

## ANALYSIS OF CLUSTER SURVEYS WITH *EPI INFO*

Another feature of *Epi Info* is a set of three programs for the analysis of cluster surveys. This is the only software program, other than more complex statistical packages such as *Stata* and *SUDAAN*, for doing this form of analysis. Over the years, I had discussed the need for such a feature with Dr. Andrew Dean of CDC for several years, emphasizing the importance to public health of cluster surveys. CDC eventually asked Professor William Kalsbeek, a statistician at the University of North Carolina at Chapel Hill, to design the statistical module. Included with the program are two data sets I created and sent to Dr. Dean at CDC to use as examples: EPI1 and EPI10, both included as views in Sample.mdb (i.e., viewEpi1 and viewEpi10). The former contains data on a two-stage cluster survey of 210 children; 30 clusters were selected with probability proportionate to size (PPS) and 7 children were sampled per cluster. The latter contains data on 2,152 children in 10 two-stage cluster surveys, all with PPS selection at the first stage, stratified by geographic location (each survey is a different stratum), and weighted to the sampled population. Both are based on series of cluster surveys done in Iran some year earlier. You will find them in C:\Epi\_Info\Sample.mdb, distributed with the *Epi Info* software.

In this section we will analyze cluster survey data with *Epi Info*. Then in the following section we will analyze the same information with *Stata*. As you will see, *Epi Info* works very well for analyzing point estimates (i.e., the occurrence of health conditions presented as proportions or percentages), and for doing simple cross-tabulations of two variables. The program does not adjust for confounding, however, and cannot be used to do multivariate analyses. For this we will use *Stata*.

■ **EPI1 and EPI10.** *Epi Info* includes data from two cluster surveys that tested if those who received prenatal care were more or less likely to receive a complete immunization series than those who did not receive prenatal care. The analysis was done with both EPI1 (a small survey in one region) and EPI10 (a much larger survey done in ten regions). The four-fold tables for this analysis are shown in Figure 1.48.

**Figure 1.48**  
Two example data sets included with *Epi Info*

		EPI1			EPI10		
		Completed vaccination			Completed Vaccination		
		Yes	No		Yes	No	
Received prenatal care	Yes	78	9	87	675	413	1088
	No	77	46	123	567	497	1064
		155	55	210	1242	910	2152

As mentioned above, the EPI10 data set is actually ten different cluster surveys. Therefore Figure 1.48 shows for EPI10 the crude analysis between PRENATAL and VACCINE. To analyze the data correctly, you need to separate the surveys by stratification into ten subgroups and measure the association between prenatal care and vaccination status in each subgroup (see Figure 1.49). The survey in *location 1* has 225 children sampled from a population of 9,870 children. The number of children in the other nine surveys and the size of the sampled population are included in Figure

1.49.

	Location 1	Location 2	Location 3	Location 4	Location 5
	VAC	VAC	VAC	VAC	VAC
	Y N	Y N	Y N	Y N	Y N
Care	Y	Y	Y	Y	Y
	N	N	N	N	N
	n = 225	n = 219	n = 212	n = 219	n = 212
	N = 9,870	N = 33,600	N = 14,130	N = 27,900	N = 12,750
	Location 6	Location 7	Location 8	Location 9	Location 10
	VAC	VAC	VAC	VAC	VAC
	Y N	Y N	Y N	Y N	Y N
Care	Y	Y	Y	Y	Y
	N	N	N	N	N
	n = 214	n = 210	n = 212	n = 217	n = 212
	N = 15,810	N = 16,050	N = 180,740	N = 9,030	N = 25,650

**Figure 1.49**  
Concept for  
analysis of  
EPI10

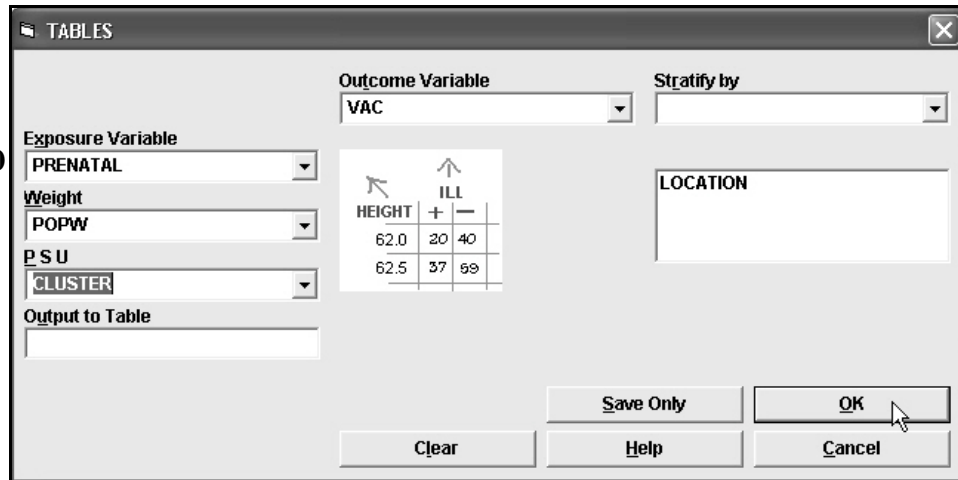
To do a stratified analysis, *Epi Info* must first know your *main* variable (i.e., the **outcome** or **dependent** variable shown here as VAC), the *crossstab* variable (i.e., the **exposure** or **independent** variable shown here as PRENATAL), the variable that identifies the *strata* (shown here as LOCATION), and the variable that identifies the number of children represented by each stratum and weights the strata accordingly (i.e., the number of children in the population that each surveyed child represents; stated in the variable, POPW). Finally because it is a cluster survey with 30 clusters of about 7 children each per survey, the program must know the variable that identifies the cluster number (i.e., CLUSTER).

■ **EPI10.** The analysis just described is the most sophisticated (or complicated) that can be done by *Epi Info*. You likely will not be doing such large surveys. To give you experience with population weights, however, I have included this data set as an example. Return to the main *Epi Info* menu and click on *Analyze Data*. In the *Analysis Commands* column, click on *Read (Import)* under *Data*. The data source should appear as: C:\Epi\_Info\Sample.mdb. Move the cursor in *Views* to *viewEpi10* and with your left mouse click *OK*. The program should indicate that you have loaded a data set with 2,152 records.

You will be determining if children who received prenatal care (the exposure variable or PRENATAL) are more or less likely to have been vaccinated (the outcome variable or VAC), taking into account the sampling strategy (the primary sampling unit [PSU] or CLUSTER, the ten strata (stratified by LOCATION) and the sampling weights (POPW). To do so, go to the *Advanced Statistics* section of the *Analysis Commands* column and click on *Complex Sample Tables*. In the *Tables* window, enter PRENATAL as the exposure variable, VAC as the outcome variable, POPW

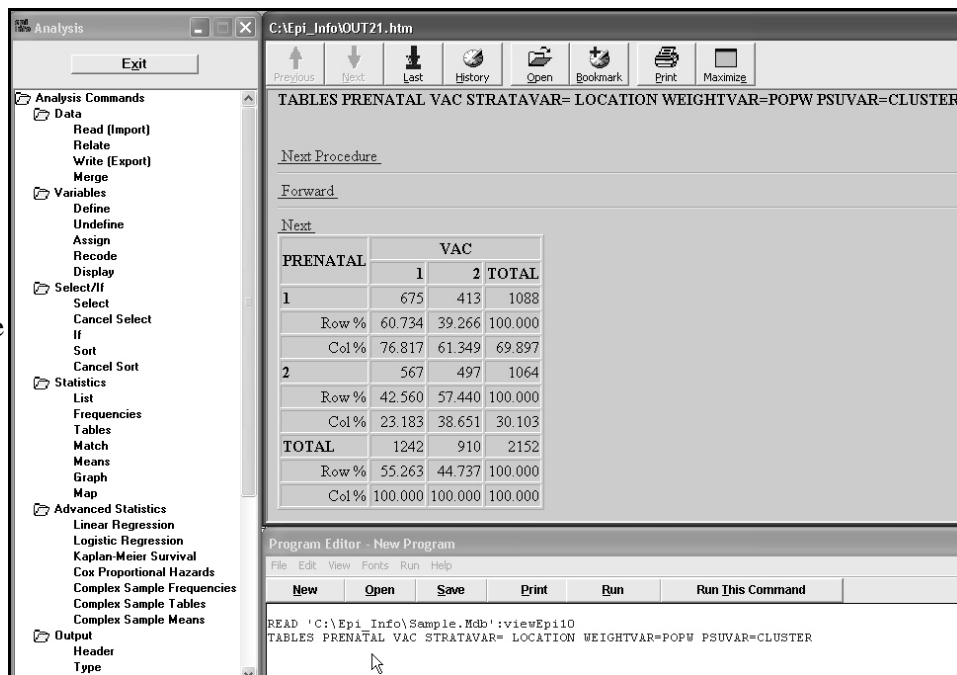
as the weight, LOCATION as Stratified by, and CLUSTER as the PSU (see Figure 1.50).

**Figure 1.50**  
Entry screen for analysis of EPI10



Specifically, our intent is to determine if mothers who received prenatal care (*PRENATAL*=1) are as likely to have their children vaccinated (*VAC*=1) as mothers who did not receive prenatal care (*PRENATAL*=2). Once done entering the variables, click *OK* and the output in Figure 1.51 should appear.

**Figure 1.51**  
Output of cluster sample analysis of EPI10



Prenatal care was received by 69.97% of the children’s mothers, while 30.1% received no prenatal care (see vertical percentages in TOTAL column). Among those who had received prenatal care, 60.7% were vaccinated (see horizontal percentage in VAC=1 column of PRENATAL=1). Conversely, vaccination only occurred among 42.6% of those whose mothers had not received prenatal care (see horizontal percentage in VAC=1 column of PRENATAL=2). Scroll down the output section and note the additional statistical calculations shown in Figure 1.52.

**Figure 1.52**  
More  
output of  
cluster sample  
analysis of  
EPI10

```

CTABLES COMPLEX SAMPLE DESIGN ANALYSIS OF 2 X 2 TABLE

Odds Ratio (OR) 2.088
Standard Error (SE) 0.307
95% Conf. Limits (1.50, 2.898 )

Risk Ratio (RR) 1.427
Standard Error (SE) 0.110
95% Conf. Limits (1.23, 1.659 )
RR = (Risk of VAC=1 if PRENATAL=1) / (Risk of VAC=1 if PRENATAL=2)

Risk Difference (RD) 0.182
Standard Error (SE) 0.040
95% Conf. Limits (0.10, 0.261 )
RD = (Risk of VAC=1 if PRENATAL=1) - (Risk of VAC=1 if PRENATAL=2)

Sample Design Included:

Weight Variable: POPW
PSU Variable: CLUSTER
Stratification Variable: LOCATION

0 records with missing values

```

The risk of being vaccinated is 1.427 times greater in the prenatal care group than those whose mothers did not receiving prenatal care. The 95% confidence interval for the risk ratio (now done correctly, taking into account the sampling design) extends from 1.23 to 1.66. The difference in the rate of becoming vaccinated between the two prenatal groups is 18.2% (i.e., 60.7 - 42.5), with a 95% confidence interval of 10 to 26%.

■ **Incorrect Analysis - Prevalence Estimates.** The material on cluster surveys so far has introduced you to the topic and given you some experience with the program. We will now return to our problem and use the data set, *AIDSAL.mdb* distributed on the Rapid Survey Course web page (i.e., <http://www.ph.ucla.edu/epi/rapidsurveys/RScourse/RSstmanual.html>). Copy *AIDSAL.mdb* to your C drive (i.e., C:\Epi\_Info\418\ ) for use in this exercise. Note: the 418 subdirectory was used for the UCLA course, EPI 418 *Rapid Epidemiological Surveys in Developing Countries*. For the Rapid Survey Course, you can save the file in a subdirectory of your choosing. The file has information on 300 men in 360 sampled households, as described earlier in this chapter (see pages 1-4 to 1-10). The questionnaire for this study was shown in Figure 1.5. We will first retrieve *AIDSAL.mdb* and analyze it incorrectly using the program listed under *Statistics* in the *Analysis Commands* column of *Epi Info*. This set of programs, like most statistical packages, assumes that the data have been gathered with each element being independent. This is not what happens with cluster surveys. We are sampling groups of households that are often close to one another and interviewing or examining the eligible persons in the sampled households. Such persons tend to be more alike than if sampled independently from throughout the region. This similarity is termed “homogeneity” by sampling statisticians. Homogeneous samples tend to have greater variances than heterogenous samples (we will discuss the reasons why during the Rapid Survey Course). The variances of cluster surveys tend to be larger than comparable-sized simple random samples. A larger variance means larger confidence limits; how much larger will vary from one survey to the next and from one variable to the next.



**Figure 1.54**  
Frequency  
distribution  
of SEXA

FREQ SEXA				
<u>Next Procedure</u>				
<u>Forward</u>				
SEXA	Frequency	Percent	Cum Percent	
1	52	17.3%	17.3%	
2	233	77.7%	95.0%	
9	15	5.0%	100.0%	
<b>Total</b>	300	100.0%	100.0%	

95% Conf Limits		
1	13.2%	22.1%
2	72.5%	82.3%
9	2.8%	8.1%

appear as in Figure 1.54.

Among the 300 men, 52 stated that they had anal intercourse during the past month. Another 15 men refused to answer the question, likely feeling it was too personal. Since we do not know if these 15 men truly had or did not have anal sex, we cannot use all 300 to estimate the percentage who had anal sex. More on that in a minute. First, however, click again on *Frequencies* but this time enter *HIV*, the outcome variable. The image shown in Figure 1.55 should appear.

**Figure 1.55**  
Frequency  
distribution  
of HIV

FREQ HIV				
<u>Previous Procedure</u> <u>Next Procedure</u> <u>Current Dataset</u>				
<u>Forward</u>				
HIV	Frequency	Percent	Cum Percent	
1	27	9.0%	9.0%	
2	267	89.0%	98.0%	
3	4	1.3%	99.3%	
9	2	0.7%	100.0%	
<b>Total</b>	300	100.0%	100.0%	

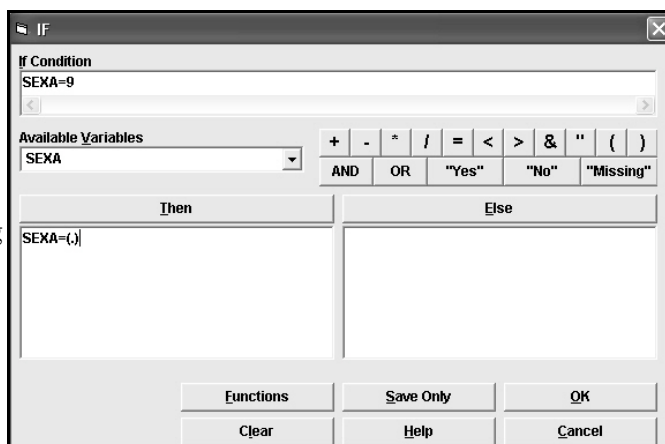
  

95% Conf Limits		
1	6.0%	12.8%
2	84.9%	92.3%
3	0.4%	3.4%

- **If-Then.** While 27 had HIV antibodies in their saliva and 267 did not, the laboratory test

was not definitive as positive or negative with 4 persons. In addition, there was no specimens collected for two subjects. The denominator for the prevalence estimate of HIV infection should thus be 300 minus 6 or 294. You can calculate the occurrence of recent anal sex or prevalence of HIV by hand or have *Epi Info* do it for you using the *If* command. Under *Select/If* in the *Analysis Commands* column, click on *If*. Enter the *If Condition* of *SEXA=9* (i.e., if *SEXA* equals “no response”) and the *Then* command *SEXA=(.)* (i.e., then *SEXA* is missing), as shown in Figure 1.56.

**Figure 1.56**  
If-then statement for removing no response to *SEXA*



This turns the 15 subjects who were coded as 9 into missing values, but not permanently. The actual data set stored on your disk is not changed. Next click with your left mouse on *Frequencies* and enter *SEXA* in *Frequency of* followed by a click on *OK*. The frequency distribution shown in Figure 1.57 should appear.

**Figure 1.57**  
Frequency distribution of *SEXA* with code 9 removed

FREQ SEXA				
<a href="#">Previous Procedure</a> <a href="#">Next Procedure</a> <a href="#">Current Dataset</a>				
Forward				
SEXA	Frequency	Percent	Cum Percent	
1	52	18.2%	18.2%	
2	233	81.8%	100.0%	
<b>Total</b>	<b>285</b>	<b>100.0%</b>	<b>100.0%</b>	
<b>95% Conf Limits</b>				
1	13.9%	23.2%		
2	76.8%	86.1%		

Now with the corrected denominator you get a factual estimate of the occurrence of recent anal sex, namely 18.2%.

We next will eliminate the indeterminant (i.e., *HIV=3*) and missing (i.e., *HIV=9*) entries for the variable *HIV*. Under *Select/If* in the *Analysis Commands* column, click on *If*. Enter the *If*

Condition of HIV=3 (i.e., if HIV equals “indeterminant”), click on **OR** and enter HIV=9 followed by the *Then* command HIV=(.) (i.e., then HIV is missing), and *OK* (see Figure 1.58).

Then click with your left mouse on *Frequencies* and enter HIV in *Frequency of* followed by a click on *OK*. The frequency distribution shown in Figure 1.59 appears.

**Figure 1.59**  
Frequency distribution of HIV with codes 3 and 9 removed

FREQ HIV				
Previous Procedure Next Procedure Current Dataset				
Forward				
HIV	Frequency	Percent	Cum Percent	
1	27	9.2%	9.2%	
2	267	90.8%	100.0%	
<b>Total</b>	294	100.0%	100.0%	
<b>95% Conf Limits</b>				
1	6.1%	13.1%		
2	86.9%	93.9%		

Notice that the prevalence of HIV infection among men who had classifiable specimens was 9.2 percent. The third variable to be considered is believing a drug is available to cure AIDS (*DRUG*), the frequency distribution for which is seen in Figure 1.60 (create this on your own).

**Figure 1.60**  
Frequency distribution of DRUG

FREQ DRUG				
Previous Procedure Next Procedure Current Dataset				
Forward				
DRUG	Frequency	Percent	Cum Percent	
1	235	78.3%	78.3%	
2	65	21.7%	100.0%	
<b>Total</b>	300	100.0%	100.0%	
<b>95% Conf Limits</b>				
1	73.2%	82.9%		
2	17.1%	26.8%		

The variable *DRUG* will be treated as a confounding variable in our coming analysis. After we have assembled the reduced data set with usable values for *SEXA*, *HIV* and *DRUG*, we will have the program derive the 95% confidence intervals for the prevalence estimates of both *SEXA* and *HIV*. There is no need to create a confidence interval for *DRUG* since it is a confounding variable, used

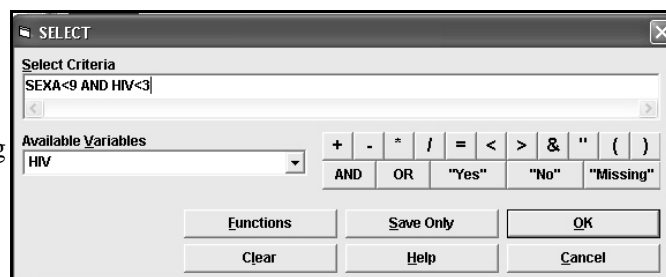
only to separate the data into two groups, DRUG=1 and DRUG=2, for further stratified (i.e., unconfounded) analysis.

- **Select.** At this point you will need to use the *Select* command (under *Select/If* of the *Analysis Commands* column) to reduce the data set to a smaller number of cases with appropriate values for each of the three variables, SEXA, HIV and DRUG. That is, we will eliminate 21 subjects (6 due to *HIV*, 15 due to *SEXA*, and 0 due to *DRUG*) so that all variables can be treated as binary or dichotomous (i.e., two outcome) variables and we can do all analyses on the same data set.

Using the *Statistics* programs in *Epi Info*, we will derive the occurrence of recent anal sex, prevalence of HIV infection, and proportion who believe there is a curative drug for AIDS. For the first two variables, we will calculate the 95% confidence intervals as well. I have labeled this the "incorrect" analysis because we are not taking into account that the data were derived from subjects in a cluster survey. Instead, the analysis assumes the data were collected in a simple random sample survey.

First, however, we will eliminate with the *Select* command 15 subjects from the SEXA analysis and 6 subjects from the HIV analysis. This will reduce the data set to 279 persons with values of 1 or 2 for SEXA, HIV and DRUG. Under *Select/If* in the *Analysis Commands* column, click on *Select*. Enter the *Select Criteria* of SEXA<9 (i.e., select only those who responded to the question) and HIV<3 (i.e., select only those who had positive or negative test results). The entry should be as in Figure 1.61.

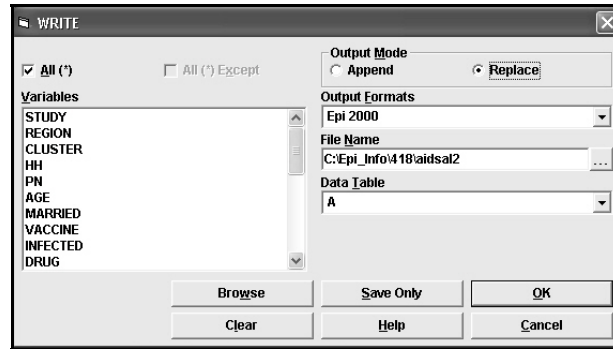
**Figure 1.61**  
Select statement for removing values of SEXA and HIV



Click on *OK* and notice that there are now 279 records instead of 300, as before.

- **Write(export).** If you feel like stopping for a while (and I suggest you do), save the data set with 279 values in a separate file. To do so, click on *Write (export)* under *Data* in the *Analysis Command* column. Use the Epi 2000 output format. Enter the output file name of: *C:\Epi\_Info\418\aidisal2* and Data Table A, as shown in Figure 1.62. It is good to get in the habit of clicking "replace" to make sure you do not add the data to another data set with the same name that was previously saved.

**Figure 1.62**  
Saving  
reduced  
file with  
new name  
*aidsal2.mdb*



If you have stopped for a while, return to the *Analyze Data* section of *Epi Info*, click on *Read (Import)*, and enter *C:\Epi\_Info\418\ aidsal2.mdb*. To find data table *A*, Show *All*, then move the cursor to *A* and click *OK*.

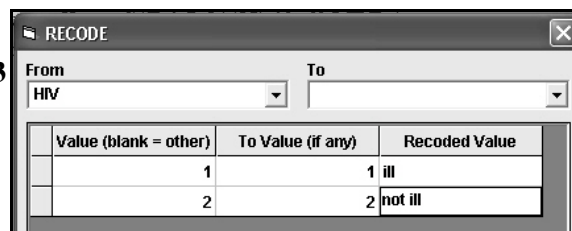
- **Recode.** Epidemiological tables comparing an exposure variable to a disease outcome variable typically have four cells (usually listed as a, b, c, and d), with exposed individuals listed in the first row and ill persons listed in the first column. *Epi Info* depends on this arrangement to do the right analysis. Thus if recoding, you need to make sure that columns and rows are in the desired place.

		Outcome Variable	
		Ill	Not ill
Exposure variable	Exp	a	b
	Unexp	c	d

For recoding, *Epi Info* creates tables with the variable labels in alphabetical or numeric order. Thus when using “exp” (for exposed) and “unexp” (for unexposed), “e” precedes “u” in the alphabet; hence the “exp” line is listed first, as shown above. If we continue to use the labels “1” (for “yes”) and “2” (for “no”), *Epi Info* would also do the correct analysis since “1” precedes “2” in numeric order. Later, however, we will be recoding “1” and “2” to “1” (i.e., yes) and “0” (i.e., no) to use for doing logistic analyses in *Stata*. For such a dataset, *Epi Info* would list the variables backwards (i.e., the unexposed row [coded as 0] would be listed first), and thus produce the wrong analysis. This point will be further discussed later in the *Software Training Manual*.

In our data set of 279 records, we will recode the outcome labels of HIV as “ill” and “not ill” and SEXA as “exp” and “unexp.” First recode HIV by clicking with the left mouse on *Recode* under *Variables* in the *Analysis Commands* column of *Epi Info*. Enter HIV in *From*, the range of values for 1 (i.e., 1 to 1) in the first row of the recode table, and the ranges for value 2 (i.e., 2 to 2) in the second row of the recode table. The recoded value for 1 becomes *ill*, while the recoded value for 2 becomes *not ill*. To insert a second line in the recode table, press [enter]. When through, the recode table for HIV should appear as in Figure 1.63, just before clicking *OK*.

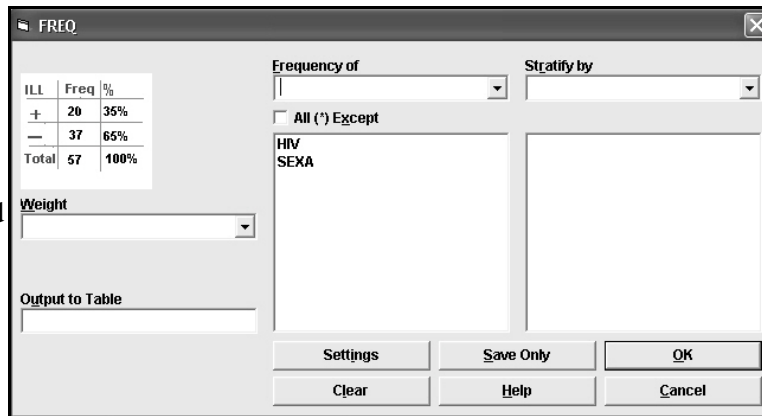
**Figure 1.63**  
Recode  
HIV



Repeat the recoding process for SEXA, changing values of 1 and 2 to values of *Exp* and *Unexp*.

• **Frequencies.** With the left mouse key, click on *Frequencies* under *Statistics* in the *Analysis Commands* column. Again obtain the frequency distribution for HIV and SEXA, but this time in a single statement, created as in Figure 1.64.

**Figure 1.64**  
Frequency  
of SEXA and  
HIV



Enter *OK*. The output should be as in Figure 1.65.

**Figure 1.65**  
Frequency  
distributions  
of HIV and  
SEXA with  
reduced  
data set and  
recoded  
labels

**FREQ HIV SEXA**

**HIV**

Forward

HIV	Frequency	Percent	Cum Percent
ill	27	9.7%	9.7%
not ill	252	90.3%	100.0%
<b>Total</b>	<b>279</b>	<b>100.0%</b>	<b>100.0%</b>

**95% Conf Limits**

ill 6.5% 13.8%

not ill 86.2% 93.5%

**SEXA**

Back Forward Current Procedure

SEXA	Frequency	Percent	Cum Percent
exp	52	18.6%	18.6%
unexp	227	81.4%	100.0%
<b>Total</b>	<b>279</b>	<b>100.0%</b>	<b>100.0%</b>

**95% Conf Limits**

exp 14.2% 23.7%

unexp 76.3% 85.8%

For the reduced data set, the prevalence of HIV infection is 9.7 percent with a 95% confidence interval of 6.5 percent to 13.8 percent (incorrect for this data set). Notice that 18.6 percent had anal intercourse during the past month, with a 95% confidence interval of 14.2 to 23.7 percent (also incorrect for this data set).

• **Tables.** You will next consider the two-by-two (or crude) relationship between SEXA (the exposure variable) and HIV (the outcome variable). Using the left mouse key, click on *Tables* under *Statistics* in the *Analysis Commands* column. Enter SEXA and HIV in the appropriate locations. The results are as shown in Figure 1.66.

**Figure 1.66**  
Cross-tabulation of SEXA and HIV

TABLES SEXA HIV			
Forward			
HIV			
SEXA	ill	not ill	TOTAL
exp	13	39	52
Row %	25.0	75.0	100.0
Col %	48.1	15.5	18.6
unexp	14	213	227
Row %	6.2	93.8	100.0
Col %	51.9	84.5	81.4
TOTAL	27	252	279
Row %	9.7	90.3	100.0
Col %	100.0	100.0	100.0

Single Table Analysis			
	Point Estimate	95% Confidence Interval	
		Lower	Upper
PARAMETERS: Odds-based			
Odds Ratio (cross product)	5.0714	2.2147	11.6132 (T)
Odds Ratio (MLE)	5.0299	2.1623	11.6805 (M)
		2.0101	12.5580 (F)
PARAMETERS: Risk-based			
Risk Ratio (RR)	4.0536	2.0288	8.0993 (T)
Risk Difference (RD%)	18.8326	6.6542	31.0110 (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	17.1668	0.0000354058	
Chi square - Mantel-Haenszel	17.1053	0.0000365330	
Chi square - corrected (Yates)	15.0799	0.0001042253	
Mid-p exact		0.0001239047	
Fisher exact		0.0002106524	

Notice that the odds ratio is 5.07 and the risk ratio is 4.05. Later you will be comparing the point estimates and confidence intervals with other analyses.

• **Frequencies.** The third variable to be considered is believing a drug is available to cure AIDS (*DRUG*), the frequency distribution output of which is seen in Figure 1.67, again in the reduced data set.

**Figure 1.67**  
Frequency  
distribution  
of DRUG  
with reduced  
data set

FREQ DRUG

Previous Procedure Next Procedure Current Dataset

Forward

DRUG	Frequency	Percent	Cum Percent
1	222	79.6%	79.6%
2	57	20.4%	100.0%
Total	279	100.0%	100.0%

95% Conf  
Limits

1 74.4% 84.1%

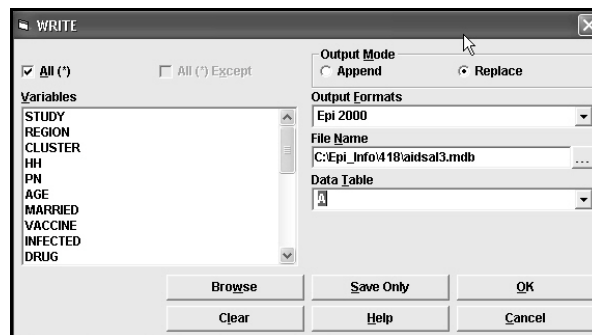
2 15.9% 25.6%

Nearly 80 percent of the quarried men believed that there is a drug available to cure AIDS.

Our intention in the final incorrect Epi Info analysis, is to view the relationship between *SEXA* and *HIV* after controlling for *DRUG*. That is, we want to determine the relationship between anal sex and HIV among those who believe in an AIDS drug and those who do not. If this was a simple random sample, we would analyze the reduced data set with the programs in the *Statistics* section of the *Analysis Commands*. Since it is a cluster survey, this analysis will likely not be correct with respect to the confidence limits. To see the nature of the error, we will go ahead and analyze the data incorrectly with the *Statistics* programs and then compare our findings (at least with respect to the odds ratio) with the same analysis done correctly with the Stata program.

- **Write(export).** This is another good point to stop, or at least to create another data set with the new values of HIV and SEXA. To do so, click on *Write (export)* under *Data* in the *Analysis Command* column. Use the Epi 2000 output format. Enter the output file name of: *C:\Epi\_Info\418\aid3.mdb* and Data Table A, as shown in Figure 1.68. Click on “replace” to make sure you do not add the data to another data set with the same name that was previously saved.

**Figure 1.68**  
Saving  
reduced  
file with  
new name  
*aid3.mdb*

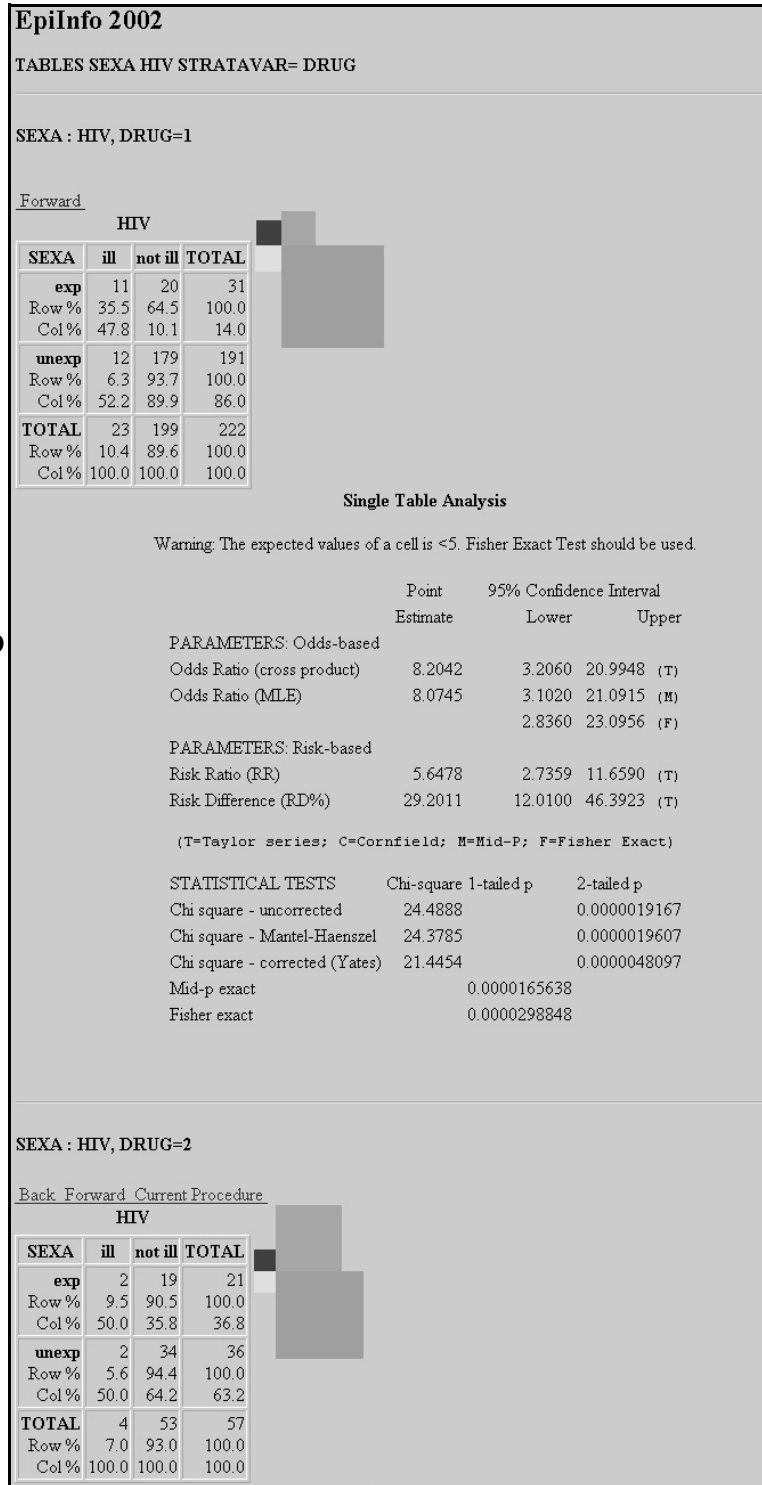


- **Incorrect Analysis - Stratification.** If you have stopped for a while, return to the *Analyze Data* section of *Epi Info*, click on *Read (Import)*, and enter *C:\Epi\_Info\418\aid3.mdb*. To find data table A, Show All, then move the cursor to A and click *OK*. This loads the reduced data set of 279 persons with recoded labels for HIV and SEXA. We will use the *Tables* command (under *Statistics* in the *Analysis Commands* column) to create a two-by-two table that compares HIV prevalence (outcome variable) among those who had recent anal sex (exposure variable; *SEXA=exp*) versus

those who did not (*SEXA=unexp*). The analysis will be divided into two strata by whether or not they believed a drug (Stratify by) is available to cure AIDS (*DRUG=1*, yes; *DRUG=2*, no). After clicking with the left mouse on *Tables*, enter *SEXA* as the exposure variable, *HIV* as the outcome variable, and stratify by *DRUG*.

The outputs should be as shown in Figure 1.69.

**Figure 1.69**  
Cross-tab  
of HIV by  
SEXA  
controlling  
for DRUG



**Figure 1.69**  
 Cross-tab  
 of HIV  
 by SEXA  
 controlling  
 for DRUG  
 (continued)

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)	1.7895	0.2330	13.7459 (T)
Odds Ratio (MLE)	1.7704	0.1727	18.1579 (M)
		0.1194	26.2632 (F)
PARAMETERS: Risk-based			
Risk Ratio (RR)	1.7143	0.2604	11.2871 (T)
Risk Difference (RD%)	3.9683	-10.6475	18.5840 (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	0.3201		0.5715595546
Chi square - Mantel-Haenszel	0.3145		0.5749533688
Chi square - corrected (Yates)	0.0008		0.9774326017
Mid-p exact		0.3038277512	
Fisher exact		0.4712918660	
<b>SUMMARY</b>			
<a href="#">Back</a> <a href="#">Forward</a> <a href="#">Current Procedure</a>			
SUMMARY INFORMATION			
	Point Estimate	95% Confidence Interval Lower Upper	
Parameters			
Odds Ratio Estimates			
Crude OR (cross product)	5.0714	2.2147,	11.6132 (T)
Crude OR (MLE)	5.0299	2.1623,	11.6805 (M)
		2.0101,	12.5580 (F)
Adjusted OR (MH)	5.7573	2.4281,	13.6515 (R)
Adjusted OR (MLE)	6.3189	2.6016,	15.4284 (M)
		2.4048,	16.7187 (F)
Risk Ratios (RR)			
Crude Risk Ratio (RR)	4.0536	2.0288,	8.0993
Adjusted RR (MH)	4.4464	2.2740,	8.6944
(T=Taylor series; R=RGB; M=Exact mid-P; F=Fisher exact)			
STATISTICAL TESTS (overall assoc)		Chi-square	1-tailed p 2-tailed p
MH Chi square - uncorrected		20.5200	0.0000
MH Chi square - corrected		18.1262	0.0000
Mid-p exact			0.0000
Fisher exact			0.0001
In the following two tests, low p values suggest that ratios differ by stratum			
Chi-square for differing Odds Ratios by stratum (interaction)		1.7675	0.1837
Chi-square for differing Risk Ratios by stratum		1.3394	0.2471

Figure 1.69 appears in two screens. Notice that both the adjusted odds ratios and risk ratios are mildly different from the crude odds ratio of 5.07 or the crude risk ratio of 4.05, suggesting that

*DRUG* is a confounding variable, although only slightly so. Also notice that the OR and RR values are much greater in stratum 1 (both highly positive) than in stratum 2 (both mildly positive). This suggests that the effect of *SEXA* on *HIV* is modified by the third variable, *DRUG*. If this is so, then *DRUGS* would be viewed as an **effect modifier** as well as a slight **confounder**. Notice also that the confidence intervals for the odds and risk ratios of the two strata are very wide. Thus the differences in the size of OR and RR between the two strata could be due to chance variation alone and thus, not be real.

The bottom portion of the analysis is shown in the continuation of Figure 1.69. Here we see the summary statistics that combine the two strata into an adjusted odds ratio or risk ratio. Notice that the crude OR of 5.07 is nearly the same as the Mantel-Haenszel adjusted OR of 5.76, and the crude RR of 4.05 is nearly the same as the Mantel-Haenszel adjusted RR of 4.45. This indicates that confounding by *DRUG* did not distort the crude association between *SEXA* and *HIV* to any noticeable extent, although *DRUG* was an effect modifier with dramatically different effects in the two strata.

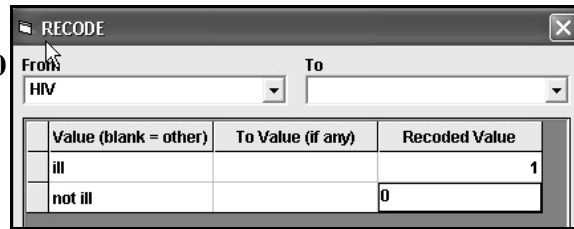
Also notice in the bottom of Figure 1.69 the chi square test which measures if the two strata differ in the magnitude of the odds or risk ratio (i.e., chi square for differing odds ratios or risk ratios [interaction]). It appears that the effect modification we observed in the **odds ratio** is not statistically significant, with an 18.4 percent chance that the difference between strata (i.e., interaction) is due to random variation. Statisticians refer to **effect modification** as *interaction* so you will see this term used as well. There might also be effect modification in the two stratum-specific **risk ratios**, although the test of interaction has a value of 0.2471 which indicates that there is a 24.7 percent chance that the difference is due to random variation inherent in the sampling process. Typically the p-value should be less than 5 percent (i.e., <0.05) before we make a big deal out of the effect modification findings, although this is not a rule that is always followed.

This ends the incorrect analysis section (incorrect because the analysis assumes simple random sampling but the data set comes from a cluster survey). We will subsequently compare the correct analysis output to what has been observed so far.

■ **Correct Analysis - Prevalence Estimates.** Earlier you did a frequency distribution of *HIV* using the inappropriate *Frequencies* command under *Statistics* in the *Analysis Command* column (see Figure 1.65). The program presented both the percent that was coded “ill” (i.e., the prevalence estimate) and the 95% confidence limits for the prevalence estimate. We will now do the same analysis, but correctly, assuming the data were derived from a cluster survey. First, however, we need to recode *HIV* and *SEXA* as 0,1 variables, since the *Complex Sample* commands do not use labels such as “ill” or “exp.”

• **Recode (note Epi Info error in this section).** Using dataset *AIDSAL3.mdb*, you need to recode *HIV* from “ill” and “not ill” to 1 and 0, and *SEXA* from “exp” and “unexp” to 1 and 0. First recode *HIV* by clicking with the left mouse on *Recode* under *Variables* in the *Analysis Commands* column of *Epi Info*. Enter *HIV* in *From*, the value *ill* in the first row of the recode table, and the value *not ill* the second row of the recode table. The recoded value for *ill* becomes 1, while the recoded value for *not ill* becomes 0. When through, the recode table for *HIV* should appear as in Figure 1.70, just before clicking *OK*.

**Figure 1.70**  
Recode  
HIV



Repeat the recoding process for SEXA, going from *exp* and *unexp* to 1 and 0, and for DRUG going from 1 (i.e., yes) and 2 (i.e., “no”) to 1 and 0. (**Note following error**). For some reason, the latest version of *Epi Info* does not accept a 0 as a recoded value, but rather enters the value as missing. The program editor at the bottom of the screen shows what is occurring, as seen in Figure 1.71.

**Figure 1.71**  
Error in  
Recode  
command  
when  
entering 0.

```

Program Editor - New Program
File Edit View Fonts Run Help
New Open Save Print Run
READ 'C:\Epi_Info\418\aida13.mdb':A LINKNAME=TMPLNK_1
RECODE HIV TO HIV
    "ill" = 1
    "not ill" = (.)
END
RECODE SEXA TO SEXA
    "exp" = 1
    "unexp" = (.)
END
RECODE DRUG TO DRUG
    1 = 1
    2 = (.)
END

```

Note in the program editor that “unexp” is being recoded as (.) [i.e., the *Epi Info* designation for “missing,” rather than 0 as was specified]. To correct this flaw, with your cursor and [backspace] key, replace each (.) with a 0, as shown in Figure 1.72.

**Figure 1.72**  
Correction of  
error in  
Recode  
command  
when  
entering 0.

```

File Edit View Fonts Run Help
New Open Save Print Run
READ 'C:\Epi_Info\418\aida13.mdb':A LINKNAME=TMPLNK_1
RECODE HIV TO HIV
    "ill" = 1
    "not ill" = 0
END
RECODE SEXA TO SEXA
    "exp" = 1
    "unexp" = 0
END
RECODE DRUG TO DRUG
    1 = 1
    2 = 0
END

```

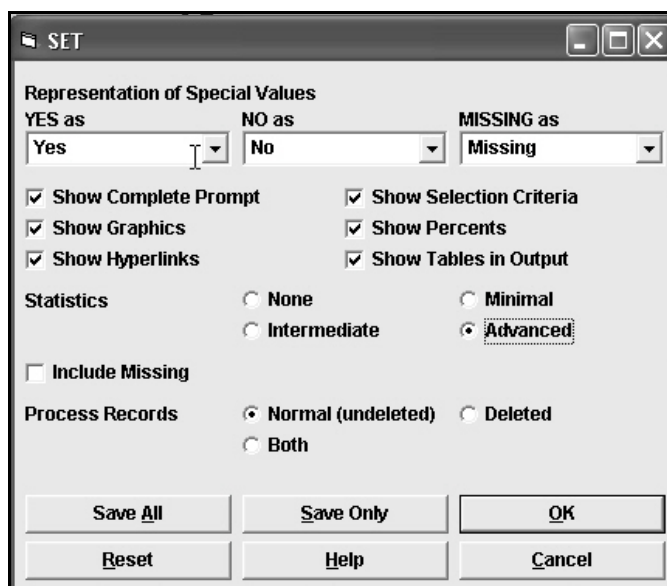
Then click on **Run** to re-run the recoding program.

- **Write(export)**. When done, create still another data set with the new values of HIV and SEXA. To do so, click on *Write (export)* under *Data* in the *Analysis Command* column. Use the *Epi 2000* output format. Enter the output file name of: *C:\Epi\_Info\418\aida14.mdb* and Data Table A. Click on “replace” to make sure you do not add the data to another data set with the same name that

was previously saved.

- **Complex Sample Means.** Be sure that *aidsal4.mdb* is loaded. You have created three binomial (i.e., two name) variables that were coded 0 or 1. The mean of a binomial variable with such codes is a proportion, or in our case, the prevalence of HIV infection and the prevalence of anal sex. When analyzing cluster survey data, you will want to display all the statistics that *Epi Info* has available, including the standard error when deriving prevalence or incidence estimates and the *design effect*, a number that compares the variance of the value analyzed as a cluster survey to the variance of the value analyzed as a simple random sample. We will be discussing the *design effect* in class. To having the program show all statistics, click with the left mouse key on *Set* under *Options* at the bottom of the *Analysis Commands* column. *Set Statistics* to *Advanced*, as shown in Figure 1.73. Click on OK. This results in the output showing all available statistics.

**Figure 1.73**  
Change  
Statistics  
to Advanced



To appreciate the coming analysis of *complex sample means*, we will first do the incorrect means analysis, assuming the study was a simple random sample with independent observations. The mean of a 0,1 variable is the proportion (or percentage if multiplied times 100) with the attribute. To do the incorrect means analysis, click on *Means* under *Statistics* in the *Analysis Commands* column. Enter HIV as the *Means of* variable, followed by OK. The output is shown in Figure 1.74.

**Figure 1.74**  
Mean of  
HIV  
coded 0,1  
with variance  
and standard  
deviation

HIV	Frequency	Percent	Cum Percent	
0	252	90.3%	90.3%	
1	27	9.7%	100.0%	
<b>Total</b>	<b>279</b>	<b>100.0%</b>	<b>100.0%</b>	

	Obs	Total	Mean	Variance	Std Dev
	279	27.0000	.0968	.0877	.2962
Minimum	25%	Median	75%	Maximum	Mode
0.0000	0.0000	0.0000	0.0000	1.0000	0.0000

Notice the variance of 0.0877 and the standard deviation of 0.2962. The formula for the variance of the HIV binomial variable, coded 0 or 1 and assuming a simple random sample, is...

$$V = pq = p(1 - p) = \frac{27}{279} \left( 1 - \frac{27}{279} \right) = 0.0874$$

This is slightly different from the value of 0.0877 shown in Figure 1.74. The variance of the mean is...

$$V_p = \frac{pq}{n} = \frac{p(1 - p)}{n} = \frac{\frac{27}{279} \left( 1 - \frac{27}{279} \right)}{279} = 0.0003133$$

We will later be comparing this variance to the variance of the mean when analyzed correctly as cluster sample. Now on to the analysis. With your left mouse, click on *Complex Sample Means* under *Advanced Statistics* in the *Analysis Commands* column. For *Means of* enter HIV and for *PSU* enter CLUSTER, then with the left mouse click *OK*. The output is shown in Figure 1.75.

**Figure 1.75**  
Mean of HIV coded 0,1 with standard error and 95% confidence limits

MEANS HIV PSUVAR=CLUSTER							
Next Procedure							
Forward							
Next							
	Count	Mean	Std Error	Confidence Limits		Minimum	Maximum
				Lower	Upper		
TOTAL	279	0.097	0.027	0.041	0.152	0.000	1.000

**Sample Design Included**

Weight Variable: None  
 PSU Variable: CLUSTER  
 Stratification Variable: None

Records with missing values: 0

Compare the outputs in Figure 1.65 (incorrect analysis) and 1.75 (correct analysis). Notice that both show that the prevalence of HIV is 9.7%; clearly this is correct. Where the two differ is in the size of the 95% confidence intervals, created from the variance of the value. In Figure 1.65 (incorrect analysis) the confidence limits extended from 6.5% to 13.8% or an interval that is about 7.3% wide (i.e., 13.8 - 6.5 = 7.3). In Figure 1.75 (correct analysis), the confidence limits extend from 4.1% to 15.2%, or about 11.1% wide, or about 52% wider than the analysis done incorrectly as a simple random sample. With wider confidence limits, the findings are deemed less precise or reliable (i.e., have a greater variance). Such increase in variance is typical of a cluster survey, and explains why you need to use special software that compensates for the larger variance in the analysis. The *Complex Sample* programs in *Epi Info* accounts for such increased variance.

Repeat the process, but for SEXA. For *Means of* enter SEXA and for *PSU* enter CLUSTER, then with the left mouse click *OK*. The output is in Figure 1.76.

**Figure 1.76**  
Mean of  
SEXA  
coded 0,1  
and 95%  
confidence  
limits

MEANS SEXA PSUVAR=CLUSTER

Next Procedure

Forward

Next

	Count	Mean	Std Error	Confidence Limits		Minimum	Maximum
				Lower	Upper		
TOTAL	279	0.186	0.035	0.115	0.257	0.000	1.000

Sample Design Included

Weight Variable: None  
PSU Variable: CLUSTER  
Stratification Variable: None

Records with missing values: 0

Again, compare the output with that in Figure 1.65 (incorrect analysis). Both show that the prevalence of SEXA is 18.6%. The point estimate remains the same, whether using the incorrect or correct software program. Instead the difference lies in the variance estimate, and the statistics that depend on the variance such as the 95% confidence interval. In Figure 1.65 (incorrect analysis) the confidence limits extended from 14.2% to 23.7% or an interval that is about 9.5% wide. In Figure 1.76 (correct analysis), the confidence limits extend from 11.5% to 25.7%, or about 14.2% wide. Hence again, the regular *Frequencies* program underestimated the variability in SEXA that was correctly noted by the *Complex Sample Means* program.

- **Complex Sample Tables.** For the next analysis, you will be doing a regular two-by-two analysis of an exposure variable (SEXA) related to an outcome variable (HIV), but this time using the correct *Tables* program for cluster survey data. Load *aidsal3.mdb* (with word labels for HIV and SEXA) rather than *aidsal4.mdb* which you have been using. Click with your left mouse on *Complex Sample Tables* under *Advanced Statistics* in the *Analysis Commands* column. Enter the variables as shown in Figure 1.77, including CLUSTER as the PSU or primary sampling unit. End by clicking *OK*.

**Figure 1.77**  
Analysis of  
crude  
association  
between  
SEXA and  
HIV

TABLES

Outcome Variable: HIV

Stratify by:

Exposure Variable: SEXA

Weight:

PSU: CLUSTER

Output to Table:

HEIGHT

	↑	ILL	
	+	-	
62.0	20	40	
62.5	37	59	

Save Only OK

Clear Help Cancel

The results of the two-by-two analysis is shown in Figure 1.78. The odds ratio of SEXA and HIV is 5.071 and the risk ratio is 4.054, the same as the non-survey data *Tables* analysis in Epi Info

(see Figure 1.66). Where the two programs differ is in the size of the confidence limits, reflecting the different variance of cluster surveys. In Figure 1.66, you earlier observed that the confidence interval for the odds ratio was 2.21 - 11.61 while for the cluster survey, Figure 1.78 shows a confidence interval of 2.33 - 11.053, or somewhat narrower than the similar statistic done with the incorrect *Tables* analysis. The same unusual finding is evident with the confidence interval for the risk ratio which was 2.03 - 8.10 with the *Tables* analysis (see Figure 1.66) versus 2.07 - 7.928 in Figure 1.78. Why? The answer lies with the nature of cross-tabulation analysis which reflects the joint variability of two variables, sometimes greater and sometimes lesser than cluster surveys.

Finally, observe the design effect, the measure of how much greater the variance of a complex survey is to a survey of the same number of subjects analyzed as a simple random sample. In Figure 1.78, the design effect is derived for the occurrence of HIV, first among those with SEXA = exp (i.e., 1.233), then those with SEXA = unexp (i.e., 1.735), and finally for the total values of HIV (i.e., 2.366). This means that the variance of the prevalence in relationship in our cluster survey is 2.366 times greater than if the data had been mistakenly analyzed as a simple random sample (larger variance means larger confidence interval).

$$SE_{p,clu} = \sqrt{V_{p,clu}} = \sqrt{\text{design effect} \times \frac{pq}{n}} = \sqrt{2.366 \times \frac{27}{279} \left(1 - \frac{27}{279}\right)} = \sqrt{0.000741} = 0.02723$$

Notice that this is the same value as shown in the bottom section of Figure 1.78 (i.e., 0.0273 = 2.723%). To derive the design effect for the odds or risk ratio in *Epi Info*, you need to do the calculations both with the incorrect analysis (i.e., using the *Statistics* commands which assume the data were derived as independent observations) and correct analysis (i.e., using the *Advanced Statistics Complex Sample* commands), then square the standard errors and compare the size of the variances (see equation below).

$$\text{Design effect} = \frac{V_{p,clu}}{V_{p,srs}} = \frac{SE_{p,clu}^2 \text{ (from Complex Sample commands)}}{SE_{p,srs}^2 \text{ (from Statistics commands)}}$$

## Epi Info

[Results Library](#)

Current View: C:\Epi\_Info418\laid3.mdb:A

Record Count: 279 Date: 4/3/2005 3:34:22 PM

### TABLES SEXA HIV PSUVAR=CLUSTER

[Next Procedure](#)

[Forward](#)

[Next](#)

SEXA	HIV		TOTAL
	ill	not ill	
<b>exp</b>	13	39	52
Row %	25.000	75.000	100.000
Col %	48.148	15.476	18.638
SE %	6.669	6.669	
LCL %	11.361	61.361	
UCL %	38.639	88.639	
Design Effect	1.233	1.233	
<b>unexp</b>	14	213	227
Row %	6.167	93.833	100.000
Col %	51.852	84.524	81.362
SE %	2.103	2.103	
LCL %	1.866	89.531	
UCL %	10.469	98.134	
Design Effect	1.735	1.735	
<b>TOTAL</b>	27	252	279
Row %	9.677	90.323	100.000
Col %	100.000	100.000	100.000
SE %	2.723	2.723	
LCL %	4.109	84.754	
UCL %	15.246	95.891	
Design Effect	2.366	2.366	

**Figure 1.78**  
Crude association between SEXA and HIV with survey data

### CTABLES COMPLEX SAMPLE DESIGN ANALYSIS OF 2 X 2 TABLE

Odds Ratio (OR) 5.071  
Standard Error (SE) 1.961  
95% Conf. Limits (2.33, 11.053)

Risk Ratio (RR) 4.054  
Standard Error (SE) 1.330  
95% Conf. Limits (2.07, 7.928)  
RR = (Risk of HIV=ill if SEXA=exp) / (Risk of HIV=ill if SEXA=unexp)

Risk Difference (RD%) 18.833  
Standard Error (SE) 6.048  
95% Conf. Limits (6.46, 31.203)  
RD = (Risk of HIV=ill if SEXA=exp) - (Risk of HIV=ill if SEXA=unexp)

#### Sample Design Included:

Weight Variable: None  
PSU Variable: CLUSTER  
Stratification Variable: None

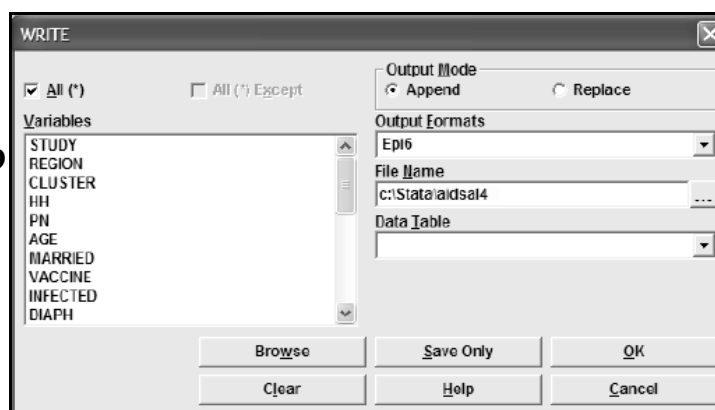
Records with missing values: 0

## ANALYSIS OF CLUSTER SURVEYS WITH STATA

When viewing the relation between more than two variables, the analysis for cluster surveys is not correct in *Epi Info*. Assume that you want to compare two variables (SEXA and HIV), controlling for the potential confounding effects of a third variable, DRUG. To do so, you might want to use the *Complex Sample Tables* programs in *Epi Info*, but you would run into problems. The program is set up the same way as the *Tables* program under *Statistics* in the *Analysis Command* column, but the designation “Stratify by” is not the same. In the *Tables* program, the term *Stratify by* refers to a potential confounding variable, to be adjusted with a Mantel-Haenszel Odds Ratio or Risk Ratio. In the *Complex Sample Tables* the term *Stratify by* refers to a third variable that unfortunately is not properly adjusted with a Mantel-Haenszel Odds Ratio or Risk Ratio. I alerted CDC to this error in their program via correspondence with Roger Friedman of CDC. He agreed that there is a problem but unfortunately his office does not have the financial resources, program personnel (for changing the *Epi Info* software) or technical writers (for updating the Help section) to make the correction at this time. Thus to derive a proper adjusted OR or RR, you will need to use *Stata*, the more sophisticated statistical program with modules for sample surveys.

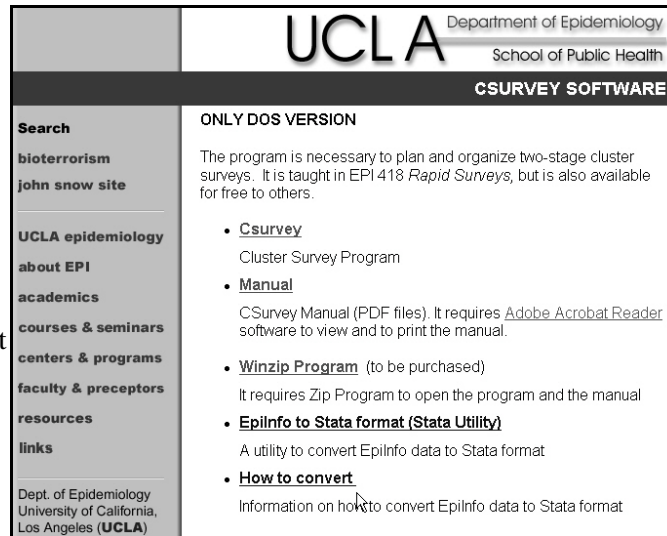
- **Create a Stata Data Set.** You will be doing a logistic regression analysis and other analyses in *Stata* which use variables coded as 0 or 1. For *aidsal4.mdb*, you recoded HIV, SEXA and DRUG as 0 and 1, so are all set, but will need to re-save *aidsal4.mdb* as *aidsal4.rec* (the file extension used by the DOS version of *Epi Info*). Then you must change *aidsal4.rec* to *aidsal4.dct* (needed to be recognized by *Stata*), and then to *aidsal4.dta* (a *Stata* dataset). To do so, load *aidsal4.mdb*, then with the left mouse click on *Write (Export)* under *Data* in the *Analysis Commands* column, entering the information as shown in Figure 1.79 followed by OK.

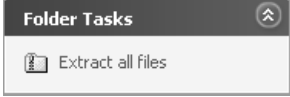
**Figure 1.79**  
Create and  
save  
*aidsal4.rec*



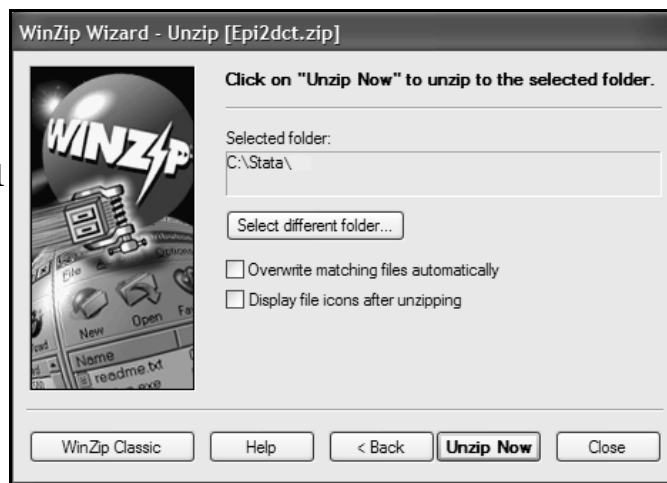
The *aidsal4.rec* file will then be placed in *C:/Stata/*, ready to be converted (in two steps) to a *Stata* file. To do this, you will need to use the *epi2dct* program found at the Epidemiology Department website at <http://www.ph.ucla.edu/epi/csurvey.html>, under *Epiinfo to Stata format* (see Figure 1.80). Click and follow the instructions.





**Figure 1.80**  
Software  
program  
to convert  
aidsal4.rec  
to aidsal4.dct



To unzip the downloaded *epi2dct* zip file (if you are using Windows XP), first use Windows Explorer to find the *epi2dct* zip file, then click on the file, and in the screen column at left, click either on  or if you are using Winzip, proceed as follows. When either the *Extraction Wizard* appears or *Winzip Wizard* appears, enter *Stata* (or whatever your Stata directory is termed), as shown in Figure 1.81.

**Figure 1.81**  
Extraction  
Wizard  
to unzip  
*epi2dct*.



If using the UCLA web instructions for *epi2dct*, be sure to use the file name *aidsal4* rather than *epi1* as in the example. Once *epi2dct* is ready for use, you will need first to click on  (bottom left of your main screen), followed by a click on **All Programs** , then click on  **Accessories** and finally, click on  **Command Prompt**. Change the director to C:/Stata (see Figure 1.82 for command – *cd\Stata*) then enter the program commands for *epi2dct*, as shown in Figure 1.82.

**Figure 1.82**  
Create  
aidsal4.dct.

```
C:\ Command Prompt
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.
C:\Documents and Settings\Owner>cd\Stata
C:\Stata>epi2dct aidsal4.rec aidsal4.dct
```

When through entering the information, press [enter], note the quick conversion, and read the following message: **Conversion complete... aidsal4.dct has been written to disk.** Thereafter, move *AIDSAL4.dct* to c:\Stata\data\.

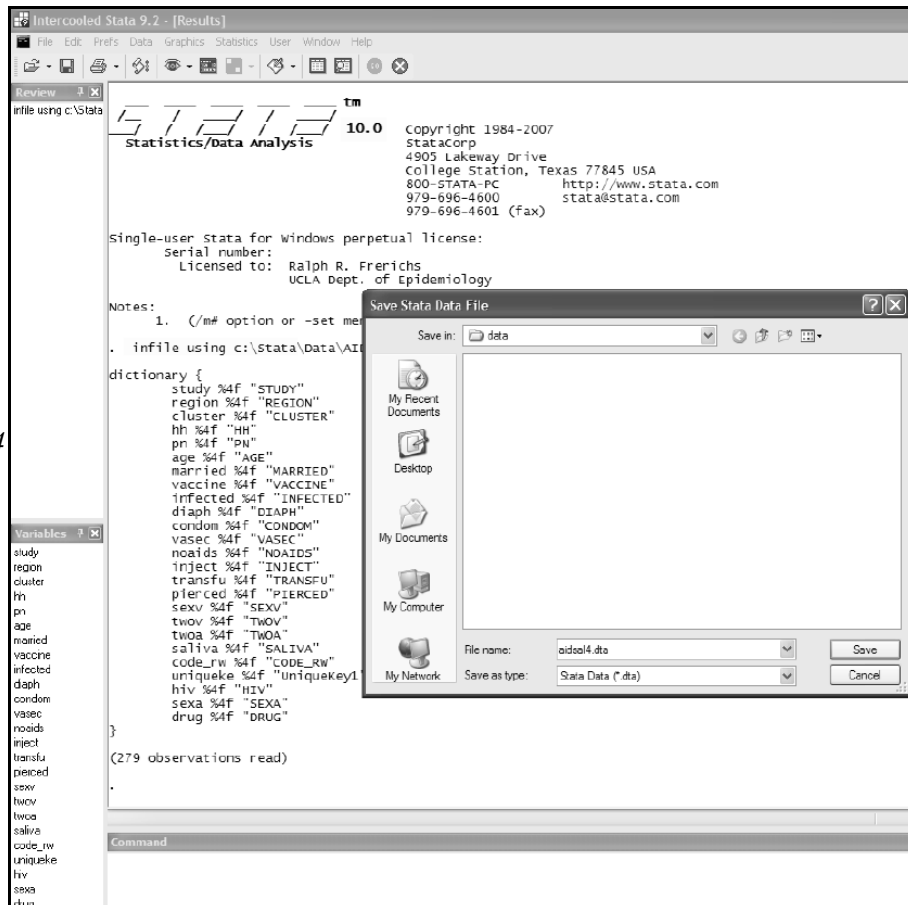
Start *Stata* and then load *AIDSAL4.dct* as shown in Figure 1.83.

**Figure 1.83**  
Stata  
command

```
Command
infile using c:\Stata\Data\AIDSAL4.dct
```

Once loaded, click with your left mouse on *File* at the top left, followed by *Save As*. In the screen that pops up, enter *aidsal4.dta* as shown in Figure 1.84.

**Figure 1.84**  
Save *aidsal4*  
in *Stata*



Once you are done, *Stata* acknowledges that all is well, by stating  

```
. save "C:\stata\data\aida14.dta"
file C:\stata\data\aida14.dta saved .
```

• **Mean Analysis in Stata.** First we will see how the *svy*mean analysis program in *Stata* compare to the *Complex Sample Means* program in *Epi Info*. Before doing the analysis, however, you need to tell *Stata* which variable (i.e., cluster) defines the primary sampling units (PSU). To do so, write *svyset cluster* in the *Stata Command* box. The program will respond

```
pweight: <none>
VCE: linearized
strata 1: <one>
SU 1: cluster
FPC 1: <zero>
```

in the *Stata Results* box, showing that it accepted the command and acted according. Next enter *svy: mean hiv sexa* to derive the proportion with HIV and the proportion who engaged in anal sex. The results are as shown in Figure 1.85.

**Figure 1.85**  
Mean  
in *Stata*  
of HIV  
and SEXA

```
. svy: mean hiv sexa
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =      1      Number of obs   =    279
Number of PSUs   =     30      Population size =    279
                                           Design df      =     29
```

	Mean	Linearized Std. Err.	[95% Conf. Interval]	
hiv	.0967742	.0272282	.0410863	.1524621
sexa	.1863799	.0346989	.1154127	.2573471

Please notice that the mean and 95% confidence interval are the same for *Stata* and *Epi Info* (see Figures 1.75 for HIV and 1.76 for SEXA). To derive the design effect which compares the variance of the cluster survey to the variance of a simple random sample of the same size, enter in the command line *estat effects, deff* as shown in Figure 1.85a.

**Figure 1.85a**  
Design effects  
for HIV and  
SEXA

```
. estat effects, deff
```

	Mean	Linearized Std. Err.	Deff
hiv	.0967742	.0272282	2.3579
sexa	.1863799	.0346989	2.20727

• **Odds Ratio Analysis in Stata (Logistic Regression).** One major strength of *Stata* is that you can determine odds ratios for cluster survey data that are adjusted for several confounding variables, as you did earlier with *Epi Info* in an incorrect analysis (i.e., assuming independence of observations, not appropriate for cluster surveys).

- **Crude Analysis.** First however, we will view the crude relationship between SEXA (the exposure or independent variable) and HIV (the outcome or dependent variable) to see how the program compares to *Epi Info*. While still in *Stata*, enter *svy: logistic hiv sexa*; the top section of Figure 1.86 should appear. Then enter *estat effects, deff* to determine the design effect for the odds ratio (in this instance, 0.809072, slightly smaller than an odds ratio derived by simple random sample). The results are shown in Figure 1.86.

**Figure 1.86**  
Odds ratio  
in Stata  
of HIV  
and SEXA

```
. svy: logistic hiv sexa
(running logistic on estimation sample)

Survey: Logistic regression
Number of strata   =      1          Number of obs   =    279
Number of PSUS    =     30          Population size  =    279
                                          Design df       =     29
                                          F( 1, 29)       =    18.17
                                          Prob > F        =    0.0002
```

	hiv	Odds Ratio	Linearized Std. Err.	t	P> t	[95% Conf. Interval]
	sexa	5.071429	1.931749	4.26	0.000	2.326995 11.05262

```
. estat effects, deff
```

	hiv	Coef.	Linearized Std. Err.	Deff
	sexa	1.623623	.3809082	.809072
	_cons	-2.722235	.3634633	1.72919

The size of the odds ratio in Figure 1.86 is the same as derived earlier with the *Tables* procedure of *Epi Info* (incorrect, since it does not take into account this being a cluster survey; see Figure 1.66) and with the *Complex Sample Tables* command of *Epi Info* (correct for cluster surveys; see Figure 1.78). In general, I favor the *Stata* analysis, but find the *Epi Info Complex Sample Tables* analysis acceptable, as long as the source is cited. The *Epi Info Tables* analysis procedure is not acceptable for cluster surveys.

**- Confounder-adjusted Analysis.** Next we will analyze the relationship between SEXA and HIV, controlling for the potential confounding effects of DRUG. That is, we will be using SEXA as the exposure variable, HIV as the outcome variable and DRUG as the confounding variable. While still in *Stata*, enter *svy: logistic hiv sexa drug* to have HIV serve as the dependent (or outcome) variable and SEXA and DRUG serve as the independent variables. Notice that the *logistic* command derives the odds ratios and 95% confidence intervals. To obtain the design effect (*deff*), enter *estat effects, deff*, as shown in Figure 1.87. By the way, this analysis was done before with the erroneous *Tables* command of *Epi Info*, as shown in Figure 1.69. This time, however, you used the survey analysis feature of *Stata* and logistic regression to correctly compute an adjusted odds ratio. The findings are as shown in Figure 1.87. Notice in *deff* with this analysis, that the variance of the odds ratio as done by a cluster survey is actually smaller than the variance of the odds ratio if done by a same-sized simple random sample. While with proportions such as prevalence or cumulative incidence estimates, the *deff* of a cluster survey is usually greater than 1.0, and sometimes much greater. Yet, when doing internal analyses of odds ratios, you never know what will happen to *deff*.

**Figure 1.87**  
Odds ratio  
in Stata  
of HIV  
and SEXA,  
by DRUG

```
. svy: logistic hiv sexa drug
(running logistic on estimation sample)

Survey: Logistic regression
Number of strata = 1
Number of PSUs = 30
Number of obs = 279
Population size = 279
Design df = 29
F( 2, 28) = 10.13
Prob > F = 0.0005
```

	hiv	Odds Ratio	Linearized Std. Err.	t	P> t	[95% Conf. Interval]	
sexa		6.421378	2.608484	4.58	0.000	2.797768	14.73821
drug		2.757895	1.809195	1.55	0.133	.7209287	10.55026

```
. estat effects, deff
```

	hiv	Coef.	Linearized Std. Err.	Deff
sexa		1.859633	.4062188	.743729
drug		1.014468	.6560059	.987953
_cons		-3.625842	.7064831	.984725

Here the adjusted odds ratio of 6.42 is similar but slightly greater than the Adjusted OR (MLE) in the *Tables* analysis of *Epi Info* (i.e., 6.32; see Figure 1.69) and much greater than the Adjusted OR (MH) in the same *Epi Info* program (i.e., 5.76; see Figure 1.69). Likely *Stata* uses a statistical procedure that creates a maximum likelihood estimate (MLE) of OR, rather than the Mantel-Haenszel (MH) version favored by epidemiologists. Finally, the confidence intervals are also different with the two programs. The erroneous *Tables* program of *Epi Info* with the *Adjusted OR (MLE)* produced a confidence interval of 2.60, 15.43 (see Figure 1.69), as compared to 2.80, 14.74 with *Stata* (see Figure 1.87). Thus the *Stata* confidence interval of the survey data is slightly narrower (as noted by a *deff* of less than 1.0 – see comment above), opposite to what was observed with prevalence estimates. This has more to do with the specific variability of the data in *aidsal4*, and cannot be generalized to other data sets.

• **Risk (or Prevalence) Ratio Analysis in Stata (Poisson Regression).**

When analyzing the relationship between an exposure variable and outcome variable, epidemiologists most often use risk ratios (i.e., risk of a disease among the exposed divided by risk of the disease among the unexposed) and next most commonly use odds ratios (i.e., odds among the exposed divided by odds among the unexposed). The *Epi Info* program derives both risk ratios (RR) and odds ratios (OR) for both regular data and data from cluster surveys. Yet for cluster surveys, the *Epi Info* program cannot be used to view the relationship between an exposure variable and an outcome variable after controlling for one or more confounding variables. To do this, you need to use *Stata*. How to derive a confounder-adjusted OR with *Stata* was presented previously. Here I will present how to derive a confounder-adjusted risk ratio (or prevalence ratio, if using prevalence data).

Earlier, as shown in Figure 1.66, you derived the relationship between SEXA and HIV using the *Tables* command (under *Statistics* in the *Analysis Commands* column). You observed that the risk ratio was 4.0536 with a 95% confidence interval of 2.0288 to 8.0993. That is, if there is no bias or additional confounding, you can be 95% confident that the true risk ratio in the sampled

population is bracketed by confidence interval. These data, however, were analyzed as if they came from a simple random sample, not a cluster survey. The correct analysis for a cluster survey was shown in Figure 1.78. Here the risk ratio was the same as with the *Tables* module (i.e., 4.054 vs 4.0536), but the limits of the confidence interval were narrower (i.e., 2.13, 7.71 versus 2.0288, 8.0993). As mentioned before, when doing point estimates for a single variable such as the prevalence of HIV infection or prevalence of anal sex, the confidence intervals for cluster surveys are usually wider than those derived for the same number of subjects in a simple random sample (SRS). When comparing one variable to another as we do in a risk ratio, however, there is no consistent pattern in variance estimates with respect to the SRS surveys versus cluster surveys.

Next, we will focus on how to use *Stata* to derive the risk ratio for SEXA as a risk factor for HIV, and for SEXA as a risk factor for HIV controlling for DRUG. To do so, you will be doing a Poisson regression analysis using *svy: poisson* to compute risk ratios or prevalence ratios.

- **Crude Analysis.** In *Stata*, click with the left mouse key on *File* and then *Open*, followed by *aidsal4.dta*. The *Review* screen should state: use "C:\Stata\data\aidsal4.dta", clear and the *Variables* screen should show the names of all the variables. In the *Stata Commands* box, enter *svy: poisson hiv sexa, irr*. Then enter *estat effects, deff* to derive the design effect. The output is shown in Figure 1.88. Notice again that *deff* is less than 1.0, suggesting our cluster survey analysis is more efficient than a simple random sample of the same size. Keep in mind, however, that you cannot generalize about *deff* when doing a risk or odds ratio.

**Figure 1.88**  
Poisson regression of SEXA and HIV

```
. svy: poisson hiv sexa, irr
(running poisson on estimation sample)

Survey: Poisson regression
Number of strata = 1          Number of obs = 279
Number of PSUs  = 30        Population size = 279
                                   Design df = 29
                                   F( 1, 29) = 18.21
                                   Prob > F = 0.0002
```

	IRR	Linearized Std. Err.	t	P> t	[95% Conf. Interval]	
hiv						
sexa	4.053571	1.329568	4.27	0.000	2.072519	7.928247

```
. estat effects, deff
```

	Coef.	Linearized Std. Err.	Deff
hiv			
sexa	1.399598	.3279992	.859538
_cons	-2.785893	.3410471	1.72919

The output indicates that the RR is 4.053571 (comparable to 4.054 and 4.0536 in *Epi Info*) and that the 95% confidence interval is 2.073 to 7.928, slightly greater than the confidence limits of 2.13, 7.71 presented with the correct *Epi Info* analysis. Why is there a difference? Likely it is due to different statistical procedures being used in the two programs. Since *Stata* is a more sophisticated software program, I suggest using the findings from it, although *Epi Info* remains acceptable, certainly for a univariate (i.e., one variable) analysis of cluster survey data and for a bivariate (i.e., two variable) analysis. Where *Epi Info* is not acceptable is for the analysis of more than two variables when using cluster survey data.

- **Confounder-adjusted Analysis.** For the final analysis, you will view the relationship between SEXA and HIV controlling for DRUG. To do so, enter *svy: poisson hiv sexa drug, irr*

followed by [enter] and *estat effects, deff*, also followed by [enter].

As observed in Figure 1.89, the adjusted risk ratio of SEXA related to HIV is 4.79 with 95% confidence limits of 2.43 and 9.43. Compare this to the Mantel-Haenszel adjusted RR shown in Figure 1.69 of 4.45 with the incorrectly derives confidence limits of 2.27 and 8.69. With the design effect being less than 1.0, we would expect the confidence interval to be narrower for the correct analysis, as occurred. Why the two point estimates for the adjusted RRs differ is explained by slight differences in the Mantel-Haenszel and Poisson regression methods. For cluster survey data, you should use *Stata*.

**Figure 1.89**  
Poisson regression of SEXA and HIV controlling for DRUG

```
. svy: poisson hiv sexa drug, irr
(running poisson on estimation sample)

Survey: Poisson regression
Number of strata =      1          Number of obs   =    279
Number of PSUs  =     30          Population size =    279
                                          Design df      =     29
                                          F( 2, 28)      =    10.98
                                          Prob > F       =    0.0003
```

	IRR	Linearized Std. Err.	t	P> t	[95% Conf. Interval]	
hiv						
sexa	4.785721	1.587555	4.72	0.000	2.428267	9.431883
drug	2.31283	1.276959	1.52	0.140	.7477104	7.154083

```
. estat effects, deff
```

	Coef.	Linearized Std. Err.	Deff
hiv			
sexa	1.565637	.3317275	.814865
drug	.8384718	.5521196	1.01297
_cons	-3.530031	.6180684	1.05453

### • Risk (or Prevalence) Difference Analysis in Stata.

So far, you have learned how to derive risk ratios and odds ratios (or if the outcome is a prevalence estimate, prevalence ratios and prevalence odds ratios). Often, however, you may want to compare the difference between one group or another, subtracting the prevalence or incidence point estimate of one finding from that of another. The risk difference is routinely derived in *Epi Info*. In this final section I will show you how to do the same in *Stata* using the *svymean* and *svylc* commands.

As before, while in *Stata* read the data file *aidsal4.dta* from the appropriate computer directory. Use *svyset* to set the primary sampling unit (PSU) as *cluster* by entering *svyset cluster*. You will be comparing the risk difference between HIV among those who reported “yes” to anal sex (i.e., SEXA=1) versus those who reported “no” to anal sex (SEXA=0). Enter *svy: mean hiv, over(sexa)*, followed by [enter] and *estat effects, deff* and [enter]. The output in Figure 1.90 should appear.

**Figure 1.90**  
HIV  
occurrence  
by SEXA

```
. svy: mean hiv, over(sexa)
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      1      Number of obs   =    279
Number of PSUs  =     30      Population size =    279
                                   Design df      =     29

      0: sexa = 0
      1: sexa = 1
```

over	Mean	Linearized Std. Err.	[95% Conf. Interval]	
<b>hiv</b>				
0	.061674	.0210337	.0186552	.1046928
1	.25	.0666891	.1136055	.3863945

```
. estat effects, deff

      0: sexa = 0
      1: sexa = 1
```

over	Mean	Linearized Std. Err.	Deff
<b>hiv</b>			
0	.061674	.0210337	1.72919
1	.25	.0666891	1.229

As you can see, there are two estimates of HIV infection, 25 percent among those who reported anal sex (i.e., the exposed group – listed as “over” with value 1) and 6.2 percent among those who did not reported anal sex (i.e., the unexposed group – listed as “over” with value 0). For the risk difference, we want to know first what is the difference between these two numbers and second is the difference statistically significant? To determine this, follow the above command by entering *lincom [hiv]1 - hiv[0]* followed by [enter] and *estat lceffects [hiv]1 -[hiv]0, deff*, also followed by [enter]. This tells the computer to compare the linear combination of HIV among those with SEXA values of 1 versus among those with SEXA values of 0, and to compute the design effect for the linear combination. The results are shown in Figure 1.91.

**Figure 1.91**  
Difference in  
HIV by SEXA

```
. lincom [hiv]1-[hiv]0
( 1) - [hiv]0 + [hiv]1 = 0
```

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
(1)	.188326	.0604821	3.11	0.004	.0646261	.3120259

```
. estat lceffects [hiv]1-[hiv]0, deff
( 1) - [hiv]0 + [hiv]1 = 0
```

	Coef.	Std. Err.	Deff
(1)	.188326	.0604821	.944123

The difference in HIV between SEXA=1 and SEXA=0 is 18.8 percent with 95% confidence limits of 6.5 to 31.2 percent. You did the same analysis earlier with the regular *Analysis Command*

of *Epi Info* (see Figure 1.66), done without consideration that the data came from a cluster survey. The value of the risk difference is the same as before but there is a slight difference in the confidence limits. Note that the risk difference derived in *Stata* is also very similar to the value derived in *Epi Info* with the *Advanced Statistics* command (see Figure 1.74). With some variables the variance of the incorrect analysis (i.e., *Epi Info* with regular *Analysis Command*) may not differ much from the variance done with the correct analysis (i.e., an analysis that takes into account the effect of *cluster*). Notice that *deff* in this analysis is 0.94, indicating that the confidence interval will be very close to that of a simple random sample of the same size. Since you do not know ahead of time if the variance will be larger or smaller than an equivalent sized simple random sample, with rapid survey data you should always use either the *Advanced Statistics* commands of *Epi Info* or the survey commands of *Stata*.

■ **Summary.** All statistical tests have assumptions that may or may not be met. Usually, the value of these tests are debated by statisticians and evaluated by statistics graduate students. Epidemiologists have long favored the Mantel-Haenszel estimators for both the odds ratio and risk ratio, especially useful when there are less than 10 subjects per stratum. The reason is that the Mantel-Haenszel estimators are more accurate over a wider range of values. Nevertheless, the Maximum Likelihood estimators are also popular and tend to be used in many statistical packages. For survey data, I suggest using *Complex Sample* modules of *Epi Info* or the *svy* programs of *Stata*. I do not suggest using the regular statistics of *Epi Info*, although the program is very useful for data entry, editing and preliminary analysis. For advanced analyses that consider more than two variables, I suggest using *Stata* rather than *Epi Info*.

## CONCLUDING REMARKS

---

The beauty of the *Epi Info* program is that it has enabled epidemiologists around the world to analyze their data and use statistics to enhance their insights into epidemiologic processes. To effectively communicate information to policy- and decision-makers, epidemiologists need to be able to convey their findings in an understandable manner. Standard errors (or even more fundamental, variances) are not understandable to most people. On the other hand, confidence intervals are very effective at conveying both findings and the uncertainty of findings. We have come far in epidemiology in our ability to simplify our research findings. This is our strength. By creating a free software program that serves the needs of epidemiologists and sampling statisticians with parameter estimates and confidence limits, CDC and the World Health Organization have done much to promote rapid surveys as instruments for gaining information in developing countries.

While good, the *Epi Info* program is not perfect for clusters surveys. The program can analyze the prevalence or incidence of disease or conditions (derived as proportions), and the odds and risk ratios relating two variables, such as risk factor to disease. It also can determine the difference between two proportions, measured as a risk difference. What the program cannot do, however, is do more complicated analyses that involve confounding or intervening variables. Fortunately, there are other programs available that do such advanced analysis. The one featured in this class is *Stata*.

Neither *Epi Info* nor *Stata* derive sample size estimates for cluster surveys or select clusters with probability proportionate to size (PPS). For these calculations you will need to use the *C survey* program, either version 1.5 (DOS) created by Dr. Iwan Ariawan of the University of Indonesia and Professor Ralph R. Frerichs of UCLA with support of the UCLA/Fogarty HIV/AIDS

Training Program, or version 2.0 (Windows) created by Muhammad N. Farid (also supported by the Fogarty program) in collaboration with Professor Frerichs. We will discuss features of this program in greater detail during the Rapid Survey Course.