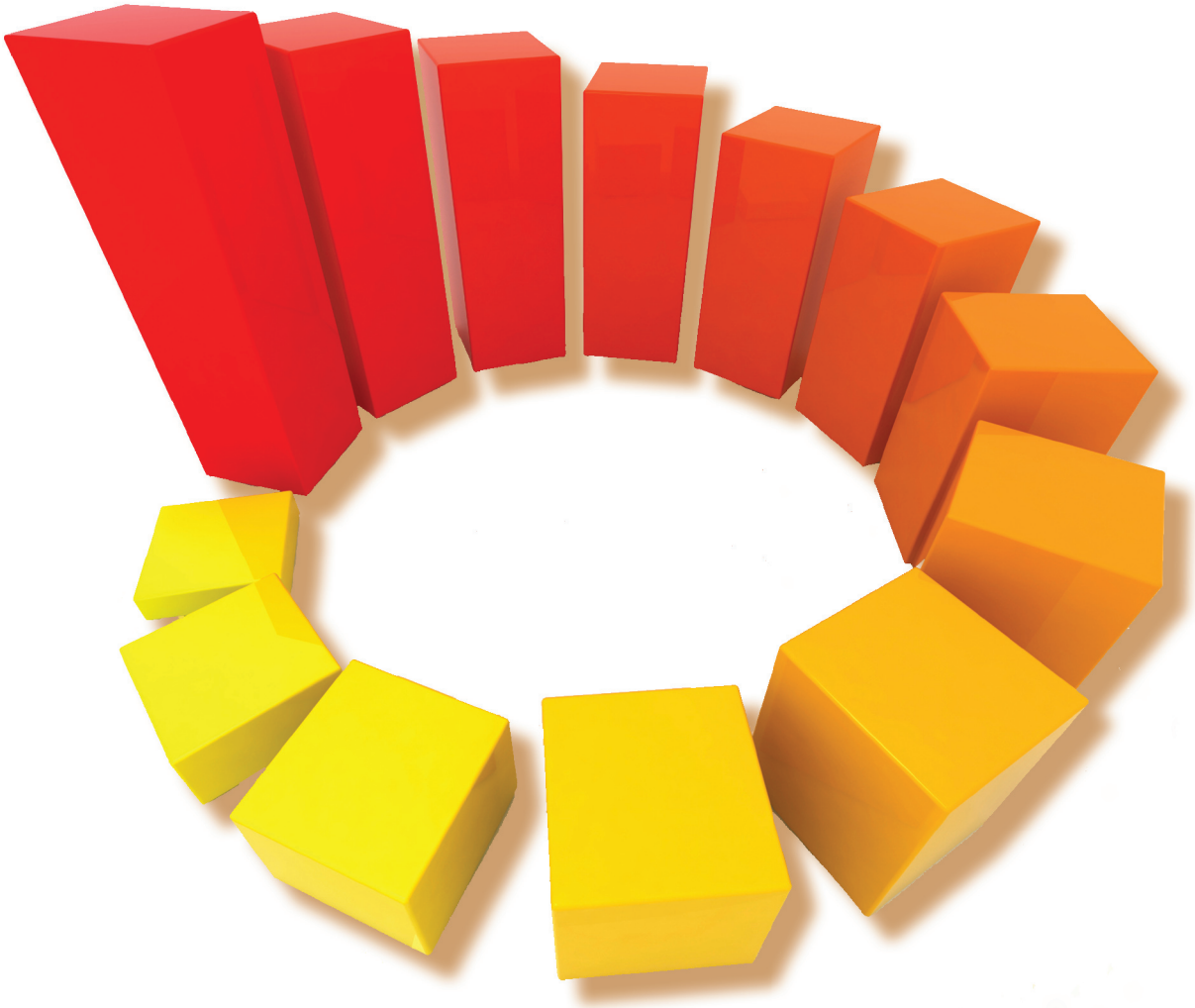
Advanced Management and Analysis of Data Using Epi Info for Windows

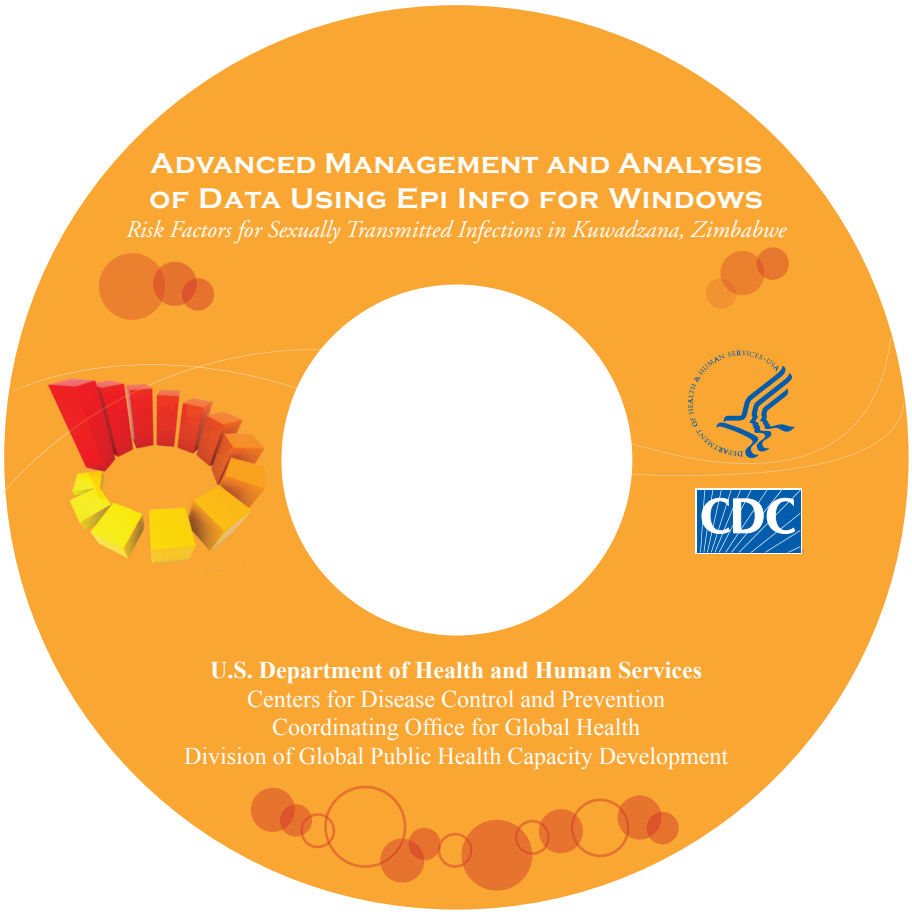
Risk Factors for Sexually Transmitted Infections in Kuwadzana, Zimbabwe



U.S. Department of Health and Human Services
Centers for Disease Control and Prevention
Coordinating Office for Global Health
Division of Global Public Health Capacity Development

On the CD

The accompanying CD contains the necessary data files referred to in this manual, scanned copies of the questionnaires for review, and a PDF version of the manual.



Atlanta, GA: U.S. Centers for Disease Control and Prevention, 2006. Reprint 2008.

Advanced Management and Analysis of Data Using Epi Info for Windows

Risk Factors for Sexually Transmitted Infections in Kuwadzana, Zimbabwe

U.S. Department of Health and Human Services
Centers for Disease Control and Prevention
Coordinating Office for Global Health
Division of Global Public Health Capacity Development

Development Team

Technical Advisor And Developer: Donna Jones, MD, MPH, CDC/COGH/DESCD

Instructional Designer: Nadine Sunderland, MEd, CDC/COGH/DESCD

Technical Review: Tsitsilina Apollo, MBChB, MPH University of Zimbabwe
Peter Nsubuga, MD, MPH, CDC/COGH/DESCD
Henry Walke, MD, MPH, CDC/COGH/DESCD
Karen Gieseker, PhD, CDC/COGH/DESCD

Acknowledgements

Dr. Mufuta Tshimanga
Dr. M. Wellington
University of Zimbabwe
Department of Community Medicine
Masters of Public Health Programme

Centers for Disease Control and Prevention
National Center for HIV/AIDS, STIs and TB
Global AIDS Program
Zimbabwe

Advanced Management and Analysis of Data Using Epi Info for Windows

Risk Factors for Sexually Transmitted Infections in Kuwadzana, Zimbabwe

U.S. Department of Health and Human Services
Centers for Disease Control and Prevention
Coordinating Office for Global Health
Division of Global Public Health Capacity Development

Development Team

Technical Advisor And Developer: Donna Jones, MD, MPH, CDC/COGH/DESCD

Instructional Designer: Nadine Sunderland, MEd, CDC/COGH/DESCD

Technical Review: Tsitsilina Apollo, MBChB, MPH University of Zimbabwe
Peter Nsubuga, MD, MPH, CDC/COGH/DESCD
Henry Walke, MD, MPH, CDC/COGH/DESCD
Karen Gieseker, PhD, CDC/COGH/DESCD

Acknowledgements

Dr. Mufuta Tshimanga
Dr. M. Wellington
University of Zimbabwe
Department of Community Medicine
Masters of Public Health Programme

Centers for Disease Control and Prevention
National Center for HIV/AIDS, STIs and TB
Global AIDS Program
Zimbabwe

Table of Contents

About This Training 7
General Instructions 9
Analytic Strategy 11
STEP 1: Problem Identification 12
STEP 2: Development of Research Question 13
STEP 3: Hypothesis Generation 13
Examining Risk Factors 13
Developing Plausible Hypotheses 15
STEP 4: Determine Study Design and Data Needed 16
STEP 5: Plan Draft Analysis 17
Structuring Analysis Tables 17
STEP 6: Proposal Development and Questionnaire Design 21
STEP 7: Study Implementation/Data Collection 21
STEP 8: Data Analysis 22
Enter and Edit Data 22
Checking for Duplicate Records 23
Delete Duplicate Record 27
Documenting Errors and Changes to the Dataset 28
Miscoded, Missing, or Out-of-Range Values 30
Consistency/Logic Checks 34
Descriptive Statistics: Univariate Analysis 41
Categorical Variables 41
Continuous Variables 44
Grouping Data 46
Simple Cross-Tabulations 54
Bivariate Analysis 54
Grouping Responses for Comparison 56
Create a New Variable from the Results of Two or more Other Variables 63
Documenting New Variables 67
Create a New Data Table 67
Measures of Association 69
Tests of Statistical Significance 72
Null Hypothesis 72
P-Value 72
Testing Data in a 2x2 Table 73
Fisher Exact Test 73
Chi-Square Test 73
Confidence Intervals 76
Confidence Intervals for Measures of Association 76
Interpreting the Confidence Interval 78
Confounding and Effect Modification 81
Stratified Analysis 82
Data-Based Approach to Effect Modification and Confounding 83

Stratification for Subgroup Analysis 88
Frequency Matching and Stratification..... 90
Multivariate Analysis..... 93
Logistic Regression..... 93
Step 9: Interpretation and Reporting..... 116

References..... 117
Appendix A: Questionnaire 121
Appendix B: Answers to Training Questions 125
Appendix C: Data Tables..... 129
Appendix D: Statistical Formulas..... 133

About This Training

Description of the Training

Using a scenario based on an actual study of factors associated with having a sexually transmitted infection (STI) in Zimbabwe and data derived from the case-control study conducted, you will learn how to manage and analyze study data. In this training you will use Epi Info for Windows software to perform data management and analysis activities commonly encountered in an analytic study and investigation.

The training is designed as a self-study activity but also works well in a classroom setting.

Time

10 hours

Instructional Goal

The goal of this course is to teach the learner to *analyze and interpret data from descriptive and analytic studies*.

Learning Objectives

1. Identify the key elements of planning a data analysis prior to data collection.
2. Use a systematic approach to data management and editing.
3. Describe how to conduct quality-control checks, including identification of duplicates and missing data.
4. Recode data variables to facilitate data analysis and document the editing process.
5. Conduct and interpret univariate analyses for a case-control study.
6. Discuss when to perform a stratified analysis.
7. Calculate a summary risk estimate using the Mantel and Haenszel test.
8. Conduct and interpret bivariate and stratified analysis.
9. Differentiate between effect modification and confounding.
10. Identify presence of confounding using stratified analysis.
11. Identify presence of effect modification in a dataset.
12. Present findings to interpret effect modification.
13. Describe how frequency matching controls to cases reduces confounding.
14. List the advantages and disadvantages of matching.
15. Calculate measures of association in a case-control-study.
16. Conduct and interpret multivariate analysis.
17. Describe the purpose of using logistic regression.
18. Describe the process to create a logistic regression model.

Training Techniques

This training presents *content, examples,* and *practice* for each learning objective through assigned reading incorporated in the material and multiple practice and problem solving exercises.

Assessment

The training will be assessed based on trainee's ability to analyze data from epidemiologic studies and related activities conducted in the field.

Materials and Equipment

Materials	1. This Training Manual 2. Questionnaire file 3. Datasets (included on CD-ROM or from internet)
Equipment	1. Computer with MS Windows software 2. Epi Info software (included on CD-ROM or can be found on the internet at http://www.cdc.gov/epiinfo.

Target Audience

The intended audience for this training activity is first-year Field Epidemiology Training Program (FETP) and Field Epidemiology Laboratory Training Program (FELTP) trainees who have previously taken a basic course on Epi Info; however, anyone with the following prerequisites will benefit from the training.

Prerequisites

You should have an understanding of

- Basic epidemiology, including study design
- Basic biostatistics, including descriptive statistics, use of rates and ratios; use of logistic regression requires some familiarity with multivariate analysis
- How to develop questions for an investigation
- Basic functions of Windows-based computers

Additional background in using Epi Info is helpful but not necessary. However, this training is not intended to teach the user basic aspects of Epi Info so you may wish to take a basic course on Epi Info before participating in this training. An example of this course is the *Using Epi Info in an Outbreak Investigation (Cholera in Rwenshama)*. This training can be found on the Centers for Disease Control and Prevention website at http://www.cdc.gov/descd/materials.html.

Resources

The list of references for the training can be found on page 117.

General Instructions

In this training certain icons are used below to illustrate an action or note.

Icon	Description
?	The question icon indicates that the trainee needs to complete a question. If answers are provided in the material (e.g., in an appendix), tell the trainee where they can be found.
🗋	The note icon is used for notes and other useful information, including Hints.
✍	The handwriting icon indicates an activity where you will need to document your response in the participant guide. Directions are included and, if there are answers provided, you will be advised where you can find them in the appendices.
FYI	Boxes with a "For Your Information" (FYI) icon provide you with additional information about the topic being discussed.

Analytic Strategy

This training follows the approach to an analytic study of risk factors for sexually transmitted infections in a country in Africa. While there are several types of analytic studies, all follow a general analytic strategy. The steps for approaching this analysis are:

STEP 1: Problem identification

STEP 2: Development of research question

STEP 3: Hypothesis generation

STEP 4: Determine study design and data needed

STEP 5: Plan draft analysis

STEP 6: Proposal development

STEP 7: Study implementation/data collection

STEP 8: Data analysis

- Enter and edit data
- Descriptive statistics
- Simple cross-tabulations
- Measures of association
- Tests of significance and confidence intervals
- Stratified analysis
- Multivariate analysis

STEP 9: Interpretation and reporting

* Dicker RC. Analyzing and interpreting data. In: Gregg MB, editor. Field Epidemiology. 2nd ed. New York: Oxford University Press; 2002. p. 132-172.

This training addresses each step in this outline, but the primary focus is on the steps as they relate to data analysis (Step 8).

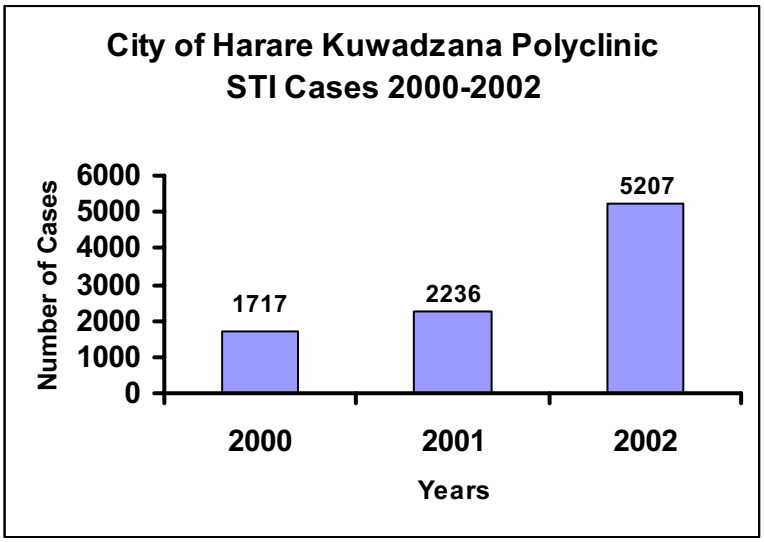
STEP 1: Problem Identification

Problem identification usually occurs because of the appearance of a new health problem or an increase in a known health problem. The following scenario sets out our problem.

Scenario

Harare City had observed a decline in the numbers of reported sexually transmitted infection (STI) cases from 1991 to 2000. However, in 2001 a total of 49,166 cases were reported representing a 2.4% increase from the previous year.

A significant increase in cases was also observed at the City of Harare Kuwadzana Polyclinic, located in the Kuwadzana suburb in the western district of Harare. From 2000 to 2002, there was a 200% increase in STI cases (from 1717 cases to 5207 cases).



The continued increase in STI cases in this area was identified as a problem. As one method of gaining further insight into the problem, it was decided to conduct a study to examine the risk factors for STI among patients attending the clinic.

Background Information on the Area

Kuwadzana suburb is located about 14 kilometers southwest of Harare city centre. It has a total population of 110,869 according to the 2002 population estimates. The actual population may be much bigger than this due to new inhabitants in Kuwadzana extension. Kuwadzana is a high-density suburb and its residents are mainly engaged in informal small-scale trade. It is generally characterized by a high unemployment rate and poverty. HIV infection is a significant problem as reflected by the antenatal care (ANC) sentinel surveillance of 2000 when 27.8% (Zimbabwe Ministry of Health & Child Welfare, 2000) of all women attending the ANC clinic during the survey period were HIV positive, similar to national figures. The health needs of the community are served by Kuwadzana polyclinic.

STEP 2: Development of Research Question

Based on the identified problem, the researchers have posed a research question which is the objective of the study:

Research Question What is the association between commonly known social and behavioral risk factors and contracting an STI for persons attending Kuwadzana Polyclinic?

STEP 3: Hypothesis Generation

"The hypothesis is a version of the research question that provides the basis for testing the statistical significance of the findings" (Hulley, p.7). Developing your research question is critical in planning your analysis and assuring that the relevant data will be appropriately collected, recorded, managed, analyzed, and interpreted. The research question you consider will drive your study design, data collection, and analyses. You will need to decide which data to collect to test the hypotheses, so this will be critical in the design of your questionnaire. If you do not ask for information about a potential risk factor or ask for it in an appropriate way, you will not be able to examine that risk factor in your analysis.

Examining Risk Factors

What do we know about risk factors for STI in general and in Zimbabwe specifically?

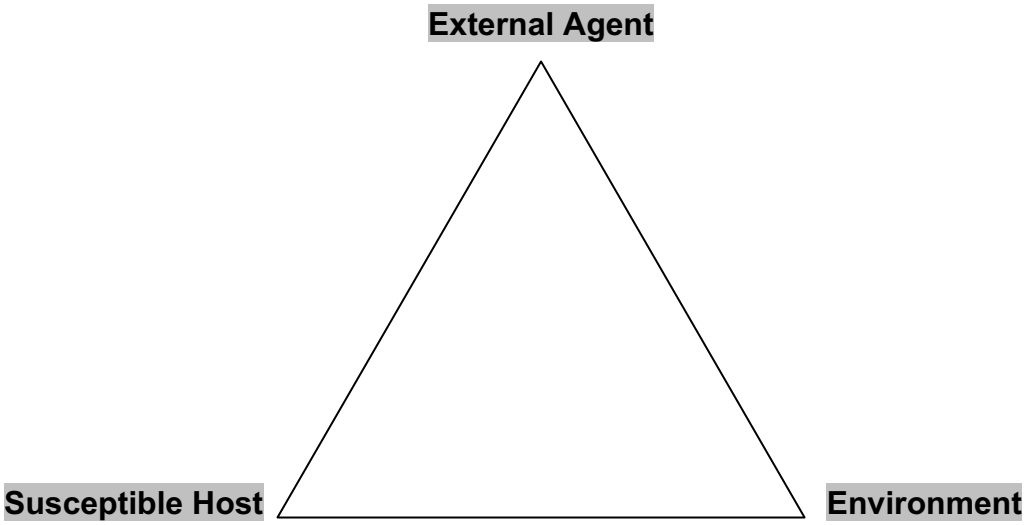
Behavioral risk factors

Several behaviors have been identified to increase the risk for STIs. In general, STIs are acquired through unprotected sex with an infected partner. Increasing the number of partners increases the risk by increasing potential exposure. The reasons for unprotected sex include alcohol and drug use, separation from regular partner, and economic need. Several studies have shown that payment for sex increases the STI risk (Wellington & Ndowa, 1997; Da Costa et al, 1985). Intimate partner violence has been associated with increasing risk of STI and HIV among women (Bauer et al, 2002).

Consistent and correct condom use (protected sex) has been promoted as a means of preventing HIV and other STIs.

Epidemiologic Triad

We can use the structure of the epidemiologic triad to help in structuring our considerations. "The epidemiologic triad is the traditional model of infectious disease causation. It has three components: an external agent, a susceptible host, and an environment that brings the host and agent together" (CDC, p. 35).



For this increase in STIs in Zimbabwe , there are likely to be several specific agents causing disease. We are looking at syndromes caused by a number of infectious agents but all known to be passed via sexual contact. We will not be focusing specifically on the agents. Because STIs result from personal behavior, we will examine the host characteristics/behaviors that increase the risks for infection by this mode. In addition, there may be environmental characteristics that make it more likely that persons will engage in the behaviors that cause risky sexual contact.

Proposed Risk Factors

Risk factors are those behaviors or exposures that are associated with an increased possibility of infection.

Below, list at least five possible risk factors that would increase an individual's risk of acquiring an STI.

1.

2.

3.

4.

5.

In the scenario we have presented, the following are some of the possible risk factors for acquiring an STI.

- Having multiple sexual partners
- Not using a condom
- Having sex with commercial sex workers
- Not living with a spouse/partner (due to increased likelihood of multiple partners)
- Being forced to have sex
- Using alcohol when associated with sexual activity

Developing Plausible Hypotheses

What possible hypotheses will you explore in your investigation?

Hypotheses for Acquiring Disease based on Proposed Risk Factors

- Having multiple sexual partners increases a person's risk of acquiring an STI.
- Not using condoms increases a person's risk of acquiring an STI.
- Paying or receiving payment for sex increases a person's risk of acquiring an STI.
- Not living with spouse/partner increases a person's risk of acquiring an STI.
- Being the victim of sexual violence increases risk of acquiring an STI.
- Use of alcohol during sexual encounters increases a person's risk of acquiring an STI.

Now that we have developed plausible hypotheses we need to further plan our analysis by clearly indicating our outcome of interest, exposures of interest, potential confounders, and variables for subgroup analysis.

Outcome of interest

The outcome of interest is the effect or outcome we are planning to study. Later, in analysis, you will see that this variable is the one we refer to as the *dependent* variable.

In this exercise, we will focus on persons diagnosed with an STI between May and July, 2003, in Zimbabwe.

Exposures of interest

The exposures of interest are the causes or determinants of disease, the outcome of interest. In analysis, they are the *independent* variables.

- Multiple sexual partners
- Non-use of condoms
- Paying for sex
- Not living with partner
- Forced to have sex
- Alcohol use

Potential Confounders and Effect Modifiers

Confounders and effect modifiers are additional variables or factors which may affect the apparent association between an exposure of interest and an outcome of interest. You will learn more about confounding and effect modification on page 81.

Example: Living with one's partner may reduce number of sexual partners.

Variables for Subgroup Analysis

Certain factors can vary significantly among subgroups of the population; therefore, it is often useful to examine the factors separately for those groups. Note that when determining sample size for a study, it is important to consider whether subgroup analysis is needed. Depending on the strength of the association between the outcome and the risk factor in subgroups, you must make sure to choose a sufficient sample size for the subgroup to obtain statistically significant results.

Example: If gender is the group of interest, you will need a sufficient number of both subgroups (male and female) to examine the effect in each group.

STEP 4: Determine Study Design and Data Needed

The next step is to design the study that can best examine our hypotheses. Conducting a study to examine factors associated with contracting an STI in this community is part of conducting analytical epidemiology. While descriptive epidemiology focuses on the distribution of disease in a population, analytic epidemiology is concerned with examining the determinants of disease through testing of hypotheses, often formulated through descriptive studies (Hennekens, p.16). The process of analytic epidemiology involves quantifying or estimating the association between exposure and disease and testing the statistical significance of the association. There are a number of study designs that can be used. These should have been covered in your epidemiology course and/or text.

For this study, a case-control design was chosen because

- It is relatively quick to perform,
- We are looking at only one outcome of interest (acquiring an STI), and
- It allows us to examine multiple factors related to that outcome.

STEP 5: Plan Draft Analysis

Structuring Analysis Tables

To structure the analysis, it is useful to create analysis tables. Analysis tables (also called table shells, empty tables, or dummy tables) are used to lay out the tables one expects to produce when the data are analyzed. Creating these tables **before** the actual data collection allows the investigator to be very clear about what is to be studied and how it will be studied. As in the actual analysis, one should move from the simple to complex. We will conduct our analysis in the following order:

- *Univariate* tables (examining single variables—descriptive frequencies),
- *Bivariate* tables (examining two variables or 2x2 tables—risk factor and outcome),
- *Stratified analysis*, and
- *Multivariate analysis* (e.g., logistic regression).

We start with basic descriptive (univariate) tables. We have chosen a case-control study design and our tables need to reflect this design. For a case-control study, always present the data for cases and controls separately.

Univariate Table Activity

Below is an example of a table shell for descriptive characteristics. Cases and controls are presented separately. Several variables that describe the cases and controls can be considered in a single table. The table includes space to identify the number of responses (n) and the percent of the total (%) that the response represents for each risk factor by case or control subgroup.

Directions: We have listed two possible demographic variables. Add three other possible demographic variables to complete this univariate analysis table.

Case-control study: Descriptive table

Descriptive variables	**Cases**		**Controls**	
	n	(%)	n	(%)
Sex: Female				
Male				
Marital status: Single				
Married				
Divorced				
Widowed				

Now we will create our bivariate or 2x2 tables.

Bivariate Table Activity

Bivariate analysis involves the comparison of two variables to each other. Usually we are comparing the outcome of interest (dependent variable) to one of the exposures of interest (independent variables). In this comparison we will note the measure of association, which is a measure showing the strength of the relationship between the exposure and outcome of interest. For case-control studies, the measure of association is the odds ratio (OR). For a cohort study, the measure of association is relative risk (RR).

We are conducting a case-control study, so we will concentrate on odds ratios in this training. Measures of association will be discussed in more depth on page 69. The table below illustrates how the 2x2 table should be created. The odds ratio and lower and upper confidence intervals (CI) should be documented for each table. It is useful to include both actual numbers in each cell and also percentages of the whole. For example, next to the actual number of exposed cases, you should include the percentage of exposed cases out of the total cases. The percentage can often be more meaningful as a comparison tool than the actual number.

2x2 Table Format

Exposure	**Case (% of cases)**	**Control (% of controls)**	**OR (95% CI, lower - upper)**
Present	n (%)	n (%)	
Absent	n (%)	n (%)	

Directions: Create examples of at least two possible 2x2 tables to examine the risk factors for contracting STIs (refer to the hypotheses for acquiring STIs on page 15).

Example

Exposure	**Case**	**Control**	**OR (95% CI, lower - upper)**
More than 1 sexual partner last year			
1 sexual partner last year			

Exposure	**Case**	**Control**	**OR (95% CI, lower - upper)**

Exposure	**Case**	**Control**	**OR (95% CI, lower - upper)**

What are the variables you would need to be able to complete your table shells?

Consider how variables will be defined. For example, do we want all possible number of sexual partners or are we going to pre-code for standard answers? For example, we could either ask an open-ended question such as, "How many partners have you had in the last year?" or a closed or yes/no question with predetermined responses such as, "Did you have more than 1 partner in the last year?"

Remember that when working with continuous variables such as age or number of partners it is generally preferable to ask an open-ended question and later recode the variable into categories if needed. Pre-coding a continuous variable into categories may lead to loss of information.

For example, if you create the age variable so that responses are categorized (<1, 1-5, 6-10, 11-15, etc.), you will not be able to analyze data for a specific age (e.g., 8 year olds) and you reduce your ability to recode the age groups if you discover that the original categories you selected were not appropriate for the intended analysis.

Suggestions:

- Be able to justify the inclusion of each variable.
- Avoid the temptation to include variables that "might be interesting."
- Do not limit data collection tools (e.g., questionnaires) to risk factors. Collect appropriate identifying information, perhaps clinical information, and other descriptive factors related to time, place, and person to be able to adequately characterize the population and assess comparability between your cases and controls.

STEP 6: Proposal Development and Questionnaire Design

At this point you would use the information we have identified to create a proposal outlining exactly how and where you would do your study, your sampling methodology, and your plan for data collection. You would include the questionnaire or other data collection instrument needed for your study. We will not be reviewing that process here.

STEP 7: Study Implementation/Data Collection

Since this training is primarily concerned with analysis of data, we will not go into great detail about study implementation. We will assume that you created a questionnaire based on the hypotheses and variables you previously defined. After completing the questionnaire, you will have conducted the study.

Assume you determined an appropriate sample size and have interviewed 113 STI cases and 113 controls from the same clinic. Cases were defined as any new patient seen at Kuwadzana Primary Care Clinic from 26 May, 2003, to 19 July, 2003, and treated for any STI (as defined by clinical protocol). Controls were defined as any new patient at Kuwadzana Primary Care 16 years and older who reported sexual activity and without a diagnosed STI during the same period. Controls were frequency matched for gender and age group. *Frequency matching* is a method of matching controls to cases on the basis of certain categories, such as age or sex, so that the proportion of controls and cases is the same in each category. This may also be referred to as category matching.

NOTE: In this study you may want to have sufficient numbers in some specific groups so that you can analyze the data for each as a specific subgroup. This is what was done for gender – a nearly equal number of male and female cases were identified to obtain sufficient numbers in each group. The researcher had prior information that male STI patients in this community are less likely to attend the public clinic for care, making it harder to examine behavioral risk factors for males if the sample size in this subgroup was too small.

Because you have frequency matched on age group and gender, these should not differ between cases and controls. This means that we will be unable to examine age or gender as possible risk factors. When you have frequency matched (also referred to as group or category matching), you do not need to do a pair-matched analysis (Hennekens, p. 215); however, it is recommended that you stratify on these factors. We will discuss this in-depth later in the analysis.

You have interviewed all the cases and controls.

STEP 8: Data Analysis

Enter and Edit Data

The questionnaire is in Appendix A. Take a few moments to review it.

The dataset is called *Zimstudytrn.mdb* and is included on the CD-ROM.

1. Create a folder called "STI" on your computer desktop.
2. Move the *Zimstudytrn.mdb* file from the CD-ROM to the STI folder.

The next step is to get to know your data and to be sure that you have a clean dataset for analysis. Errors may have occurred in collection, coding, or data entry. Data editing is the process of identifying values in your data that are unusual or unexpected and examining them to decide if the data is correct or if there is an error.

We will use Epi Info to examine our dataset for possible problems.
We can start by looking at the most basic description of our study – how many records do we have and does it match the number we expect to have?

Checking for Duplicate Records

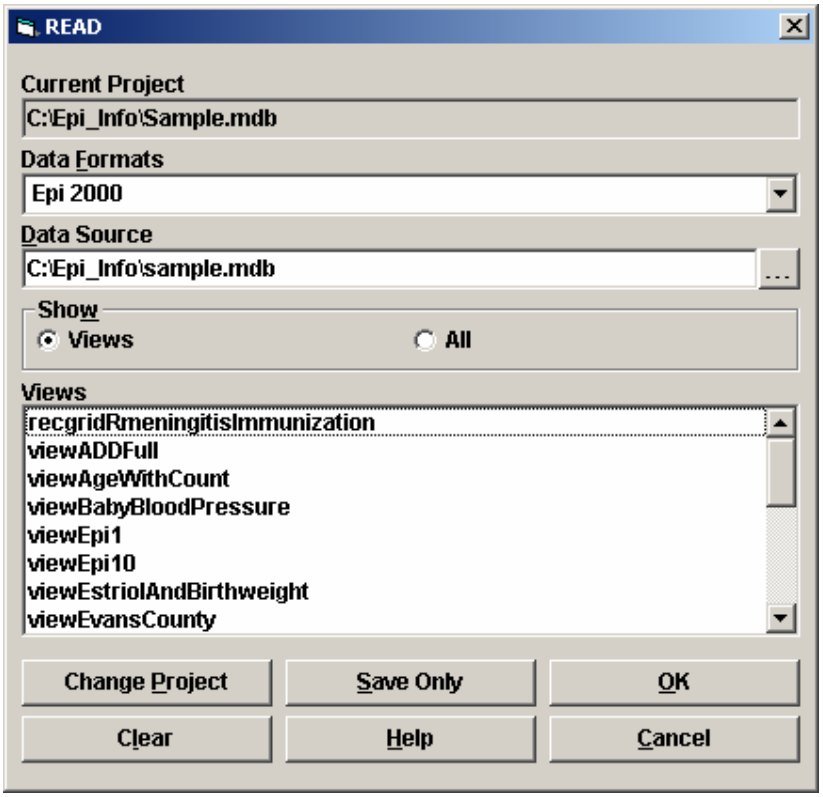
The first step in checking for duplicate records is to be sure that the number of records in your database matches the number of questionnaires. How many records are in our database? How many are there supposed to be?

NOTE: Remember to be sure that your Epi Info software is updated to at least version 3.3.2 before you begin. You can check the version of your Epi Info software by opening up the Main Menu and checking the version at the bottom center of the screen. If you received this material on a CD-ROM, you should have the latest version of the software included. If you do not have a CD-ROM of this training material, you can download the Epi Info software online at http://www.cdc.gov/epiinfo .

Read a Database File

1. **Open Epi Info Analysis.**
2. **READ the database file.**

   a. **Click on the READ command.**



   b. **If project listed is not correct, click on CHANGE PROJECT button. Select the correct file, and click OPEN.**

Go to the STI folder (on the computer desktop) and select the database *Zimstudytrn.mdb*.

NOTE: Remember that the Epi Info Project, or database, (the Microsoft DataBase or .mdb file) is made up of Views (which are created in MakeView) and, once data is entered, also contains tables with data. Additionally, an Epi Info Project can consist of only a table of data if the data were imported from another database file.

c. Select the table to be opened. Click OK.

Click on *basicdata* and then click OK.

3. Check the number of records versus the number of questionnaires created.

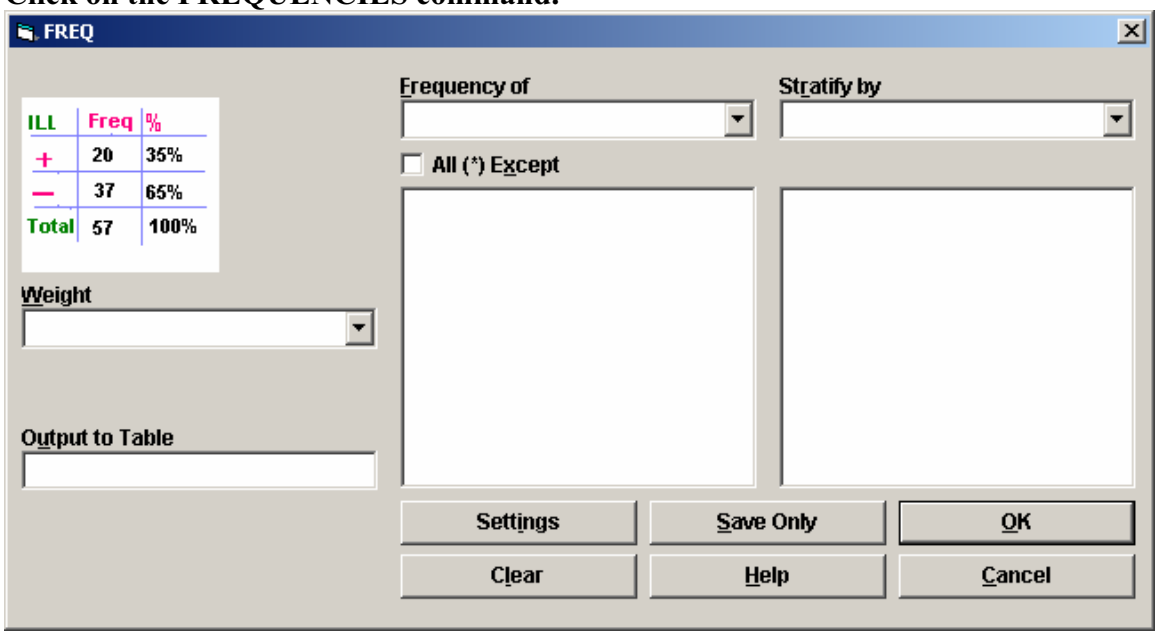
In the Analysis Output window, check the Record Count. The number of records should match the number of questionnaires. There were 226 questionnaires.

? Question 1: Does the number of records match the number of questionnaires?
Yes / No
If not, how many records are there? _____
(The answers to these questions are in Appendix B on page 125.)

Each questionnaire was given a unique number (*IdNumber)*. We will look for duplicates by examining the frequency of the variable *IdNumber*. Since each questionnaire has a unique number, no two records should have the same *IdNumber*.

Analyze the FREQUENCY of a Variable

1. Click on the FREQUENCIES command.



2. **From the Frequency Of drop-down box, select the variable you wish to analyze.**

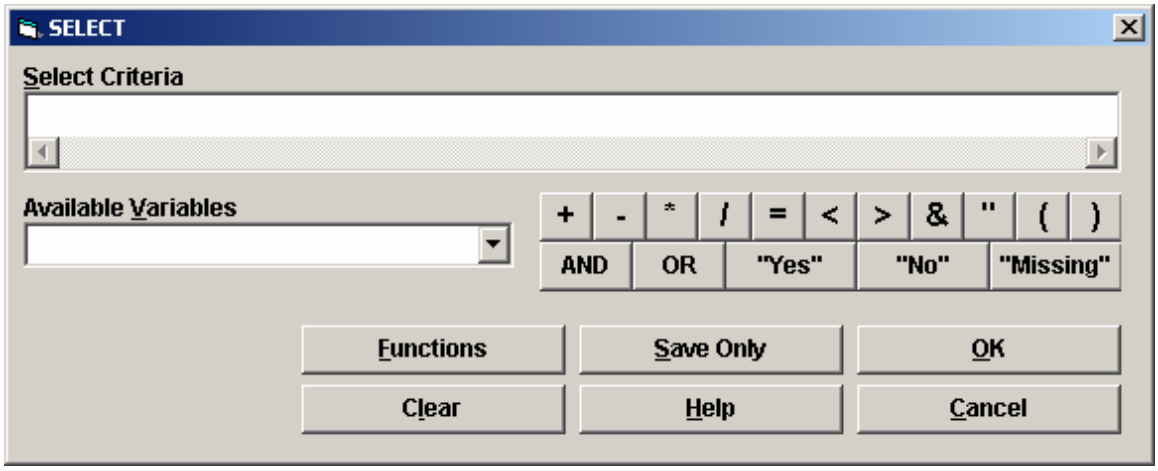
   Select *IdNumber*.

3. **Click OK.**

Scroll through the output on your screen. When we do this we see that there are two records with an identical *IdNumber* (51). We now need to select these records and examine them to see if they are identical (a duplicate record) or whether an *IdNumber* was entered incorrectly.

Create a Subset of Data (Use the SELECT Command)

1. **Click on SELECT command.**



2. **From the Available Variables, choose the variable to select.**

   Select the *IdNumber* variable from the variable drop-down list.

3. **Write the criteria for selection.**

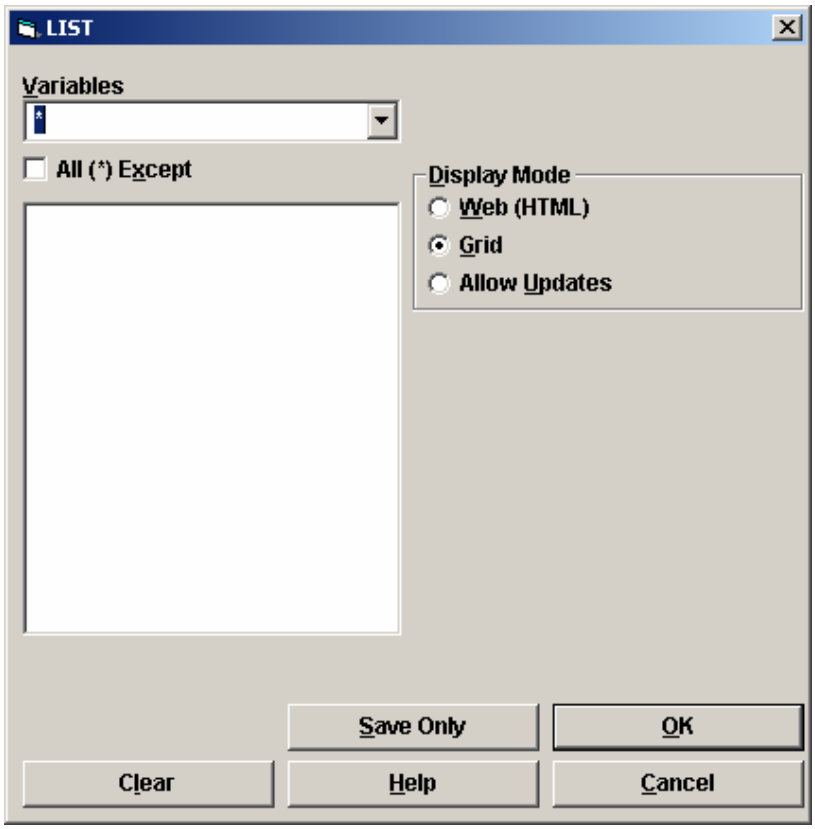
   You want to select the records with a questionnaire number (*IdNumber)* equal to 51. Click the = button, then type in "51".

4. **Click OK.**

   You will see that the record count is listed as two. In order to see these records and all the variables contained in them, you will need to LIST them.

Create a Line Listing

1. **Click on the LIST command.**



2. **Choose variables to display.**

   We want to see all the variables, so use the default (*).

3. **Choose the Display Mode.**

   Use the default, Grid.

4. **Click OK.**

We can see that they are nearly identical except for one variable, *A1Age*. We will need to find questionnaire 51 and examine it to determine which record is accurate (if either is). You will find the questionnaires in the *Questionnaires.pdf* file on your CD-ROM. To find a specific questionnaire, open the PDF file and look for the Bookmarks tab to the left. If bookmarks are not already displayed, click on the tab. You will now see a list of all the questionnaires by number. Click on the questionnaire you wish to review. Each questionnaire has five pages, so you may need to scroll through to find the information you need.

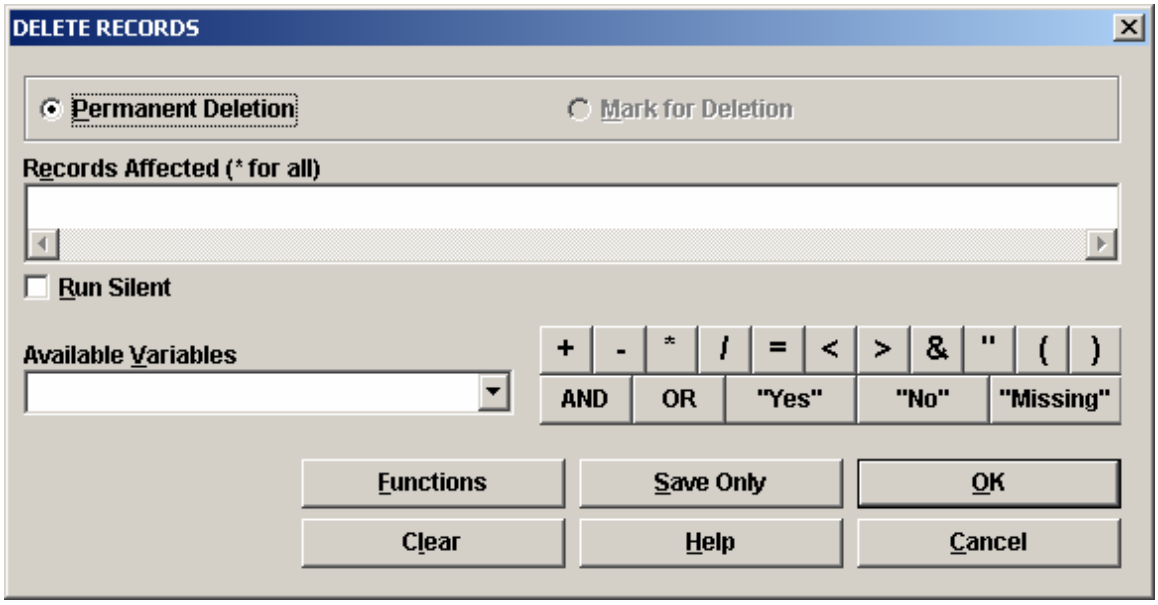
Review Questionnaire 51 in the *Questionnaires.pdf* file and determine which record it matches.

? Question 2: What is the correct age for Questionnaire 51? _____

Delete Duplicate Record

When we have identified the correct record we then need to remove the inaccurate duplicate from our dataset so it will not be a problem in our analysis. To do this we will need to use the DELETE RECORDS command.

1. **Click on the DELETE RECORDS command.**



2. **Under Records Affected, write the statement that identifies the records to be deleted.**

   We need to use two criteria to delete this record. There are two questionnaires with the number 51, so we will also need to use the age criteria to delete the incorrect record.

   a. **From the Available Variables, select the variable to use in the delete criteria.**

      Select *IdNumber*.

b. Write the delete criteria.

Type = and then type 51.

c. If there is more than one variable needed to delete the incorrect variable, select AND and repeat steps a and b.

We need to select the questionnaire numbered 51 that includes the age of 23, which is incorrect. From the Available Variables drop-down, select *A1Age*. Type = and then type "23".

3. Click OK.

You will see a dialog box that asks you if you want to delete the records. It includes the number of records to be deleted in parentheses. In this case, we can tell that we have selected correctly because we only wanted to delete one record. Click OK to accept the deletion. This deletes the record permanently from the database. Keep in mind that once you delete a record, you can not retrieve it, so always keep a backup of the original dataset.

If you LIST the table again, you will now see that there is only one record with the *IdNumber* equal to 51.

Remember to click on CANCEL SELECT.

Documenting Errors and Changes to the Dataset

A key principle of the process of data management is to document everything. Document

- Changes to the data set and
- Decisions about how to assess certain fields.

This documentation will help assure you make consistent decisions and provides a reference for those who have questions later about your analysis.

Another important principle is to fix an error as soon as it occurs (or you become aware of it). It is difficult to remember to return to a problem once you have moved on, so fix it when you find it. This can be particularly important when you are entering data while you are in the field.

Use a table such as the one below to document all changes made to the dataset:

Variable with Errors	Describe Needed Changes	Identify Questionnaires Associated with Change

NOTE: Because these edits permanently change the database, it is very important to maintain a backup copy of the original dataset. This way, if you make any mistakes, you can access the original untouched dataset and start over.

Miscoded, Missing, or Out-of-Range Values

In addition to checking for duplicate records, there are three other major types of edits. These are "range" edits, miscoding edits, and missing value edits. To begin this detection process, the first step is to examine the data in each variable.

We will do this by conducting a FREQUENCY of each variable.

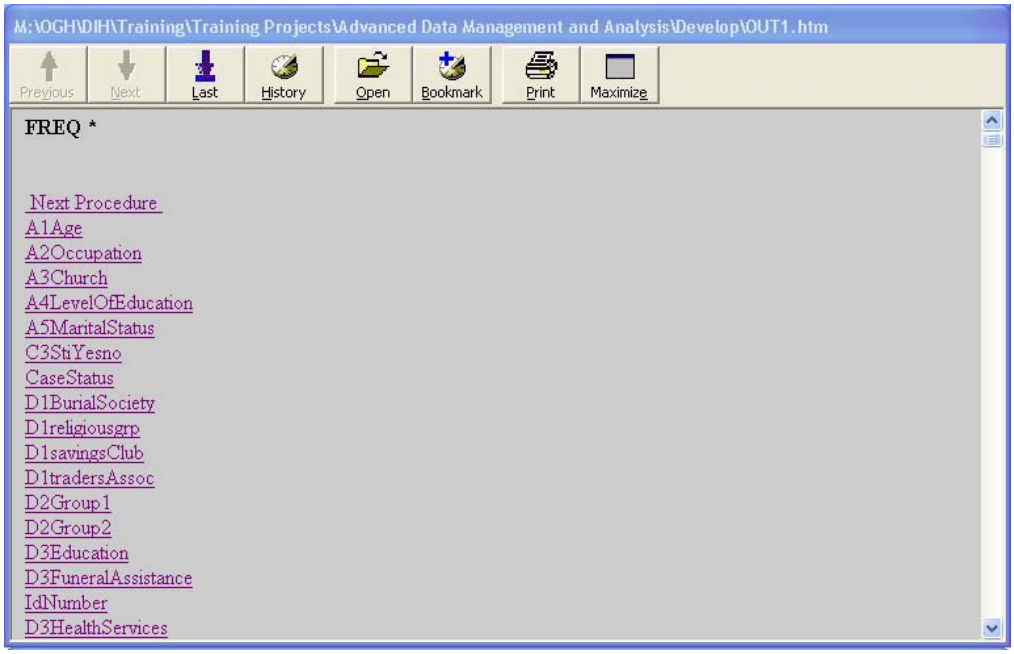
1. **Click on the FREQUENCIES command.**
2. **From the Frequency Of drop-down box, select the variable you wish to analyze.**

   We want to select all variables. Under the Frequency Of drop-down box, select the asterisk (*). The asterisk indicates you are selecting all variables. It may take several moments for Epi Info to run multiple frequencies.

NOTE: Depending on the size or characteristics of your dataset you may get an error message when you try this. In that case, you will need to either select each variable you want to look at, or alternatively, select each variable you do not want to examine and ask to do a frequency of all except those variables (select the variables for which you do not want to perform a frequency and click the All (*) Except checkbox).

3. **Click OK.**

   You should see output that looks similar to this. All the variables for which a FREQUENCY was performed are listed.



Now we can review the individual variables to look for values that are out-of-range or inconsistent with other data in the record or where data is missing. Either scroll through

to find the frequency you wish to view or link to it directly by clicking on the variable name.

Identify Records with Missing Values

One way of reviewing variables is to identify variables that you expect to have a certain number of responses and those for which you specifically do not expect to see missing values. For example, you might expect responses to all the basic demographic questions from each of the persons interviewed in your study; therefore, in this study, there should be 226 responses to each demographic question.

Look at the frequency of the variable *Sex*. You should see the following results:

Sex

Back Forward Current Procedure

Sex	Frequency	Percent	Cum Percent	
Female	107	47.8%	47.8%	
Male	117	52.2%	100.0%	
Total	224	100.0%	100.0%	

95% Conf Limits

Female 41.1% 54.5%

Male 45.5% 58.9%

Note that there are only 224 responses rather than 226. You will want to find the questionnaires that are missing this value and check to see if the data was indeed missing from the questionnaire or was not entered.

1. **SELECT the records of interest.**

   Use the SELECT command to select all those records with *Sex* equal to "Missing".

2. **Verify that the number of records now active is the same as expected from the frequency distribution.**

   There were two records with "Missing" in the S*ex* field.

3. **Run the LIST command to identify the records to review.**

When running the LIST command, select only the *IdNumber* and S*ex* variables. Identify the questionnaires to review from the *IdNumber* variable.

We can now go back to the questionnaires and determine if the correct information is available.

Please answer the questions below. The answers can be checked in Appendix B.

? Question 3: Which questionnaires had the missing information? _____, _____
? Question 4: Was the correct response Male or Female? _____ , _____

Correcting the Dataset

In this case, we have already SELECTed the records we wish to update. One of the quickest ways to update is to use one of the other LIST options.

1. **Choose the LIST command.**
2. **Choose the variables you wish to update (and the unique identifier).**

   In this case, we want to select *IdNumber* since it is our unique identifier for our questionnaires and the variable *Sex*.

3. **Choose Allow Updates as the Display Mode. Click OK.**
4. **To correct data, click inside each cell in the grid you wish to update and change the incorrect data.**

   You will need to type the responses just as all the others have been typed.

   - For male, type: Male
   - For female, type: Female

   Your grid should look similar to this once you have made the updates.

	IdNumber	Sex
	48	Male
	213	Female

5. **Click the X to the upper right to close the Allow Updates window.**

   Once you close the Allow Updates window, you should see the Analysis Output window underneath.

When you update a dataset in Allow Updates mode, you are making permanent changes to the database. When you leave Analysis, the changes you have made will remain. The next time you READ the dataset, it will contain the corrections.

Now CANCEL SELECT. Use the FREQUENCY command again, this time only selecting the *Sex* variable. You should now see 226 records for *Sex*.

If your dataset also contained a View, you could also update the records in Enter Data.

Document the changes you made in the table on page 28.

Identify Records with Miscoded Values

Another problem that can occur is to have values which have been miscoded. This can often be avoided if you use a data entry screen with legal values for text variables or range checks (for numeric variables). By reducing the opportunity to enter incorrect data, you will reduce the need to check for miscoded values.

Create a FREQUENCY of the *CaseStatus* variable and note the responses. Be aware that 1 represents cases and 2 represents controls.

Review the original questionnaire (Appendix A) to see what types of responses are possible for this variable.

? **Question 5:** Which value stands out as incorrect? _____

Use the SELECT command to identify the questionnaire with the incorrect value. Review this questionnaire to determine what the value was meant to be.

? **Question 6:** Which questionnaire has this incorrect value? _____
What value should it be? ____

Change this value following the directions given for *Correcting the Dataset* (page 32).

Again, remember to document this change on page 28.

Identify Records with Out-of-Range Values

Other variables may contain values that seem out-of-range compared to the responses from the other participants in the study. These are often numerical values that may have been incorrectly coded.

A set of questions was asked about the number of sex partners an individual had.

- How many sexual partners have you had in the past 3 months? (*N6NoSexpartn3mnth*)
- How many sexual partners have you had in the past 1 year? (*N7Nosexpart1year*)
- How many sexual partners have you had in your lifetime? (*N8Nosexpartlifetime*)

Do a FREQUENCY of only these three variables and review the values in them. Are there any values which seem out-of-range?

Question 7: Which variable has a value out-of-range? _____ Which questionnaire needs to be reviewed? _____ What is the correct value (see the questionnaire in *Questionnaire.pdf* file)? _____

Make the needed change and document on page 28.

Resolving Data Problems

When you have data that is an outlier (e.g., age of 98) that cannot be resolved by looking at the questionnaire, you must decide whether to try and verify the data or leave it as entered. This will depend on the effort required, the importance of that particular variable and the overall size of the data set. For a very small data set and a key variable, it is probably worth the effort to get it right. In another circumstance, having a "missing" value for such a variable may be acceptable. **It is never okay to change a value just because it does not seem valid.**

Consistency/Logic Checks

Next we need to look for logic or consistency errors. A logic check compares responses in two or more fields and identifies those that are inconsistent.

Example: If the answer to the question, "Have you ever had an HIV test?" is "No," there should not be an answer recorded in the field, "Date of HIV Test".

Ideally, if the data entry form was created to address most of these logic and consistency issues, you would not need to spend a significant amount of time checking the logic of responses after data is entered. The Check Code function in Epi Info is used to verify the data entered in one field against data in other fields to avoid these errors.

This current dataset did not make use of all the check code it could have. You will practice identifying potential logic errors and developing logic statements that can be used to create logic checks for a dataset with errors. This skill is very useful when you are asked to analyze a dataset that you did not prepare yourself.

Logic Error Activity

Look over the questionnaire in Appendix A. Identify and list below three (or more) potential logic errors that you can use to check the accuracy of data.

These are some examples of logic checks we could create:

- A person should not have had more partners over a shorter, recent time period (e.g. past 3 months) than a longer period of time (e.g. past year).

  Past 3 months <= past year or lifetime
  Past year <= lifetime

- The age of sexual debut should be at least the same age as person is now or younger. It cannot be older than person is now.
- If a person does not have a regular partner then they cannot be living with a regular partner.

Select a Subgroup and Check Status of a Variable

Let's look at the issue regarding a record indicating a person lived with a regular partner if the record indicated the person did not have a regular partner. We want to find out if there are any persons who reported no regular partner but who answered 1 (Yes) to the question about living with a regular partner.

1. **SELECT only those records that would contain the potential error.**

   In this case, those records with the type of error we are looking for would be persons who report no regular partner (*N14Doyouhave*). Use the SELECT command to select responses to variable *N14Doyouhave* that are equal to 2 (No).

NOTE: In the "Select Criteria" section, do not select the "No" button. This variable uses the number 2 to designate a No response. The "Select Criteria" section should read: *N14Doyouhave*=2

---

2. **For this selected group, run the FREQUENCY command on the variable of interest.**

   We want to look at the variable regarding living with the regular partner for these persons. Run a FREQUENCY for the variable *N15LivingTogether*.

3. **Identify any records that may contain incorrect data and make the required changes.**

   If persons answered "No" to having a regular partner, then we would expect the response to this question about living with a regular partner would be "Missing" and therefore would not appear in the results. In this case, we received three results. That means that there are three records where the person is listed as not having a regular partner but where the question regarding whether they had been living together in the past 6 months with their regular partner has been completed. Of these three persons, one response indicates he/she is living together with a regular partner (1 = living together) and two responses indicate they are not living with the partner (2 = not living together).

The question regarding living together may have been answered when it should have been skipped or the incorrect answer could have been entered for the question regarding having a partner. Therefore, you may be correcting either of the two variables, *N14Doyouhave* or *N15LivingTogether*.

NOTE: An alternate way of looking at two variables that together may indicate an inappropriate response is to use the TABLE command first rather than using SELECT and FREQUENCIES. In the TABLE command you will use one variable as the exposure and the other variable as the outcome. You will need to determine which cell values may indicate incorrect responses. Only if you find incorrect responses will you need to use SELECT to identify the specific questionnaires. If you decide to try this now, you will need to CANCEL SELECT and then use the TABLE command.

Identify the three questionnaires that appear to have incorrect responses and review the questionnaires in the *Questionnaires.pdf* file.

? Question 8: Which records answered "2-No" to the question about having a regular partner, but also answered the question about living together in the past 6 months (either 1-Yes or 2-No)? _____

Now, make the appropriate changes to the dataset.

CANCEL SELECT to return to the full dataset.

FYI: What Does an Unchecked Checkbox Mean

Although we are not using the Checkbox field type in this dataset, this type of field can cause certain problems in regard to the meaning of a non-response (no check). The Checkbox field is essentially the same as a Yes/No question. In a Yes/No question missing data indicates no response to the question. However, in a Checkbox field, any time the checkbox is not selected the response indicates a No. Does it really mean No or is the value Missing?

Checkboxes are very useful so you should incorporate them in your questionnaires when meaningful. If you must select multiple responses for a question, such as in an outbreak investigation when studying items of food eaten, checkboxes are the most efficient method of data entry.

What is important is to find a method to determine if these checkboxes may have been missed. In the example of checking multiple food items in an outbreak investigation, you could create a selection that determines if all responses to food items eaten equaled Missing. This may indicate that these responses were not entered since everyone – both cases and controls – would likely have eaten at least one of the food items listed.

Compare Two Numbers

We will examine whether anyone's recorded age is currently younger than the reported age of sexual debut.

To examine this we need to compare the current age (*A1Age*) to the age of first sex (*N2SexDebut*). If the age of first sex (*N2SexDebut*) is greater than the current age (*A1Age*), then the difference will be greater than zero. All responses greater than zero will need to be checked.

1. **Click on the SELECT command.**
2. **Create the selection criteria.**

   a. **Click on the open parentheses button (.**
   b. **From the Available Variables drop-down box, select the first variable.**

      Select *N2SexDebut*.

   c. **Click on the minus (-) button.**
   d. **From the Available Variables drop-down box, select the variable to be subtracted from the first variable.**

      Select *A1Age*.

   e. **Click on the close parentheses button) .**
   f. **Select result will be equal to, greater than, or less than a certain number.**

      If the age of sexual debut is listed as older then the age of the individual than we would expect the difference to be greater than zero. Click on the *greater than* (>) button.

   g. **Type the number to be compared.**

      Type 0.

      You should see the following in the "Select Criteria" data entry area: *(N2SexDebut –A1Age)>0.*

3. **Click OK.**

? **Question 9:** Were there any records for which the age of sexual debut was greater than the age of the participant? ____
If so, for which questionnaires? _____ Make any changes necessary.

CANCEL SELECT when you are done.

More Examples

Let's look at the questions regarding number of sexual partners. We want to find out if there are any persons who reported more sex partners in the past 3 months than in the past year.

In this case, records with the type of error we are looking for would be those with the number of partners in the past 3 months (*N6sexpartn3mnth*) greater than the number of partners in past year (*N7sexpart1year*).

Use the SELECT command to choose those records where *N6sexpartn3mnth* > *N7sexpart1year*.

? Question 10: Did any persons report more sex partners in past 3 months than past year? _____ If so, for which questionnaires? _____

Change any of the responses, if needed. Remember to CANCEL SELECT when you are finished.

FYI: Calculate the Difference Between Two Dates

We have just indicated how to compare two numbers – you can use regular mathematical functions, such as subtraction, or greater than and less than. However, for comparing dates, Epi Info uses a specific function to calculate the number of days, months, or years between dates.

For example, if you wanted to calculate how long someone had been ill before reporting to a clinic, you could find the difference between the two dates using the DAYS function. This function will subtract one date from the other and return the difference in number of days.

STEP 1: Create a new variable.
STEP 2: Assign a value to the new variable using the DAYS function. This function follows the following format:

DAYS (variable name 1, variable name 2)

The first date is the earlier date (e.g., date of onset) and the second date is the later date (e.g., date of report). The first two steps might look like this in the Program Editor:

```
DEFINE NumberOfDays
ASSIGN NumberOfDays=DAYS(DateOfOnset, DateOfReport)
```

Note that if the date recorded in the first variable were later than the date in the second variable, the value returned would be negative.

If you wanted to calculate years or months of someone's age on a certain date, the structure of the method stays the same but you use the MONTHS or YEARS function. The date of birth would be the first date and the date on which you wish to calculate age (e.g., such as date of onset) would be the second variable. The value returned will be either the years or months of age. See an example of the function format below.

MONTHS (variable name 1, variable name 2)

YEARS (variable name 1, variable name 2)

Our data set does not have any dates so we cannot use an example here, but this is an important function for data sets that include date variables.

Descriptive Statistics: Univariate Analysis

Now we are ready to start doing our initial analysis. As stated earlier, we start from the simple and move to the complex. The first table we are going to create is our descriptive or demographic data by case and control (see page 17). Some key descriptive variables from our study include age, sex, marital status, education level, and occupation. We have looked at these variables already in the process of editing our data set. We now want to examine the results and organize them so they can be put into a table for a clear comparison of cases and controls.

NOTE: At this point you can either choose to continue to use the *Zimstudytrn.mdb* file that you have already cleaned or, if you are had any difficulty with the previous section or did not finish it, you can use the clean dataset we provided in the Back Up folder called *CleanData_Zimstudytrn.mdb*. If you use this new dataset, copy the file from the Back Up folder to the main STI folder. Then use the READ command and select "Change Project" to choose the new file. Also, if you are copying this material from a CD-ROM and have trouble opening the database you may need to look at the "Properties" of the file (right-click on the file icon and select Properties) and make sure the Read-only checkbox is unchecked so that you can open the database file in Epi Info.

Categorical Variables

For our categorical variables – those that have clear categories (*Sex* – male, female; *A4LevelofEducation* – none, primary, secondary, tertiary; etc) – we will analyze the data first by looking at the FREQUENCY of responses for each variable stratified by case status.

1. **Click on the FREQUENCY command.**
2. **Select the variable of interest from the Frequency of drop-down box.**

   Choose the following four variables: *A2Occupation*, *A3Church*, *A4LevelofEducation*, *A5Marital Status*.

3. **Select the variable to stratify by from the Stratify by drop-down box.**

   In this case you will select *CaseStatus* as your variable to stratify by so that you will receive results for cases and controls separately.

4. **Click OK.**

   You will see two tables for each variable – one for cases and one for controls. The example for *A2Occupation* appears on the following page:

A2Occupation, CaseStatus=1

Forward

A2Occupation	Frequency	Percent	Cum Percent	
1 unemployed	45	39.8%	39.8%	
2 informal	16	14.2%	54.0%	
3 formal	45	39.8%	93.8%	
4 student	7	6.2%	100.0%	
Total	113	100.0%	100.0%	

A2Occupation, CaseStatus=2

Back Forward Current Procedure

A2Occupation	Frequency	Percent	Cum Percent	
1 unemployed	31	27.4%	27.4%	
2 informal	29	25.7%	53.1%	
3 formal	46	40.7%	93.8%	
4 student	7	6.2%	100.0%	
Total	113	100.0%	100.0%	

Frequency of Categorical Variables Activity

Directions: Fill in the results for the categorical variables, using the provided table shells on the next page.

> **FYI: Use the TABLES Command to Find Frequencies**
>
> Alternatively, you can use the TABLES command, choosing the variable of interest as the "Exposure" and *CaseStatus* as the "Outcome". This will give you a single table of the variable with values for cases and controls. When completing the descriptive tables, be sure that you report the column percents (Col %) from this output. If you were to report the row percents (Row %), you would be indicating the proportion of Men who were cases, when what you want to report is the percent of cases who were Men.
>
> Note that with this method you must select each variable individually.

Table 1: Demographic data

Variable	Case - # (%)	Control - # (%)
A2Occupation		
1 Unemployed		
2 Informal		
3 Formal		
4 Student		
A3Church		
2 Apostolic		
3 Methodist		
4 Anglican		
5 Pentecostal		
6 Atheist		
7 Roman Catholic		
8 Other		
A4LevelofEducation		
1 None		
2 Primary		
3 Secondary		
4 Tertiary		
A5MaritalStatus		
1 Single		
2 Married		
3 Cohabiting		
4 Divorcee		
5 Widowed		

You can check your results in Appendix C on page 129, Table 1.

Continuous Variables

We also have continuous numeric variables, such as age (*A1Age*) that will produce such a large number of response categories that it would be difficult to make sense of by FREQUENCY alone (as well as being tedious to report). In addition to age (*A1Age*), we have at least four other continuous variables.

1. Age at first sex (*N2SexDebut*),
2. Number of sexual partners in past 3 months (*N6Nosexpartn3mnth*),
3. Number in the past year (*N7Nosexpart1year*), and
4. Age of sexual partner (*N16HowOldIs*).

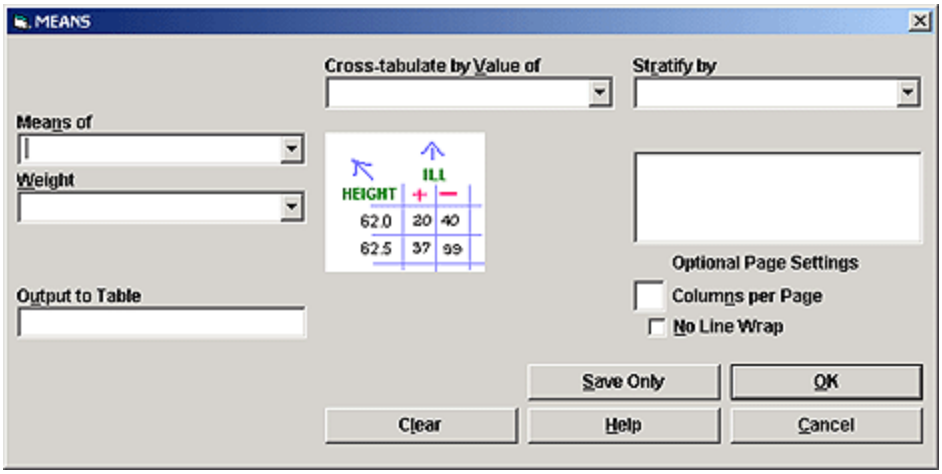
There are two primary ways we can look at this type of variable.

One is to look at the data in a summary form by calculating the measures of central location (mean, median, mode) and dispersion (range, quartiles, standard deviation, variance).

Let's do this first for *A1Age*.

1) Click on the MEANS command.

The MEANS command is under the Statistics folder.



2) In the Means Of drop-down box, select the main variable of interest.

Select *A1Age*.

3) In the Stratify By drop-down box, select the variable by which you want to stratify your results.

In this instance, we want to stratify our results by cases and controls. Select *CaseStatus*.

4) Click OK.

You should see the list of the frequencies by *A1Age*, *CaseStatus*=1 (i.e., age of cases). Scroll down until you see the Total, Mean, Variance, etc. data.

Obs	Total	Mean	Variance	Std Dev
113	3146.0000	27.8407	51.3851	7.1683

Minimum	25%	Median	75%	Maximum	Mode
17.0000	23.0000	26.0000	32.0000	63.0000	23.0000

The mean age of the cases is 27.84 years, the median is 26.0 years, and the mode is 23.0 years. While Epi Info provides answers to four decimal places, there is rarely the need for this level of detail. In general two decimal places is the most you will need. You may have to follow your best judgment when it comes to rounding. Ask yourself how many decimal places add value to understanding the number. This guide will generally round to two decimal places because as we refer to comparisons, that added detail is useful to have.

Mean of Continuous Variables Activity

Directions: Complete the table below for cases and controls for *A1Age*, as well as *N2SexDebut* and Number of partners in the past year (*N7Nosexpart1year*).

Table 2: Continuous Variable

Variable	Case	Control
A1Age	**Mean =** 27.8	**Mean =**
	Median = 26.0	**Median =**
	Mode = 23.0	**Mode =**
N2SexDebut	**Mean =**	**Mean =**
	Median =	**Median =**
	Mode =	**Mode =**
N7Nosexpart1year	**Mean =**	**Mean =**
	Median =	**Median =**
	Mode =	**Mode =**

You can check your results in Appendix C, Table 2.

Grouping Data

While the mean, median, and mode of continuous numeric variables provide useful information, there are times we prefer to group the continuous variable data into logical intervals (categories) and then compare the frequency distributions of these new categories. To do this we need to create the appropriate intervals.

We should keep these guidelines in mind when creating intervals:

1. Create intervals that are mutually exclusive and include all of the data.
2. Use a relatively large number of narrow intervals initially. You can combine intervals later after you have looked carefully at the distributions.
3. Use natural or biologically meaningful intervals when possible. For example, look at standard or frequently used age groupings when considering age.
4. Create a category for unknowns if relevant.

If there is no clear natural or standard interval, there are several strategies available for creating intervals.

- Divide the data into groups of equal size.
- Base the intervals on mean and standard deviation.
- Divide the range into equal class intervals.

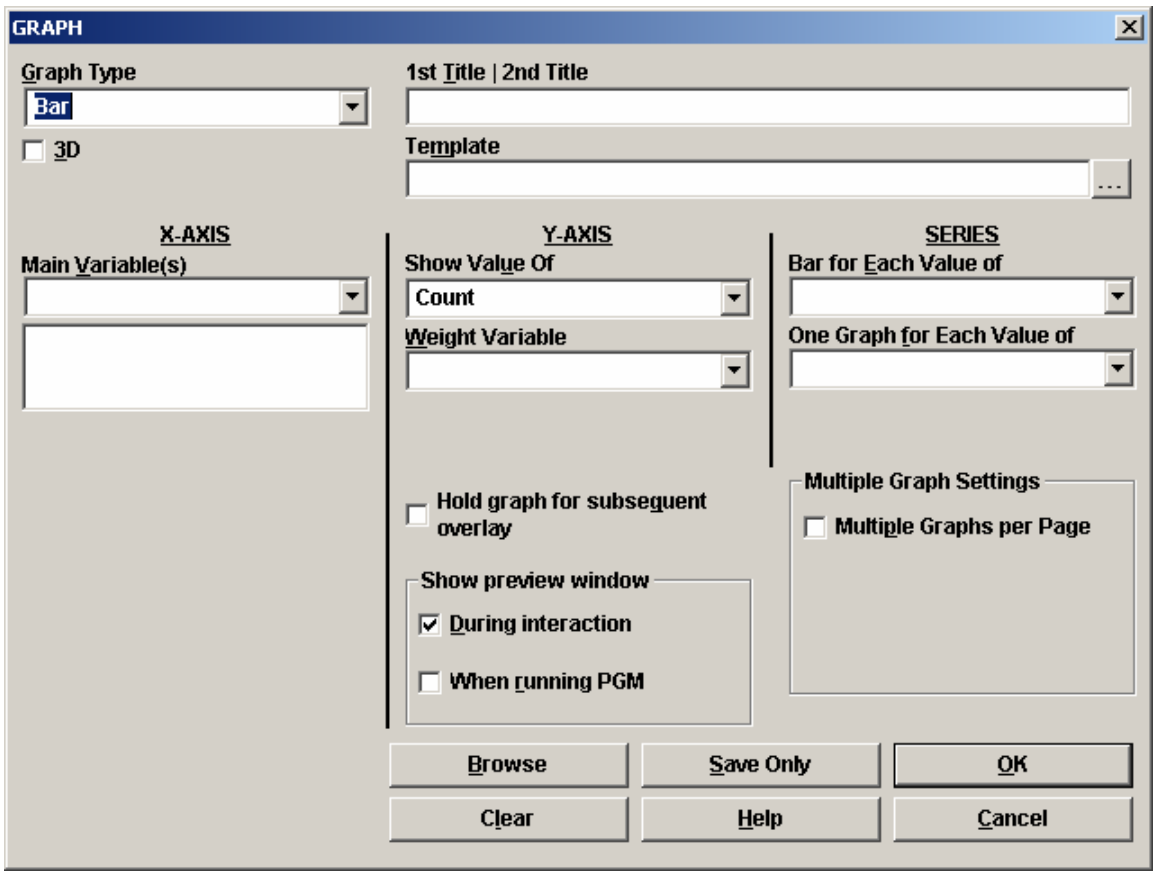
Details and examples for accomplishing these strategies can be found in Principles of Epidemiology (pp. 218-224).

Reviewing the Distribution of Continuous Data

A good first step in determining the intervals to use as categories is to take a look at the distribution of your data. We are going to group our age data. Let's take a look at how the data is distributed using a bar graph.

1. **Click on the GRAPH command.**

   The GRAPH command is under the Statistics folder.



2. **Under Graph Type, select type of graph you would like to create.**

   Choose *Histogram* from the drop-down box. The histogram will allow us to see a distribution of ages and will include values of zero.

3. **Under 1st Title | 2nd Title, write a page title for the graph.**

   Type "Age of Cases and Controls".

4. **Select the variable you wish to graph from the X-Axis (Main Variables) drop-down box.**

   Select *A1Age*.

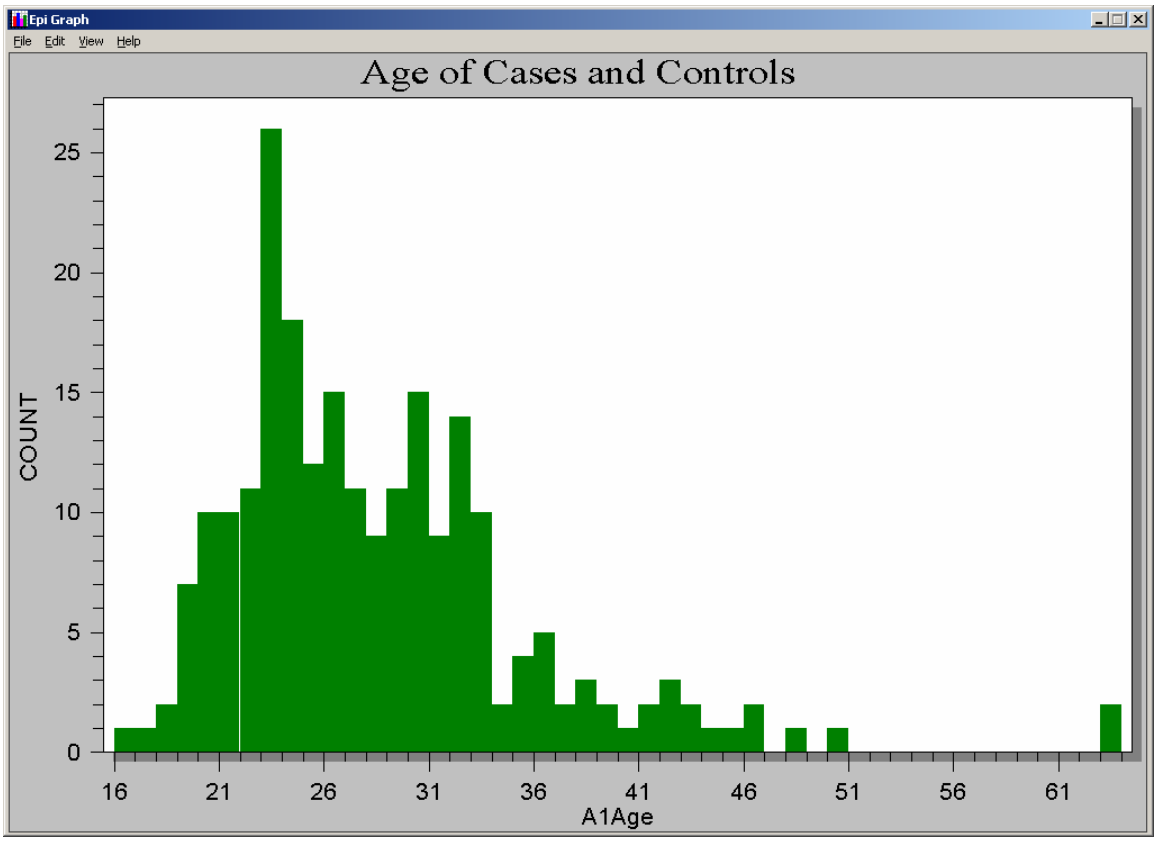
5. **Select the value you want to show from the Y-Axis (Show Value of) drop-down box.**

   Use the default, *Count*.

6. **Choose the Interval and/or First Value.**

For a histogram, you can choose an interval. No changes are required, as *Auto* is the default.

7. **Click OK.**



As you can see there is a particularly heavy concentration of cases and controls between the ages of 19 and 33, and there are very few over the age of 40. This suggests it might be good to use five year age intervals up to the age of 40 and then group all those over 40 together.

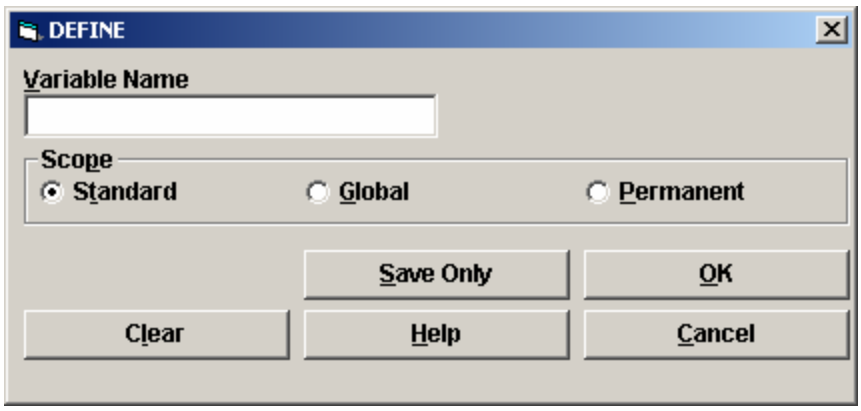
Grouping the *A1Age* Variable

We decide to use the groups < 21 years, 21-25 years, 26-30 years, 31-35 years, 36-40 years, and 41plus years. In order to create this grouping, we will need to create a new variable that will store the new data and then recode the original *A1Age* variable data.

CREATE a new variable

1. **Click on the DEFINE command under the Variable folder.**

   Make sure you do not choose DEFINE COMMAND, which is found under the User-Defined Commands folder.



2. **Under Variable Name, type the name of the new variable.**

   Name the new variable *Agegrp*.

3. **Select the Scope of the variable.**

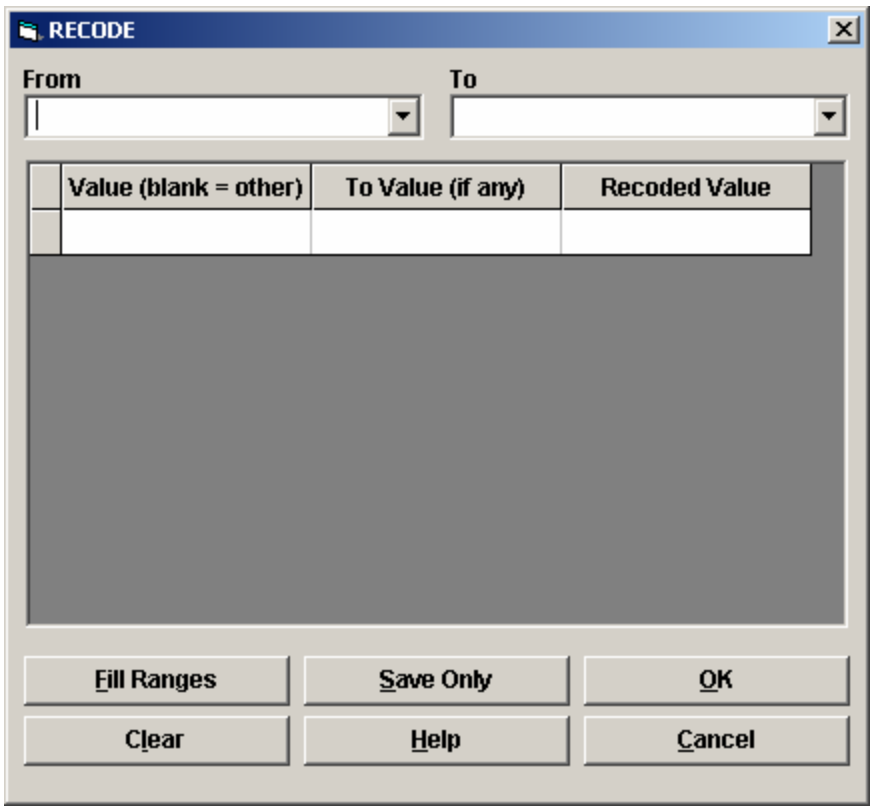
   Use the default *Standard* as we only want the defined variable to be available until our next READ statement.

4. **Click OK.**

RECODE a variable

To now give values to this new variable we will need to RECODE from *A1Age* to A*gegrp*.

1. **Click on the RECODE command.**



2. **Select the variable to be recoded from the From drop-down box.**

   Select the existing age variable (*A1Age*).

3. **Select the variable the new data will be recoded to from the To drop-down box.**

   Select the new variable you created (*Agegrp*).

4. **List the first Value in the range you wish to recode and type it in the Value (Blank = Other) column.**

   The first group you wish to recode is all those less than 21 years old. In recoding, you can designate the lowest value by typing "LOVALUE".

   For each subsequent value, you can type the number representing the first number in the range.

   Press Enter to proceed to the next column.

5. **Select the last value in the range you wish to recode and type it in the To Value (if any) column.**

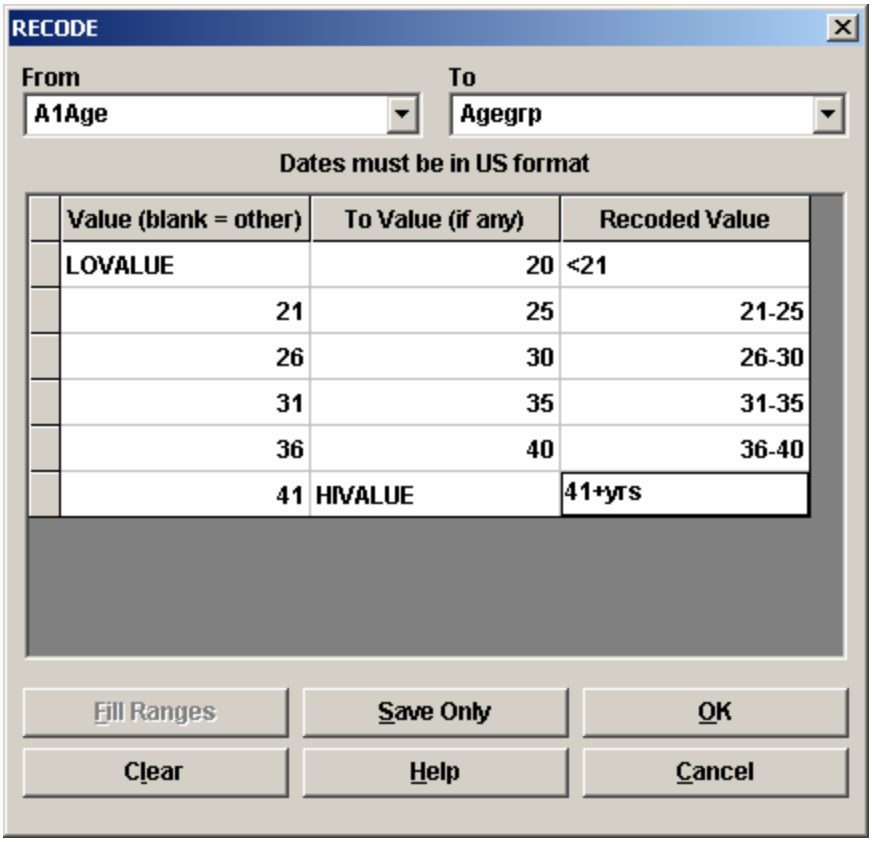
The last value in the range is 20 (since this is the next value less than 21). Type "20".

6. **List the Recoded Value in the last column.**

   Type "<21".

7. **Repeat steps 4, 5, and 6 until all ranges have been filled.**

   To create a new row, click Enter and then fill in the new row with information for the next range of numbers and the recoded value. In recoding, you can designate the highest value by typing "HIVALUE". Your finished recode values for *Agegrp* should look similar to the table below.



8. **Click OK.**

NOTE: Recoded values can consist of numbers, letters, and a certain few symbols such as "<" and ">". However, we could not recode the 40 and higher age range as "41+" because the "+" sign cannot be the last symbol. However, you can add letters after the "+" sign ("yrs" in this case) and the recoding will work. If you receive an error message after you click OK, check all the values in the table and determine if any of them might be illegal.

You can review the code you created in the Program Editor. There is a scroll bar to the right. Read the code and, if you think you know where your error occurred, you can change it in the Program Editor by retyping over the original code. To rerun a single command, place your cursor in front of the command and click the Run This Command button.

After you have completed the recode, do a FREQUENCY of your new variable (*Agegrp*) to be sure that all ages have been recoded. You should always complete a FREQUENCY when you recode to be sure that the recoding is done correctly.

Frequency of Regrouped Variables Activity

Directions: Do a FREQUENCY of *Agegrp* and stratify by *CaseStatus.* Complete the table below for *Agegrp*. Also, recode *N2SexDebut* using the categories noted below and complete the table.

Table 3: Continuous Variables (Regrouped)

Variable	Case # (%)	Control # (%)
Agegrp (previously *A1Age*) < 21		
21-25		
26-30		
31-35		
36-40		
41+yrs		
AgeFirstSex (previously *N2SexDebut*) 1) <17 years		
2) 17+ years		

You can check your results in Appendix C, Table 3.

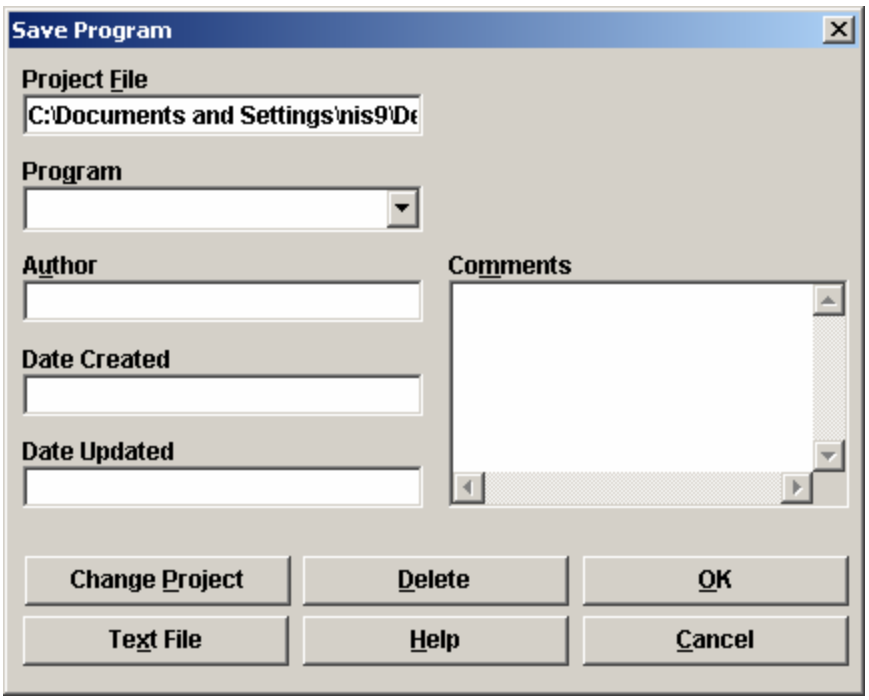
Save Program Files

If you were to leave your database at this point and exit Epi Info, your new variables would not be saved.

NOTE: In order for the new variables to become a permanent part of the data set, you will need to WRITE the data set. We will not do that now as we are going to create other new variables. Be aware when you are doing your own analyses you need to WRITE the database to save these new variables before you exit if you do not want to repeat your work.

To save the coding that you have already completed:

1) In the Program Editor window, click on SAVE.



2) Under Program, name the program.

Choose an appropriate name for the program so that you can recognize it later (e.g., type "*Analysis of STI data")*.

3) Create author name and comments.

You can type your name in as author and create any comment to describe the program you are saving.

4) Click OK.

We do not need to access this program file now, but in order to do so in the future, you would go to the Program Editor, select OPEN, and then select your file from a drop-down list under Program and click OK.

Simple Cross-Tabulations

We now need to take our analysis to the next step of comparing our cases to our controls.

Bivariate Analysis

In many epidemiologic studies, exposure and the health event under study (here STI) can be characterized as binary variables (Yes or No, 1 or 2, etc.). The relationship between exposure and disease can then be cross-tabulated in a 2x2 table (bivariate) as both the exposure and disease have just two categories. You want to be sure that your variables are organized such that the Case will come first in the table. To do this be sure that you use either a true Yes/No variable or if using a number or text be sure that Case is the first to be listed (lowest number or alphabetically).

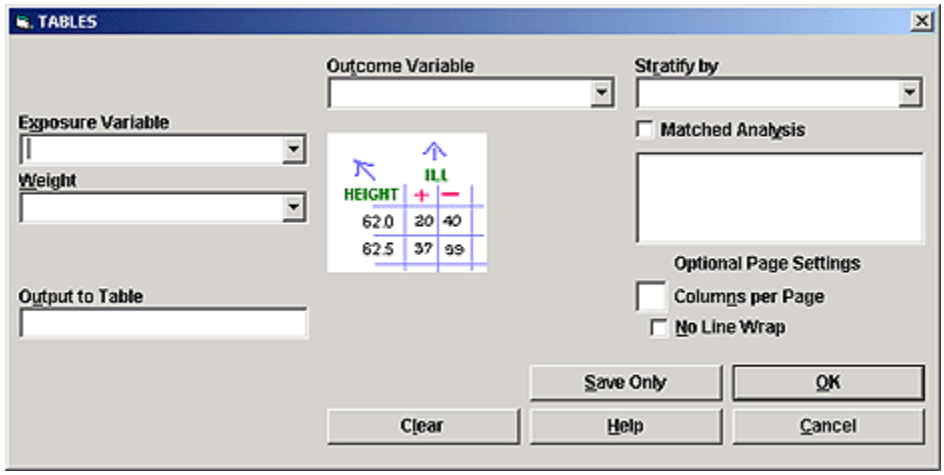
To compare the cases and controls we use the TABLE command. In this case our various possible risk factors (age at first sex, number of sexual partners, paying for sex, etc.) are all possible exposure variables.

Our outcome is whether the person was ill (case) or not (control). This is our *CaseStatus* variable.

Let's analyze the association between having paid or been paid for sex and having an STI.

1) Click on TABLES command.

The Tables command is under the Statistics folder.



2) In the Exposure Variable drop-down box, select the risk factor variable.

Select *N10givereceiveforsex*.

3) In the Outcome Variable drop-down box, select the outcome variable.

Select *CaseStatus*.

4) Click OK.

You will see the following 2x2 table:

CASESTATUS			
N10givereceiveforsex	**1**	**2**	**TOTAL**
1 yes	8	3	11
Row %	72.7	27.3	100.0
Col %	7.1	2.7	4.9
2 no	105	110	215
Row %	48.8	51.2	100.0
Col %	92.9	97.3	95.1
TOTAL	113	113	226
Row %	50.0	50.0	100.0
Col %	100.0	100.0	100.0

And the following statistics:

Single Table Analysis

	Point Estimate	95% Confidence Interval Lower	Upper	
PARAMETERS: Odds-based				
Odds Ratio (cross product)	2.7937	0.7216	10.8152	(T)
Odds Ratio (MLE)	2.7817	0.7398	13.2645	(M)
		0.6456	16.7134	(F)
PARAMETERS: Risk-based				
Risk Ratio (RR)	1.4892	1.0114	2.1927	(T)
Risk Difference (RD%)	23.8901	-3.2640	51.0442	(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	2.3890		0.1221923012
Chi square - Mantel-Haenszel	2.3784		0.1230217192
Chi square - corrected (Yates)	1.5290		0.2162689041
Mid-p exact		0.0684988319	
Fisher exact		0.1074750377	

In this instance, our exposure and outcome variables each only had two categories, so the creation of the 2x2 table was very straightforward. However, for a number of our risk

factor (exposure) variables, there may be more than two categorical responses or exposure variable may be a continuous variable. In that case, we will need to create new variables with only two possible responses so that we can analyze the data using 2x2 tables.

Once we create these new variables, we will discuss the interpretation of the statistics (p. 69).

Grouping Responses for Comparison

Using Recode to Modify Categorical and Continuous Variables

In order to create 2x2 tables for bivariate analysis, we will need to determine how to group both categorical variables with more than two responses and variables that are continuous.

As an example, if we wanted to look at marital status as a risk factor for STI, it would be difficult to compare all of the five possible responses we currently have (Married, Single, Co-habiting, Widowed, Divorced) to each other. What we want to compare are those who are married (Married) to the other categories of those who are not married (Single, Co-habiting, Widowed, Divorced). The group Co-habiting presents a bit of a dilemma. Do we think they are more like Married persons in terms of their behaviors and thus risks or do we consider them unmarried, since technically they are unmarried? Based on our knowledge of behavior, we would consider co-habiting more similar to married than the other options so we plan to group Co-habiting and Married together in one category and the others together in a second category.

1. **DEFINE a new variable.**

   Use the DEFINE command to create a new marriage variable called *HabitationStatus*.

2. **RECODE the old variable to the new variable to make it a two-part response.**

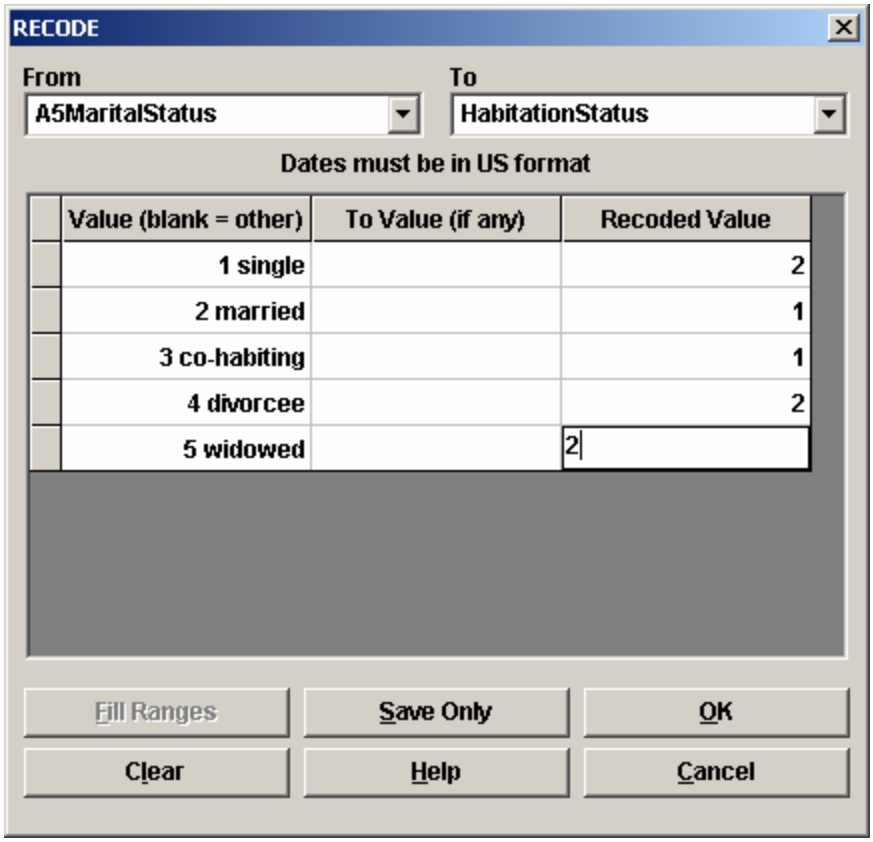
   Recode *A5Maritalstatus* to *HabitationStatus* by recoding

   - Marital status, "2 married" and "3 co-habiting", to 1 and
   - All other values to 2.

A5MaritalStatus
1 single
2 married
3 co-habiting
4 divorcee
5 widowed

   Follow the steps starting on page 49 to recode these responses. Responses to *A5Maritalstatus* appear to the right. See the example below to assist you with completing the recode.

NOTE: When recoding a categorical variable, which does not include a "range" of values, you do not need to include a To Value in the recode table.



Now do a FREQUENCY of your new variable (*HabitationStatus*) stratified by *CaseStatus*.

? Question 11: How many total non-cohabiting persons (response of 2) do you have for cases and for controls (_____Cases _____ Controls)?

Try It!

Do the same type of recoding for the following variables.

- Recode *A2Occupation* to *Unemployed*.

Value	Recoded Value
1 unemployed	1
2 informal	2
3 formal	2

For the recoded value, 1 will stand for unemployed and 2 for employed. Remember to document when you make changes and/or additions to your dataset.

- Recode *A4LevelOfEducation* to *Education*.

  For the recoded value, 1 will stand for none/primary level of education and 2 for secondary/tertiary level of education.

Value	Recoded Value
1 none	1
2 primary	1
3 secondary	2
4 tertiary	2

Using Recode to Eliminate Responses from Analysis

Sometimes you may choose to eliminate certain responses in your analysis in order to create a two-part response. For example, a question that was originally coded to include Yes, No, and Don't Know responses is a three-part response. If you have very few Don't Know responses, you may choose to eliminate them. Because Epi Info ignores Missing responses in Analysis, changing Don't Know to Missing will allow you to analyze the data in a 2x2 table because the only responses will be Yes or No. You should be careful when eliminating responses as you are losing information. If there are a large number of a certain response (such as Missing or Don't Know) it may not be appropriate to eliminate that information. Consider this choice carefully. In our case, the very small number of lost (eliminated) responses is an appropriate choice to improve our interpretation of this variable.

1. **Do a FREQUENCY of the variable in question to determine if you can eliminate a response.**

   Do a FREQUENCY of the variable for use of alcohol at last sexual encounter (*N13TakenAlcohol*).

   You only have two Don't Know responses, so we could choose to eliminate these responses to create a two-part response.

2. **DEFINE a new variable.**

   Use the DEFINE command to create a new variable called *AlcoholUse*.

3. **RECODE the old variable to the new variable to make it a two-part response.**

Recode *N13TakenAlcohol* to *AlcoholUse* using the information in the table below. In this case, the recoded value will still equal 1 for yes and 2 for no, but we will have eliminated the Don't Know responses simply by not including them in the recode.

Value	Recoded Value
1 yes	1
2 no	2

For the new variable, those records containing a value of Don't Know will now contain a Missing response, which will not be included in the analysis of data.

Look at the FREQUENCY of *AlcoholUse*. You will now see that there are only a total of 224 responses rather than 226 since the two Don't Know responses are missing.

Using IF-THEN statements to Group Responses

Another way of grouping responses is to use the If command. This will work with variables where either

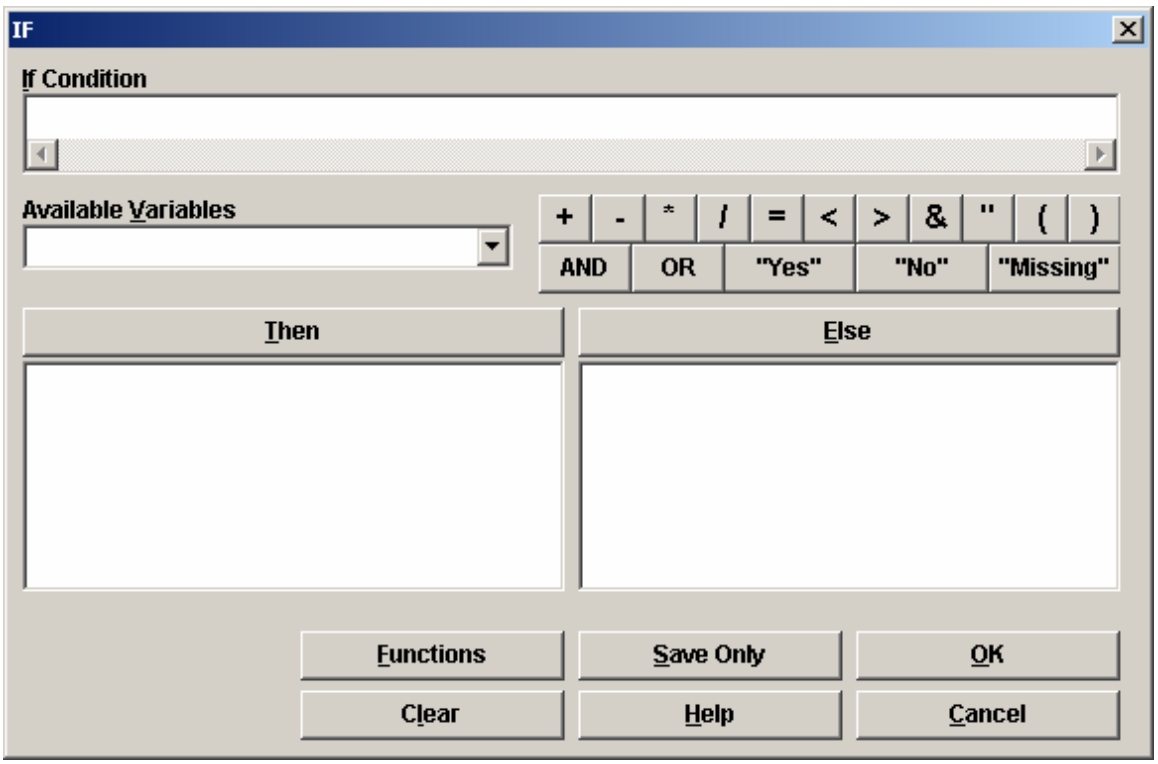
- A numeric variable can be regrouped into more than or less than a certain value (e.g., group the age variable into two categories, <15 and 15+), or
- A variable with multiple responses can be regrouped so that one response takes on one value and all other responses taken on a second value. For example, if a certain occupation were considered a risk factor, the occupation variable may be regrouped so that the at-risk occupation is given one value in a new variable, and all other occupations are given a second value in the new variable.

For the variable for number of sexual partners in the last year (*N7Nosexpart1year)*, we may decide that having less than two partners (in other words, either only one or no partners) in the past year is a different risk than having two or more partners in the past year.

1. **DEFINE a new variable.**

   Use the DEFINE command to create a new variable called *SexPartner1year*.

2. **Choose the IF command.**



3. Enter the If condition.

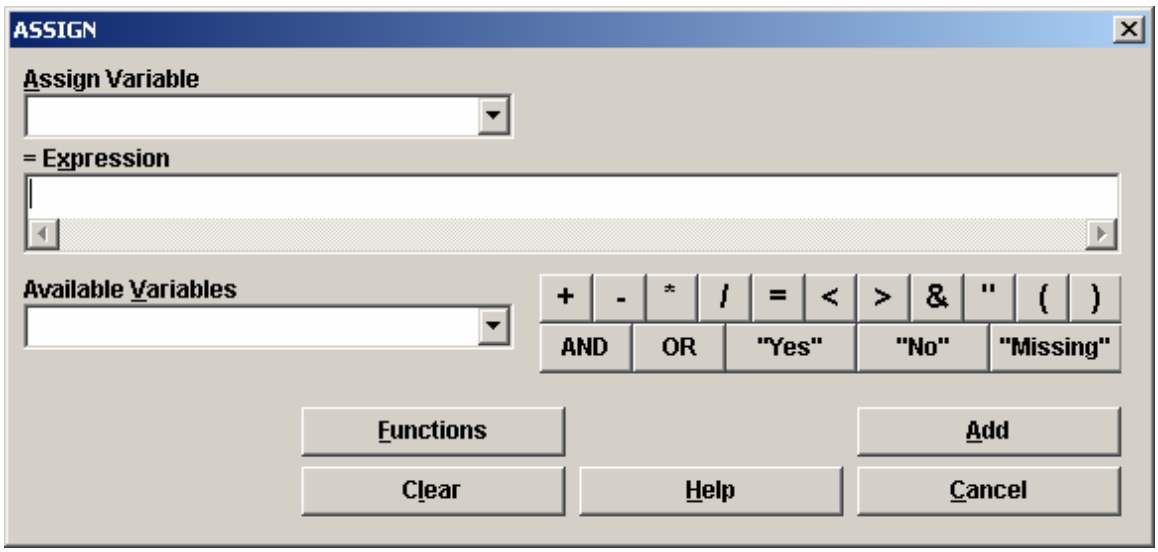
From the Available Variables drop-down box, choose the *N7Nosexpart1year* variable, choose the less than button (<) and then type 2.

NOTE: Once you become comfortable with the correct phrasing that Epi Info uses, you can type out the expressions instead of using the mouse. However, Epi Info will not recognize the expression as valid if the spacing is incorrect. Your typed expression must be exactly as it would appear if you used the mouse.

a. Write the THEN statement.

- **Click on the THEN button.**
- **Create the action that will occur when the If condition is met.**

We are going to ASSIGN a value to the new variable. Click on the ASSIGN command.



From the Assign Variable drop-down box, select *SexPartner1year*. Place the cursor under "= Expression" and click on the "No" button. Click Add.

4. **Write the Else statement.**

   - **Click on the Else button.**
   - **Create the action that will occur when the If condition is *not* met.**

     Click on the ASSIGN command. From the Assign Variable drop-down box, select *SexPartner1year*. Place the cursor under = Expression and click on the "Yes" button. Click Add.

5. **Click OK.**

If you run the FREQUENCIES command of the *SexPartner1year* variable, you will note that the Yes variable (which in this case refers to all those who have had two or more partners in the past year) is listed first in the FREQUENCIES table. This is because our assignment of values to this variable has created a Yes/No type variable and, in epidemiologic analysis, the Yes response is always listed first. We will discuss this more in the section on direction of association (p. 65).

Try It!

Just as we created a risk factor variable for number of sex partners in 1 year that we can use in our bivariate analysis, we can create a risk factor variable for the number of sex partners in the past 3 months. Use the same If-Then process to regroup *N6Sexpartn3mnth* to *SexPartner3month*, using the same criteria of <2 partners as a No response and 2 or more partners as a Yes response.

One of the questions in our study asked what the person's relationship was to the last person he or she had sex with. Multiple responses were possible so we also need to create a two-part response to this question in order to analyze the risk factor. We are going to use an If-Then statement to create a new variable for relationship to last sexual partner (*N9Relationship*) where we compare all those whose last partner was a spouse or live-in partner to all other responses.

1. **DEFINE a new variable.**

   Use the DEFINE command to create a new variable called *LastPartnerSpouse*.

2. **Choose the If command.**
3. **Enter the If condition.**

   From the Available Variables drop-down box, choose the *N9Relationship* variable, choose the equals button (=) and then type (in quotes) "1 spouse".

   Note here that while the response that was created in the database refers to "spouse," the original response to the *N9Relationship* variable in the questionnaire was "1 Husband/Wife/live in partner". So, those who are married and those who co-habit are considered the same for this question.

b. **Write the THEN statement.**

   - **Click on the THEN button.**
   - **Create the action that will occur when the If condition is met.**

     Click on the ASSIGN command.

     From the Assign Variable drop-down box, select *LastPartnerSpouse*. Place the cursor under = Expression and click on the "Yes" button. Click Add.

4. **Write the Else statement.**

   - **Click on the Else button.**
   - **Create the action that will occur when the If condition is *not* met.**

     Click on the ASSIGN command. From the Assign Variable drop-down box, select *LastPartnerSpouse*. Place the cursor under = Expression and click on the "No" button. Click Add.

5. **Click OK.**

Create a New Variable from the Results of Two or more Other Variables

Another type of variable creation involves taking the results from two or more variables and combining them to create a new variable that can be analyzed more easily.

One risk factor this study examined was a person's social support in the community. A number of questions were used to determine if the person/household had some level of social support. (See questions D1-D3 in the sample questionnaire, Appendix A.) We now want to create a summary variable that will allow us to classify persons as having support or not. In looking at this section we can see that questions D1 – "Does anyone in your household belong to any of these (social support) groups?" – and D3 – "Do the groups help your household to access any of the following services?" – are each made up of four unique Yes/No questions.

How do we create one variable out of the results from four variables? First, let's consider "belonging" to an organization.

1. **DEFINE a new variable.**

   We will call it *Belong*.

   To give values to this variable we will use the IF command to Assign values based on the four contributing variables. Essentially, if any of these four variables indicates that the person (or household) does belong to a group (i.e., that the answer is 1 to any of the questions) then the new variable will be assigned the value of Yes. If none of the responses to the four questions is 1, then the value assigned to the new variable will be No.

2. **Choose the IF command.**
3. **Enter the If Condition.**

   a. **From the "Available Variables" drop-down box, choose the variable and the value.**

      Select the *D1tradersAssoc* variable, choose the equals (=), and then type 1. One (1) is the value that indicates that a person belongs to that organization.

   b. **If there are additional variables, choose AND or OR,**

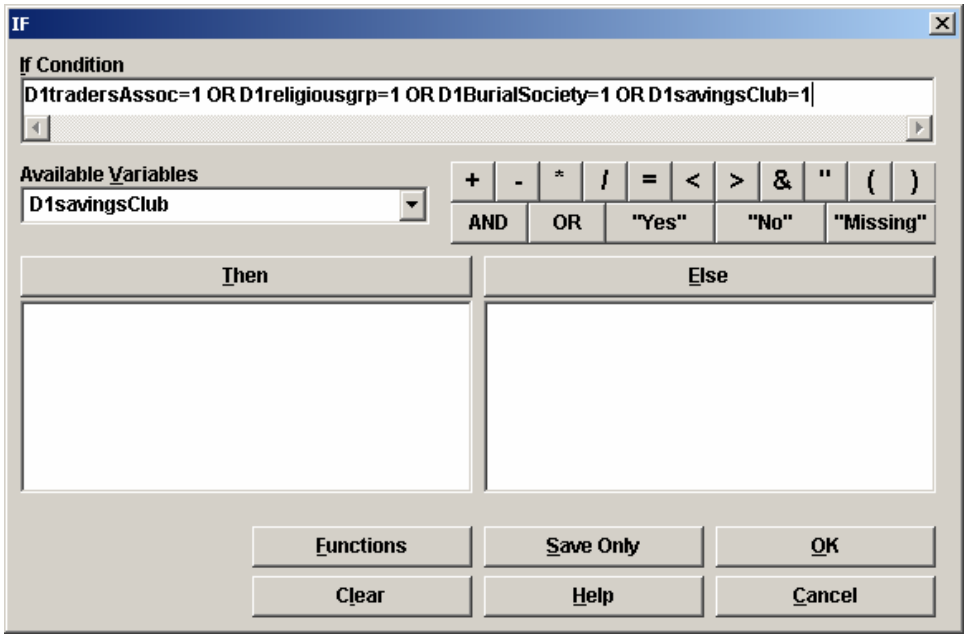
      Choose OR since a yes answer to any of the variables is what we are looking for.

   c. **Repeat steps *a* and *b* for all remaining variables.**

      D1religiousgrp =1

D1BurialSociety=1
D1savingsClub=1

It should look like this:



4. **Write the THEN statement.**

   - **Click on the THEN button.**
   - **Create the action.**

     Click on the ASSIGN command. From the Assign Variable drop-down box, select *Belong*. Place the cursor under =Expression and click on the "Yes" button. Select Add.

5. **Write the ELSE statement.**

   - **Click on the Else button.**
   - **Create the action.**

     Click on the ASSIGN command. From the Assign Variable drop-down box, select *Belong*. Place the cursor under =Expression and click on the "No" button. Select Add.

6. **Click OK to close the IF box.**

Do a FREQUENCY on your new variable, *Belong* (without stratifying).

? Question 12: How many of the respondents belong to at least one organization? _____

In this instance, we used OR between our variables as we wanted everyone who belonged to any organization to be counted as a Yes response in our new variable. If we had mistakenly used AND in our phrase to connect the variables, we would only have gotten those persons who belonged to all of the organizations.

Try It!

We can use the same process to create a new variable for persons/households who have received help from any of the organizations. The new variable name will be *ReceiveHelp*. The four variables you will check for a value of 1 are *D3Education*, *D3FuneralAssistance*, *D3HealthServices*, and *D3receivecredit*.

? Question 13: Do a FREQUENCY of *ReceiveHelp* without stratifying. How many respondents receive help? _____

Recoding for Direction of Association

One thing to consider in coding our variables for analysis are the codes used to represent the exposed and the outcome of interest, e.g. "illness" or "case". Epi Info uses the item that comes first in the "Exposure" side of the table as the exposure for analysis and in the calculation of measures of association. What will come first in the table is based on how the variable is coded. If you have assigned legal values to the variable when you created it, you could assign the order. If not, Epi Info uses certain simple rules:

- Numeric Variables - ascending order (1,2,…n),
- Text Variables - alphabetic order (married, unmarried),
- Yes/No Variables - Yes comes first.

However, if you had simply used a Text type field for the variable for the question, "Was your last illness an STI?" and typed in, "Yes" and "No", then No will come first as it is treated as a text variable and the responses are listed in alphabetic order.

When you are doing analysis, you need to decide how you want to examine the risk factors. The particular coding used will have an impact on the direction of the association, or whether the association is causal or protective. In our study we could look at an increasing number of sex partners as a causal risk factor and calculate that measure

of association. Conversely, we could look at a fewer number of sex partners as a protective factor and look at that association.

As an example, the variable *N8Nosexpartlifetime* has been coded in the questionnaire so that the value 1 is given to having fewer than three lifetime sexual partners while the value 2 is given to having three or more lifetime sex partners. We can either analyze this as

- fewer sexual partners is protective against an STI (i.e., associated with a lower probability or risk of an STI), or
- more sexual partners is causative for STIs (i.e., associated with a higher probability of an STI).

In this case, it probably makes the most sense to have the behavior (rather than lack of behavior) be the risk factor. To analyze the variable so that having more partners is considered the exposed category of the risk factor, the response related to having a greater number of partners needs to come first in the table of exposure. Thus we will need to recode the variable so that having three or more partners is considered the exposure.

Create the new variable using *SexPartnerLife3* as your new variable. We will give you a couple of hints, but try to do this yourself.

- Use the FREQUENCIES command to check the values of the original variable, *N8Nosexpartlifetime.*
- Use Epi Info's rules of assigning order.

? **Question 14:** What programming code did you create? See Appendix B, Question 14 for an example.

Documenting New Variables

As we mentioned earlier, you should document all the changes you make to a dataset; this includes the creation of new variables. When you create new variables, write down in a log exactly how you have defined each of these variables. We have provided a sample log table below. This will be essential for explaining your dataset later. The commands that you have written to create your variables need to be noted as verification to explain your values.

New Variable Name	Description	Values	Recoding

SexPartnerLife3	# of partners in a lifetime grouped as 3 or more or less than 3	• Yes (3 or more)
		• No (less than 3)
			`DEFINE SexPartnerLife3`
			`RECODE N8Nosexpartlifetime TO SexPartnerLife3`
			`"1 Less than 3" = (-)`
			`"2 3 or more" = (+)`
			`END`

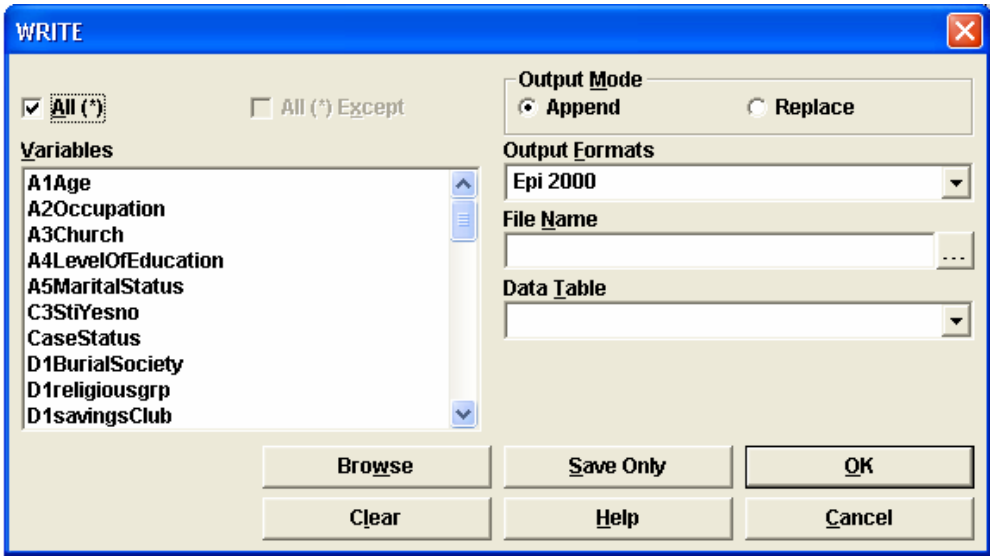
Remember that you can save your programming code (see Saving Program Files on p. 52).

Create a New Data Table

Remember that all the new variables you have created are only temporary for this analysis session. To keep these variables for later analysis, use the WRITE command to save your database including all the new variables (as well as the original ones).

NOTE: Before you create a new data table, always be sure that you CANCEL SELECT so that your new data table includes the entire dataset. The SELECT command will be discussed in more detail later in the analysis.

1) Click on WRITE command.



2) Select variables to be included in the new file.

You can select all or some of the variables in the original file to be included in your new file. We want to include all the possible variables. Just above the Variables list, you can see a checkbox next to All. Make sure that box is checked.

3) Select Output Mode (Append or Replace).

If you are going to add one file to another, select Append. If you are going to replace any original file data with new data or create an entirely new file or table, select Replace.

We are going to create a new table within the same .mdb file, so select *Replace*.

4) Select Output Format.

You can create a file in multiple formats. We have been working with Epi Info 2000 files (in the .mdb format). Choose the default format, *Epi 2000*.

5) Select or create File Name.

You will either write your data to a current file or create a new file. We are going to create a new table within the same file.

Click on the button next to File Name in the dialog box, find the file you have been working in (*CleanData_Zimstudytrn.mdb*). Click *Save*.

6) Name the table.

Under Data Table, type: "STIdataCreatedVariables"

7) Click OK.

Now you need to READ the new data table. After you select the READ command, click on All so that you can see all of the data tables. Select *STIdataCreatedVariables*, and click OK.

If you would like to check that your new variables are there, click on the DISPLAY command and click OK. All the variables are now listed in alphabetical order, so the new and old variables are combined.

Next, we will review how to interpret the results of the analysis of our new variables.

Measures of Association

We want to look at the association between the exposure and the disease. For this we use a measure of association.

"A *measure of association* quantifies the strength or magnitude (size) of the statistical association between exposure and the health problem of interest" (Dicker, p.142). Many epidemiologic studies are designed to determine whether the risk of disease or other health problem (injury, obesity, etc.) among persons exposed to some factor differs from the risk among persons not exposed. Thus the data from these studies can be fit into the 2x2 table of exposure and outcome.

In this study, we are looking at whether the risk of having an STI differs for persons exposed to various factors, such as having multiple sexual partners, not living with your spouse, etc. In the study type represented here (the case-control study), the *odds ratio* is the appropriate measure of association. "In cohort studies, the measure of association most commonly used is the *relative risk*…In cross-sectional studies, either a prevalence ratio or a prevalence odds ratio may be calculated" (Dicker, p.142).

NOTE: At this point you can either choose to continue to use the *CleanData_Zimstudytrn.mdb* file in which you have already created the new variables table (STIdataCreatedVariables) or, if you had any difficulty with the previous section or did not finish it, you can use the revised dataset we provide in the Back Up folder called *CreatedVariables_Zimstudytrn.mdb*. If you use this new dataset, copy the file from the Back Up folder to the main STI folder. Then use the READ command and select Change Project to choose the new file. Remember to select the new table, STIdataCreatedVariables.

Also, if you are copying this material from a CD-ROM, don't forget that you may need to look at the Properties of the file (right-click on the file icon and select Properties) and make sure the Read-only checkbox is unchecked so that you can open the database file in Epi Info.

Let's start with creating the 2x2 table for one possible risk factor – having more than one sexual partner in the last 3 months. As noted above, the command is TABLE, the exposure variable *SexPartner3month* and the outcome variable is *CaseStatus*.

We get an output that looks like the one shown on the next page:

CASESTATUS

SexPartner3month	1	2	TOTAL
Yes	19	3	22
Row %	86.4	13.6	100.0
Col %	16.8	2.7	9.7
No	94	110	204
Row %	46.1	53.9	100.0
Col %	83.2	97.3	90.3
TOTAL	113	113	226
Row %	50.0	50.0	100.0
Col %	100.0	100.0	100.0

The resulting table will allow us to determine not only the proportion of cases and controls that were exposed to a given risk factor – in this case having more than one sexual partner in the last 3 months – but also the measure of association for this risk factor (exposure).

Scroll down to see the information under the title "Single Table Analysis". We see a number of figures.

Single Table Analysis

	Point Estimate	95% Confidence Interval Lower	Upper	
PARAMETERS: Odds-based				
Odds Ratio (cross product)	7.4113	2.1268	25.8269	(T)
Odds Ratio (MLE)	7.3556	2.2937	32.0538	(M)
		2.0715	40.0036	(F)
PARAMETERS: Risk-based				
Risk Ratio (RR)	1.8743	1.5000	2.3419	(T)
Risk Difference (RD%)	40.2852	24.3970	56.1734	(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	12.8913		0.0003312880
Chi square - Mantel-Haenszel	12.8342		0.0003415053
Chi square - corrected (Yates)	11.3302		0.0007637295
Mid-p exact		0.0001318400	
Fisher exact		0.0002352530	

Epi Info has calculated measures of association and the related statistical tests. Epi Info does not know what type of study you have done. It is your responsibility to understand your study and to use the appropriate statistical test.

Look first at the measure of association. This is a case-control study, so the *Odds Ratio (cross-product ratio)* is the appropriate measure of association. The odds ratio of 7.4 (rounded) indicates that the odds of having more than one sexual partner in the last 3 months are 7.4 times higher among cases than among controls. It is also reasonable to say that the odds of having an STI were 7.4 times higher among those exposed to more than one sexual partner than among those not exposed.

Odds Ratio Interpretation

- Odds ratio > 1 (Exposure increases risk for the outcome)
- Odds ratio = 1 (Risk is equivalent between exposed and unexposed)
- Odds ratio < 1 (Exposure decreases risk for outcome – protective effect)

Measures of association (odds ratio, relative risk, prevalence ratio)

- Reflect the strength of the association between an exposure and a disease
- Are generally independent of the size of the study
- Are considered the "best guess" of the true degree of association in the source population
- Provide no indication of how reliable a guess it may be

So measures of association observed in our sample data are not enough to indicate there is an association between exposure and disease in the whole population. "Tests of significance (and confidence intervals) provide an indication of how likely the observed association may be due to chance, if exposure was not actually related to disease" (Dicker, p. 147). Thus, we use tests of significance and confidence intervals to determine how reliable is our "best guess" of the association between exposure and disease.

Tests of Statistical Significance

First, we will review concepts related to and tests of statistical significance.

Null Hypothesis

For statistical testing, one assumes that the study population is a sample from an underlying or source population. You also assume that, in the source population, the incidence (or rate) of disease is the same for those exposed or not exposed to a particular risk factor. In other words, there is no association between exposure and disease. This assumption is known as the "null hypothesis" (Dicker, p. 147).

P-Value

We have calculated an appropriate measure of association. Now, to test statistical significance, we calculate the probability of finding an association as strong as (or stronger than) the one you would have observed by chance if the null hypothesis (no association) were really true. This probability is called a *p-value*. A very small p-value means that you would be unlikely to observe such an association if the null hypothesis were true. In other words, a small p-value indicates that the null hypothesis is implausible (unlikely) given the data. If this p-value is smaller than some predetermined cutoff (usually 0.05 or 5%), you can discard (reject) the null hypothesis and accept the alternative hypothesis that exposure and disease are associated. The association is then said to be "statistically significant" (Dicker, p. 147-8).

The null hypothesis, as stated here (no association), is considered a "non-directional" hypothesis. This means that the alternative hypothesis includes the possibility that exposure may either **increase** or **decrease** the risk. A non-directional hypothesis is tested by a "two-tailed" test of statistical significance. Epidemiologists use a two-tailed test in most situations. The one-tailed test is reserved for those situations where there is substantial prior knowledge about the relationship of the risk factor and disease to indicate direction.

Type I and Type II Errors

In reaching a decision about the null hypothesis, be alert to two types of error. In a Type I error (also called alpha error), the null hypothesis is rejected when in fact it is true. In a Type II error (also called the beta error), the null hypothesis is not rejected when in fact it is false. (Dicker, p. 148)

Hypothesis Testing Outcomes

	Null Hypothesis (H_0)	
	True	False
Reject H_0	Type I Error (Alpha or α)	Correct Decision (Power)
Fail to Reject H_0	Correct Decision ($1 - \alpha$)	Type II Error (Beta or β)

* Modified, Blair and Taylor, p. 174

Testing Data in a 2x2 Table

Two different tests, each with some variations, are used for significance testing for data in a 2x2 table (or a 2xn table). These two tests are the *Fisher exact test* and the *chi-square test*. These tests are not specific to any particular measure of association. The same test can be used regardless of whether you are interested in risk ratio, odds ratio, or attributable risk. These tests are calculated by Epi Info when you use the TABLES command.

We will not review how to do the various statistical tests; see your epidemiology and biostatistics texts for this. However, we will discuss interpreting the statistical results provided by Epi Info.

Fisher Exact Test

The Fisher exact test is considered the gold standard for a 2x2 table and is the test of choice when the numbers in the 2x2 table are small. The Fisher exact test involves computing the probability of observing an association in a sample equal to or greater than the one observed (Dicker p. 148). As a rule of thumb, the Fisher exact test is the test of choice when the expected value in any cell in the 2x2 table is less than five. The expected value is calculated by multiplying the row total by the column total and dividing by the table total (see Appendix D for the formula).

A reasonable approximation to the Fisher exact test when there are large numbers is the chi-square test.

Chi-Square Test

When you have at least 30 subjects and the expected value in each cell of the 2x2 table is at least five, the chi-square test provides a reasonable approximation to the Fisher exact test. The chi-square test provides a test statistic that corresponds to a two-tailed p-value. Epi Info provides both the chi-square and the resultant p-value. In fact, it calculates the chi-square for the 2x2 table using the three different formulas in common use as well as the "exact" p-value.

STATISTICAL TESTS	Chi-square	1-tailed p	2-tailed p
Chi square - uncorrected	12.8913		0.0003312880
Chi square - Mantel-Haenszel	12.8342		0.0003415053
Chi square - corrected (Yates)	11.3302		0.0007637295
Mid-p exact		0.0001318400	
Fisher exact		0.0002352530	

The three formulas are the

- uncorrected (also known as the Pearson uncorrected),

- the Mantel-Haenszel, and
- the corrected (or Yates corrected).

How do you decide which result to use?

- **Uncorrected**

  For a given set of data in a 2x2 table, the uncorrected chi-square formula gives the largest chi-square value and hence the smallest p-value. This p-value is often somewhat smaller than the p-value calculated by the Fisher exact test. This uncorrected (or Pearson) chi-square is thus more likely to lead to a Type I error, concluding that there is an association when there is not (Dicker, p. 149).

- **Mantel-Haenszel**

  The Mantel-Haenszel formula yields a p-value which is slightly larger than that from the uncorrected formula but often smaller than the Yates corrected and the Fisher exact p-value.

- **Corrected**

  The Yates corrected chi-square gives the largest p-value of the three formulas, sometimes even larger than the corresponding Fisher exact p-value. The Yates corrected is preferred by those epidemiologists who want to minimize their likelihood of making a Type I error, but it increases the likelihood of making a Type II error.

See the formula for all three in Appendix D. In summary, each of the three formulas has its advocates among epidemiologists and Epi Info will provide all three. Many field epidemiologists prefer the Yates corrected as noted above. (Dicker, p.149)

Let's go back to our "Single Table Analysis" for *SexPartner3month*. What are the chi-square values and corresponding two-tailed p-values?

	Chi –Square	**2-tailed p-value**
Uncorrected		
Mantel-Haenszel		
Corrected (Yates)		

? Question 15: Using the corrected chi-square, what can we say about the interpretation of the measure of association (the odds ratio) and its significance?

Chi-Square Test for 2xn table

The chi-square statistic can also be calculated for tables other than 2x2 tables. When you do this in Epi Info, only one statistic is calculated. You will see the chi-square, the degrees of freedom ((number of rows – 1) x (number of columns-1)) and the corresponding p-value.

The Yates corrected chi-square can not be used when it is not a 2x2 table. The interpretation of a chi-square in this instance is also different and less clear. Do a TABLES analysis of the original variable for the number of sexual partners in past 3 months (*N6NoSexpartn3mnth*) by case-control status (*CaseStatus*).

There are six values for number of partners and two values for case-control status. The degrees of freedom are therefore ((6-1) x (2-1) = 5). The calculated chi-square is 16.3 (rounded) and the p-value is 0.0061.

Single Table Analysis

Chi-squared	df	Probability
16.2953	5	0.0061

In this instance, this chi-square tells us that there is a relationship between the number of sex partners and case control status that does not appear to be due to chance. However, it tells us nothing about the association – not the strength of the association, nor the direction. Thus the use of the chi-square statistic alone is not sufficient for interpreting the relationship.

Although the chi-square test statistic is influenced by both the magnitude of the association and the study size, it does not distinguish the contribution of each. Thus, together, the measure of association and the test of significance (or confidence interval, see below) provide complementary information.

Confidence Intervals

Another measure of the statistical variability of the association is the confidence interval.

The chi-square test and the confidence interval are closely related. The *chi-square test* uses the observed data to determine the probability (p-value) under the null hypothesis, and you reject the null hypothesis if the probability is less than some preselected value, called alpha (α). Usually this value is 5% (0.05).

The *confidence interval* uses a pre-selected probability value, also called alpha (α), to determine the limits of the interval. As with the chi-square test, epidemiologists traditionally choose an alpha level of 0.05 or 0.01. The *confidence coefficient* is a function of alpha, calculated as *100 x (1-alpha %)*.

Typically, a confidence interval is referred to as 95% confidence interval or 99% confidence interval. The 95% or 99% is the confidence coefficient. Once you calculate the confidence interval, you can reject the null hypothesis if the interval does not include the null association value (i.e., 1). Both indicate the precision of the observed association; both are influenced by the magnitude of the association and the size of the study group.

Statisticians define a 95% confidence interval as the interval that, given repeated sampling of the source population, will include the true association value 95% of the time. The confidence interval from a single study may be roughly interpreted as the range of values that, given the data at hand and in the absence of bias, has a 95% chance of including the true value in the population. The confidence interval can be thought of as the range of values that is consistent with the data in your study.

Confidence Intervals for Measures of Association

Unlike the chi-square, the calculation of the confidence interval is a function of the particular measure of association. That is, each association measure has its own formula for calculating confidence intervals (Dicker, p. 152).

Let's return again to the analysis of the *SexPartner3month* variable. (You can redo the TABLE analysis of this variable or scroll up until you find your previous output.)

As you can see from the Parameters section of the Single Table Analysis, two different results are calculated for odds ratio and three for confidence interval.

Single Table Analysis

	Point Estimate	95% Confidence Interval Lower	Upper	
PARAMETERS: Odds-based				
Odds Ratio (cross product)	7.4113	2.1268	25.8269	(T)
Odds Ratio (MLE)	7.3556	2.2937	32.0538	(M)
		2.0715	40.0036	(F)

The first odds ratio is the familiar cross-product ratio (7.4) and the confidence interval is calculated using the Taylor (T) series method for calculating standard error (2.12, 25.83). This is the most well known and commonly used and, in general, should be the statistics used to report results for case-control studies or prevalence odds ratios from cross-sectional studies.

FYI: The Taylor Series

The Taylor series confidence interval, as calculated by Epi Info, is not based on the chi-square test. As a result, the Taylor series may exclude 1 when some chi-square statistics (particularly the "corrected ones") yield a p-value larger than 0.05 or conversely the interval may include 1 when a p-value is smaller than 0.05. This should only happen when the p-value is very close to alpha (usually p = 0.05).

The other odds ratio of 7.36, referred to as MLE (maximum likelihood estimate), is calculated using several iterations on the computer. The confidence interval is marked by (M) in Epi Info. We will not usually use this calculation.

The third set of confidence intervals (2.07, 40.00) are the Fisher exact confidence intervals (F) are considered more appropriate than the Taylor series (T) when numbers in the cells of the 2x2 table are < 5.

You will also see that if the appropriate measure of association for your study were a risk ratio (relative risk) or risk difference (see below), then a different set of confidence intervals would be calculated. You must be sure to use the confidence intervals appropriate for your study type.

PARAMETERS: Risk-based				
Risk Ratio (RR)	1.8743	1.5000	2.3419	(T)
Risk Difference (RD%)	40.2852	24.3970	56.1734	(T)

Interpreting the Confidence Interval

Calculation of a measure of association, such as an odds ratio, plus calculation of a confidence interval provides the "best guess" of the true association as well as an index of how precise or variable that "best guess" is. The width of a confidence interval (i.e., the values included) reflects the precision with which a study can pinpoint an association. A wide confidence interval reflects a large amount of variability or imprecision. A narrow confidence interval reflects little variability and high precision. Usually, given a larger number of subjects or observations in a study, the narrower the confidence interval, the greater the precision.

"Since a confidence interval reflects the range of values consistent with the data in a study, one can use the confidence interval to determine whether the data are consistent with the null hypothesis. Since the null hypothesis specifies that the relative risk (or odds ratio) equals 1.0, a confidence interval that includes 1 is consistent with the null hypothesis. This is equivalent to deciding that the null hypothesis cannot be rejected. On the other hand, a confidence interval that does not include 1.0 indicates that the null hypothesis should be rejected as it is inconsistent with the study results. Thus the confidence interval can be used as a test of statistical significance" (Dicker, p. 153).

Now we need to look at our findings and interpret our results.

Odds Ratio (cross product)	7.4113	2.1268	25.8269	(T)

For *SexPartner3month,* we can see that the interval does not include 1. Our statement can now be phrased, "Having more than one sexual partner in the last 3 months *is significantly associated with being an STI case (OR 7.4, 95% CI 2.13,25.83)."*

We now can complete the analysis for the other exposures of interest.

Analysis of Risk Factors Activity

Directions: Using TABLES command, complete the table below. For p-Value, use chi-square corrected (Yates). Circle significant p-values. Answers are in Appendix C, Table 4. Results reported in the appendix may be rounded.

Table 4: Risk Factor Table

Variable	Cases N (Col %)	Controls N (Col %)	OR (95% CI)	P value
Marital status (*HabitationStatus*)				
Married/Cohabiting (1)				
Single/Div/Widowed (2)				
Level of education (*Education*)				
None/Primary (1)				
Secondary/Tertiary (2)				
Employment (*Unemployed*)				
Unemployed (1)				
Employed (2)				
Age at First Sex (*AgeFirstSex*)				
< 17 years				
17+ years				
> 2partner last 3 months (*SexPartner3month*)				
Yes				
No				
> 2 partner last year (*SexPartner1year*)				
Yes				
No				
3 or more partners in lifetime (*SexPartnerLife3*)				
Yes				
No				
Living w/ partner (*N15LivingTogether*)				
Yes				
No				
Relationship last partner (*LastPartnerSpouse*)				
Yes				
No				
Alcohol use last partner (*AlcoholUse*)				
Yes				
No				
Paid/rcvd money for sex (*N10givereceiveforsex*)				
Yes				
No				
Received help from at least one group (*ReceiveHelp*)				
Yes				
No				
Forced to have sex (Use Select to choose women only) (*N18WereYou*)				
Yes				
No				

Interpretation of Risk Factor Analysis

Having completed the table of bivariate analysis of risk factors we can now draw preliminary conclusions about our findings.

? **Question 16:** What factors do not appear to be associated with having an STI in bivariate analysis? _____

? **Question 17:** What factors are statistically significantly associated with having an STI?
_____

? **Question 18:** What factors might be considered to be of borderline significance?
_____

? **Question 19:** Which among the significant and borderline factors could be considered risk factors and which factors could be considered protective against having an STI?
Risk Factors: _____
Protective Factors: _____

Summarize what you know about risk factors and protective factors from this analysis.

_____

_____

_____

_____

_____

_____

_____

_____

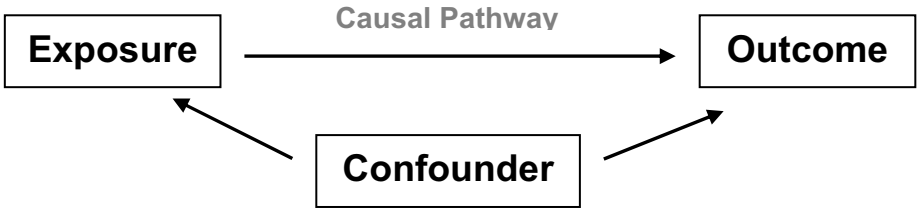
We have now completed the univariate and bivariate analysis. Be sure to CANCEL SELECT to return to the original dataset.

Confounding and Effect Modification

Many epidemiologic studies require more sophisticated analyses than the simple 2x2 tables we have done so far. Even if we are only interested in the association of particular exposures and one particular outcome, other factors (covariates) often complicate the association. The two principal types of complications are *confounding* and *effect modification*.

Confounding is a "mixing" of effects that occurs when a third factor distorts the true association between the exposure and disease. This is a type of bias and we need to control for it in our analysis, if it exists. Like other types of bias, confounding results in a mistaken estimate of an exposure's effect on the risk of the outcome (e.g. disease). However, unlike most types of bias, we can sometimes control for it in our analysis. To be a confounder, three criteria must be met.

1. A variable must be a *risk factor for the outcome*.
2. A variable must be *associated with the exposure*.
3. A variable must *not* be *in the causal pathway* between the exposure and outcome.



As an example, look at the effect of alcohol (the exposure) on developing lung cancer (the outcome). This relationship could be confounded by smoking. Let's see if smoking fits the three requirements to be a confounder.

1. Smoking is a *risk factor* for lung cancer even in the absence of alcohol.
2. Smoking is *associated* with alcohol use (i.e. drinkers are more likely to smoke than the general population).
3. Smoking is not caused by use of alcohol (i.e., smoking is *not in the causal pathway* between use of alcohol and lung cancer).

Thus, smoking meets the criteria for being a possible confounder.

Effect modification (also called statistical interaction) is when a third factor interacts with an exposure to affect the magnitude of an association between exposure and disease. Essentially, the degree of association between an exposure and outcome differs in different subgroups of the population. This is not a bias; it is a true effect and can be informative. It may provide information on particularly high risk or resistant subgroups. We want to understand and report effect modification.

An example would be the effect of use of oral contraceptives (the exposure) on myocardial infarction in women. Smoking in this case has been shown to be an effect modifier. There is a difference in risk of having a myocardial infarction between those women who use oral contraceptives and *do not* smoke and those who use oral contraceptives and *do* smoke. Note that smoking lies within the causal pathway between oral contraceptives and myocardial infarction.

Stratified Analysis

Stratified analysis is one method of dealing with confounding and effect modification and involves examining the association of interest within each category (stratum) of the variable or variables of interest.

In this study, we have found a number of factors associated with STIs – both factors that increase the risk and decrease the risk (were protective). In our analysis, we found that

- The number of sexual partners in the past 3 months, year, and lifetime,
- Alcohol use, and
- Forced sexual intercourse

were all *risk factors* for acquiring an STI.

We also found that

- Living with your regular sexual partner,
- Having your spouse/live-in partner as your last sexual partner, and
- Receiving assistance from certain organizations

were *protective* against acquiring an STI.

However, many of these variables are related to each other. Those who live with their partners may be more likely to have sex with that partner than those live apart from each other. This may also have an effect on the number of partners, which has been identified as a risk factor. Because these variables are so closely related to each other, one variable may change the effect that the other variable has on the outcome through effect modification or confounding.

How do we begin to look for these various issues in our data?

Data-Based Approach to Effect Modification and Confounding

Step 1	Perform crude analysis: calculate measure of association; perform hypothesis testing; calculate confidence intervals and/or p-values.
Step 2	Stratify the data by potential confounders and effect modifiers; calculate stratum-specific measures of association.
Step 3	Evaluate for effect modification.
Step 4	If effect modification is present, show stratum-specific results.
Step 5	If effect modification is absent, assess and control for confounding as necessary.

Step 1: Perform Crude Analysis

In our case, we have already conducted Step 1. Doing this step we identified those factors significantly associated with the disease or outcome (see Analysis of Risk Factors Activity, page 79). The odds ratios we have calculated would be considered crude as they are not adjusted for any other variables.

Step 2: Stratify the Data

To conduct Step 2 we need to stratify by other variables that are also significantly associated with the outcome. In stratification, all the subjects in a given stratum have the same value for the stratification variable. For example, if we stratify by gender, all the persons in one stratum are female and all the persons in the other stratum are male. Thus, the associate between the exposure and the outcome cannot be confounded by gender.

We identified that having your last sexual relationship with your spouse/live-in partner was weakly protective against STIs (i.e., the OR was slightly less than 1.0). We also found that living together with your regular sexual partner was statistically significant for a protective effect. How might the protective effect of one's last sexual encounter being with a spouse/live-in partner be affected by whether one lived with one's regular sexual partner? *This may seem confusing, but consider that one might live with one's spouse and have had one's last sexual encounter with a person other than one's spouse.*

We will examine the risk for last sexual encounter with spouse/live-in partner (*LastPartnerSpouse*) stratified by whether one lived with one's regular sexual partner (*N15LivingTogether*).

1. **Click on the TABLES command.**
2. **Choose the Exposure variable.**

   Choose *LastPartnerSpouse*.

3. **Choose the Outcome variable.**

   Choose *CaseStatus*.

4. **Choose the variable to Stratify By.**

   Choose *N15LivingTogether*.

5. **Click OK.**

🗋 **NOTE:** The analysis output will include a table and statistics for *N15LivingTogether* = 1, a table and statistics for *N15LivingTogether* = 2, and a Summary table. Make sure to scroll through the output to find the answers.

Fill in the blanks, based on the results as presented. Round to the nearest tenth.

? **Question 20:** What is the odds ratio for *LastPartnerSpouse* among those who are living together with their partner/spouse (N15LivingTogether = 1)? _____

? **Question 21:** What is the odds ratio for *LastPartnerSpouse* among those who are not living together with their partner/spouse (N15LivingTogether = 2)? ____

? **Question 22:** What is the crude odds (cross product) ratio? _____ (See the Summary table for this information)

Note that the crude odds ratio for *LastPartnerSpouse* in the stratified analysis output (0.70) differs from the odds ratio for *LastPartnerSpouse* that we analyzed in the bivariate table (0.58). This is because the number of persons who answered the question *N15LivingTogether* (n = 191) was less than all those who answered *LastPartnerSpouse* (n = 226) and only those 191 persons with complete information for both variables are considered in this analysis.

Step 3: Evaluate for Effect Modification

To identify effect modification we want to determine if the stratum specific measures differ from one another. We are trying to answer two questions:

1. **Is the range of associations wide enough to be of importance?**

   In order for a difference to be meaningful, it should have implications for disease burden or control.

2. **Does the range of associations just represent normal sampling variation?**

> We can evaluate this qualitatively ("eyeballing" or scanning the data), by looking to see if the stratum-specific or confidence intervals overlap, or quantitatively using statistical tests for heterogeneity.

In our example, we can see that the stratum-specific measures of association (the odds ratios) range from less than one (statistically significant protective effect) to greater than one (statistically significant risk factor); thus, they appear wide enough to be important.

We can see that, among those who are living together with their regular sexual partner, having one's most recent sexual partner be one's spouse/live-in partner was significantly protective (OR = 0.06, 95% CI = 0.01, 0.47).

LastPartnerSpouse : CaseStatus, N15LivingTogether=1

	Point Estimate	95% Confidence Interval Lower	Upper	
PARAMETERS: Odds-based				
Odds Ratio (cross product)	0.0556	0.0066	0.4693	(T)
Odds Ratio (MLE)	0.0574	0.0024	0.3861	(M)
		0.0012	0.4670	(F)

However, among those who did not live with their regular sexual partner, having sex with one's spouse/live-in partner was not statistically related to the risk of STI (OR = 2.05, 95% CI = 0.90,4.66).

LastPartnerSpouse : CaseStatus, N15LivingTogether=2

	Point Estimate	95% Confidence Interval Lower	Upper	
PARAMETERS: Odds-based				
Odds Ratio (cross product)	2.0490	0.9007	4.6612	(T)
Odds Ratio (MLE)	2.0344	0.8901	4.7004	(M)
		0.8309	5.0520	(F)

The **effect** of having last sex with your spouse/live-in partner is **modified** by whether you live with your regular sexual partner. The protective effect of having your last sexual encounter with your spouse is only realized when you actually live with that person. In statistical terms, there is an **interaction** between living together and last sexual encounter when assessing the risk of STIs. In addition, we have information from the statistical test to confirm that they are statistically different.

The *Chi-square for differing Odds Ratios by stratum (interaction)* in the Summary section of the output, which calculates a chi-square and p-value, has been done for us and indicates that the odds ratios differ significantly by stratum (Chi-Square statistic = 9.56, p-value = 0.002).

In the following two tests, low p values suggest that ratios differ by stratum		
Chi-square for differing Odds Ratios by stratum (interaction)	9.5631	0.0020
Chi-square for differing Risk Ratios by stratum	23.1411	0.0000

Step 4: Present Stratum-Specific Results When Effect Modification is Present

If there is a significant difference between the measure of effect (OR, RR) between the strata, effect modification has occurred. In our example, we cannot report the adjusted or the crude odds ratio as they do not reflect what is happening. When effect modification is identified, it is important to present the stratum specific results.

STI and Last Sexual Partner as Spouse/Live-In Partner, Stratified by Living Together with Regular Sexual Partner

* Results Rounded

Covariate		LastPartnerSpouse		Stratum-specific OR (95% CI)	Chi-square for interaction p-value
		Case (%)	Control (%)		
Crude (unadjusted) OR=0.7 95% CI=(0.4, 1.4)					
Living Together with Last Sexual Partner	Yes	75.0	98.2	0.06 (0.001, 0.5)* *Fisher exact	0.002
	No	72.1	55.8	2.05 (0.9, 4.7)	

Thus, at this point we would stop, as we do not need to report Mantel-Haenszel odds ratios in the presence of effect modification.

Step 5: Assess for Confounding When Effect Modification is Not Present

Let's look at another example. We found that having one's last sexual partner be one's spouse/live-in partner was almost a protective factor. Is this independent of the risk of having more than 1 sexual partner in the past year?

Analyze the data using the TABLE command. Choose the Exposure variable as *LastPartnerSpouse,* the Outcome Variable as *CaseStatus*, and Stratify By *SexPartner1year*.

Examine the results and complete the following questions.

? **Question 23:** What is the odds ratio for those having last sex with spouse among those who had more than one partner in the last year (*SexPartner1year* = Yes) ? _____

? **Question 24:** What is the odds ratio for those having last sex with spouse among those who did not have more than one partner in the last year (*SexPartner1year* = No)? _____

? **Question 25:** Examining these odds ratios and the Woolf test for heterogeneity (chi-square for differing odds ratios by stratum or interaction) in the summary statistics, do you think there is important effect modification? Yes No

? **Question 26:** Looking at the summary statistics, compare the crude odds ratio _____ to the adjusted odds (MH) ratio _____ (round to nearest hundredth).

To look for confounding (Dicker, p.158):

Step 5a	Look at the smallest and largest values of the stratum-specific measures of association and compare them with the crude value. If the crude value does not fall within the range between the smallest and largest, confounding may be present.
Step 5b	Calculate a summary adjusted measure of association. This is a weighted average of the stratum-specific measures usually the Mantel-Haenszel summary odds ratio. The Mantel-Haenszel test is used for testing the overall association between an exposure and an outcome in a stratified analysis.
Step 5c	Compare the summary adjusted measure to the crude measure to see if they are "appreciably different". If they are different, then the crude estimate is confounded.

In our case, for Step 5a we are comparing the smallest value (Question 24 answer = 0.80) to the largest value (Question 23 answer = 1.24). The crude value (Question 26 answer = 0.58) does *not* fall within the range of the smallest and largest values, so confounding is present.

The next step, 5b, is to calculate a summary adjusted measure of association (second answer in Question 26 = 0.88). Then, in Step 5c, we compare this summary adjusted measure of association value (0.88) to the crude measure (0.58) to see if they are different. There are no hard and fast rules or statistical tests to determine what constitutes appreciable difference. In practice, we usually assume that the adjusted value is a preferred summary of the study odds ratio or risk ratio. The question becomes, does the

crude value adequately approximate the adjusted value or would the crude value be misleading to the reader?

Confounding is an issue of distortion or misrepresentation of an association. (see previous discussion p. 81). If a crude measure of association is different enough from an adjusted one to be misleading, the adjusted measure should be presented.

In our example, we see that the two stratum specific odds ratios are 1.24 and 0.80. When we look at the Woolf test for heterogeneity or *Chi-square for differing Odds Ratios by stratum (interaction*), we see it has a chi-square of 0.30 and p-value of 0.59. We conclude that the differences between stratums are consistent with sampling variation. Thus there is no effect modification.

Considering confounding, we note that our maximum (1.24) and minimum (0.80) values do not include the crude odds ratio (0.58). Comparing the adjusted odds ratio to the crude (0.58 vs. 0.88), we can conclude that they are different and the crude value overestimates the protection of having last sex with spouse.

Stratification for Subgroup Analysis

We have presented the data analysis to this point using cases and controls as single groups. That is, all the cases have been compared to all the controls. This is how data is analyzed in the situation where cases and controls are identified appropriately but are not matched in anyway. We have done this to clearly illustrate the basic points in comparing cases and controls in an analysis.

However, in this study choices were made in the selection of cases and controls that have implication for an appropriate analysis. As you recall, an equal number of males and females were selected as cases which did not reflect the proportion of men and women seeking care for STIs in this public clinic. This has the advantage of allowing us to look at risk factors by gender.

In this case, gender cannot be considered a true confounder as it is not an independent risk factor for STI. However, it is very likely that certain potential risk factors (e.g., behaviors or personal characteristics) vary by gender and that those factors impact the exposure variables that one faces. Although gender itself may not be a confounder, it could indirectly impact other variables. For this reason doing a subgroup analysis of these behaviors/characteristics by gender may provide gender-specific information that will be useful in the interpretation of our results.

Effect Modification Activity

The following variables are ones that you previously computed in your bivariate analysis and now you will analyze them again with the addition of stratification by gender (*Sex*). Evaluate for effect modification and comment on your findings using the table below.

- Having a regular sexual partner (*N14DoYouHave*)
- Living with the regular sexual partner (*N15LivingTogether*)
- Relationship with last sexual partner (*LastPartnerSpouse*)

? Question 27: Complete the table below.

	N14DoYouHave	***N15LivingTogether***	***LastPartnerSpouse***
What is the OR for females?			
What is the OR for males?			
What are the Chi-square and p-value for the *Chi-square for differing Odds Ratios by stratum (interaction)*?			
Does it appear that there is effect modification?			

Frequency Matching and Stratification

The second method used in our study to control for confounding that affects analysis is *frequency matching* in the identification of controls. The controls were matched by gender and five year age group to the cases. Why did we do that? The most common reason for frequency matching (also known as category matching) is to have controls that are similar to cases on factors that may be associated with the outcome of interest but are not being targeted for study. In fact, age and sex are very frequently related to an outcome, but are not the factors of concern to investigators in this study. Thus, matching helps to control for confounding that might occur if these factors varied significantly between cases and controls. However, matching in design alone does not control for confounding; it must be accompanied by stratification in the analysis.

As you may recall from your epidemiology course, if you have done pair-wise matching (where each case has a specifically matched control), you need to do a different type of analysis called matched pairs analysis (Hennekens, p. 296).

In the case of frequency matching, as we have done, you do not need to change the analysis method; however, you do need to *stratify* on the factors you have matched on and, to be accurate, you need to report the Mantel-Haenszel adjusted odds ratio (Hennekens, p. 304). In Epi Info, this statistic is reported as the "Adjusted OR (MH)". We will now demonstrate this for risk factor *SexPartnerLife3* (number of lifetime sexual partners greater or less than 3) by age group (*Agegrp*) and gender (*Sex*).

For this example we are going to look at males only.

1. **Use the SELECT command to choose only one of the strata.**

   Select only the males.

2. **Use the TABLES command to select the variable of interest and stratify by the second strata.**

   In TABLES, choose *SexPartnerLife3* as the exposure variable, *CaseStatus* as the outcome variable, and stratify by *Agegrp.*

3. **Click OK.**

You will notice that you get a series of 2x2 tables – one for each age group. As you scroll through, notice the odds ratios for each strata. Some are zero as there are cells in the table with no values.

Scroll down to the summary table until you see the statistics for odds ratios (see next page).

C:\Documents and Settings\NIS9\Desktop\STI\OUT10.htm

Previous Next Last History Open Bookmark Print Maximize

SUMMARY

Back Forward Current Procedure

SUMMARY INFORMATION

Parameters	Point Estimate	95%Confidence Interval Lower	Upper	
Odds Ratio Estimates				
Crude OR (cross product)	4.3714	1.9810,	9.6461	(T)
Crude OR (MLE)	4.3128	1.9701,	9.7605	(M)
		1.8538,	10.4793	(F)
Adjusted OR (MH)	4.3817	1.8923,	10.1459	(R)
Adjusted OR (MLE)	4.4023	1.9252,	10.5377	(M)
		1.8020,	11.4235	(F)

Note that the crude odds ratio for males is 4.37. Next we will look at the Mantel-Haenszel (MH) adjusted odds ratio.

Formula for Mantel-Haenszel Odds Ratio (adjusted)

$$\text{OR} \quad \frac{\sum[\frac{a_i d_i}{N_i}]}{\sum[\frac{b_i c_i}{N_i}]} \text{ for each stratum}$$

If we look at the data from the Epi Info summary table we see that the crude odds ratio that we previously calculated of 4.37 is not importantly different from the MH adjusted OR of 4.38. Thus, in this case, age groups are not confounding the relationship between the number of lifetime sex partners and we can present either the crude or the adjusted odds ratio.

If there is little confounding as demonstrated by little difference in the crude and adjusted odds ratios, as is often the case in frequency matching or the matching for convenience on age and sex, then the risk estimates (odds ratios and risk ratios) may be virtually identical in which case the unmatched (or crude) results may be presented (Hennekens, p.304).

NOTE: Although you will need to use your best judgment, a general guideline to follow when assessing similarity is 20%. If there is a difference of <20%, then we can say that the numbers are similar.

Assessing Confounding Activity

Directions: Complete the following table on these risk factors for men indicating the crude and adjusted odds ratios for these risk factors. In the comment section, note whether the crude and adjusted odds ratios are *similar* (and do not indicate confounding by age group) or if there is enough dissimilarity to indicate *possible confounding* by age group. See the results in Appendix C, Table 5.

Table 5: Stratified analysis for men to check for confounding by age group

Variable	Crude OR	95% CI	MH Adjusted OR	95% CI	Comment
LastPartnerSpouse					
N15LivingTogether					
SexPartner1year					
Education					
AlcoholUse					

See the results of these same risk factors for females on the following page. (We have provided the answers to 4 decimals for comparability with the output from Epi Info.)

Results for Females

Variable	Crude OR	95% CI	MH Adjusted OR	95% CI	Comment
LastPartnerSpouse	**3.1977**	0.95-10.77	**3.2625**	0.96-11.03	similar
LivingTogether	**0.39**	0.17-0.92	**0.3598**	0.15-0.88	similar
SexPartner1year	**9.2174**	1.11-76.49	**9.4463**	1.13-78.64	similar
Education	**1.6**	0.53-4.85	**1.9599**	0.55-6.98	Similar – but possible confounding
AlcoholUse	**1.5114**	0.68-3.36	**1.4946**	0.68-3.27	similar

Having done this exercise you will have realized that so many of the tables have cells that are very small or zero, making calculation of the stratum specific odds ratios unspecified. This leads us to the case where we need to consider another method to control for these factors and still get reliable results.

Multivariate Analysis

In many studies, univariate, bivariate, and stratified analyses are sufficient to provide the necessary information to make meaningful conclusions about the risk factors for a given outcome. In some studies, however, we have too many potential confounders. Stratified analysis is unable to control simultaneously for even a moderate number of potential confounders.

In a study such as ours, where we have multiple significant risk factors that may be confounding or interacting with other factors, controlling for several risk factors at the same time would give us many strata with few or no individuals, making our analysis unreliable or even impossible.

Multivariate analysis allows for the efficient estimation of measures of association while controlling for a number of confounding factors simultaneously, even in situations where stratification would fail because of insufficient numbers (Hennekens, p.315).

It is beyond the scope of this exercise to describe all the various types of multivariate analyses. We will focus on applying a specific multivariate technique that is commonly used and available in Epi Info. Logistic regression involves the construction of a mathematical model to describe the association between exposure and disease and other variables that may confound or modify the effect of exposure. Modeling is a technique of fitting the data to particular statistical equations (Dicker p. 160). In this model the outcome (disease) is a function of

- Exposure variables,
- Confounders, and
- Interaction terms (effect modifiers).

Logistic Regression

Logistic regression is used when the outcome is binary such as ill/well, case/control, alive/dead etc. In logistic regression, a dependent (binary outcome) variable is modeled as a function of independent variable or variables. The independent variables should include the exposure or exposures of primary interest and may include confounders and more complex interaction terms.

In logistic regression:

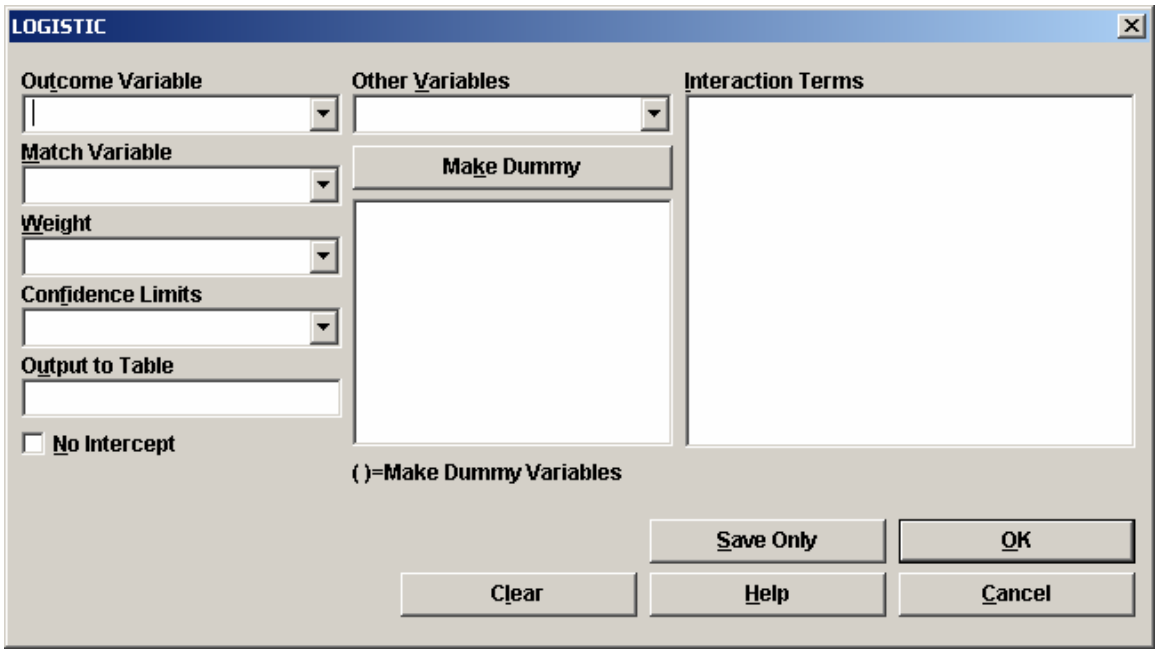
Dependent Variables include:	**Independent Variables (Covariates) include :**
• Binary outcome variables	• Exposure variables • Confounders • Complex interaction terms

Epi Info will calculate the odds ratio for each independent variable. If the model includes only the outcome variable and one exposure variable, the results should equal the odds ratio that can be calculated from the 2x2 table. When other variables are included, the odds ratio calculated is adjusted for all the other variables.

Logistic regression can be applied to case-control, cohort, and cross-sectional data. All types of variables (categorical and continuous) can be included in a logistic regression model, although categorical (coded) variables make results easier to interpret because the number of possible responses has been limited (i.e., all possible responses for continuous variables can be numerous). As mentioned, the outcome variable must be binary or dichotomous. The purpose of logistic regression is to produce a mathematical equation that relates the probability of an outcome to the particular value of risk factor variables. In our case we will be modeling the probability of having an STI given various risk factors. For example,

Risk of STI = Number of sexual partners + Living with spouse + Last sex with spouse + etc.

Look at the command window that appears when we choose "Logistic Regression":



- The **Outcome Variable** represents the dependent variable. As noted before, this must be a dichotomous 0/1 variable (or Yes/No).
- The **Other Variables** represents the independent variable(s) that can be a numeric variable or a non-numeric variable.
- **Interactions Terms** are created by selecting at least two variables from the Other Variables drop-down box, clicking on each of them as they appear in the box below (which will cause a new button called Make Interaction to replace the Make Dummy button), and then clicking on the Make Interaction button. The

variables will then appear in the box below Interaction Terms specified by an asterisk (*) between variables. This will be demonstrated in the exercise to follow.

- **Confidence Limits** are 95% by default. You can select 90% or 99% as well.

Dummy Variables
Where categorical variables include more than two values, a dummy variable (dichotomous coded variable) is created for all but one level of the variable so that the levels can be compared to that one level. This program will create dummy variables for those categorical variables with more than two values (as well as any variable you designate as a dummy variable by clicking on the Make Dummy button).

For example, in our data set we have a categorical variable with more than two levels that needs to be included in the model. This is our age group variable; because we frequency matched on five year age groups, we need to keep age group in the model. The program will create dummy variables for each of the values of age group except the lowest age group. It then compares each age group level to the youngest age group. We will look at this later in this section.

For more information you can review dummy variables in the Help section of Epi Info (see page titled "How Can I Use Explanatory Categorical Variables with More Than 2 Values?").

If there are continuous variables in the dataset (age, or duration of a treatment, for example), it is possible to include them as continuous variables in the logistic regression model. The results are expressed as the risk for each unit change of the continuous variable (each year of age or each day of a treatment, for example) and the risk for the number of units desired will have to be computed.

Let's do an exercise with a very simple model. Let us look at just one Other Variable first. The first thing we must do is change our *CaseStatus* variable so that it is a Yes/No variable.

Recode a Numeric Variable to a Yes/No Variable

1. **DEFINE a new variable.**

   Use the DEFINE command to create a new variable called *CaseLR*.

2. **RECODE the Number variable to a new Yes/No variable.**

   We will walk through the steps for recoding a numeric variable with a single response to a Yes/No variable.

   a. **Click on the RECODE command.**

b. In the From drop-down box, select the original variable you are going to convert.

Select *CaseStatus*.

c. In the To drop-down box, select the newly defined variable.

Select *CaseLR*.

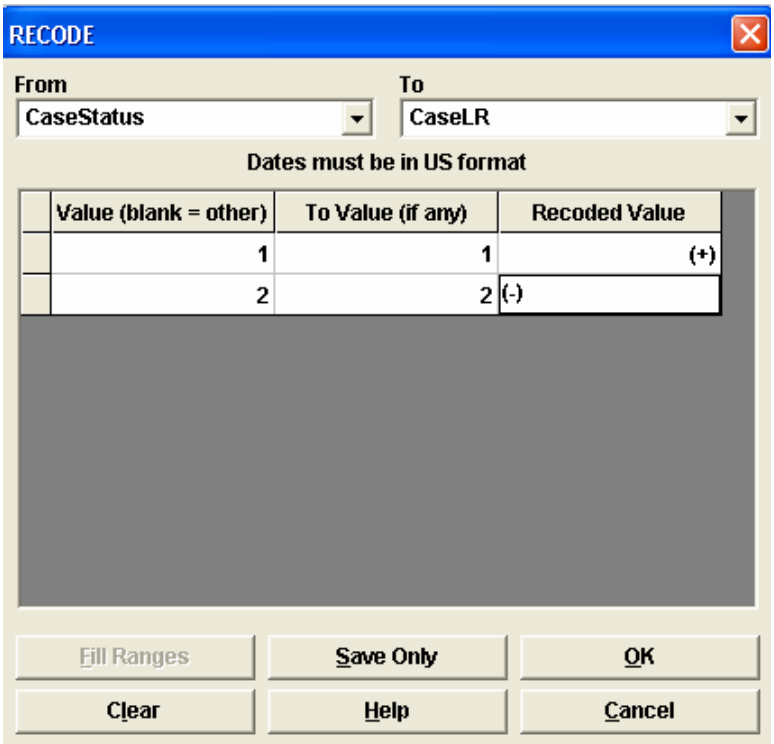
d. In the Value and To Value columns, enter in the range of numeric values from the original variable.

Although in this case we only have one value for each range (e.g., 1 for case), we will need to enter in values in both columns. Type 1 in the "Value" and in the "To Value "columns.

e. In the Recoded Value column, enter the value to give to the new variable.

In the Recoded Value column, enter in the code for Yes: (+)

Repeat steps d. and e. for the No variable. See the example below.



f. Click OK.

Next we will do a Logistic Regression using *CaseLR* as our outcome variable and use only *LastPartnerSpouse* as the only other variable.

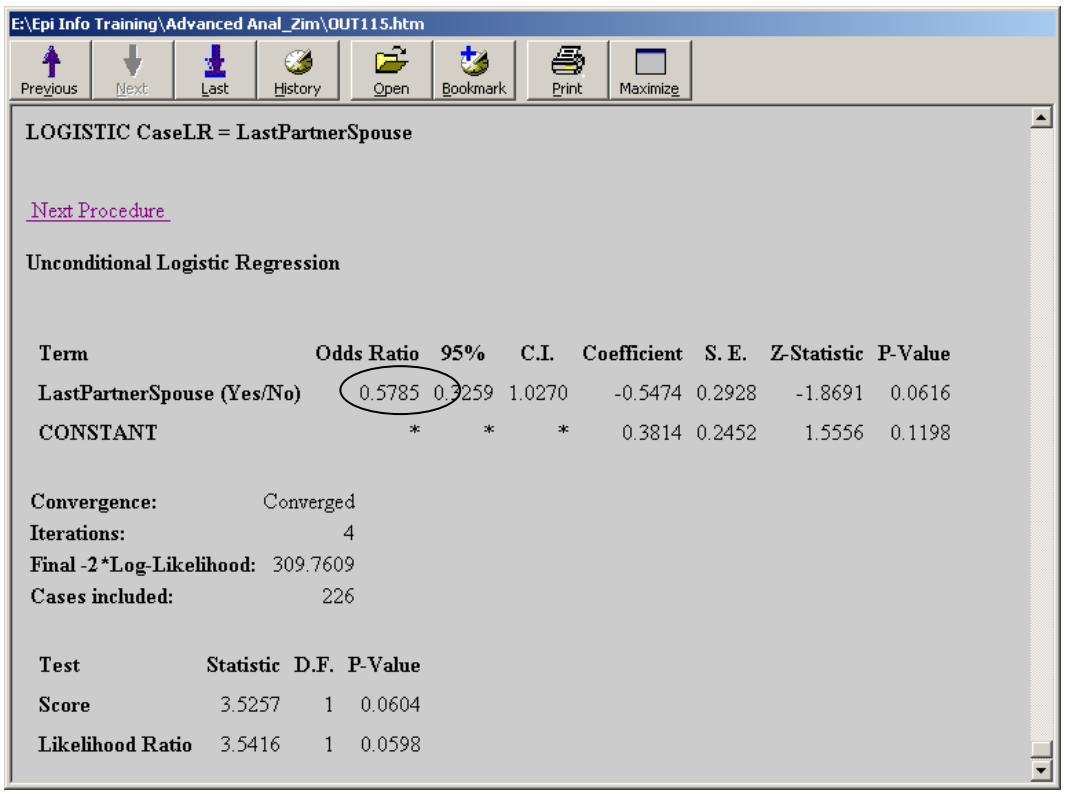
1. **Choose the Logistic Regression command.**
2. **Choose the Outcome Variable.**

   Choose *CaseLR.*

3. **Choose the Other Variables.**

   Choose *LastPartnerSpouse*.

4. **Click OK.**



NOTE: There are multiple elements to the output from logistic regression analysis. It is not possible here to address each element. We will limit our discussion to the minimum we need to identify risk factors. For additional information on logistic regression, please consult a biostatistics text.

Looking at our output, you will see that for each variable Epi Info has provided

- an odds ratio,
- the 95% confidence interval,
- the coefficient,
- standard error (S.E.),

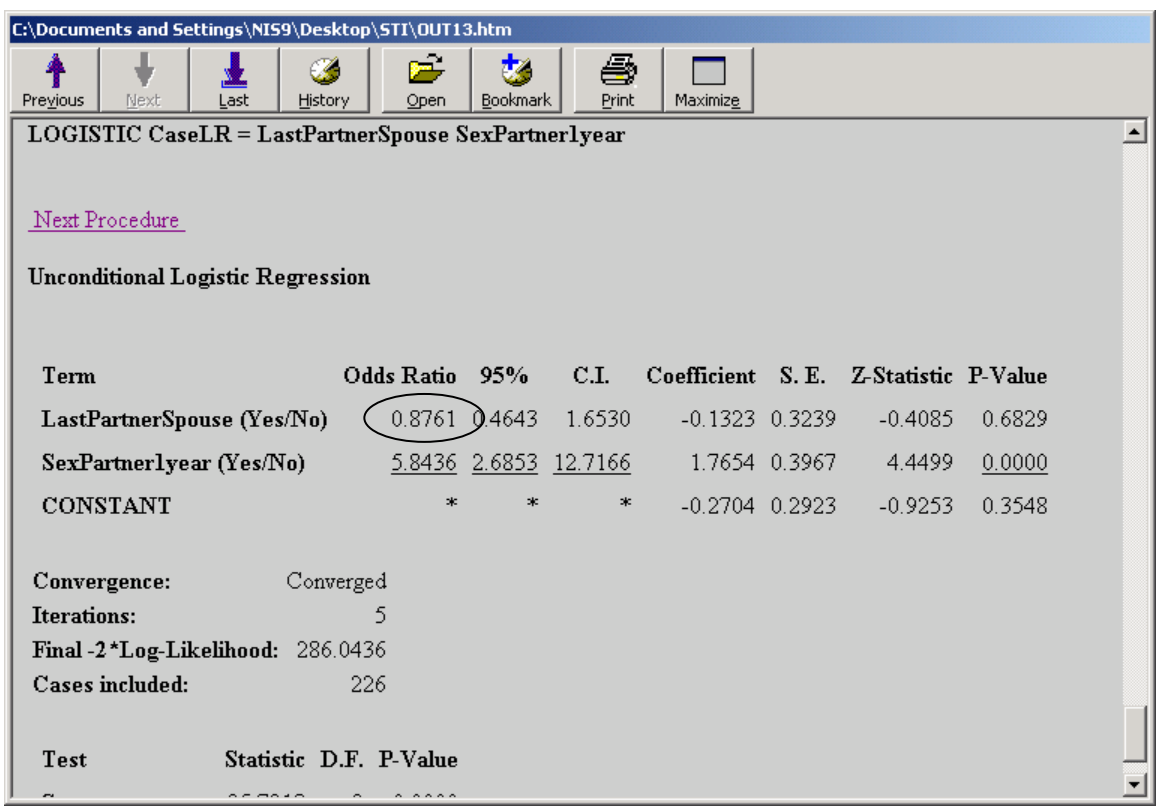
- Z-statistic and
- the p-value (Note that significant p-values are underlined; 0.0000 indicates a p-value < 0.0001).

When we previously analyzed *LastPartnerSpouse* using the tables command, we got 0.5785 as our odds ratio. We can see that the calculated odds ratio for logistic regression is the same as our "Tables" function calculated.

The p-Value is calculated based on the Wald test. The Wald statistic is calculated from the coefficient divided by the standard error. This is a Z-statistic that, when squared, is a chi-squared statistic with one degree of freedom for which the p-value can then be determined. This test and its accompanying p-value tests whether there is a significant effect of that variable on the outcome variable controlling for the other variables in the model. In this case there are no other variables.

Try It!

Using the same steps, do a logistic regression with the same outcome variable (*CaseLR*) and, for the Other Variables, choose both *LastPartnerSpouse* and the variable that was confounded by this variable (*SexPartner1year*).



Previously, we conducted a stratified analysis using *LastPartnerSpouse* as our Exposure variable and *SexPartner1year* as the stratification variable. Our adjusted odds ratio was 0.8763. We see that the odds ratio that appears in our logistic regression output (0.8761) is essentially the same as our adjusted odds ratio in our stratified analysis.

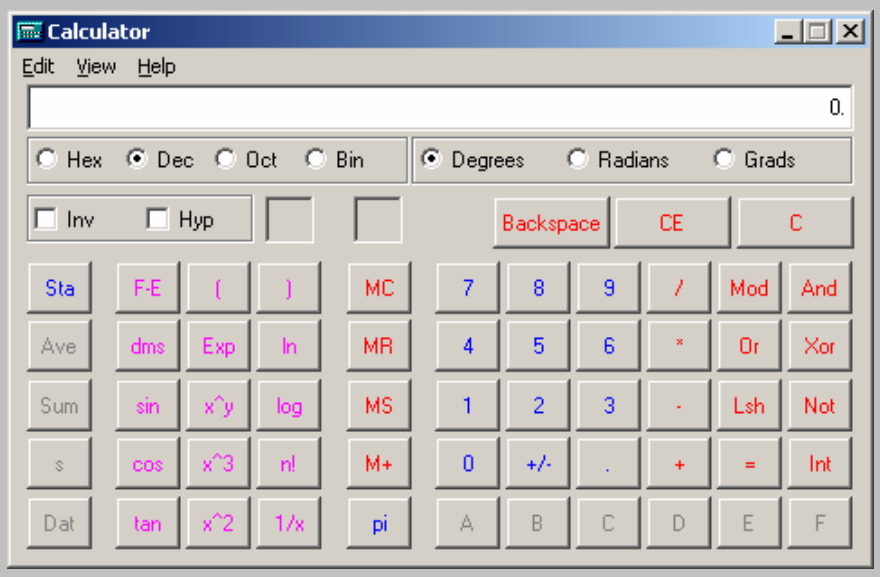
The odds ratio is calculated as the exponent of the coefficient (referred to as the Beta). For example, you can see that the odds ratio for *LastPartnerSpouse* is 0.9 and the coefficient is -0.1. If you were to calculate the natural antilog of the coefficient $e^{(-0.1)}$, you would see that it equals 0.9, or the odds ratio. (See the next page for directions on using the Windows scientific calculator.)

FYI: Windows Scientific Calculator

Most Windows computers have a calculator automatically installed. It has both a basic and a scientific calculator. To find the calculator,

1. Click on the Start button (usually at the bottom left of your computer).
2. Go to Programs, then Accessories.
3. Select the Calculator.

If you do not find the calculator, you may try using the Run function. Click on Start, click Run, and then type: calc (or calculator), and click OK. Typically, you will see a standard calculator with basic functions. To choose the scientific calculator, click on View and select Scientific.



To find the exponent of a number,

1. Type in the number (use the +/- key if it is a negative number).
2. Click on the **Inv** checkbox.
3. Click on the **ln** button.

Try this using the coefficient for *LastPartnerSpouse* (-0.1323) and see if your result is the same as odds ratio (rounded).

We will now develop the model for our logistic regression analysis.

Model Building Strategy: Including Variables in a Logistic Regression Model

To conduct a logistic regression, you must have a strategy for how to include variables in your model. There are several approaches to creating the appropriate final model. However, a critical step is ALWAYS to have done univariate, bivariate, and stratified analysis first so that you understand your data sufficiently. The stratified analysis allows you to ensure that you understand the confounders and effect modifiers (interaction) and assists in choosing which variables to include in the model.

Logistic Regression Model Building Strategy

1. Identify the variables to be considered through analysis of univariate, bivariate, and stratified analysis.
2. Evaluate the interaction.
3. Evaluate confounding.
4. Choose the final model.

Step 1: Identify the variables

Include variables that may be risk factors or control variables (e.g., variables that have been used to frequency match). Many suggest including all variables for which the p-value of the chi square, Fisher exact, or t-test is less than 0.25 (not the usual threshold of 0.05) in the Epi Info TABLES or MEANS commands.

Look back at our bivariate risk factor analysis (Appendix C, Table 4) and identify those variables that were significantly associated with STI at the p = 0.25 level.

? **Question 28:** Which variables were significantly associated with STI at the p = 0.25 level? _____

_____

_____

Age group and gender are variables on which we frequency matched our controls so they also must be included.

Step 2: Evaluate interaction

As we have discussed interaction means that the odds ratio for the association between a risk factor variable and an outcome variable varies with the value of another variable. Interaction needs to be considered in constructing your logistic regression model.

In our stratified analysis we identified interaction between *LastPartnerSpouse* and N15*LivingTogether*; thus, we will want to consider these interaction terms related to those two factors as we create our model.

Step 3: Evaluate confounding
Confounding is evaluated by examining the coefficients in the model as variables are added.

Step 4: Choose the final model
Some people start with all possible variables in the model and then work to simplify by removing insignificant values. Others work up by starting with a simple model with a single exposure variable and then looking at increasingly complex models. We will demonstrate the second or "step-up" procedure.

We will begin the process by starting with a simple model and building up until we reach what we consider our best and final model, taking into consideration the factors we have mentioned in the first three steps.

Other factors to consider:

- For most epidemiological research you want to explain your data with as few variables as possible.
- The model should include the most number of cases.
- The model should make biologic sense.

You will very likely need to work with a statistician to optimize your choice of models – but you should be sure that you fully understand your data through your stratified analysis before you consider logistic regression analysis.

Create a Model for Logistic Regression Analysis

Now, we will create a model for our logistic regression analysis.

Although the variables that we created for bivariate analysis are dichotomous, some of the original variables are not simple Yes/No variables (or coded 0/1). In any dataset you would need to recode these variables just as we did for *CaseStatus*. In this case, we are supplying the completed data set (*zimstudytrnlr.mdb*), so this has already been done for you.

There are some variable name changes in this process which are noted below. The new variable names will be indicated through the rest of the material.

Old Variable Name	New Variable Name
AlcoholUse	AlcoholUseLR
N10givereceiveforsex	MoneyforSex
N15LivingTogether	LiveTogether
N18WereYou	ForcedSex
Unemployed	UnemployedLR

Due to the level of interaction due to gender (*Sex*), which you found in your stratified analysis, and the fact that a significant risk factor for women (*ForcedSex*) was not studied for men, we have chosen to conduct our logistic regression analyses separately for men and women.

Let us first set up a model for the men. We will begin with a simple model that includes only one of the variables that were significant in our bivariate analysis as well as the categorical variable for age (*Agegrp*). This needs to be included because we frequency matched on age categories and thus must control for age group in our analysis. We will build up by adding other variables to obtain a final model.

Logistic Regression Model (for Males)	STI = *Agegrp* + *LastPartnerSpouse*

Logistic Regression Analysis

First, **READ the *zimstudylr.mdb* file.** If you have not moved it yet from the CD-ROM to the STI folder, do so now and then READ the file. Do not forget to click on the Change Project button and select the file there first. Once you have changed the project, select the table called *STIdataLogisticRegression*.

1. **SELECT any variable you wish to analyze separately.**

   Select *Sex*= "Male"

2. **Choose the Logistic Regression command.**
3. **Choose the Outcome Variable.**

   Choose *CaseLR.*

4. **Choose the Other Variables.**

   Choose *Agegrp*, *LastPartnerSpouse*.

5. **If a dummy variable is needed, click on the variable and then click the Make Dummy button.**

   Epi Info automatically creates dummy variables for text and yes/no type variables. The *Agegrp* variable, a text variable, has six possible nominal responses (< 21, 21-25, 26-30, 31-35, 36-40, and 41+ yrs). You will see in the output that a dummy variable is created for five of the possible responses and are each compared to the remaining response; in this case, to the < 21 response.

   The dummy variable will automatically be created for us, so we do not need to follow this procedure. You should know, however, that you may choose to create a dummy variable for variable types other than text or yes/no types.

6. **Click OK.**

C:\Documents and Settings\NIS9\Desktop\STI\OUT16.htm

Previous | Next | Last | History | Open | Bookmark | Print | Maximize

LOGISTIC CaseLR = Agegrp LastPartnerSpouse

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
Agegrp (21-25/<21)	2.3661	0.3548	15.7818	0.8613	0.9682	0.8896	0.3737
Agegrp (26-30/<21)	14.8642	1.6435	134.4334	2.6990	1.1235	2.4022	0.0163
Agegrp (31-35/<21)	12.2483	1.2673	118.3740	2.5054	1.1574	2.1647	0.0304
Agegrp (36-40/<21)	5.3829	0.3445	84.0972	1.6832	1.4024	1.2002	0.2301
Agegrp (41+yrs/<21)	18.0121	1.5728	206.2774	2.8910	1.2440	2.3240	0.0201
LastPartnerSpouse (Yes/No)	0.0742	0.0205	0.2692	-2.6007	0.6574	-3.9560	0.0001
CONSTANT	*	*	*	-0.4435	0.9005	-0.4925	0.6224

Convergence:	Converged
Iterations:	6
Final -2*Log-Likelihood:	137.0353
Cases included:	118

Test	Statistic	D.F.	P-Value
Score	23.4866	6	0.0006
Likelihood Ratio	26.5474	6	0.0002

This indicates that, controlling for age, *LastPartnerSpouse* is significantly associated with STI and, in fact, is protective. We determine this by looking at the p-value calculated

using the Wald test. Now we want to see if this remains true when we add other variables.
Note that while we have included the age group to control for the frequency matching by 5 year age groups in our design, we are not actually going to be interpreting the odds ratios for the different age groups. Because we controlled for age it must be included in the model, but it cannot be examined in the analysis.

We want to add another significant variable from our univariate analysis. We will add *LiveTogether*. Recall that we identified an interaction between two variables (*LastPartnerSpouse* and *LiveTogether*), so this will also be included.

NOTE: Although we demonstrated finding this interaction for men and women combined – the interaction was also demonstrated for men alone in a stratified analysis.

In setting up the model, list the single variables first. Then include the interaction term.

Logistic Regression Model (for Males)	STI = *Agegrp* + *LastPartnerSpouse* + *LiveTogether* + *LastPartnerSpouse*LiveTogether*

1. **Choose the Logistic Regression command.**
2. **Choose the Outcome variable.**

   Choose *CaseLR.*

3. **Choose the Other Variables.**

   Choose *Agegrp*, *LastPartnerSpouse*, *LiveTogether*.

4. **Click on Interaction Terms to highlight them.**

   Choose *LastPartnerSpouse* and *LiveTogether*.

5. **Click on the Make Interaction button.**
6. **Click OK.**

C:\Documents and Settings\NIS9\Desktop\STI\OUT16.htm

Previous | Next | Last | History | Open | Bookmark | Print | Maximize

LOGISTIC CaseLR = (Agegrp) LastPartnerSpouse LiveTogether LastPartnerSpouse*LiveTogether

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
Agegrp (21-25/<21)	1.6793	0.1324	21.3011	0.5184	1.2961	0.4000	0.6892
Agegrp (26-30/<21)	13.9045	0.7813	247.4596	2.6322	1.4689	1.7919	0.0731
Agegrp (31-35/<21)	12.8159	0.6794	241.7697	2.5507	1.4986	1.7020	0.0888
Agegrp (36-40/<21)	5.4272	0.1739	169.4270	1.6914	1.7556	0.9634	0.3353
Agegrp (41+yrs/<21)	13.5807	0.6009	306.9417	2.6087	1.5908	1.6398	0.1010
LastPartnerSpouse (Yes/No)	0.2087	0.0332	1.3127	-1.5670	0.9383	-1.6700	0.0949
LiveTogether (Yes/No)	3.0206	0.2677	34.0821	1.1055	1.2364	0.8941	0.3713
LastPartnerSpouse (Yes/No) * LiveTogether (Yes/No)	0.1105	0.0075	1.6318	-2.2024	1.3735	-1.6034	0.1088
CONSTANT	*	*	*	-0.7417	1.2074	-0.6143	0.5390

Convergence:	Converged
Iterations:	6
Final -2*Log-Likelihood:	108.0776
Cases included:	99

In this model we see that none of the variables are significant (i.e. Wald test has a p-value of < 0.05), including the interaction term.

Try It!

Let's redo the logistic regression model but, this time, do not include the interaction term. Follow all the steps above except for creating the final interaction term (steps 5 and 6). Our results now have two individual variables, *LastPartnerSpouse* and *LiveTogether* plus the variable we are controlling for, *Agegrp*.

C:\Documents and Settings\NIS9\Desktop\STI\OUT16.htm

Previous | Next | Last | History | Open | Bookmark | Print | Maximize

LOGISTIC CaseLR = (Agegrp) LastPartnerSpouse LiveTogether

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
Agegrp (21-25/<21)	2.0329	0.1597	25.8796	0.7095	1.2980	0.5466	0.5847
Agegrp (26-30/<21)	27.3886	1.5408	486.8571	3.3101	1.4683	2.2544	0.0242
Agegrp (31-35/<21)	25.2088	1.3693	464.1069	3.2272	1.4862	2.1714	0.0299
Agegrp (36-40/<21)	11.1907	0.4088	306.3695	2.4151	1.6887	1.4302	0.1527
Agegrp (41+yrs/<21)	28.0821	1.2906	611.0202	3.3351	1.5715	2.1223	0.0338
LastPartnerSpouse (Yes/No)	0.0810	0.0165	0.3968	-2.5138	0.8109	-3.0999	0.0019
LiveTogether (Yes/No)	0.5349	0.1964	1.4568	-0.6256	0.5112	-1.2240	0.2210
CONSTANT	*	*	*	-0.7241	1.2131	-0.5969	0.5506

Convergence:	Converged
Iterations:	6
Final -2*Log-Likelihood:	110.9630
Cases included:	99

Test	Statistic	D.F.	P-Value
Score	21.2908	7	0.0034
Likelihood Ratio	25.4609	7	0.0006

Now we see that *LastPartnerSpouse* is significant, and *LiveTogether* is not significant. Therefore, as we continue to develop our model, we will keep *LastPartnerSpouse* but eliminate *LiveTogether*.

Try It!

Let's redo the logistic regression model, this time taking out *LiveTogether* and putting in the variable *SexPartner1year* (i.e., include *LastPartnerSpouse, SexPartner1year,* and *Agegrp.*

C:\Documents and Settings\NIS9\Desktop\STI\OUT16.htm

Previous | Next | Last | History | Open | Bookmark | Print | Maximize

LOGISTIC CaseLR = (Agegrp) LastPartnerSpouse SexPartner1year

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
Agegrp (21-25/<21)	1.5645	0.1999	12.2450	0.4476	1.0498	0.4263	0.6699
Agegrp (26-30/<21)	8.8926	0.8744	90.4335	2.1852	1.1834	1.8466	0.0648
Agegrp (31-35/<21)	7.2999	0.6056	87.9950	1.9879	1.2701	1.5651	0.1176
Agegrp (36-40/<21)	5.8388	0.2731	124.8171	1.7645	1.5624	1.1293	0.2588
Agegrp (41+yrs/<21)	23.6102	1.7802	313.1377	3.1617	1.3189	2.3972	0.0165
LastPartnerSpouse (Yes/No)	0.0839	0.0216	0.3266	-2.4781	0.6934	-3.5735	0.0004
SexPartner1year (Yes/No)	7.0202	2.5547	19.2913	1.9488	0.5158	3.7785	0.0002
CONSTANT	*	*	*	-0.8334	0.9734	-0.8562	0.3919

Convergence:	Converged
Iterations:	6
Final -2*Log-Likelihood:	120.8157
Cases included:	118

Test	Statistic	D.F.	P-Value
Score	36.6701	7	0.0000
Likelihood Ratio	42.7671	7	0.0000

Our results now indicate that, controlling for age group, both *LastPartnerSpouse* and *SexPartner1year* are significant in predicting being an STI case.

Try It!

Now let's consider other variables from our bivariate analysis. Add *SexPartnerLife3* into the model, so that you have *LastPartnerSpouse, SexPartner1year,* and *SexPartnerLife3* in your model, while controlling for *Agegrp*.

Create Regression Model Activity

Now let's test by adding all the additional variables from the bivariate analysis that were significant at the p = 0.25 level, as we discussed above: *AlcoholUseLR, SexPartner3month, UnemployedLR,* and *ReceiveHelp*.

Directions: Determine what your final model will be. You will need to add each of the additional variables you just noted into your model one at a time, noting those that are significant and those that are not. As soon as you find one that is not significant, leave it out of your model. Remember, you will include *Agegrp* and make it a dummy variable each time. Complete the table. For all the variables that are significant, also note whether the variable is protective or a risk factor (OR < 1 is a protective factor, OR >1 is a risk factor).

Variable	Significant?	Protective or Risk Factor?
LastPartnerSpouse	Yes	Protective
SexPartner1Year	Yes	Risk Factor
SexPartnerLife3	Yes	Risk Factor
AlcoholUseLR		
SexPartner3month		
UnemployedLR		
ReceiveHelp		

? **Question 30:** What are the variables in your final model? _____

_____

? **Question 31:** Write two or three sentences that summarize your findings for men. You can compare your response to the one in the answer key.

_____

_____

_____

_____

_____

Presenting Results
The adjusted odds ratio, 95% confidence intervals, and p-value are presented as in the table below.

Table 6: Risk factors for STI among men

Variables	Odds Ratio	95% Confidence Interval	P-value
Last sexual partner was spouse/regular partner (*LastPartnerSpouse*)	0.07	[0.02, 0.32]	< 0.001
More than one sex partner in last year (*SexPartner1year*) - dichotomous variable	4.80	[1.52, 15.01]	0.01
More than 3 sex partners in lifetime (*SexPartnerLife3*) - dichotomous variable	3.83	[1.14, 12.92]	0.03
Currently unemployed (*UnemployedLR)*	7.17	[1.02, 50.37]	0.05

Now we will look at a model for the women. Remember to first CANCEL SELECT and then Select Sex =“Female” before starting your analysis.

We can start the same way we did for the men, by beginning with *LastPartnerSpouse* and controlling for *Agegrp* using the Make Dummy button.

C:\Documents and Settings\NIS9\Desktop\STI\OUT17.htm

Previous | Next | Last | History | Open | Bookmark | Print | Maximize

LOGISTIC CaseLR = (Agegrp) LastPartnerSpouse

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
Agegrp (21-25/<21)	0.6653	0.1885	2.3484	-0.4076	0.6435	-0.6333	0.5265
Agegrp (26-30/<21)	0.7697	0.2057	2.8794	-0.2618	0.6732	-0.3889	0.6974
Agegrp (31-35/<21)	0.7698	0.1884	3.1458	-0.2616	0.7182	-0.3643	0.7157
Agegrp (36-40/<21)	0.7628	0.1080	5.3878	-0.2707	0.9974	-0.2714	0.7861
Agegrp (41+yrs/<21)	0.3132	0.0224	4.3888	-1.1609	1.3469	-0.8619	0.3888
LastPartnerSpouse (Yes/No)	3.5668	1.0114	12.5787	1.2717	0.6430	1.9776	0.0480
CONSTANT	*	*	*	-0.8039	0.7235	-1.1112	0.2665

Note that having the last sexual partner as a spouse for a woman appears to be a significant risk factor rather than significantly protective as it was for men.

Next we add *LiveTogether* and again create the interaction term between *LastPartnerSpouse* and *LiveTogether* using the Make Interaction button. When we do this we get the message

> Error
> Matrix Tolerance Exceeded

and no results appear.

Because of the high degree of correlation between variables and a number of cells with no cases we are not able to include any interaction terms in this model.

This time, include *LiveTogether* but do not create the interaction term.

C:\Documents and Settings\NIS9\Desktop\STI\OUT17.htm

Previous Next Last History Open Bookmark Print Maximize

LOGISTIC CaseLR = (Agegrp) LastPartnerSpouse LiveTogether

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
Agegrp (21-25/<21)	1.0736	0.2636	4.3730	0.0711	0.7165	0.0992	0.9210
Agegrp (26-30/<21)	1.6183	0.3647	7.1800	0.4814	0.7602	0.6332	0.5266
Agegrp (31-35/<21)	1.9435	0.3610	10.4629	0.6645	0.8589	0.7737	0.4391
Agegrp (36-40/<21)	1.1949	0.1478	9.6599	0.1781	1.0663	0.1670	0.8674
Agegrp (41+yrs/<21)	0.5471	0.0268	11.1695	-0.6031	1.5390	-0.3919	0.6951
LastPartnerSpouse (Yes/No)	8.6596	1.5276	49.0880	2.1587	0.8852	2.4386	0.0147
LiveTogether (Yes/No)	0.2694	0.1026	0.7070	-1.3117	0.4923	-2.6643	0.0077
CONSTANT	*	*	*	-1.5556	0.9430	-1.6496	0.0990

Convergence:	Converged
Iterations:	5
Final -2*Log-Likelihood:	114.8416
Cases included:	92

Test	Statistic	D.F.	P-Value
Score	12.0202	7	0.0999
Likelihood Ratio	12.5236	7	0.0846

We find both variables remain significant while controlling for age group, with *LastPartnerSpouse* being a risk factor for women and *LiveTogether* with regular partner being protective.

Logistic Regression Activity

Directions: Now proceed to add in the other variables from the bivariate analysis and determine your final model. Also use the *ForcedSex* variable which was asked of women.

Variable	Significant?	Protective or Risk Factor?
LastPartnerSpouse	Yes	Risk Factor
LiveTogether	Yes	Protective
SexPartner1Year		
SexPartnerLife3		
AlcoholUseLR		
SexPartner3month		
UnemployedLR		
ReceiveHelp		
ForcedSex		

Final model

Logistic Regression Model (for Females)	STI = LastPartnerSpouse + LiveTogether + SexPartner1year + ForcedSex

Logistic Regression Activity 2

Directions: Complete the analysis and present your results in the table below. See the results in Appendix C, Table 7.

Table 7: Risk Factors for STI Among Women

Variables	Odds Ratio	95% Confidence Interval	P-value
Last partner was spouse (*LastPartnerSpouse*) - dichotomous variable			
Lives together with spouse or partner (*LiveTogether*) – dichotomous variable			
More than one sex partner in last year (*SexPartner1year*) - dichotomous variable			
Forced to have sex in the past year (*ForcedSex*) – dichotomous variable			

Step 9: Interpretation and Reporting

The final step is interpretation and reporting. We have been interpreting our data analysis throughout the training. The interpretation of data is used to draw conclusions and make evidence-based recommendations. In essence, we are answering our research question. As you summarize the outcome of your logistic regression analysis from the previous page below, think of some recommendations you might make based upon your findings.

? **Question 32:** Write two or three sentences to summarize your findings for women.

_____

_____

_____

_____

Your next step would be to report your data. You would need to decide what and to whom you would report. Sometimes your job requires that you report specific data analysis and interpretation to your superiors, but there are other purposes and other interested parties to whom you might report. For example, the results of this study were presented both to decision makers at the city health department and to peers at an international conference.

For what other purposes would you report findings from a study such as this one?

_____

To whom might you report the results of a study you conducted? _____

_____

The reporting of data is one way that scientists share information with each other. It is how we support any recommendations for policy change to decision makers. It is how we decide what interventions will be the most effective. It is this final step that allows our research to have a purpose.

Congratulations!

You have completed the Advanced Management and Analysis of Data training.

References

Bauer HM, Gibson P, et al Intimate partner violence and high-risk sexual behaviours among female patients with sexually transmitted diseases. Sex Trans Dis. 2002 Jul; 29(7):411-6

Breslow NE, Days, NE. Statistical methods in cancer research, Vol. 1: The analysis of case-control studies. Lyon, France: IARC Scientific Publication No. 32; 1981.

Centers for Disease Control and Prevention. Principles of Epidemiology. 2nd ed; 1992.

Da Costa LJ, Plummer FA, Boumer I et al, Prostitutes are a major source of sexually transmitted diseases in Nairobi, Kenya. Sex Trans Dis 1985; 12:64-67

Dicker RC. Analyzing and interpreting data. In: Gregg MB, editor. Field Epidemiology. 2nd ed. New York: Oxford University Press; 2002. p. 132-172.

Hennekens CH, Buring JE. Epidemiology in Medicine. Philadelphia: Lippincott William and Wilkins;1987.

Hosmer DW, Lemeshow S. Applied logistic regression (Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics Section). New York: John Wiley; 1989.

Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB. Designing Clinical Research: An Epidemiologic Approach, Second Edition. Piladelphia: Lippincott William & Wilkins; 2001.

Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: Principles and quantitative methods. New York: Van Nostrand Reinhold; 1982.

Kleinbaum DG, Kupper LL, Muller KA. Applied regression analysis and other multivariate methods. Second edition. Boston: Duxbury Press; 1987.

Kleinbaum DG. Statistics in the health sciences: Logistic regression. New York: Springer-Verlag; 1994.

Mack TM, Pike MC, Henderson BE, Pfeffer RI, Gerkins VR, Arthur M, Brown SE. Estrogens and endometrial cancer in a retirement community. N Engl J Med 1976 Jun 3;294(23):1262-7.

Ministry of Health & Child Welfare [Zimbabwe] (2000); Zimbabwe National HIV/AIDS Estimates, 2000

Paz-Bailey G, Kilmarx PH, et al. Risk factors for sexually transmitted diseases in northern Thai adolescents: an audio-computer-assisted self-interview with noninvasive specimen collection. Sex Trans Dis. 2003 Apr;30(4):320-6.

Wellington M. Ndowa F. Risk factors for sexually transmitted diseases in Harare: a case-control study. Sexually Transmitted Diseases. 24(9):528-32, 1997 Oct.

Appendices

Appendix A: Questionnaire
Appendix B: Answer Sheet
Appendix C: Data Tables
Appendix D : Statistical Formulas

Appendix A: Questionnaire

QUESTIONNAIRE: PART 1

I am a public health officer from the city health department. I am carrying out a study to determine the risk factors for contracting STIs in Kuwa. The study findings will be used to make recommendations for interventions. No names will be taken. All data will be handled with strict confidence.

HEALTH SEEKING BEHAVIOUR
SECTION A

Personal information:

Identification number _ _ Sex 1) Male 2) Female

Case Status: 1 Case 2 Control Type of STI_____

A1 How old are you? __ __

A2 What is your occupation? 1 Unemployed 2 Informal employment
3 Formal employment 4 Other (specify)_____

A3 Which church do you go to?
1 Catholic 2 Apostolic 3 Methodist 4 Anglican
5 Pentecostal 6 Atheist 7 Roman Catholic 8 Other (specify)_____

A4 What is the highest level of education you attained?
1 None 2 Primary 3 Secondary 4 Tertiary

A5 What is your marital status?
1 Single 2 Married 3 Co-habiting 4 Divorcee
5 Widowed 6 Other(specify)_____

C2 When were you last ill? 1 never
2 <1 year ago
3 1-3 years ago
4 >3 years ago
5 Can't remember
6 I don't know

C3 Was this last illness an STI? Yes No

Social Capital

I would like to ask you about the groups, or organizations to which you or any member of your household belong.

D1 Does anyone in your household belong to any of these groups. Tick the appropriate.

	Yes	No
1 Traders or business association	1	2
2 Religious groups	1	2
3 Burial society	1	2
4 Savings club	1	2

D2 Which two groups are the most important in your household?

1 Group 1_____

2 Group 2_____

D3 Do the groups help your household to access any of the following services?

	1-Yes	2-No
1 Education		
2 Health services		
3 Receive credit		
4 Funeral assistance		

For cases only

E8 Why do you think you have this illness?

_____

_____

RISKY SEXUAL BEHAVIOR
(For both cases and controls)

2. How old were you when you first had sex?

__ __ Years

3. Have you had a sexually transmitted infection in the past 3 months?

1 Yes 2 No

What did you do?

	Mentioned	Prompted	
		Y	N
1 Buy the prescribed antibiotics			
2 Complete course of			

5. Does your current sexual partner have an STI?

1 Yes 2 No 3 Don't know

6. How many sexual partners have you had in the past 3 months?
   Number of partners_____

7. How many sexual partners have you had in the past 1 year?
   Number of partners____

8. How many sexual partners have you had in your lifetime?
   1 Less than three (3) 2 Three (3) or more

9. What was your relationship to the last person you had sex with?
   1 Husband/Wife/live in partner
   2 Fiancee/lover
   3 Friend
   4 Occasional partner
   5 Just met/stranger
   6 Prostitute
   7 Family member/relative
   8 Other (specify _____
   9 Refused

10. Did you give or receive money or goods in exchange for sex?
    1 Yes 2 No
    3 Don't remember 4 No response

11. Did you or your partner use a condom the last time you had sex ?
    1 Yes 2 No 3 Don't know

12. Do you use a condom every time you have sex?
    1 Yes 2 No 3 Don't remember

13. Had you or your partner taken alcohol the last time you had sex?
    1 Yes 2 No 3 Don't know

14. Do you have a regular sexual partner? (Regular partner: had sex with more than 3 times in the past year)
    1 Yes 2 No [*If no, skip to question 16*]

15. Have you been living together in the past 6 months? 1 Yes 2 No

16. How old is your regular sexual partner? _____(years)

17. Do you suspect that your regular partner is having a sexual relationship with somebody else?
    1 Yes 2 No 3 Don't know

[*18 for women only*]

18. Sometimes women are forced to have sex and end up getting STIs. Were you forced to have sex in the past 1 year?
    1 Yes 2 No 3 I don't remember 4 Refused to answer 5- N/A

19. Which is your usual area of residence? (suburb)
    _____

19b. How long have you been living in this place for?
    _____(years) _____(months)

[*19c,d for those who live outside the city only*]

19c. When did you come to city? (year)__ __

19d. What was the purpose of coming to city?

Appendix B: Answers to Training Questions

Question 1: No, the number of records does not match. There are 227 records in the current MDB file, but there should only be 226 records.

Question 2: 21. This means that you will need to eliminate the record with age 23 listed.

Question 3: 48; 213

Question 4: 48 is male; 213 is female

Question 5: The responses should only be 1 (for case) or 2 (for control). A response of 3 is incorrect.

Question 6: 31; 1

Question 7: N7Nosexpart1yr; questionnaire139; answer should be 4 not 444.

Question 8: 11, 17, 226

Question 9: Yes, 121

Question 10: Yes; 1, 52

Question 11: 31 non-cohabiting cases; 37 non-cohabiting controls

Question 12: 158

Question 13: 155

Question 14: You could either use RECODE or an IF statement to create the new variable *SexPartnerLife3*.

```
DEFINE SexPartnerLife3
IF N8Nosexpartlifetime="2 3 or
more" THEN
      ASSIGN SexPartnerLife3= (+)
ELSE
      ASSIGN SexPartnerLife3= (-)
END
```

```
DEFINE SexPartnerLife3
RECODE N8Nosexpartlifetime TO
SexPartnerLife3
      "1 Less than 3" = (-)
      "2 3 or more" = (+)
END
```

Question 15: The odds of having more than one sexual partner in the last 3 months are 7.4 times higher among STI cases than among controls. This association is statistically significant at the p< 0.01 level.

Question 16: *HabitationStatus*, *Education*, *AgeFirstSex*, *N10givereceiveforse*x. Note that *Unemployed* and *LastPartnerSpous*e could be considered borderline significant, but including them here would be considered correct (see question 18).

Question 17: *SexPartner3month*, *SexPartner1year*, *SexPartnerLife3*, *N15LivingTogether*, *AlcoholUse*, *ReceiveHelp*, *N18WereYou* (for women)

Question 18: *Unemployed*, *LastPartnerSpous*e

Question 19:

Risk factor – *SexPartner3month*, *SexPartner1year*, *SexPartnerLife3*, *AlcoholUse*, *N18wereyou*, *Unemployed*
Protective factor – *N15LivingTogether*, *ReceiveHelp*, *LastPartnerSpouse*

Question 20: 0.0556 Report: 0.06

Question 21: 2.0490 Report: 2.05

Question 22: 0.6974 Report: 0.70

Question 23: 1.2391 Report: 1.24

Question 24: 0.7997 Report: 0.80

Question 25: No

Question 26: 0.5785; 0.8763 Report: 0.58; 0.88

Question 27: See table below.

	N14DoYouHave	***N15LivingTogether***	***LastPartnerSpouse***
What is the OR for females?	3.86 Report: 3.9	0.3913 Report: 0.39	3.1977 Report: 3.2
What is the OR for males?	0.2338 Report: 0.23	0.4048 Report: 0.40	0.2407 Report: 0.24
What are the Chi-square and p-value for the *Chi-square for differing Odds Ratios by stratum (interaction)*?	Chi-square 9.3949 P-Value 0.0022 Report: x^2=9.4, p=0.002	Chi-square 0.0001 P-Value 0.9548 Report: x^2=0.0001, p=0.95	Chi-square 12.4079 P-Value 0.0004 Report: x^2=12.4, p=0.0004
Does it appear that there is effect modification?	Yes	No	Yes

Question 28: The variables are: *Unemployed*, *AgeFirstSex*, *SexPartner3month*, S*exPartner1year*, *SexPartnerLife3*, *N15LivingTogether*, *LastPartnerSpouse*, *AlcoholUse*, *N10givereceiveforsex*, *ReceiveHelp*, and for women only *N18WereYou* (forced to have sex).

Question 29: All three variables (*LastPartnerSpouse, SexPartner1year,* and *SexPartnerLife3)* were significant, controlling for age group.

Question 30: *LastPartnerSpouse*, *SexPartner1year, SexPartnerLife3,* and *UnemployedLR* with *Agegrp* as the controlled variable.

Question 31: Having one's last sexual partner be one's spouse/live-in partner was significantly protective for males. Having more than 2 sex partners in the past year, more than 2 partners in one's lifetime, and being unemployed were significant risk factors for males.

Question 32: Among women, after controlling for age group, the number of sexual partners in the past year and being forced to have sex remained significant risk factors for STI. Living together with one's spouse or regular partner was significantly protective against STI. Having last sexual relations with one's spouse or regular partner was an independent risk factor for STI, however we know from our stratified analysis that this only reflects those women who do not live with their partner as there were no women in the dataset who lived with their partner/spouse who had last sexual relations with someone other than spouse/partner – thus the relationship could not be examined for these women.

Appendix C: Data Tables

Table 1 : Demographic data

Variable	Case - # (%)	Control - # (%)
A2Occupation		
1 Unemployed	45 (39.8)	31 (27.4)
2 Informal	16 (14.2)	29 (25.7)
3 Formal	45 (39.8)	46 (40.7)
4 Student	7 (6.2)	7 (6.2)
A3Church		
2 Apostolic	26 (23.0)	26 (23.0)
3 Methodist	9 (8.0)	4 (3.5)
4 Anglican	7 (6.2)	9 (8.0)
5 Pentecostal	18 (15.9)	18 (15.9)
6 Atheist	25 (22.1)	20 (17.7)
7 Roman Catholic	22 (19.5)	19 (16.8)
8 Other	6 (5.3)	17 (15.0)
A4 LevelofEducation		
1 None	1 (0.9)	0 (0.0)
2 Primary	14 (12.4)	12 (10.6)
3 Secondary	91 (80.5)	97 (85.8)
4 Tertiary	7 (6.2)	4 (3.5)
A5MaritalStatus		
1 Single	25 (22.1)	18 (15.9)
2 Married	74 (65.5)	75 (66.4)
3 Cohabiting	8 (7.1)	1 (0.9)
4 Divorcee	4 (3.5)	12 (10.6)
5 Widowed	2 (1.8)	7 (6.2)

Table 2: Continuous Variable

Variable	**Case – Mean Median Mode**	**Control – Mean Median Mode**
A1Age	27.8 26.0 23.0	28.3 27.0 24.0
N2SexDebut	18.9 19.0 20.0	19.4 20.0 20.0
N7Nosexpart1year	1.7 1.0 1.0	1.2 1.0 1.0

Table 3: Continuous Variables (Regrouped)

Variable	**Case # (%)**	**Control # (%)**
Agegrp (previously *A1Age*)		
< 21	10 (8.8)	11 (9.7)
21-25	38 (33.6)	39 (34.5)
26-30	33 (29.2)	28 (24.8)
31-35	20 (17.7)	19 (16.8)
36-40	5 (4.4)	8 (7.1)
41+yrs	7 (6.2)	8 (7.1)
AgeFirstSex (previously *N2SexDebut*)		
1) <17 years	25 (22.9)	16 (14.4)
2) 17+ years	84 (77.1)	95 (85.6)

Table 4: Risk Factor Table

Variable	Cases N (Col %)	Controls N (Col %)	OR (95% CI)	P value
Marital status (*HabitationStatus*) Married/Cohabiting (1) Single/Div/Widowed (2)	82 (73) 31	76 (67) 37	1.3 (0.7-2.3)	0.5
Level of education (*Education*) None/Primary (1) Secondary/Tertiary (2)	15 (13) 98	12 (11) 101	1.3 (0.6-2.9)	0.7
Employment (*Unemployed*) Unemployed (1) Employed (2)	45 (43) 61	31 (29) 75	1.8 (1.0-3.2)	0.06
Age at First Sex (*AgeFirstSex*) < 17 years 17+ years	25 (23) 84	16 (14) 95	1.76 (0.9-3.5)	0.15
> 2partner last 3 months (*SexPartner3month*) Yes No	19 (17) 94	3 (3) 110	**7.4** (2.1-25.8)	<0.001
> 2 partner last year (*SexPartner1year*) Yes No	41 (36) 72	11 (10) 102	**5.3** (2.5-11.0)	<0.0001
3 or more partners in lifetime (*SexPartnerLife3*) Yes No	61 (54) 52	31 (27) 82	**3.1** (1.8-5.4)	<0.0001
Living with partner (*N15LivingTogether*) Yes No	32 (34) 61	55 (56) 43	**0.4** (0.2–0.7)	0.002
Relationship last partner (*LastPartnerSpouse*) Yes No	72 (64) 41	85 (75) 28	0.6 (0.3-1.0)	0.08
Alcohol use last partner (*AlcoholUse*) Yes No	41 (37) 70	26 (23) 87	**1.9** (1.1-3.5)	0.03
Paid/rcvd money for sex (N10givereceiveforsex) Yes No	8 (7) 105	3 (3) 110	2.8 (0.7-10.8)	0.2
Received help from at least one group (*ReceiveHelp*) Yes No	70 (62) 43	85 (75) 28	**0.5** (0.3-0.95)	0.04
Forced to have sex (women only) (*N18WereYou*) * Yes No	16 (30) 38	5 (9) 49	**4.1** (1.4-12.3)	0.015

Table 5: Stratified analysis for men to check for confounding by age group

Variable	Crude OR	95% CI	MH Adjusted OR	95% CI	Comment
LastPartnerSpouse	**0.24**	0.11-0.52	**0.079**	0.02-0.29	Possible confounding
LivingTogether	**0.42**	0.19-0.94	**0.36**	0.15-0.86	similar
SexPartner1year	**8.07**	3.3-20.0	**6.5**	2.7-15.3	Possible confounding
Education	**1.0**	0.30-3.4	**1.0**	0.24-4.43	similar
AlcoholUse	**2.6**	1.1-6.3	**2.8**	1.1-6.9	similar

Tables 7: Risk Factors for STI Among Women

Variables	**Odds Ratio**	**95% Confidence Interval**	**P-value**
Last partner was spouse (*LastPartnerSpouse*) - dichotomous variable	17.4	[1.8-169.5]	0.01
Lives together with spouse or partner (*LiveTogether*) – dichotomous variable	0.31	[0.11-0.87]	0.026
More than one sex partner in last year (*SexPartner1year*) - dichotomous variable	20.4	[1.1-363.1]	0.04
Forced to have sex in the past year (*ForcedSex*) – dichotomous variable	4.1	[1.18-14.2]	0.026

Appendix D: Statistical Formulas

For each of the statistical tests in this appendix, note the notation in the following 2x2 table:

	Ill	**Well**	**Total**
Exposed	a	b	h_1
Unexposed	c	d	h_0
Total	v_1	v_0	t

Fisher Exact Test

The probability that the value in cell "a" is equal to the observed value, under the null hypothesis, is

$$\Pr(a) \quad \frac{(v_1)!(v_0)!(h_1)!(h_0)!}{t!a!b!c!d!}$$

where k! ("k factorial") = 1 * 2x…*k (Dicker, p. 170)

Pearson Uncorrected Chi-Square

$$\text{Pearson uncorrected } X^2 = \frac{t(ad-bc)^2}{(v_1)(v_0)(h_1)(h_0)}$$

Mantel-Haenszel Chi-Square

$$\text{Mantel-Haenszel } X^2 = \frac{t\left(|ad-bc| - \left(\frac{t}{2}\right)\right)^2}{(v_1)(v_0)(h_1)(h_0)}$$

Yates Corrected Chi-Square

$$\text{Yates Corrected } X^2 = \frac{(t-1)(ad-bc)^2}{(v_1)(v_0)(h_1)(h_0)}$$

Division of Global Public Health Capacity Development
Centers for Disease Control and Prevention
1600 Clifton Road, N.E.
Mailstop E-93
Atlanta, Georgia 30333, USA

Visit us at www.cdc.gov/cogh/dgphcd