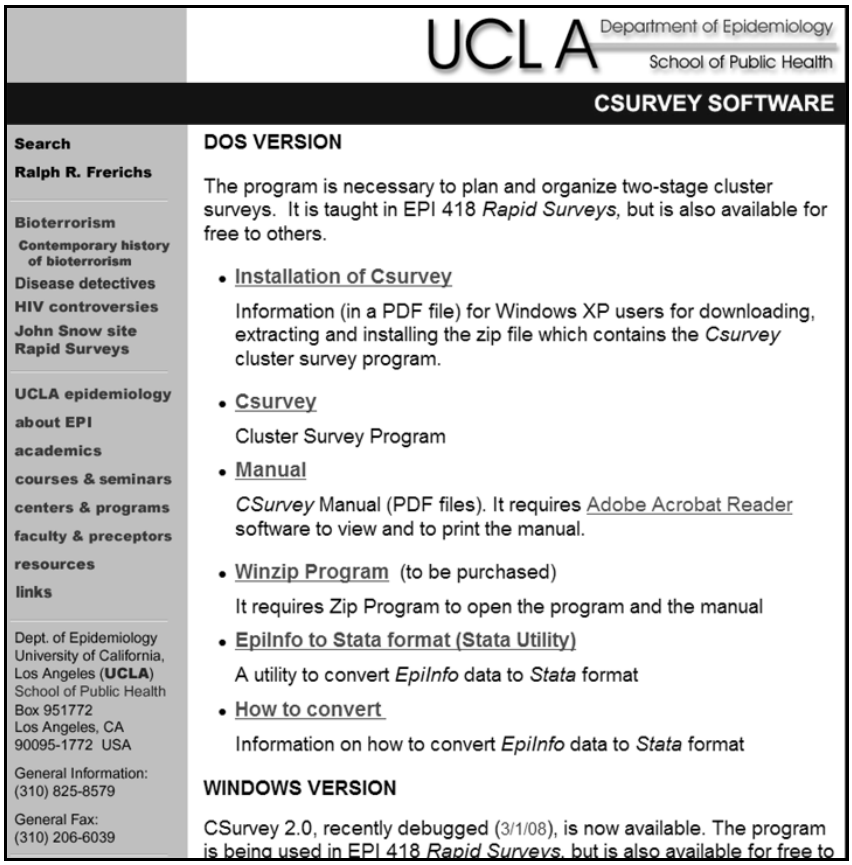


Figure 1.4 Csurvey and Epi2dct.exe programs

Epi2dct.exe. This small program allows you to convert data that are entered in *Epi Info* into a file format that is compatible with *Stata*. It is found in *Epi Info to Stata Format* section of the software linkage at the UCLA website: http://www.ph.ucla.edu/epi/csurvey.html (see Figure 1-4).

Stata. This program does multivariate analyses well beyond the capacity of the *Epi Info* program. *Stata* has a set of survey modules that permit the analysis of two-stage cluster surveys, like those featured in the Rapid Survey Course. The program and user manuals can be purchased from Stata Corporation. More details are presented on the Rapid Survey Course website: http://www.ph.ucla.edu/epi/rapidsurveys/RScourse/RSstmanual.html.

INTRODUCTION

This exercise requires both imagination and patience. Imagine that a community-based survey was done in the rural regions of a developing country to obtain information for an AIDS intervention program. With patience, proceed through the pages of this teaching exercise and try to learn the strengths and weaknesses of the *Epi Info* program for entering, editing and analyzing the survey findings.

Assume that a two-stage cluster survey was done last September of knowledge of AIDS, occurrence of injection practices and various forms of sexual activity, and the prevalence of HIV infection as measured by HIV antibodies in saliva.[1] Three hundred men, aged 20 through 39 years, were included in a sample of 360 housing units from a population of 93,250 housing units, then interviewed and asked for saliva specimens. The investigators who created the present study were interested in learning what people believe about AIDS and AIDS prevention; the prevalence of high-risk injection practices, sexual activity and HIV infection; and the association between current infection and various risk factors. They reasoned that with this information, they would 1) have some idea as to how quickly HIV infection is spreading through the population, 2) be able to provide information for planning a health education program, and 3) have baseline information to evaluate HIV control measures.

QUESTIONS TO BE ANSWERED

Specifically the investigators were interested in answering the following questions:

1. Do young and middle-aged men at the village level know that friends and neighbors could be infected with the AIDS virus but not have the AIDS disease, that there is no vaccine to prevent AIDS infection, and there is no drug available to cure a person with AIDS disease?

2. How effective do men feel are various devices or methods for preventing AIDS infection? Included are the use of a diaphragm or condom, having a vasectomy, and limiting sexual intercourse to two people who do not have the AIDS virus.

3. What percentage of men during the past year were injected with a needle, received a blood transfusion, or had their skin pierced for some other reason such as acupuncture or a tattoo?

4. What proportion of men during the past month had vaginal and rectal sexual intercourse with either a single partner or two or more partners?

5. What is the prevalence of HIV infection based on HIV antibodies in saliva?

6. Does sexual behavior and injection practices predict the prevalence of HIV antibodies?

---

[1] Frerichs, R.R., Htoon, M.T., Eskes, N. and Lwin, S.: Comparison of saliva and serum for HIV surveillance in developing countries. *The Lancet* 340: 1496-1499, 1992.
Frerichs, R.R., Eskes, N. and Htoon, M.T.: Validity of three assays for HIV-1 antibodies in saliva. *Journal of Acquired Immune Deficiency Syndrome* 7(5), 522-524, 1994.
Frerichs, R.R., Silarug, N. Eskes, N. Pagcharoenpol, P., Rodklai, A. Thangsupachai, S. and Wongba, C.: Saliva-based HIV antibody testing in Thailand. *AIDS* 8: 885-894, 1994.

■ **Complete Data Set.** The data file, *aidsal.mdb*, with information on all 300 men in the 360 households, is available at: http://www.ph.ucla.edu/epi/rapidsurveys/RScourse/RSstmanual.html. This is a realistic data set but does not contain real data. Instead it is intended only for teaching purposes. Since this is a rapid survey, the questionnaire is limited to 24 variables that can be listed on two pages. You will soon see that even two pages contain a substantial amount of information which requires time to analyze. By understanding how long everything takes, you will be more effective at convincing those seeking information that "less is more." That is, they will have more useful information readily available for decision making, if only they can limit the number of questions being asked.

In the coming pages, I will first present the questionnaire used in the survey (see Figure 1.5). You will then use a shortened version of the questionnaire to program the *Epi Info* software to enter and analyze survey findings. Next you will enter data for 20 subjects, followed by the analysis of several questions. Following that, you will use the program's statistics calculator to analyze entered numbers. Finally, you will analyze data in the *aidsal.mdb* using the cluster and regular analysis features of *Epi Info*.

Figure 1.5
HIV/AIDS risk factor questionnaire

```
Department of Epidemiology
University of California at Los Angeles
Los Angeles, California

                    AIDS RISK FACTOR CLUSTER SURVEY
_____________________________________________________________________

 Complete for all men, aged 20-39, now living in the household.  Tell each:

  1) that some of the questions are about his personal life so you will want
     to speak to him in private,
  2) the information will be used to help plan services for his community, and
  3) no one will know his identity since his name will not be written on the
     interview form.

 1. Study No. __ ___ __   2. Region No. __ __ __   3. Cluster No. __ __

 4. Household No. __ __    5. Subject No. in HH  __ __

 6. Age __ __ years (99 if Unk.)

 7. Married with wife in household    [1] Yes   [2] No   [9] Unknown or No response

 REPEAT FOR QUESTIONS 8-10   Do you believe...

 8. there is a vaccine available that protects a person from getting the AIDS virus?

                [1] Yes      [2] No       [3] Don't know     [9] No response

 9. person can be infected with the AIDS virus and not have the disease AIDS?

                [1] Yes      [2] No       [3] Don't know     [9] No response

10. there is a drug available that can cure a person with AIDS disease?

           [1] Yes      [2] No       [3] Don't know     [9] No response
```

Figure 1.5 HIV/AIDS risk factor questionnaire (continued)

AIDS RISK FACTOR CLUSTER SURVEY (continued)

REPEAT FOR QUESTIONS 11-14

How effective do you think is ... for preventing AIDS disease through sexual activity?

11. using a diaphragm

[1] Very effective [2] Somewhat effective [3] Not at all effective

[4] Don't know how effective [5] Don't know method [9] No response

12. using a condom

[1] Very effective [2] Somewhat effective [3] Not at all effective

[4] Don't know how effective [5] Don't know method [9] No response

13. having a vasectomy

[1] Very effective [2] Somewhat effective [3] Not at all effective

[4] Don't know how effective [5] Don't know method [9] No response

14. sexual intercourse only between two people who do not have the AIDS virus

[1] Very effective [2] Somewhat effective [3] Not at all effective

[4] Don't know how effective [5] Don't know method [9] No response

REPEAT FOR QUESTIONS 15-17 During the past year...

15. Have you received an injection with a needle in your muscle, vein or skin?

[1] Yes [2] No [3] Don't know [9] No response

16. Have you received a transfusion of blood or blood components (platelets or plasma)?

[1] Yes [2] No [3] Don't know [9] No response

17. Not counting injections or transfusions mentioned previously, have you had any part of your body pierced - by acupuncture, by tatoo, or having your ears, nose or nipples pierced, or something like that?

[1] Yes [2] No [3] Don't know [9] No response

REPEAT FOR QUESTIONS 18-21 During the past month...

18. Have you had sexual intercourse during which you put your penis in your partner's vagina?

[1] Yes [2] No [3] Don't know [9] No response

19. If "yes," have you done this during the past month with more than one partner?

[1] Yes [2] No [3] Don't know [9] No response

20. Have you had sexual intercourse during which you put your penis in your partner's rectum?

[1] Yes [2] No [3] Don't know [9] No response

21. If "yes," have you done this during the past month with more than one partner?

[1] Yes [2] No [3] Don't know [9] No response

22. Was a saliva specimen collected from this subject?

[1] Yes [2] No

23. Results of HIV antibody assay(laboratory findings)

[1] Positive [2] Negative [3] Indeterminant [9] No specimen

Figure 1.5
HIV/AIDS risk factor questionnaire (continued)

```
                    AIDS RISK FACTOR CLUSTER SURVEY (continued)
_____________________________________________________________________

 ┌──────────────────────────────────────────────────────────────────┐
 │ This concludes the interview.  Thank you for taking the time to participate │
 └──────────────────────────────────────────────────────────────────┘

24. Code number of interviewer  __ __  (in unknown, enter 99)
```

This will be our first survey so the Study Number will be 001. The target population is all men, aged 20-39 years in Region 234 of the country. Based on existing census records, we estimated that there are 548,529 people in 510 communities or villages (termed *clusters*) potentially accessible to our interviewers. These people live in 111,900 households, with an average of 4.90 persons per household. We further estimated that about 83 percent of the households have at least one man, aged 20-39 years. At the first stage of our two-part sampling process, we sampled 30 of the 510 clusters with probability proportionate to the number of households in the cluster. This is termed *probability proportionate to size* (PPS) sampling, and will be further explained in the workshop. In each cluster, we randomly select 12 households and interview all men, aged 20-39, living in these households. Included in the sample was 300 men in the 360 selected households.

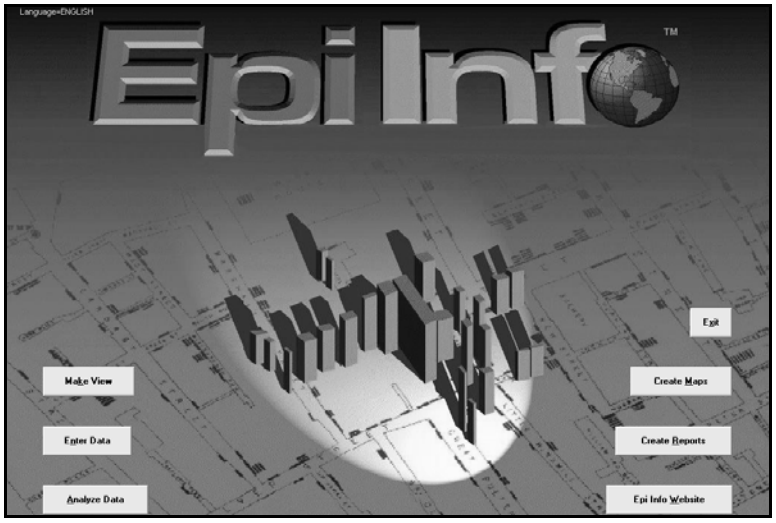
Look over the questionnaire. <u>All</u> variables to be entered into the computer must have a number and name. You also should give thought to how you want to present the findings. With *Epi Info* you will be making an entry screen, entering some data, and with the complete *aidsal.mdb* data set (to be provided), doing the initial analysis.

■ Overview of Epi Info.

Epi Info tends to be self-explanatory with many helpful message appearing here and there. To start the program, click on the screen icon 

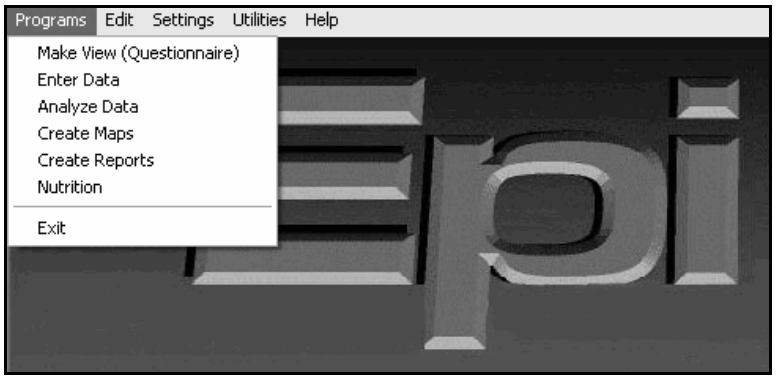
and the screen in Figure 1.6 should appear. The top row shows the various components of the program. We will briefly explore each.

Figure 1.6
Initial menu



Move the cursor with your mouse and click on *Programs*. You should see the menu shown in Figure 1.7.

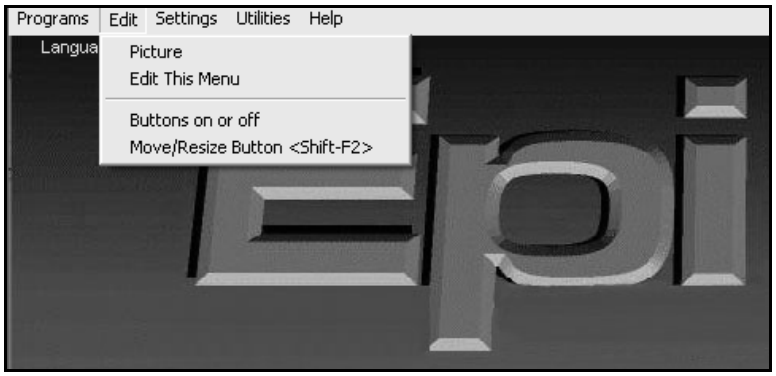
Figure 1.7 Programs menu



In this exercise, you will be using *Make View*, *Enter Data* and *Analyze Data*, but not until after we have looked at some of the other features in this program. You will return to this menu, showing the main programs, many times.

Next move the cursor to *Edit* by pressing the right arrow key [→] and the menu in Figure 1.8 appears.

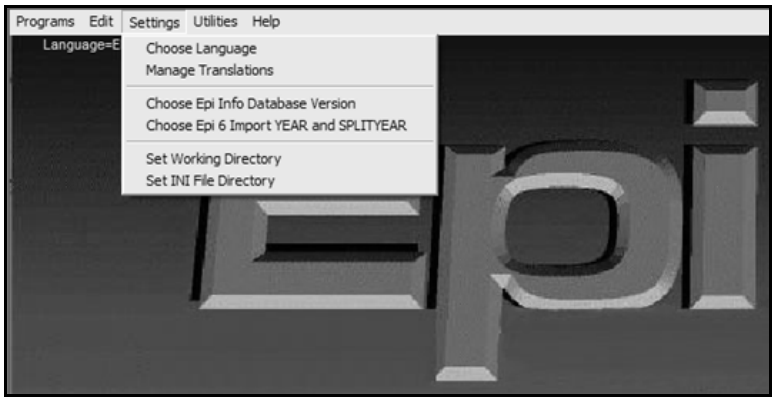
Figure 1.8 Edit menu



This provides editing functions that you will later explore on your own, once you become more familiar with the program.

Now move the cursor to *Settings*, either with your mouse or by pressing the right arrow key [→], and the menu in Figure 1.9 appears.

Figure 1.9 Settings menu



This menu gives you the option of choosing a Epi Info database version. To do so, move the cursor to *Choose Epi Info Database Version* and make sure that the option shown in Figure 1.10 is selected.

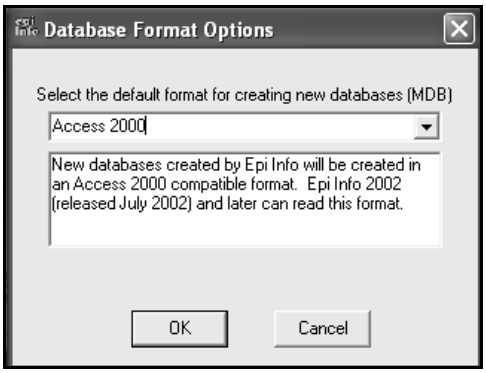


Figure 1.10 Settings menu

Create a subdirectory in your computer under *c:\Epi_Info* named *418*. This will become your working directory for the course. Once the subdirectory has been created, click on *Settings* and then *Set Working Directory*, and move the cursor to *418* as shown in Figure 1.11. Click *OK* when done.

Figure 1.11 Settings menu

The next set of programs in *Epi Info* are the utilities. Move the cursor to *Utilities* and the screen shown in Figure 1.12 should appear.

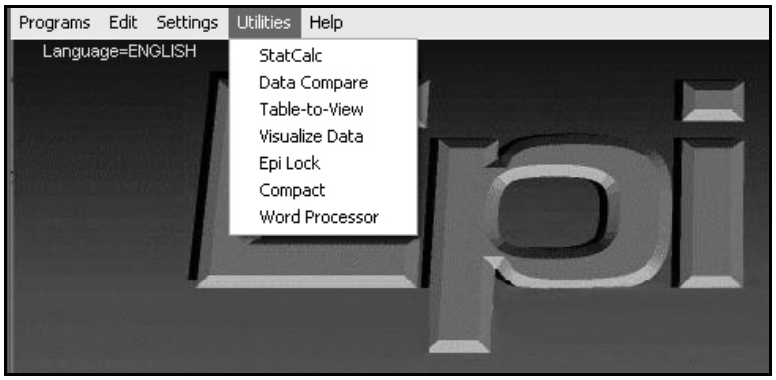


Figure 1.12 Utilities menu

Here are two programs that we will be using in this manual, namely *StatCalc* and possibly *Word Processor*, although regarding the latter, it is more likely that you will want to use a regular word

processing program of your own choosing. More will come later on the use of a word processor and on *StatCalc*.

Finally, move the cursor to *Help*, as shown in Figure 1.13.

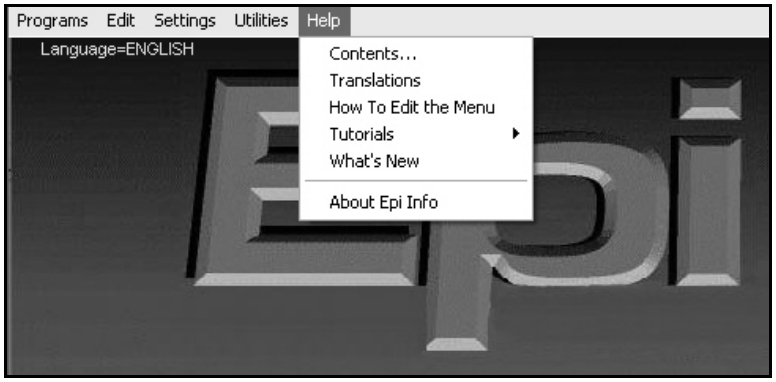


Figure 1.13 Help menu

The *Contents* tells you all about *Epi Info*, including overviews of the different components of the program. In this regard it is like a manual, but in your computer rather than in a book. While we will be using the English version of *Epi Info*, other languages are either available for planned, as explained in the *Translations* section. Besides the example of a cluster sample featured in this manual, there are three other tutorials in *Epi Info*. To see them, click on *Tutorials*. The first is for an acute outbreak investigation of a food-bourne pathogen occurring in Oswego County, New York. The second is also an outbreak investigation, but in a hospital setting following open-heart surgery. The third tutorial is for a surveillance system, showing how case records are computerized and tallied. Note that none of the three tutorials deal with cluster surveys, which are the subject of this *Software Training Manual*.

CREATING THE QUESTIONNAIRE

When doing an interview, you will need to have several pages before you with all of the questions clearly presented, along with options for the answers. To create such a questionnaire you typically use a word processing program, or if you have no favorite program available, the *Word Processor* in *Epi Info*. Once the information is collected, you will want to transfer the data to a computer using a data entry screen. To this end, you will using *Make View* create a shorter version of the questionnaire, appropriate for data entry.

First if doing a field survey and wanting to use the *Epi Info* word processor, you would return to the Utilities menu and click on *Word Processor*. Then you would enter the questionnaire text shown in Figure 1-5, as presented in Figure 1.14. You would typically print this for the field staff as the survey instrument.

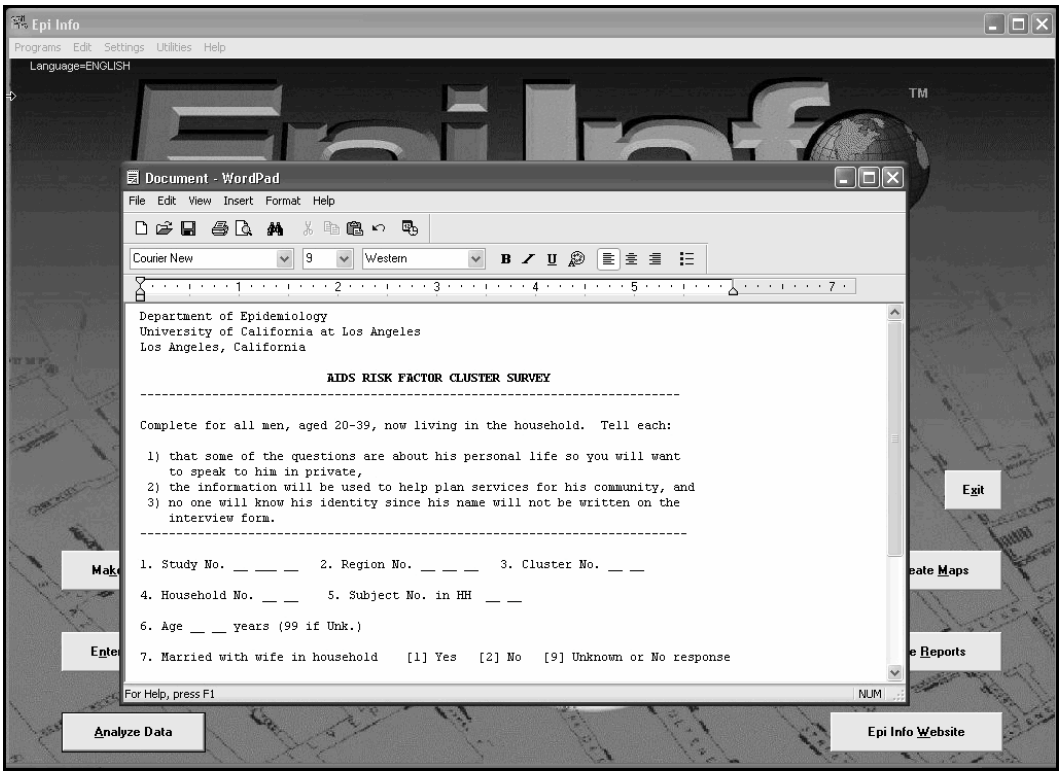


Figure 1.14 Create questionnaire for field use.

■ **Abbreviated Data Set.** Rather than starting with the larger data set, we will begin with data on only a few questions, and limited to men in the 13 sampled households in Clusters 1 and 2. The abbreviated questionnaire is shown in Figure 1.15.

```
Department of Epidemiology
University of California at Los Angeles
Los Angeles, California

                    AIDS RISK FACTOR CLUSTER SURVEY

1. Cluster No. __ __   2. HH No. __ __   3. Person No. __ __   4. Age __ __ yrs.

5. Married with wife in household    [1] Yes   [2] No   [9] Unknown or No response

REPEAT FOR QUESTIONS 6-8.   Do you believe...

6. there is a vaccine available that protects a person from getting the AIDS virus?

        [1] Yes      [2] No       [3] Don't know     [9] No response

7. a person can be infected with the AIDS virus and not have the disease AIDS?

        [1] Yes      [2] No       [3] Don't know     [9] No response

8. there is a drug available that can cure a person with AIDS disease?

        [1] Yes      [2] No       [3] Don't know     [9] No response
```

Figure 1.15 Complete text for short questionnaire

The short names of the eight variables and their characteristics for Epi Info's *Make View* program are shown in Table 1.1. You will be using the data shown in Table 1.2. First, however, we need to create the data entry screen using *Make View*.

Table 1.1 Data labels and characteristics for Make View program

No.	Short description	Name	Digits	Font	Size
	AIDS RISK FACTOR CLUSTER SURVEY			Arial	12 Bold
1	Cluster Number	Cluster	2	Arial	12 Regular
2	Household Number	HH	2	Arial	12 Regular
3	Person Number	PN	2	Arial	12 Regular
4	Age	Age	2	Arial	12 Regular
5	Married with wife in HH	Married	1	Arial	12 Regular
	Do you believe...			Arial	12 Bold
6.	available vaccine	vaccine	1	Arial	12 Regular
7.	infected but no disease	infected	1	Arial	12 Regular
8.	available drug to cure	drug	1	Arial	12 Regular

Table 1.2 Data for Make View entry screen

CLUSTER	HH	PN	AGE	MARRIED	VACCINE	INFECTED	DRUG
1	1	1	23	1	1	2	2
1	2	1	37	1	2	1	2
1	3	1	27	1	1	1	1
1	4	1	23	1	2	3	1
1	5	0					
1	6	1	25	2	1	2	1
1	7	1	26	1	1	2	1
1	8	0					
1	9	1	39	1	2	1	2
1	10	1	35	1	2	2	1
1	11	0					
1	12	1	35	1	2	1	1
1	13	1	27	1	2	1	1
2	1	1	37	1	1	2	2
2	2	1	34	2	3	2	3
2	3	0					
2	4	1	36	1	1	1	2
2	5	0					
2	6	1	28	1	1	3	1
2	7	1	26	1	1	1	2
2	8	1					
2	9	1	28	1	1	2	2
2	10						
2	11	1	26	1	1	1	2
2	12	1	28	1	1	1	1
2	13	1	39	1	1	1	3
2	13	2	20	2	1	2	2

DATA ENTRY

To enter the data shown earlier in Table 1.1, you need an entry screen. This is created using the *Make View* program of *Epi Info*. You will first enter an abbreviated version of the questionnaire for data entry. The intention here is to have enough words showing to remind person entering the data of the variable field, but not so many to clutter up the entry screen. First you should enter the title and then a short name for the various items or questions, with just enough information to remind the person entering the data which field is to be considered.

To start, click on *Make View*, either on the box at the left side of the screen or in *Programs* at

the top of the screen. When the *Make/Edit View* screen appears, click at the top on *File* and then *New*. Create a file name *aidsex1*, which should be stored in c:\Epi_Info\418\ as shown in Figure 1.16. This file will hold a database, *aidsex1.mdb*, once you have entered the data.

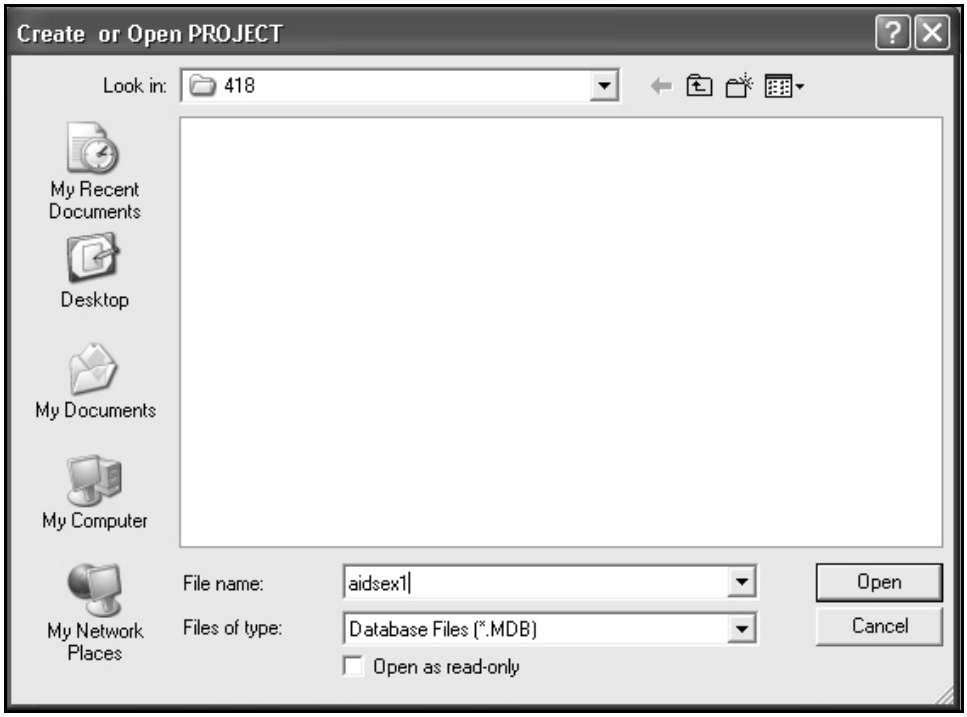


Figure 1.16 Create data entry file.

Every page in *Make View* is termed a *view*. We will be using only one page, but it still needs to be named. For our example, name the view A as seen in Figure 1-17. Click OK to continue.

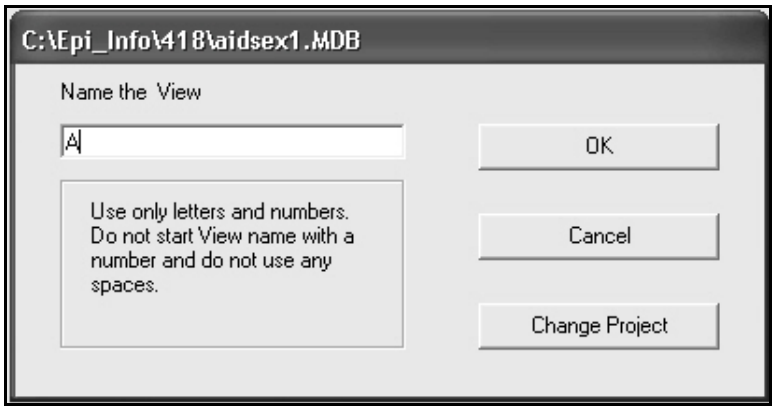


Figure 1.17 Create data entry file.

The first field that you will be entering is not a variable but rather, a label which presents the study name. The screen should read **Right click to create a field**. Towards the left border of the screen, click on the right side of the mouse. Enter the title of our survey, as shown in Figure 1.18, making the font Arial 12 (click on *Font for Prompt*) and the style of the field as *Label/Title*. Since we will not be entering information using this line, it is considered merely as a label or a title. Enter OK when done. Move the title with your mouse (hold down the left mouse key) and move it to the upper left corner, as far as it will go.

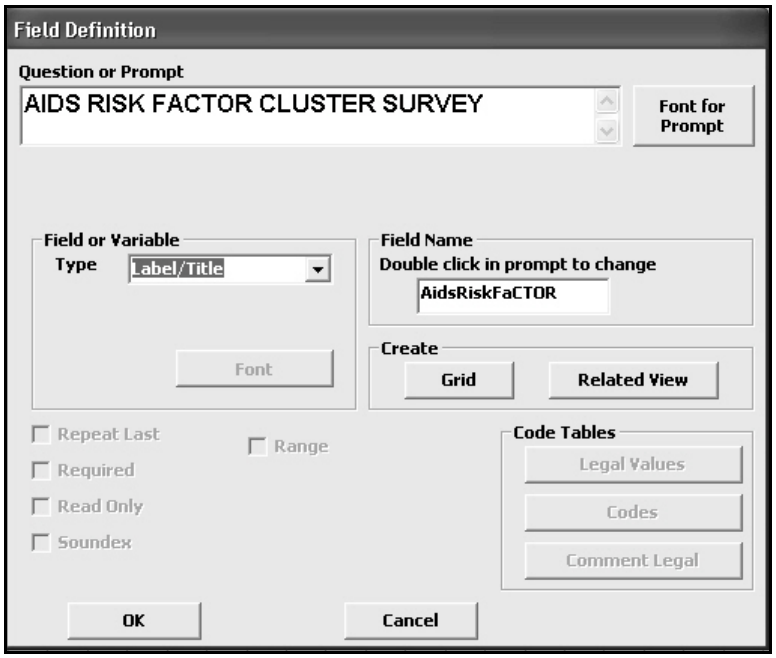


Figure 1.18 Create first entry as a label or title.

The first data field that you will be entering is the cluster number which requires two digits. The variable is to be named *cluster* for the data set but identified as *1. Cluster Number* for the data entry screen, as seen in Figure 1.19. Notice that the number field has two digits, signified by ##. The variable name is *cluster* and the font should be Arial, 12 point, regular (see Table 1.1).

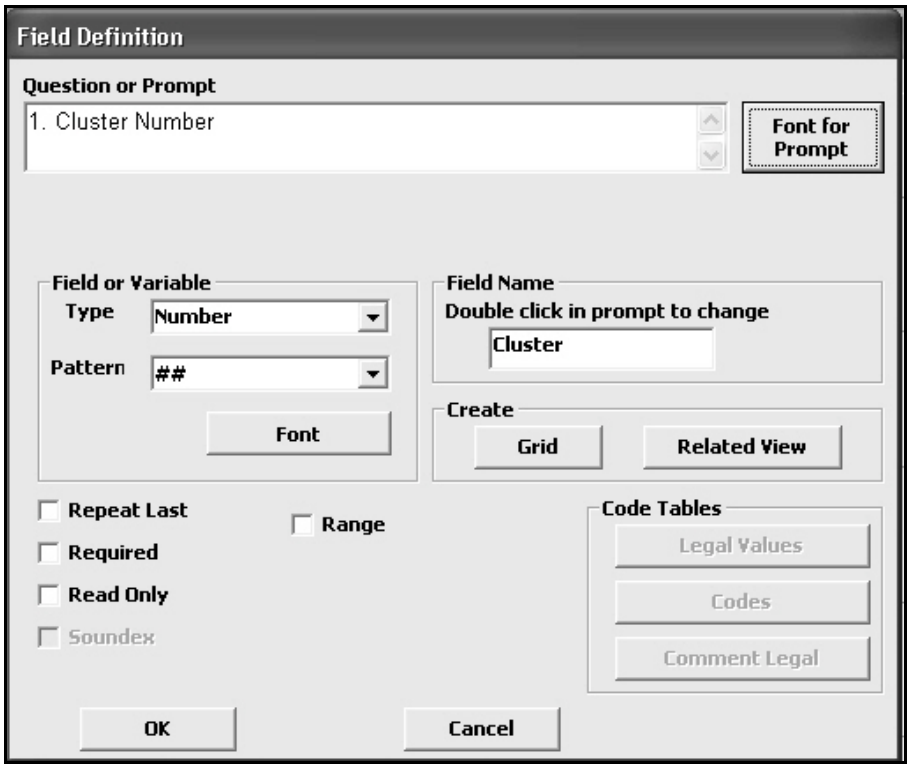


Figure 1.19 Create entry for first variable.

Continue to enter the information for the seven remaining variables and the second label, as presented earlier in Table 1.1. When through your *Make View* screen should resemble Figure 1.20.

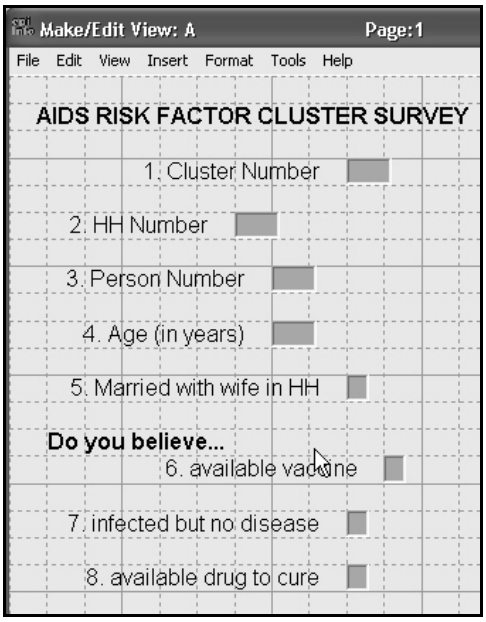


Figure 1.20 Created data fields for data entry.

While all the information is there, the entry screen looks somewhat jumbled. To arrange in better order, hold down the left mouse key, and place the ensuing box around the 10 lines of information. Release the mouse key and move to the top of screen, click *format*, then *alignment*, followed by *vertical*. The Make View screen should now appear as in Figure 1.21.

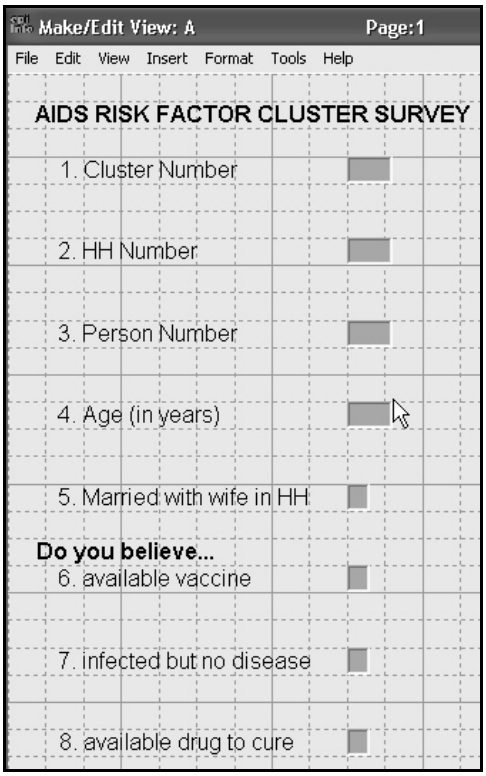


Figure 1.21 Aligned data fields for data entry.

Notice in Figure 1.21 that four of the variables have space for two digits and four have space for only one digit. If this is not so with your *Make View* screen, go back and straighten the variable fields out before continuing. When satisfied, click on *File* and then *Save* to save the *Make Screen* file *aidsex1.mdb*.

■ **Abbreviated Data Set.** Rather than starting with the larger data set, we will begin with data on only a few questions, and limited to men in the 13 sampled households in Clusters 1 and 2. The abbreviated questionnaire was shown in Figure 1.15.

Return to the initial *Epi Info* menu (see Figure 1.6) and click on *Enter Data*, followed by *File* (see the top line of the screen) and *Open*. If you had properly set the program so that it opens in C:\Epi_Info\418\, then the screen in Figure 1.22 should appear.

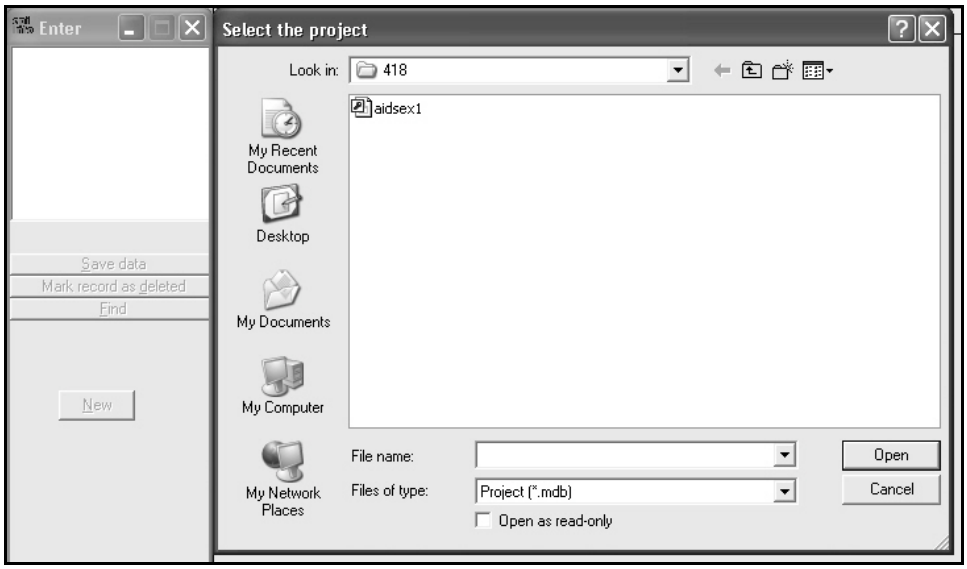


Figure 1.22 Open file for data entry.

Click with your left mouse on *Open* and on table *A*, followed by *OK*. The same screen that was presented in Figure 1.21 should now appear, ready for data entry.

Return for a moment to Table 1.2 and note the information on the first sampled household:

Table 1.2 Data for Make View entry screen

CLUSTER	HH	PN	AGE	MARRIED	VACCINE	INFECTED	DRUG
First household in cluster 1...							
1	1	1	23	1	1	2	2

Remember that *cluster* has two digits. Thus when you enter *1*, it will appear as *01*. Enter each of the numbers into the appropriate fields on the screen, followed each time by *[Enter]* (i.e., the "Enter" key). Stop after entering 2 for *Drug* but before tapping the *[Enter]* key. Your screen should appear as in Figure 1.23.

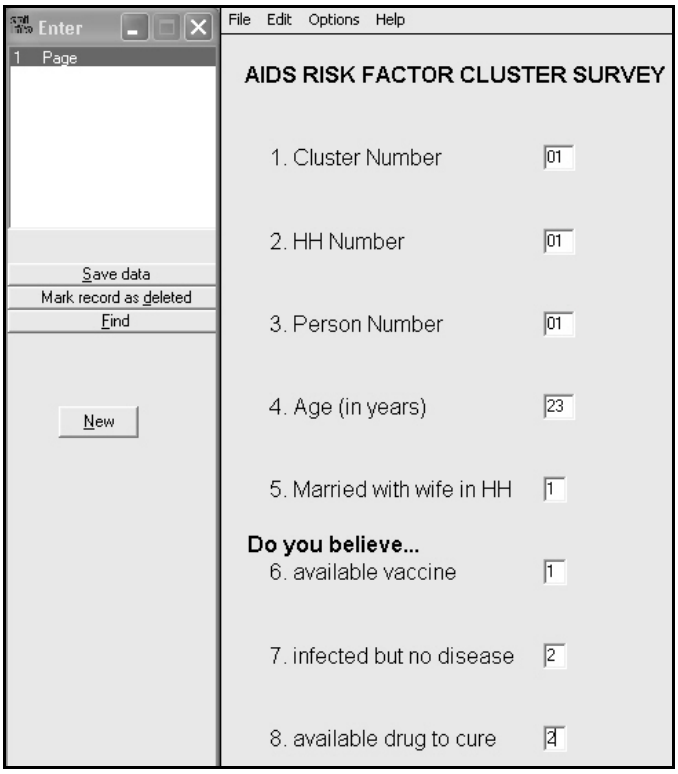


Figure 1.23 Data for first subject.

Press *[Enter]* and the data for the first household are entered into the computer, followed by a blank entry screen, ready for data for the next subject. Notice that some of the households did not have eligible subjects. Thus the data fields for them are left blank. The first such HH with no eligible subject is number 5 which should be keyed as 1, 5, 0 and then blanks. Proceed to enter the remaining data in Table 1.2 until you get to the last field of the last household.

Table 1.2 Data for Make View entry screen

CLUSTER	HH	PN	AGE	MARRIED	VACCINE	INFECTED	DRUG
Last household in cluster 2...							
2	13	2	20	2	1	2	2

Notice if your lose track of where you are, the record number is shown at the bottom left corner of the screen; for example, here is what it looks like for record 6 

.

Just before entering the last value for the last HH in cluster 2 (i.e., subject #27), stop again; do not press *[Enter]*. The screen should appear as in Figure 1.24.

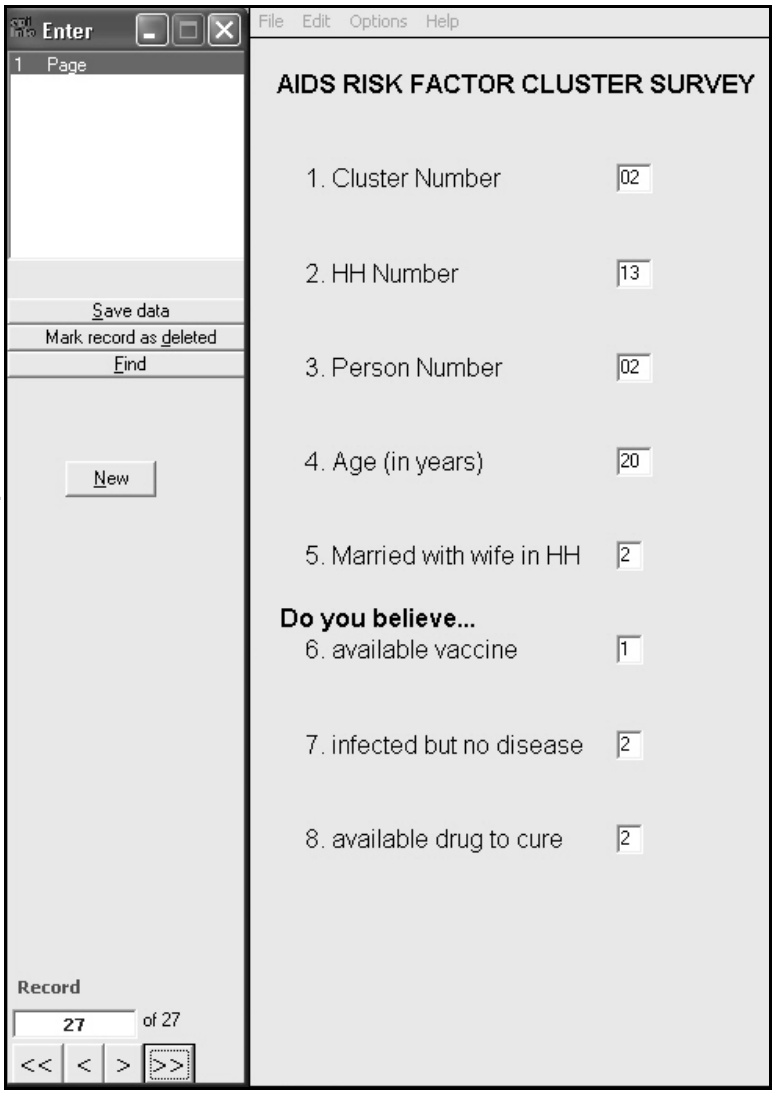


Figure 1.24 Data for first subject.

If your screen shows that you are entering data for the 27th subject and the values are as shown, press *[Enter]*. Save the data by clicking with your left mouse as shown in Figure 1.25.

Figure 1.25 Save data on 27 subjects.

To make sure that you entered the data correctly, or want to make changes, click on << in the bottom left of the screen to return to record 1 as shown here 

. Scroll through the various entry screening by pressing > and make the necessary changes, if any. When done, click on [x] at the top left of the screen, thereby closing the *Enter Data* program.

Return to the main menu to proceed to the analysis.

ANALYSIS WITH EPI INFO

The data analysis module in *Epi Info* is very flexible and allows you to do many things. We will explore only a few options here. In the main menu click with the left mouse on *Analyze Data*, then in the column at left, click on *Read (import)*. Change the Data Source by clicking with your left mouse on ... , then enter: *Epi_info\418\aidsex1.mdb*. Finally, click in Views on *ViewA*, as shown in Figure 1.26.

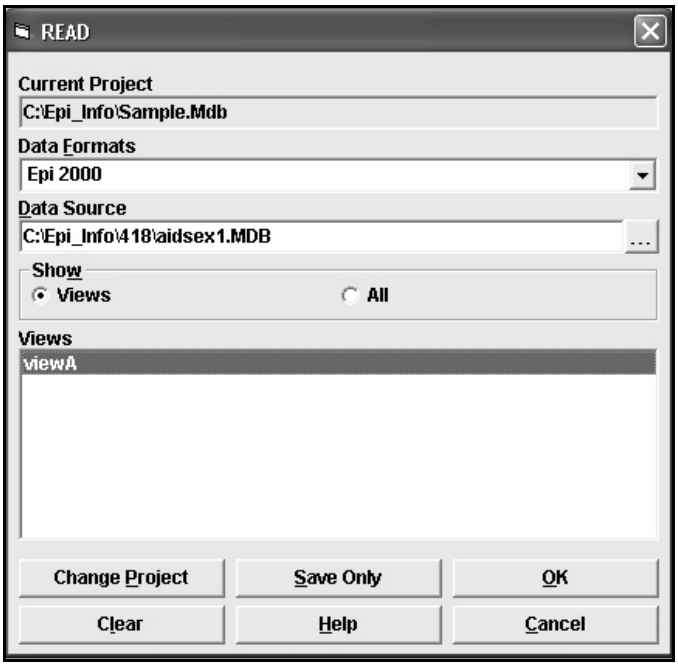
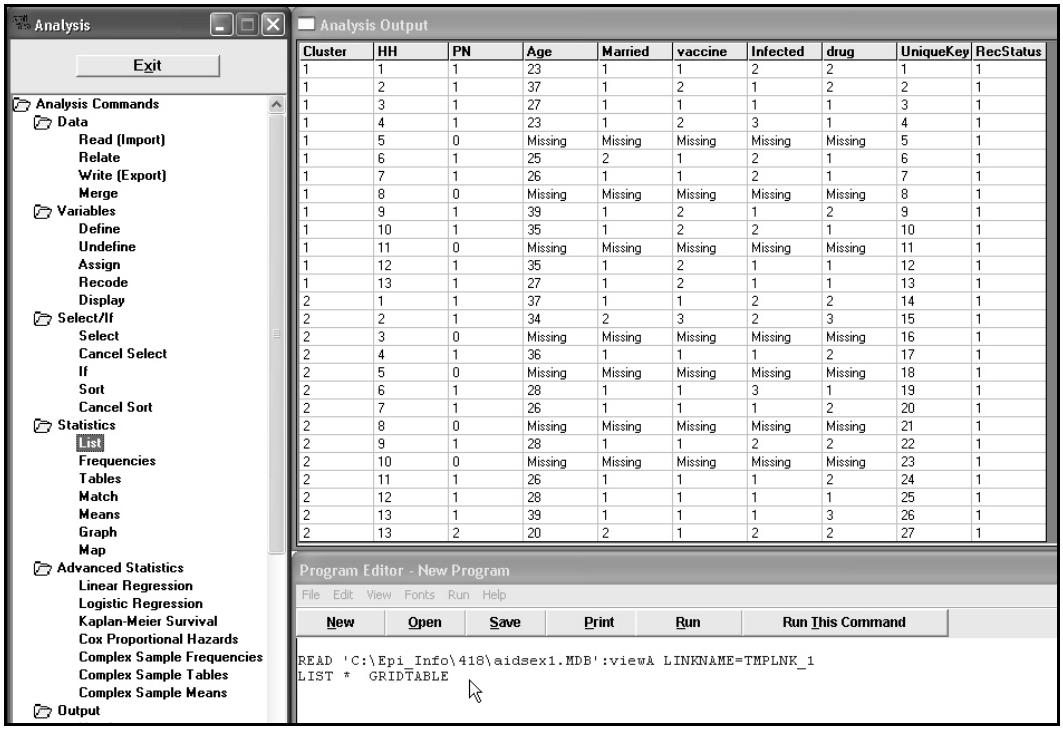


Figure 1.26 Read file with data for analysis.

A screen appears that mentions a temporary link and shows TMPLNK1. Click *OK*. Your screen should now state that you have 27 records in *C:\Epi_Info\418\aidsex1.MDB:viewA*. The program editor at the bottom right of the screen should show that your entered the instruction *READ* followed by details of the command. As you proceed with your analysis, each step will be recorded in the *Program Editor*.

■ **List data.** In the *Statistics* section, the first thing that we will do is list the data to make sure that they have been properly entered. To do so, click with the left mouse key on *List*. In the box that appears, click ☑ All (*) Except followed by *OK*. The screen should now show a grid table with all of the data, as seen in Figure 1.27.



Cluster	HH	PN	Age	Married	vaccine	Infected	drug	UniqueKey	RecStatus
1	1	1	23	1	1	2	2	1	1
1	2	1	37	1	2	1	2	2	1
1	3	1	27	1	1	1	1	3	1
1	4	1	23	1	2	3	1	4	1
1	5	0	Missing	Missing	Missing	Missing	Missing	5	1
1	6	1	25	2	1	2	1	6	1
1	7	1	26	1	1	2	1	7	1
1	8	0	Missing	Missing	Missing	Missing	Missing	8	1
1	9	1	39	1	2	1	2	9	1
1	10	1	35	1	2	2	1	10	1
1	11	0	Missing	Missing	Missing	Missing	Missing	11	1
1	12	1	35	1	2	1	1	12	1
1	13	1	27	1	2	1	1	13	1
2	1	1	37	1	1	2	2	14	1
2	2	1	34	2	3	2	3	15	1
2	3	0	Missing	Missing	Missing	Missing	Missing	16	1
2	4	1	36	1	1	1	2	17	1
2	5	0	Missing	Missing	Missing	Missing	Missing	18	1
2	6	1	28	1	1	3	1	19	1
2	7	1	26	1	1	1	2	20	1
2	8	0	Missing	Missing	Missing	Missing	Missing	21	1
2	9	1	28	1	1	2	2	22	1
2	10	0	Missing	Missing	Missing	Missing	Missing	23	1
2	11	1	26	1	1	1	2	24	1
2	12	1	28	1	1	1	1	25	1
2	13	1	39	1	1	1	3	26	1
2	13	2	20	2	1	2	2	27	1

Figure 1.27 List of 27 records in data file.

Notice that the data set contains 26 households, 7 of which have no eligible men (i.e., aged 20-39), leaving 19 households with eligible men. One household (cluster 2, household 13) had two eligible men. Thus the total number of records is 27 [i.e., (25 x 1) + (1 x2)] and the total number of records with data for the different variables is 20 [i.e., (18 x 1) + (1 x2)].

■ **Frequencies.** Next you will do a frequency distribution of the responses to Question 5 on marital status. The program command is *Frequencies* in the column at left, under Statistics. When clicking on this program, a panel appears that asks which variable is to be included. Click on ▼, then move the cursor to *Married*, as presented in Figure 1.28, and click with the left mouse. *Married* should appear in the selected box.

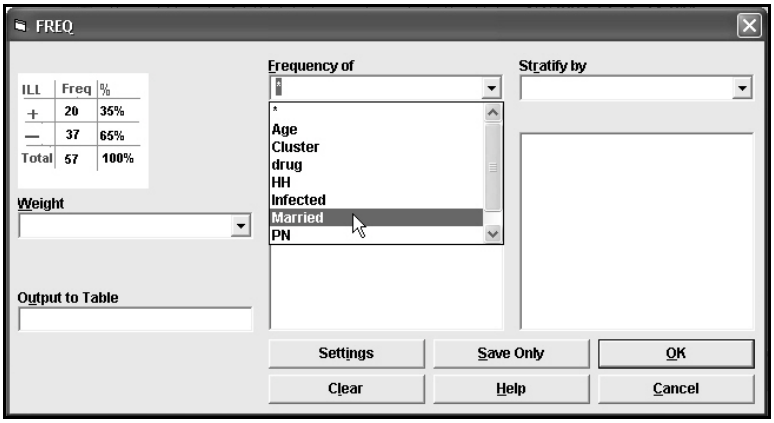


Figure 1.28 Request for frequency of variable "married."

Press OK and Figure 1.29 appears. Notice in the bottom box by the mouse arrow that the *Epi Info* command for frequencies is *FREQ* followed by the variable *married*. This is the same command structure as in the DOS version of *Epi Info*.

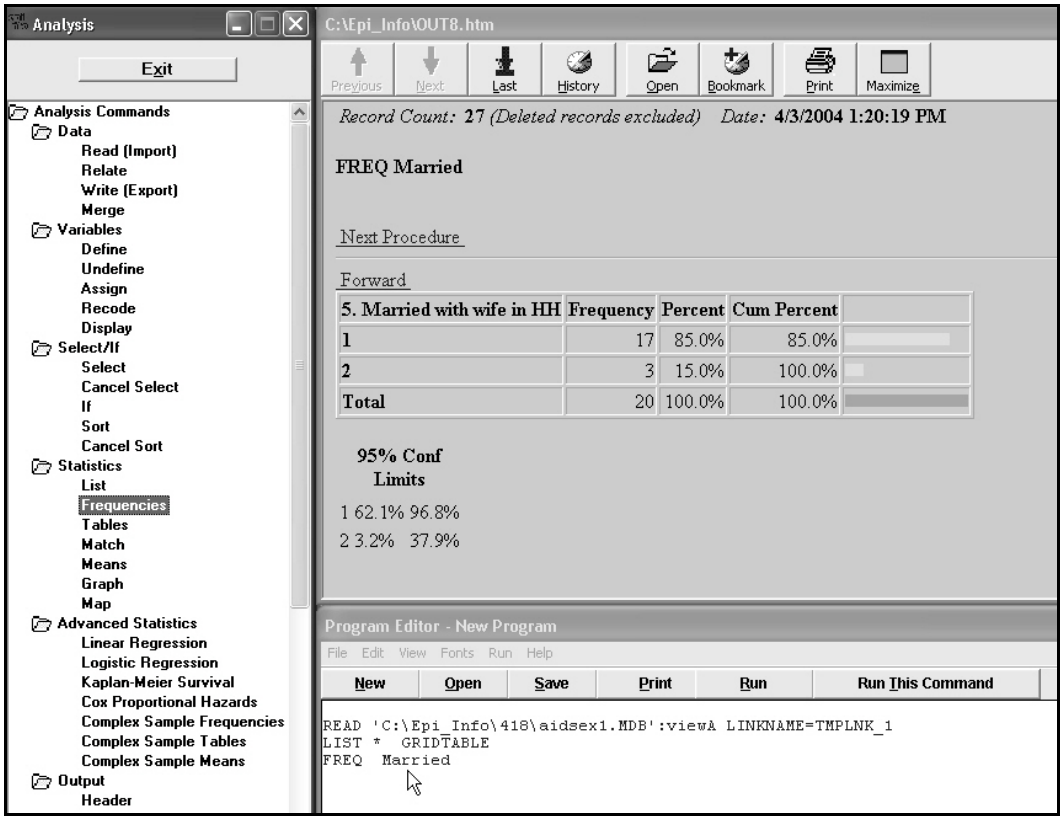


Figure 1.29 Frequency distribution of *married*.

Eighty-five percent of the 20 men in the 26 households were married with a wife present, while 15 percent were not. None refused or did not answer. The frequency distribution includes a 95% confidence interval for both percent *married* (i.e., 62.1%, 96.8%) and *not married* (i.e., 3.2%, 37.9%) . **Disregard this information**. The confidence intervals in the FREQ program assume the data were collected in a survey featuring simple random sample rather than two-stage cluster sampling. For the latter, the confidence intervals tend to be much wider, as you will learn later. The frequency distribution, however, is applicable for all kinds of sampling.

Next do a frequency of the variable *vaccine* to see how the 20 men responded to the question, *Do you believe there is a vaccine available that protects a person from getting the AIDS virus?* As before click on *Frequencies*, then in the *Frequencies of* box select *vaccine*. The results should be as shown in Figure 1.31. This time there are three categories of outcome: [1] *Yes*, [2] *No*, and [3] *Don't know*. A fourth category, [9] *No response*, was not used by any of the respondents. Only 30 percent (i.e., 6) of the 20 subjects recognized that a vaccine was not available to protect against AIDS.

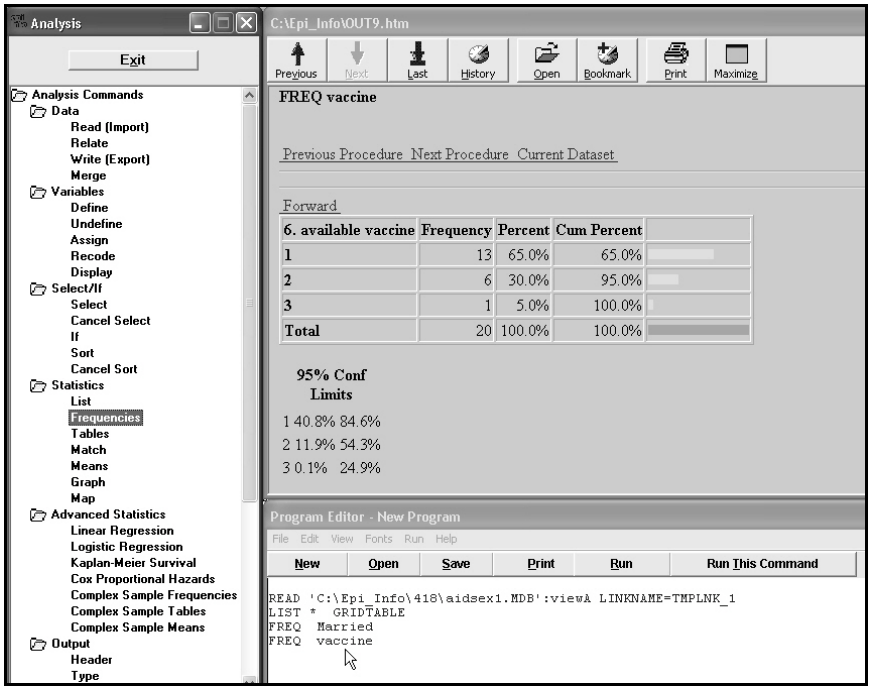


Figure 1.30 Frequency distribution of *vaccine*.

■ **Tables.** The question arises, are single men less knowledgeable about an AIDS vaccine than married men? The appropriate analysis to answer this question is a cross-tabulation of *married* and *vaccine*. To create such cross-tabulation table, select under *Statistics* the program *Tables*. In this instance the exposure variable is *married* and the outcome variable is *vaccine*. That is, we want to determine if "exposure" to marriage has an effect on the "outcome" of belief that there is a vaccine. For the findings, see Figure 1.31.

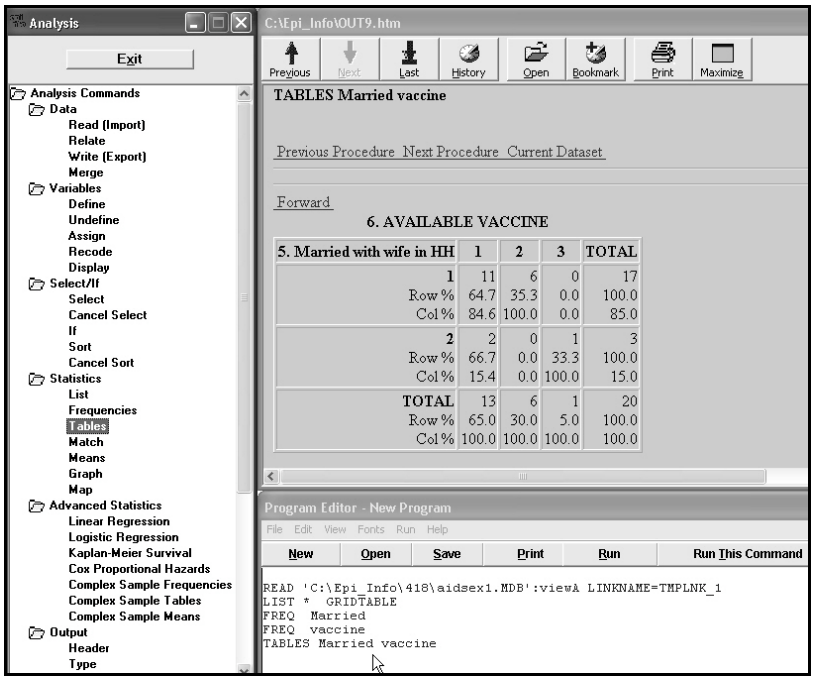


Figure 1.31 Cross-tabulation of *Married* and *vaccine*.

■ **If-then.** As seen in Figure 1.31, there was one person who responded *I don't know* to the vaccine

question. If we want to limit the analysis to those who had a definite opinion (i.e., either responded *yes* or *no*), we need to temporarily remove the code [3] response to *vaccine* from the data. *Epi Info* lets you do this with various recoding statements, one of which is an *if-then statement*. The structure is, "if *vaccine* is equal to 3, then *vaccine* should be recoded as missing ." To create an if -then statement, click on *Select/if* in the Analysis Commands column, then click on *if*. Click on available variables and select *vaccine*. Next click on = and end by entering *3*. In the box labeled *Then*, enter *vaccine=(.)* as shown in Figure 1.32.

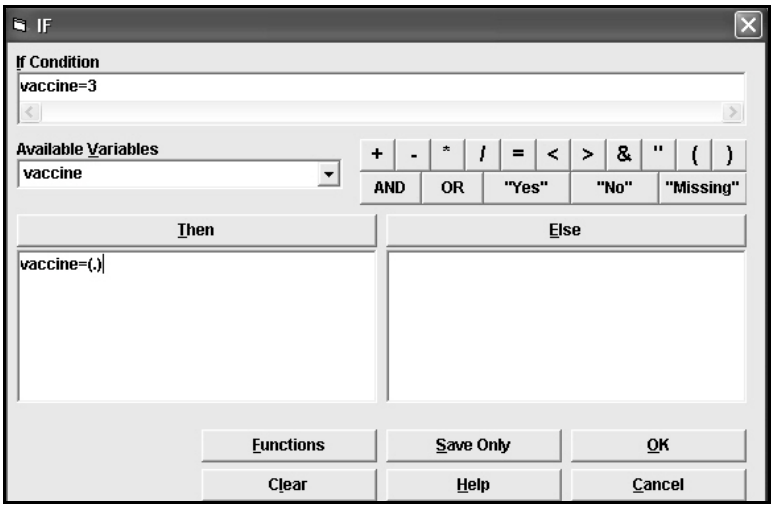


Figure 1.32 Create if-then statement to limit *vaccine* to "yes" or "no.

Click *OK*. Note that program statement has been added to the *Program Editor* box. With vaccine limited to "yes" or "no" responses, you will run the tables program again. Click on *Tables* under *Statistics* in the *Analysis Commands* column and enter *Married* and *vaccine* as before. The new table is shown in Figure 1.33.

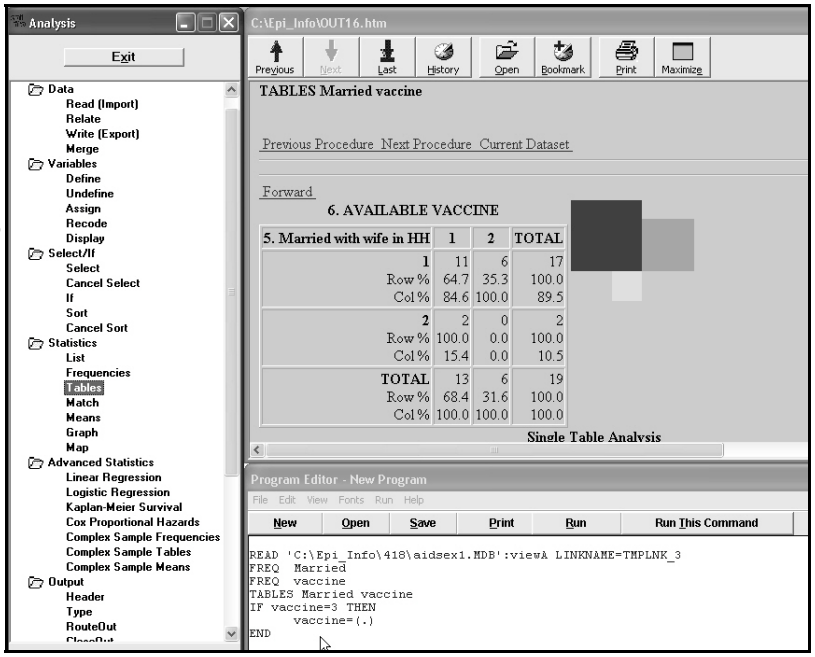


Figure 1.33 Knowledge of *vaccine* among *married.*

■ **Odds and Risk ratios.** Notice that by comparing two dicotomous (i.e., two category) variables,

married and *vaccine*, you have created a four-fold table and the analysis program derives various epidemiological statistics. These statistics are revealed, when scrolling down the output page, as shown in Figure 1.34.

Figure 1.34 Odds and risk ratios for association between *married* and *vaccine*.

Single Table Analysis

Warning: The expected values of a cell is <5. Fisher Exact Test should be used.

	Point Estimate	95% Confidence Interval Lower	Upper	
PARAMETERS: Odds-based				
Odds Ratio (cross product)	0.0000	Undefined	Undefined	(T)
Odds Ratio (MLE)	0.0000	0.0000	7.6742	(M)
		0.0000	11.8762	(F)
PARAMETERS: Risk-based				
Risk Ratio (RR)	0.6471	0.4555	0.9192	(T)
Risk Difference (RD%)	-35.2941	-58.0113	-12.5769	(T)

(T=Taylor series; C=Cornfield; M=Mid-P; F=Fisher Exact)

STATISTICAL TESTS	Chi-square	1-tailed p	2-tailed p
Chi square - uncorrected	1.0317		0.3097665756
Chi square - Mantel-Haenszel	0.9774		0.3228483885
Chi square - corrected (Yates)	0.0448		0.8324138365
Mid-p exact		0.2280701754	
Fisher exact		0.4561403509	

Since one of the cells contained a zero, the odds ratio is also zero. The risk ratio of 0.65 indicates that married men are 35 percent less likely to believe that an AIDS vaccine is available than single men. The 95% confidence interval and the various statistical tests are inappropriate with our data set since the information comes from a two-stage cluster survey with different variance estimates. The statistical tests in this section of *Epi Info* assume the data were collected in a simple random sample, with each subject being independent from others. This assumption is not valid in cluster surveys, although the risk and odds ratios are valid.

■ **Means.** For the final analysis, you will determine if those who believe in the availability of a vaccine (i.e., answered *yes*) are different in age from those who responded *no*. Age is a continuous variable. Therefore rather than requesting a table, as is done for categorical data, you should use the *means* command. To do so, click on *Means* in the *Statistics* section of the *Analysis Commands* column and enter Means of *Age* cross-tabulated by *vaccine*. The results in the long analysis section are shown in Figure 1.35.

Figure 1.35 Statistics with *means* output for *Age* and *vaccine*.

Means age vaccine

Previous Procedure Next Procedure Current Dataset

Forward

6. AVAILABLE VACCINE

4. Age (in years)	1	2	TOTAL
20	1	0	1
Row %	100.0	0.0	100.0
Col %	7.7	0.0	5.3
23	1	1	2
Row %	50.0	50.0	100.0
Col %	7.7	16.7	10.5
25	1	0	1
Row %	100.0	0.0	100.0
Col %	7.7	0.0	5.3
26	3	0	3
Row %	100.0	0.0	100.0
Col %	23.1	0.0	15.8
27	1	1	2
Row %	50.0	50.0	100.0
Col %	7.7	16.7	10.5
28	3	0	3
Row %	100.0	0.0	100.0
Col %	23.1	0.0	15.8
35	0	2	2
Row %	0.0	100.0	100.0
Col %	0.0	33.3	10.5
36	1	0	1
Row %	100.0	0.0	100.0
Col %	7.7	0.0	5.3
37	1	1	2
Row %	50.0	50.0	100.0
Col %	7.7	16.7	10.5
39	1	1	2
Row %	50.0	50.0	100.0
Col %	7.7	16.7	10.5
TOTAL	13	6	19
Row %	68.4	31.6	100.0
Col %	100.0	100.0	100.0

Descriptive Statistics for Each Value of Crosstab Variable

	Obs	Total	Mean	Variance	Std Dev
1	13	369.0000	28.3846	31.2564	5.5907
2	6	196.0000	32.6667	39.0667	6.2503

	Minimum	25%	Median	75%	Maximum	Mode
1	20.0000	26.0000	27.0000	28.0000	39.0000	26.0000
2	23.0000	27.0000	35.0000	37.0000	39.0000	35.0000

ANOVA, a Parametric Test for Inequality of Population Means

(For normally distributed data only)

Variation	SS	df	MS	F statistic
Between	75.2740	1	75.2740	2.2434
Within	570.4103	17	33.5535	
Total	645.6842	18		

T Statistic = 1.4978

P-value = 0.1525

Bartlett's Test for Inequality of Population Variances

Bartlett's chi square= 0.0841 df=1 P value=0.7718

A small p-value (e.g., less than 0.05) suggests that the variances

are not homogeneous and that the ANOVA may not be appropriate.

Mann-Whitney/Wilcoxon Two-Sample Test (Kruskal-Wallis test for two groups)

Kruskal-Wallis H (equivalent to Chi square) = 1.3150

Degrees of freedom = 1

P value = 0.2515

Persons who believe in the availability of an AIDS vaccine are 4.3 years younger than men who do not believe that such a vaccine exists (i.e., mean of 28.4 years versus mean of 32.7 years). If this had been a simple random sample, the analysis of variance (ANOVA) statistics would have been appropriate, suggesting the difference is not statistically significant. Since the findings come from a cluster survey, however, the statistical tests in this section of *Epi Info* should not be used. The means, however, are valid.

■ **Statistics Calculator.** Another analytic feature of the *Epi Info* program is the *Statcalc* program. This has long been one of my favorite components of the program, and is useful for analyzing data a wide variety of epidemiologic data. Go to the Utilities menu of *Epi Info* as shown in Figure 1.36 and click with the left mouse on *Statcalc*.

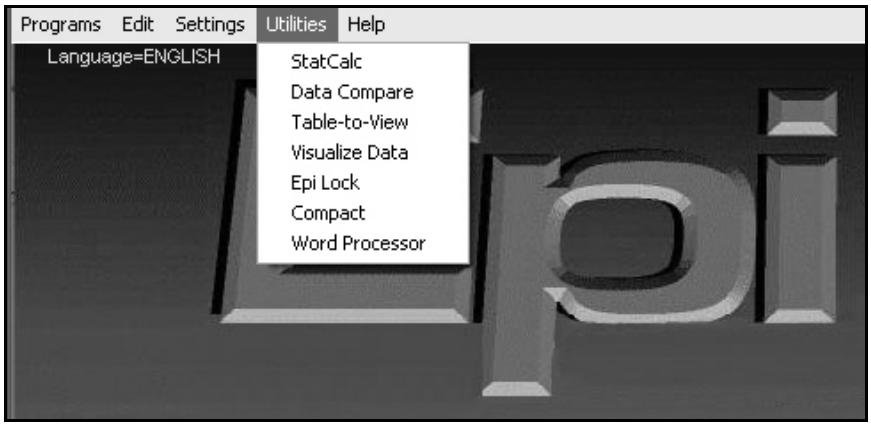


Figure 1.36 Program menu showing *StatCalc* program

Assume that you have available the following numbers for an analysis relating the *drug* question (i.e., Do you believe there is a drug available that can cure a person with AIDS disease?) to the *condom* question (i.e., How effective do you think is using a condom for preventing AIDS disease through sexual activity?), stratified by marital status.

		Married				Single		
		Believed Effectiveness of Condoms for Preventing AIDS						
		Effect	Other			Effect	Other	
Drug available	Yes	86	70	156	Yes	19	17	36
	No	27	27	54	No	11	7	18
		113	97	210		30	24	54

Rather than going through the involved steps of entering the data on 264 persons into the computer and doing the analysis as before, all you want is a simple calculation of measures of association for the available data. As you will see next, *StatCalc* is very useful for this. To use the program, press [**Enter**] and Figure 1.37 appears.

Figure 1.37 *StatCalc* opening menu

Move the cursor to *Tables (2 x 2, 2 x n)* and press [**Enter**] to start the program. Figure 1.38 should appear with an empty table for cross-tabulations. Notice that the outcome (or dependent) variable is listed as *disease* and the risk (or independent) variable is listed as *exposure*. In our example, *Condom* is the *disease* variable and *drug* is the *exposure* variable.

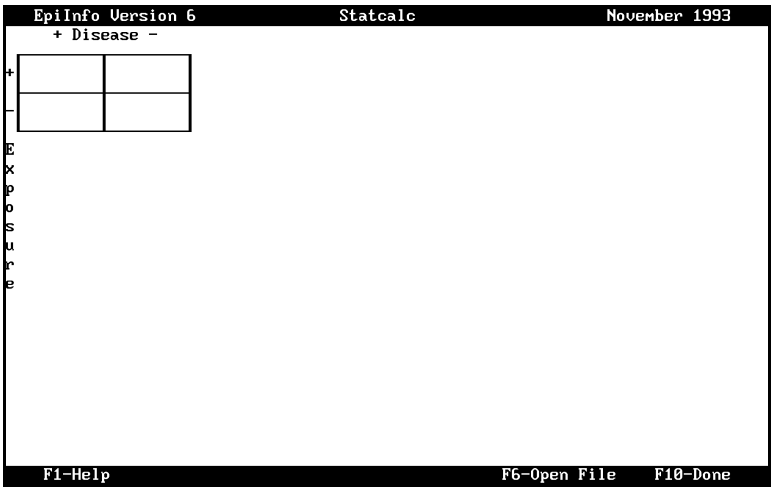


Figure 1.38 *StatCalc* cross-tabulation

First enter the numbers for those who are married (i.e., stratum one) as shown in Figure 1.39.

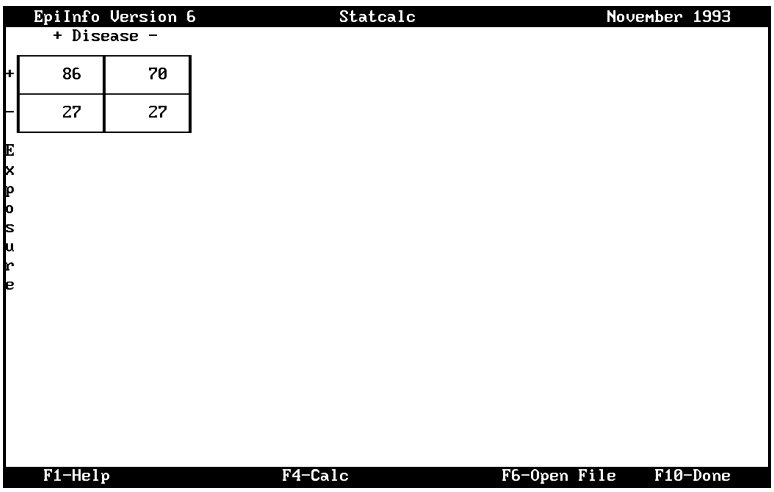


Figure 1.39 *StatCalc* entries for stratum 1

After the numbers are entered, press [F4 - Calc] and Figure 1.40 appears.

```
EpiInfo Version 6                Statcalc                    November 1993
   + Disease -                          Analysis of Single Table
                                     Odds ratio = 1.23 (0.63 <OR< 2.39)
+   86    70    156                 Cornfield 95% confidence limits for OR
                                     Relative risk = 1.10 (0.82 <RR< 1.49)
-   27    27     54               Taylor Series 95% confidence limits for RR
                                   Ignore relative risk if case control study.
E  113    97    210
x                                                  Chi-Squares   P-values
p                                                  ___________   ________
o                                   Uncorrected    :    0.42     0.5147289
s                                   Mantel-Haenszel:    0.42     0.5157316
u                                   Yates corrected:    0.24     0.6219113
r
e                                 F2 More Strata; <Enter> No More Strata; F10 Quit

   F1-Help    F2-Stratum                F5-Print    F6-Open File    F10-Done
```

Figure 1.40 *StatCalc* calculations for stratum 1

This is the interim analysis of the stratum one. To enter stratum two for the single men, press [**F2**] (see code line at bottom of screen). Enter the next set of numbers as shown in Figure 1.41.

```
EpiInfo Version 6                Statcalc                    November 1993
   + Disease -                          Stratified Analysis, Table 2
+   19    17
-   11    7
E
x
p
o
s
u
r
e

   F1-Help                                F6-Open File    F10-Done
```

Figure 1.41 *StatCalc* numeric entries for stratum 2

When done entering the numbers, the program calculates the measures of effect for stratum two (see Figure 1.42).

```
EpiInfo Version 6                Statcalc                    November 1993
   + Disease -                       Odds ratio = 0.71 (0.19 <OR< 2.60)
                                    Cornfield 95% confidence limits for OR
+   19    17     36                Relative risk = 0.86 (0.53 <RR< 1.40)
                                  Taylor Series 95% confidence limits for RR
-   11     7     18              Ignore relative risk if case control study.
E   30    24     54                                 Chi-Squares   P-values
x                                                    ___________   ________
p                                   Uncorrected    :    0.34     0.5612758
o                                   Mantel-Haenszel:    0.33     0.5649240
s                                   Yates corrected:    0.08     0.7714538
u
r                                 F2 More Strata; <Enter> No More Strata; F10 Quit
e

   F1-Help    F2-Stratum                F5-Print    F6-Open File    F10-Done
```

Figure 1.42 *Statcalc* calculations for stratum 2

Since there are no more strata, press [**Enter**] and the program derives the summary statistical measures, as shown in Figure 1.43.

Figure 1.43 *StatCalc* summary calculations for both strata

```
EpiInfo Version 6                Statcalc                    November 1993
     + Disease -
                                          ***** Stratified Analysis *****
+   | 19 | 17 |   36                            Summary of 2 Tables
-   | 11 |  7 |   18                  Crude odds ratio for all strata =  1.08
                                    Mantel-Haenszel Weighted Odds Ratio =  1.08
E     30    24    54                        Cornfield 95% Confidence Limits
x                                                 0.61 <  1.08 <  1.94
p                                   Mantel-Haenszel Summary Chi Square =  0.02
o                                              P value = 0.87735274
s
u                                         Crude RR for all strata =  1.04
r                                     Mantel-Haenszel Weighted Relative Risk
e                                        of Disease, given Exposure =  1.04
                                       Greenland/Robins Confidence Limits =
                                                0.80 < MHRR <  1.34

                                          <Enter> for more; F10 to quit.

F1-Help                                   F5-Print    F6-Open File    F10-Done
```

But there is still more. The confidence intervals for the summary odds ratio is an *estimate* rather than an *exact* value. Sometimes the estimate is very close to the exact value. Other times, however, the two might vary. The *StatCalc* program can calculate the exact value for you. To do so, press [**Enter**] and Figure 1.44 appears.

Figure 1.44 Start exact calculations

```
EpiInfo Version 6                Statcalc                    November 1993
     + Disease -
+   | 19 | 17 |   36      Press "E" for Exact Confidence Limits or <Enter>
```

Press [**E**] and the program starts to calculate the exact confidence interval. This usually takes a few moments, so the program tells you to be patient, as shown in Figure 1.45.

Figure 1.45 Ruminating

```
EpiInfo Version 6                Statcalc                    November 1993
     + Disease -
+   | 19 | 17 |   36      . . . .  Ruminating - please be patient
```

Once the calculations are done, the screen appears with the answers (see Figure 1.46).

Figure 1.46
Exact confidence interval for stratified odds ratio

```
EpiInfo Version 6                Statcalc                    November 1993
     + Disease -
+ |  19  |  17  |   36          ***Exact Confidence Limits***
- |  11  |   7  |   18            Mehta CR, Patel NR, Gray R,
                                J. Am. Stat. Assoc.,1985,78,969-973.
E    30     24     54        Pascal program by ELF Franco & N Campos-Filho
x                             Luduig Cancer Institute, Sao Paulo, Brazil
p
o                              Exact Lower 95% Confidence Limit =  0.61
s                             Mantel-Haenszel Weighted Odds Ratio =  1.08
u                              Exact Upper 95% Confidence Limit =  1.93
r
e                                   <Enter> to continue.....

  F1-Help                          F5-Print    F6-Open File    F10-Done
```

Press [**Enter**] one more time and you return to the calculation screen entry of another set of numbers (see Figure 1.47).

Figure 1.47
Entry screen for new calculations

```
EpiInfo Version 6                Statcalc                    November 1993
     + Disease -
+ |      |      |
- |      |      |
E
x
p
o
s
u
r
e

  F1-Help                                      F6-Open File    F10-Done
```

The next section features an analysis of two data sets included with the *Epi Info* software and a rapid survey of 300 men in 360 households, described earlier in this chapter.