The Young Epidemiology Scholars Program (YES) is supported by The Robert Wood Johnson Foundation and administered by the College Board.

Cross-Sectional Study Design and Data Analysis

Chris Olsen
Mathematics Department
George Washington High School
Cedar Rapids, Iowa

and

Diane Marie M. St. George
Master's Programs in Public Health
Walden University
Chicago, Illinois

Contents

Lesson Plan

TITLE: Cross-Sectional Study Design and Data Analysis

SUBJECT AREA: Statistics, mathematics, biology

OBJECTIVES: At the end of this module, students will be able to:

- Explain the cross-sectional study design
- Understand the process of questionnaire construction
- Identify several sampling strategies
- Analyze and interpret data using *Epi Info* statistical software

TIME FRAME: Two class periods and out-of-class group time

PREREQUISITE KNOWLEDGE: Advanced biology; second-year algebra level of mathematical maturity.

MATERIALS NEEDED:

- *Epi Info* software (freeware downloadable from the Internet).
- High-speed Internet connection is useful.
- Youth Risk Behavior Survey (YRBS) sample datasets (student and teacher versions accompanying this module).
- Abbreviated YRBS Codebook (included as an appendix to the module).

Please note that teachers are **not** required or expected to download the entire YRBS dataset or the YRBS Codebook. Those files have already been downloaded and formatted for use with the module, and we would recommend that teachers make use of them. However, if teachers should choose to download the YRBS dataset from the Web site, please be advised that the dataset will not be in *Epi Info* format and will require manipulation in order to be used with the *Epi Info* software.

PROCEDURE: Teachers should ask the students to read Sections I–V at home, and then in class the teacher should review the major concepts contained therein. The teacher should cover Section VI during the class period, using the worked example as a guide as needed. The groups should then assemble and begin to work together in class on the group project. This allows them to have teacher input while designing their research questions and beginning to learn the software. They should then complete the group projects as homework.

ASSESSMENT: At end of module. There are four options provided, one of which includes suggested answers.

LINK TO STANDARDS:

This module addresses the following mathematics standards:

The Standard	The Grades 9–12 Expectations
Data Analysis and Probability	
Instructional programs from prekindergarten through grade 12 should enable all students to:	
• Formulate questions that can be addressed with data and collect, organize and display relevant data to answer them.	• Understand the differences among various kinds of studies and which types of inferences can legitimately be drawn from each; know the characteristics of well-designed studies, including the role of randomization in surveys and experiments; understand the meaning of measurement data and categorical data, of univariate and bivariate data, and of the term variable; understand histograms, parallel box plots, and scatter plots and use them to display data; compute basic statistics and understand the distinction between a statistic and a parameter.
• Select and use appropriate statistical methods to analyze data.	• For univariate measurement data, be able to display the distribution, describe its shape, and select and calculate summary statistics; for bivariate measurement data, be able to display a scatter plot, describe its shape, and determine regression coefficients, regression equations, and correlation coefficients using technological tools; display and discuss bivariate data where at least one variable is categorical; recognize how linear transformations of univariate data affect shape, center and spread; identify trends in bivariate data and find functions that model the data or transform the data so that they can be modeled.
• Develop and evaluate inferences and predictions that are based on data.	• Use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions; understand how sample statistics reflect the values of population parameters and

	use sampling distributions as the basis for informal inference; evaluate published reports that are based on data by examining the design of the study, the appropriateness of the data analysis, and the validity of conclusions; understand how basic statistical techniques are used to monitor process characteristics in the workplace.
• Understand and apply basic concepts of probability.	• Understand the concepts of sample space and probability distribution and construct sample spaces and distributions in simple cases; use simulations to construct empirical probability distributions; compute and interpret the expected value of random variables in simple cases; understand the concepts of conditional probability and independent events; understand how to compute the probability of a compound event.

Problem Solving

Instructional programs from prekindergarten through grade 12 should enable all students to:

- Build new mathematical knowledge through problem solving
- Solve problems that arise in mathematics and in other contexts
- Apply and adapt a variety of appropriate strategies to solve problems
- Monitor and reflect on the process of mathematical problem solving

Communication

Instructional programs from prekindergarten through grade 12 should enable all students to:

- Organize and consolidate their mathematical thinking through communication
- Communicate their mathematical thinking coherently and clearly to peers, teachers, and others
- Analyze and evaluate the mathematical thinking and strategies of others
- Use the language of mathematics to express mathematical ideas precisely

Connections

Instructional programs from prekindergarten through grade 12 should enable all students to:

- Recognize and use connections among mathematical ideas

- Understand how mathematical ideas interconnect and build on one another to produce a coherent whole
- Recognize and apply mathematics in contexts outside of mathematics

Representation

Instructional programs from prekindergarten through grade 12 should enable all students to:

- Create and use representations to organize, record, and communicate mathematical ideas
- Select, apply and translate among mathematical representations to solve problems
- Use representations to model and interpret physical, social, and mathematical phenomena

This module also addresses the following science standards:

Science As Inquiry

- Abilities necessary to do scientific inquiry

Unifying Concepts and Processes

- Evidence, models and explanation

Bibliography

Aday L. *Designing & Conducting Health Surveys*. 2nd ed. San Francisco: Jossey-Bass Publishers; 1996.

Biemer, P. P., & Lyberg, L. E. *Introduction to Survey Quality*. Hoboken, NJ: John Wiley & Sons; 2003.

Centers for Disease Control and Prevention. 2001 Youth Risk Behavior Survey Results, United States High School Survey Codebook. Available at: www.cdc.gov/nccdphp/dash/yrbs/data/2001/index.html

Converse J, Presser S. *Survey Questions: Handcrafting the Standardized Questionnaire*. Thousand Oaks, CA: Sage Publications; 1986.

Fowler F. *Improving Survey Questions: Design and Evaluation*. Thousand Oaks, CA: Sage Publications; 1995.

Schuman H, Presser S. *Questions & Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Thousand Oaks, CA: Sage Publications; 1996.

Sudman S, Bradburn N. *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco: Jossey-Bass Publishers; 1982.

Sudman S, Bradburn N, Schwarz N. *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass Publishers; 1996.

Tourangeau R, Rips L, Rasinski K. *The Psychology of Survey Response*. New York: Cambridge University Press; 2000.

Section I: Introduction to the Cross-Sectional Study

Epidemiologists are public health researchers. Some of the most popular examples of epidemiology in action are related to research surrounding the causes of infectious disease outbreaks and epidemics. When we first began to hear about SARS (severe acute respiratory syndrome) in late 2002, the unsung heroes were those epidemiologists attempting to determine what caused the outbreak. Similarly, about 20 years ago when AIDS (acquired immunodeficiency syndrome) was first identified, albeit not by this name, epidemiologists were busy at work collaborating with basic scientists to attempt to determine what was causing the disease.

However, epidemiologists are also behind the scenes, acting as medical and health detectives and conducting research to determine causes of chronic diseases as well. Through epidemiologic studies, we learned that smoking causes lung cancer, that high-fat diets contribute to the development of heart disease and that fluoridation of water can reduce the occurrence of dental caries.

The tools or research study designs used by epidemiologists are varied. However, there is a thought process or reasoning they use that is consistent throughout: If a factor X causes a disease Y, then there will be proportionately more diseased people among the group with X than among the group that does not have X. Think about it this way: If it were true that shaving caused one's hair to grow back thicker, would you expect to find thicker hair among your classmates who shaved or among your classmates who did not shave? Among the shavers, right? In epidemiologic lingo, we would say that such a finding would mean that shaving is associated with hair thickness or that shaving is related to hair thickness.

The study designs all use the same basic reasoning, but they do it in different ways. Some designs gather information about X and then follow people over time to see who develops Y. Some designs gather information from people with Y and without Y and then see who was exposed to X in the past. And the examples could go on.

One of the most common and well-known study designs is the **cross-sectional study** design. In this type of research study, either the entire population or a subset thereof is selected, and from these individuals, data are collected to help answer research questions of interest. It is called cross-sectional because the information about X and Y that is gathered represents what is going on at only one point in time. For instance, in a simple cross-sectional study an epidemiologist might be attempting to determine whether there is a relationship between television watching and students' grades because she believed that students who watched lots of television did not have time to do homework and did poorly in school. So the epidemiologist typed up a few questions about number of hours spent watching television and course grades, and then mailed out the sheet with questions to all of the children in her son's school.

What she did was a cross-sectional study, and the document she mailed out was a simple questionnaire. In reading public health research, you may encounter many terms that appear to be used interchangeably: cross-sectional study, survey, questionnaire, survey questionnaire, survey tool, survey instrument, cross-sectional survey. Although many of those terms are indeed used interchangeably, they are not all synonymous. This module will use the term cross-sectional study to refer to this particular research design and the term **questionnaire** to refer to the data collection form that is used to ask questions of research participants. Data can be collected using instruments other than questionnaires, such as pedometers, which measure distances walked, or scales, which measure weight. However, most cross-sectional studies collect at least some data using questionnaires.

Section II: Overview of Questionnaire Design

A questionnaire is a way of collecting information by engaging in a special kind of conversation. This conversation, which could actually take place face to face, by telephone or even via the mail, has certain rules that separate the questionnaire from usual conversations. The researcher decides what is relevant to his or her study and may ask questions, possibly personal or even embarrassing questions. These questions should be both understandable and relevant to the purpose of the research. The respondent in turn may refuse to participate in the conversation and may refuse to answer any particular question. But having agreed to participate in the study, the respondent has the responsibility to answer questions truthfully.

Section III: Question Construction

We would now like to discuss some issues related to the design of questions. In many health studies researchers attempt to measure knowledge, attitudes and behaviors relating to risk factors and health events in the lives of individuals. In such studies both the sampling method and the design of the questionnaire itself are critical to obtaining reliable information. The design of the questionnaire refers to the directions or instructions, the appearance and format of the questionnaire and, of course, the actual questions.

Questionnaires have been around for a very long time, and they are likely to remain fixtures in our everyday lives for a very long time. Questions may be designed for different purposes. Some questions attempt to measure attitudes:

Do you feel your local hospital services are sufficient for your city?

To what extent do you favor federal funding of care for elderly citizens?

Other types of questions are designed to elicit facts, such as:

How many times have you visited your physician during the past 24 months?

In what month and year did you last have a mammogram?

Epidemiologists gather information by asking questions of individuals and evaluating their responses. It might seem at first glance that creating a questionnaire would be very easy to do. The epidemiologist is interested in some attitude, belief or fact. He or she writes a few relevant questions and administers the questionnaire to a random sample of people. Their responses are recorded, and the data are analyzed. However, it turns out that writing and administering a questionnaire are not easy at all. Designing questions, interpreting answers and finally analyzing the data must be done very carefully if one is to extract good information from a questionnaire.

Both the respondent and the researcher must give some thought to the questionnaire process, but the respondent has a more difficult role. Let's consider the situation of the respondent.

The Respondent's Tasks

The respondent is confronted with a sequence of tasks when asked a question. These tasks are comprehension of the question, retrieval of information from memory and reporting the response.

Task I: Comprehension

The first task of the respondent is to understand the directions and then each question as it is asked. Comprehension is the single most important task facing the respondent, and fortunately it is the characteristic of a question that is most easily controlled by the interviewer. Comprehensible questions are characterized by:

1. A vocabulary appropriate for the target population
2. Simple sentence structure
3. Little or no ambiguity and vagueness

Vocabulary is often a problem. The researcher usually knows a great deal about the topic of the questionnaire, and it may be difficult to remember that others do not have that special knowledge. In addition, researchers tend to be very well educated and may have a more extensive vocabulary than people responding to the questionnaire. As a rule, it is best to use the simplest possible word that can be used without sacrificing clear meaning. A dictionary and thesaurus are invaluable in the search for simplicity.

Simple sentence structure also makes it easier for the respondent to understand the questions. A very famous example of difficult syntax occurred in 1993 when the Roper Organization created a questionnaire related to the Holocaust, the Nazi extermination of Jews during World War II. One question in this questionnaire was:

> Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?

The question has a complicated structure and a double negative—"impossible" and "never happened"—that could lead respondents to give an answer opposite to what they actually believed. The question was rewritten and given a year later in an otherwise unchanged questionnaire. The reworded question was:

> Does it seem possible to you that the Nazi extermination of the Jews never happened, or do you feel certain that it happened?

This question wording is much clearer.

Keeping vocabulary and sentence structure simple is relatively easy compared with stamping out ambiguity in questions. In part, this is because precise and unambiguous language may be difficult to comprehend, as evidenced by definitions we see in mathematics books; they are precise but sometimes difficult to comprehend. Even the most innocent and seemingly clear questions can have a number of possible interpretations. For example, suppose you are asked, "When did you move to Chicago?" This would seem to be an unambiguous question, but some possible answers might be:

1. In 1992
2. When I was 23
3. In the summer

The respondent must decide which of these, if any, is the appropriate response. It may be possible to lessen the ambiguity with more precise questions:

1. In what year did you move to Chicago?
2. How old were you when you moved to Chicago?
3. In what season of the year did you move to Chicago?

One way to find out if a question is ambiguous is to field test the question and ask the respondents if they were unsure how to answer a question.

The table below presents ambiguities identified in the process of debriefing respondents.

Question	**Ambiguity**
1. Do you think children suffer any ill effects from watching programs with violence in them?	1. The word children was interpreted to mean everyone from babies to teenagers to young adults in their early twenties.
2. What is the number of servings of eggs you eat in a typical day?	2. It was unclear to the respondents what a serving of eggs was, as well as what the term typical day meant.
3. What is the average number of days each week you consume butter?	3. Respondents were unclear about whether margarine should count as butter.

Ambiguity is not only a characteristic of individual questions in a questionnaire. It is also possible for a question to be ambiguous because of its placement in the questionnaire. Here is an example of ambiguity uncovered when the order of two questions differed in two versions of a questionnaire on happiness. The questions were:

(i) [Considering everything], how would you say things are these days: would you say that you are very happy, pretty happy, or not too happy?

(ii) [Considering everything], how would you describe your marriage: would you say that your marriage is very happy, pretty happy, or not too happy?

The proportions of responses to the general happiness question differed for the different question orders, as follows:

General Happiness

	General-Marital	Marital-General
Very Happy	52.4%	38.1%
Pretty Happy	44.2%	52.8%
Not Too Happy	3.4%	9.1%

If the goal in this questionnaire was to see what proportion in the population is "generally happy," these numbers are quite troubling—they cannot both be right. What seems to have happened is that question (i) was interpreted differently depending on whether it was asked first or second. When the general happiness question was asked after the marital happiness question, the respondents apparently interpreted it to be asking about their happiness in all aspects of their lives *except* their marriage. This was a reasonable interpretation because they had just been asked about their marital happiness—but a different interpretation from when the general happiness question was asked first. The lesson here is that even very carefully worded questions can have different interpretations in the context of the rest of the questionnaire.

Task II: Retrieval from Memory

Once a question is understood, the respondent must retrieve relevant information from memory in order to answer the question. This is not always an easy task and not a problem limited to questions of fact.

Psychologists do not agree completely on how memory works, but most believe that memory is made up of stored representations of events in the lives of individuals. Some memories are particularly clear, such as those of wedding events, where one was at the time of a presidential assassination, or a tragedy such as the *Space Shuttle*'s exploding. Other events—the more daily typical memories—seem to be stored generically. For example, it is unlikely that one remembers every trip to the drug store. Instead one has a general idea of a typical trip stored in memory. Thus unless a question is about a particularly salient event, the respondent will probably reconstruct events by piecing together memories of typical events that are suggested by the question. For instance, consider this seemingly elementary factual question:

> How many times in the past five years did you visit your dentist's office?
>
> (a) No times
>
> (b) Between 1 and 5 times

(c) Between 6 and 10 times

(d) Between 11 and 15 times

(e) More than 15 times

It is very unlikely that many people will remember every single visit to the dentist. Generally people will respond to such a question with answers consistent with the memories and facts they are able to reconstruct given the time they have to respond to the question. For example, they may have a sense that there are usually about two trips a year to the dentist's office.

There is no option presented as, "I think usually about two trips a year," so the respondent may extrapolate the typical year and get 10 times in five years. Then there may be a memory of a root canal in the middle of last winter. Thus, the best recollection is now 13, and the respondent will answer (d), between 11 and 15—perhaps not exactly correct, but the best that can be reported under the circumstances.

What are the implications of this relatively fuzzy memory for those who would construct questionnaires about facts? First, the investigator should understand that most factual answers are going to be approximations of the truth. Second, events closer to the time of a questionnaire will be easier to recall. A question about visits to the dentist in the past year will probably be answered more accurately than a question about visits in the past five years. Third, memories of events will be cued by the questions that are asked in a questionnaire. The more carefully events of interest can be described in the question, the better the chance that the question will cue the right memories. Particularly emotional, important and distinctive events will be more easily recalled.

Task III: Reporting the Response

The third task of the respondent to a questionnaire is to actually formulate and report a response. In general if an individual agrees to respond to a questionnaire, he or she will be motivated to answer truthfully. Therefore, if the questions aren't too difficult (taxing the respondent's knowledge or memory) and there aren't too many of them (taxing the respondent's patience and stamina), the answers to questions will be as accurate as possible. However, it is also true that the respondents will wish to present themselves in a favorable light. This can be especially true when people are asked about health-related events and behaviors. This desire leads to what is known as a **social desirability bias**. Some questions may be sensitive or threatening, such as those about sex or drugs or illegal behavior. In this situation, a respondent not only will want to present a positive image but will certainly think twice about admitting illegal behavior. In such cases, the respondent may shade the actual truth or even lie about particular activities and behaviors.

The role of the interviewer can also influence responses. Who admits to their dentist that they aren't flossing? Or suppose that English teachers are administering a questionnaire about

the reading habits of their students. Might students suddenly develop an apparent interest in reading or report they read for pleasure more than is the exact truth?

It is clear that constructing questionnaires and writing questions can be a daunting task. Three guidelines to keep in mind are:

1. Questions should be understandable to the individuals in the population being studied. Vocabulary should be of appropriate difficulty, and sentence structure should be simple.

2. Questions should as much as possible recognize that memory is a fickle thing in humans. Questions that are specific will aid the respondent by providing better memory cues. The limitations of memory should be kept in mind when interpreting the respondent's answers.

3. As much as possible, questions should not create opportunities for the respondent to feel threatened or embarrassed. In such cases the responses may be subject to social desirability bias, the degree of which is unknown to the interviewer. This can compromise conclusions drawn from the questionnaire data.

Section IV: Sampling

The purpose of a questionnaire is to gain important knowledge about a population. It is almost never feasible and is never necessary to administer the questionnaire to everyone in the population. Instead the methods of sampling and statistics are used in epidemiologic studies. The methods of statistics depend crucially on how data are gathered, and statistical inferences about a population are only as good as the sampling procedures.

When researchers perform a sample survey, usually a statistician is consulted for expert assistance. When students administer a questionnaire to other students, however, a statistician is not usually available. In most cases those students are selecting a convenience sample—that is, the questionnaire is given to whoever happens to be available. The good news about this sampling technique is that it is convenient. The bad news is that absolutely no conclusions about the population can be made.

To be able to generalize results from a sample to a population, a probability-based sample must be taken. We will outline some common sampling techniques here, but if you anticipate actually doing a cross-sectional study, you should find a statistics book and study these methods in more detail.

In the discussion below we will represent the sample size by the letter *n*.

A SIMPLE RANDOM SAMPLE (SRS). A SRS is a sample taken in such a way that each combination of *n* individuals in the population has an equal chance of being selected. The SRS is the simplest sampling plan to execute if one has a list of the population. For example, suppose that you had a list of students at your school. You could write each student's name on a slip of paper, put the names in a giant barrel, shake it up and then select *n* slips of paper. The lucky winners are your SRS. (You don't actually have to use a barrel. You could assign each student a number—1, 2, 3, etc.—and use your calculator to generate random integers for the sample.)

A SYSTEMATIC RANDOM SAMPLE. A systematic sample is designed to be an easy alternative to the SRS. If one has a list of students, numbered 1, 2, 3, 4, and so on, a systematic random sample is taken by deciding on what fraction of the population is to be sampled. For example, suppose one wanted to sample 5% of the student body. To accomplish this, one would pick a random starting point from the first 20 students in the list and then take every twentieth student in the list. The chief advantages of

this method are that it gives results like those of an SRS, and it is easy to actually do. (No barrels or calculators needed!) However, the systematic sample has a clear disadvantage. If there is some known or unknown order to the list, picking every twentieth student may introduce a bias into the sample.

A STRATIFIED RANDOM SAMPLE. When doing a cross-sectional study, important subgroups of people may have different views or life experiences or health-related behaviors. For example, males and females may have different health issues and different views on how health services should be delivered. As another example, non-English speakers may rate hospital services differently because of the problems inherent in communicating with English-speaking hospital staff. So when gathering information about a diverse population, care must be taken to ensure that the relevant subgroups are adequately represented in the study sample. Which groups are relevant for a particular study may be challenging to determine, but without representation from them the results could be inaccurate. Taking a stratified random sample is easy once the subgroups are identified: Take a simple random sample from each subgroup.

These are the three basic methods for taking a sample from a population. However, please do remember that should you decide to take a sample, consult a statistics book for more detail about these methods.

Section V: Questionnaire Administration

Questionnaire design is only one step in the process that ultimately leads to generating answers to research questions of interest. After the questionnaire is designed, researchers should run a pilot test of the questionnaire to make sure it is understandable and acceptable to the intended audience. That process will ideally involve administering the questionnaire to a small group of persons from the intended target group and then following up to get feedback on the questions (e.g., how they were worded, whether the respondents understood them, whether the respondents felt comfortable answering them) and on the questionnaire itself (e.g., whether it was too long, potential barriers to getting good responses). Pilot testing also involves evaluation of other attributes, namely, precision (reliability) and accuracy (validity). Those attributes are critical to developing a questionnaire whose results are reproducible and that provides the researcher with a good measurement of the phenomenon or phenomena of interest.

After incorporating feedback from the pilot test, the questionnaire is ready to be administered to a sample from the target population. As mentioned in the section above, the process of responding to interviewer-administered questionnaires depends in part on the respondent, the interviewer and the interaction between the two. To have reliable findings, it is important to have well-trained interviewers. All interviewers should understand the research study and the questionnaire. They should be consistent in the way in which they ask questions, provide prompts and interact with the respondents. Not only should an interviewer be consistent from respondent to respondent but also the questionnaire administration process should be consistent from one interviewer to the next.

Section VI: Secondary Analysis of Data

The process of designing one's own questionnaire is often time-consuming and may become quite expensive. Moreover, there are several questionnaires conducted by others, such as the federal government, that may be helpful in answering public health–related questions. For these and other reasons, epidemiologists will often use existing questionnaire data and analyze them in order to find the answers they seek. For instance, suppose you wanted to know about the nutritional habits of U.S. teenagers who exercise regularly. To answer that question, you could design a questionnaire that asks about nutrition and exercise, give a pilot test of the questionnaire to make sure the questions are worded correctly, revise the questionnaire, hire people to administer the questionnaire, pay for photocopying the questionnaire, and then hope that the respondents will fill out the questionnaire in a timely fashion, and if not, you would have to follow up with them—I think you get the point! This process can get to be very lengthy, complicated and costly. However, if you were told that a group of epidemiologists had already administered such a questionnaire, wouldn't it be easier just to get the information from them? Absolutely. Although it would certainly be easier, researchers collect data in a way that answers the questions they are interested in, not necessarily the ones you might be interested in. Fortunately it is often possible to use their data and manipulate the data in such a way as to answer the questions in which you are interested. This process—taking existing data and reanalyzing them to answer a new question—is called **secondary data analysis** and is quite common in epidemiologic research.

The next part of this module will allow you to gain experience in conducting a secondary data analysis by analyzing the data from an existing federal government dataset. The federal government, specifically the U.S. Public Health Service, has a very large collection of periodic surveys that are used to monitor the health of the population. These surveys are generally very large, expensive, complicated and well-executed endeavors and routinely serve as the source of secondary data for many agencies and individual researchers. Although there are many such surveys, in this module we will work with one that may be of most interest to you: the Youth Risk Behavior Survey (YRBS). For detailed information, you may wish to refer to the Centers for Disease Control and Prevention Web site, available at: http://www.cdc.gov/nccdphp/dash/yrbs/about_yrbss.htm.

The YRBS is a biennial survey of ninth- to twelfth-grade students across the United States that asks questions about the following health behaviors:

- Tobacco use
- Unhealthy dietary behaviors
- Inadequate physical activity

- Alcohol and other drug use
- Sexual behaviors that contribute to unintended pregnancy and sexually transmitted diseases, including human immunodeficiency virus (HIV) infection
- Behaviors that contribute to unintentional injuries and violence

The YRBS has been in operation for over 10 years, and so several years of data are available. Of course, the YRBS researchers have already done analyses of those data. However, there may be several opportunities for secondary data analysis to answer questions as yet unanswered. In this module you will work in groups to go through the process of answering a question of interest to you, as follows:

1. Assemble in teams of four to six students.
2. Each team should work with the class teacher to decide on a research question of interest. The team should consider:
   - A primary research question that evaluates the relationship between two key variables of interest.
   - At least three secondary research questions that provide supplemental information to help understand the main relationship of interest. Examples include how the main relationship of interest may differ among demographic subgroups.
   - The available data. In deciding on your secondary data analysis you must consider both your scientific interests and the available data, because you want to ensure that the question you wish to answer is indeed possible given the data available to you. For example, a team may wish to answer the question, "Do youth from Mississippi drink more milk than California youth? Although this is a legitimate question that may be of importance, it is not possible to answer it given the YRBS data. As you will see from the Codebook (Appendix 1), state data are not available.
3. Each team will get the questionnaire data in an electronic file. If you had done your own questionnaires, you would have to enter the data from the questionnaire forms into a dataset before you could begin to analyze the data. However, this step has already been done for you by the YRBS staff. All you need to do in order to conduct the analysis is to get a copy of the dataset. These are public-access data, so they are freely distributed by the U.S. government for use by researchers such as you.
4. Your class teacher will provide you with a file that is a subset of the data from the 2001 YRBS. With approximately 100 questions and more than 13,000 student respondents, the full dataset is quite large, so the dataset you will use for this module contains only selected questions from the dataset. The dataset includes the following questionnaire items: 1–7, 10–12, 16, 29, 30, 32, 33, 41, 42, 70, 73–79, GREG and METROST. Please refer to your Data Documentation/Codebook (Appendix 1) for details about these questions.

5. Student teams should decide on a plan for analyzing the data based on the nature of their research question. For instance, if you would like to answer the question "Is fasting to lose weight more common among males or females?" you would need to consider Q70 (about fasting) and Q2 (gender). You would want to create a 2×2 contingency table (two rows and two columns) that displays proportions and calculate a Chi-square test to compare the significance of the difference in proportions. Your table would look like the following:

	Fast: Yes	**Fast: No**	**Total**
Gender: Male	Number of males who fasted	Number of males who did not fast	Total number of male respondents
Gender: Female	Number of females who fasted	Number of females who did not fast	Total number of female respondents
Total	Total number of youth who fasted	Total number of youth who did not fast	Total number of respondents

6. Now certainly you could print out all of the data and then manually count the number of male fasters, female fasters, male nonfasters and female nonfasters. Then you could put those counts in their respective cells and calculate the Chi-square statistic by hand. However, that would not be an efficient method. You can conduct all of those operations in less than a minute with the use of statistical analysis software. For this module, you will use the *Epi Info* software package to analyze the data. The instructions for using the software are given in the following section.

7. After analyzing the data, each team should write a short report. The text should be one to two typewritten pages, with extra space allowed for graphs and tables as needed. Scientific reports have a standard format. A typical report could include the following information:

   - *Introduction:* background, rationale, purpose of the study, and research question or questions. Most researchers base this section on a thorough review of the literature. They use past research on a topic as the impetus and rationale for their own work.
   - *Methods:* brief description of the YRBS study, the variables you used, and the statistical analyses you performed.
   - *Results:* your findings in text, tabular and graphic representations. You may wish to use histograms, pie charts or line graphs to present your data. This can be done with *Epi Info* or alternatively you can save the output from *Epi Info* and input it in *Microsoft Excel*® if you prefer.
   - *Conclusions:* what you learned and the implications of your findings. This is where you will state the answers to your research questions and explain to the report readers why what you did was important and how it can be useful for planning future research, crafting health policy, designing health education programs and so forth.

Section VII: Using *Epi Info*™ to Analyze YRBS Data

Epi Info™ Version 3.2 (February 2004) the most recent version of the free *Epi Info* software package. *Epi Info* is in the public domain, so it may be copied and shared at will.

Accessing and Installing *Epi Info*

The software that you need is available from the U.S. Centers for Disease Control and Prevention (CDC) via their Web site at http://www.cdc.gov/epiinfo/index.htm. To use this *Windows*®-based software, you need the following capabilities:

- *Windows 95, 98, ME, NT 4.0, 2000* or *XP*
- 32 MB of RAM; at least 64 MB is recommended for *Windows NT 4.0* and *2000;* 128 MB needed for *Windows XP*
- 200-MHz processor; 300 MHz for *XP*
- 260 MB on your hard drive to install

To download the software, access the Web site and click on Download. You are then provided with two options for downloading; select either Web Install or Download setup.exe. Note that this is a large file and if you are downloading it through a 56K modem, it will take a very long time to download. So if possible, download the software using a high-speed connection.

Using *Epi Info*

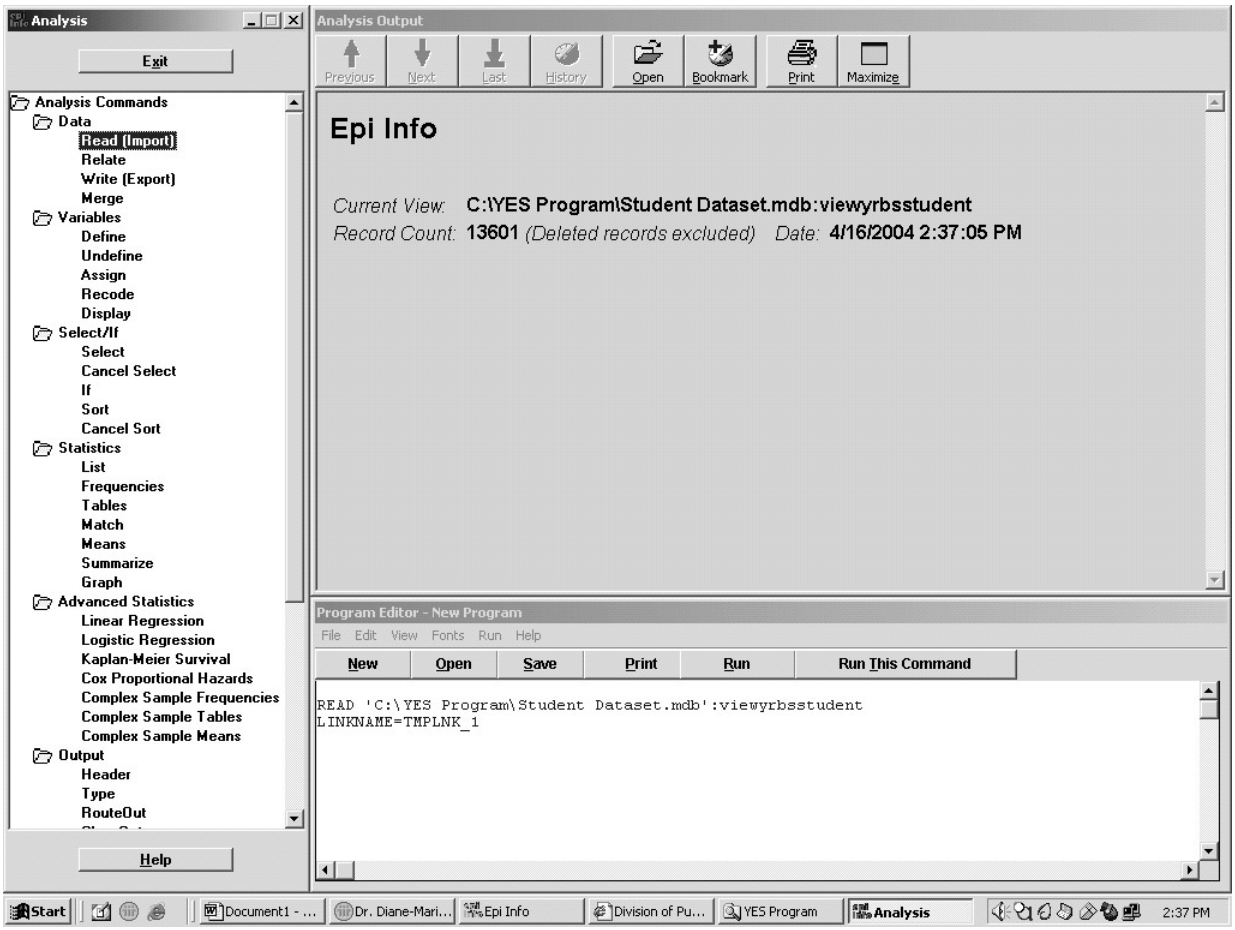
To use *Epi Info,* follow these steps:

1. Double-click on the *Epi Info* icon to open the program.
2. The program will open with a graphic in the background, the *Epi Info* logo on the top of the page and several buttons on the bottom. The buttons that may be of most interest to you are:
    - *MakeView*. This is used by those who have designed their own questionnaires and are going to enter the data and create an analysis dataset themselves. This button accesses the parts of the program you will use to create the structure of your questionnaire in *Epi Info*. It is necessary to complete this step before data entry can begin. You will not use this feature for this module because you have not designed your own questionnaire.

- *Enter Data*. This puts the program in data entry mode. *Epi Info* will create fields for each question in your questionnaire (which it does using MakeView) and then it asks you to enter the responses in those fields (in Enter). You will not use this feature in this module because you are using secondary data—the data have already been entered by the YRBS staff.
- *Analyze Data*. This is the feature that will be most relevant to you for this module. This is where you submit commands to *Epi Info,* directing it to summarize the data and conduct various statistical tests as necessary to answer your research questions.
- *Epi Info Web Site*. Clicking on this button will take you to the *Epi Info* Web site if the computer has an active Internet connection.

3. To begin your analysis, click on the Analyze Data button.
4. A screen with three distinct parts will open:
   - On the left is a menu of all available operations. To execute a given command, you must click on it and then a dialog box will open, asking for further information. For instance, if you click on List to list out all responses to a given question, the dialog box that will appear will ask for the name of the variable (question) that you wish to list.
   - On the bottom right is the Program Editor box in which the code is written. This screen allows you to keep a log of all of the commands that have been sent. After you become more familiar with the program, it will be possible for you to type in your own code rather than using the list of commands from the left-hand menu. This is analogous to using Ctrl-P to print in *Microsoft Word*® as opposed to pulling down the File menu and then highlighting Print. They are just two ways of doing the same thing. Another use of the Program Editor box is to keep a log of all of your commands for saving and reusing at a later time. You may click on Save to save the contents (called the program) of the Program Editor box. Then the next time you use *Epi Info*, you can open the saved program and resubmit it by simply clicking on Run to resubmit the entire program or Run This Command to resubmit just one operation or command.
   - At the top right is the Output screen, where the results of your commands will be displayed.
5. Before analyzing data, the first thing that must be done is to tell *Epi Info* what dataset you will be analyzing. This is done using the Read command. Click on Read and then a dialog box will appear. Keep the default Data Format (Epi 2000), and then click on the dots to the right of the Data Source box to browse and select the dataset (Student Dataset) wherever you have stored it on your computer. Click on viewyrbsstudent. Then click OK. *Epi Info* will tell you that it is creating a temporary link, click OK.

6. If you have saved your dataset on your hard drive (c:\) in a subfolder titled YES Program, your screen should look as shown in the following screen. Recall that the Command menu is on the left, the Output menu is on the upper right, and the Program Editor menu is on the lower right.



7. You are now ready to analyze your data. To do this, you click on the desired command in the left-hand menu and then the dialog box appears, asking you to select the variable or variables to use for the operation.

8. As you can see from the left-hand menu, there are several analysis commands in *Epi Info*. The ones that you are most likely to use are listed below:

   - *List*. This command is a line-by-line listing of all responses. You click on List, and then in the dialog box you select your variable name and then click OK. You may look at the listing for one variable or more than one. If you want to include more than one, simply pick

multiple variable names. Each time you should see the name show up in the box. Going back to our example, you may wish to look at the responses to the question about fasting. Does this output of responses answer your research question? Well, this output is not very informative because the data are not summarized in any way, but why not take a look at the listing for Q70 and see what happens.

- *Frequencies*. This command provides univariate frequency distributions for selected variables. To execute the Frequencies command, click on the command in the left-hand menu and then when the dialog box appears, select your variable name from the list in the box that says Frequency of. This is the command that you might wish to use to look at the responses for fasting (as in our previous example). Try this and now see what happens. Then do it again for the gender variable. Do you have the answer to your question?

- *Tables*. This command is useful for contingency (2×2) tables. To use this command, click on Tables and then in the dialog box, again identify the variables you wish to use. The exposure variable is the independent, or predictor, variable. It will form the rows of the table. The Outcome variable is the dependent variable, which will be shown in the columns. Now try to create the table we suggested, using Q2 (gender) and Q70 (fasting). Again, look at the output. This output gives you a table showing the numbers of male fasters, female fasters, male nonfasters and female nonfasters. It also gives you row percentages and column percentages. Looking below, you will see various statistics listed, including the Chi-square and the associated p-value. Now, does *this* provide the answer to your question?

- *Defining New Variables*. Your team may decide that the response categories available in the dataset do not adequately capture the ones that are of interest to you. For instance, the age variable (Q1) has the following categories: 12 years or younger; 13; 14; 15; 16; 17; and 18 or older. If your group wanted to look at differences in weight between 18- and-19-year-olds, you would not be able to do that. The dataset has 18- and 19-year-olds collapsed into one category, and you cannot separate them out. Suppose instead that your group wanted to compare weights among students aged 16 and over with weights of younger students (i.e., those aged 15 and under)—that you can do. Using a few simple commands, you can collapse categories, and instead of having seven categories as in the original dataset, you can create two categories and then conduct the comparison of interest to you. This is how:

  - Click on the Define command and create a new variable named Age (leave as standard). Then click OK.

  - Click on the Recode command.

  - Select Q1 as the “from” variable and Age as the “to” variable.

- In the first column insert 1, in the second column insert 4, and in the third column insert 1. Then press the Enter key on your keyboard, and a new line will appear.
- In the first column put 5, in the second column put 7, and in the third column put 2.
- Click OK.

If you run a frequency distribution on Q1 and Age, you should be able to see whether your procedure worked. What you have just done is to create an Age variable: Age = 1 if the student is 15 or younger and Age = 2 if the student is 16 or older. Check the frequency table to make sure this is true.

9. There are many, many more features to *Epi Info*, but the ones listed above are the most commonly used. Feel free to play around with the software and learn it. There is a help function and a downloadable manual that can provide you with additional assistance when needed.

10. When you have finished your *Epi Info* session, you may save your program so that you do not have to start all over again next time. To do this, click Save in the Program Editor box, click on the Text file button, and then save it on your hard drive or diskette. It will be a fairly small file with a *.pgm extension. You do not really need to save your output because with the saved program you merely rerun it and easily generate the output again. However, the output is actually being saved by *Epi Info* in the same folder in which the *Epi Info* software is stored, using a *.htm format by default.

Worked Example for Teachers

Your teacher dataset includes all of the questions in the student dataset, in addition to Q92–95, which you may use for this worked example or another example of your own choosing. Please note that we have included a short primer on Chi-square (Appendix 2) should you wish to refer to it in planning your class demonstration.

The first part of the process is the identification of a research question, e.g., What sociodemographic factors are related to serious sports injury among U.S. high school students?

Students should be reminded of the following:

- A research question should be clear, concise and answerable.
- Determination of the research question can be based on any one or more of the following: scientific curiosity, unanswered questions from one's own or someone else's prior research, hypotheses raised by observation or anecdote, request by an external stakeholder, such as a sports equipment manufacturer.
- All components of the research question should have clear operational definitions. For instance, one should define serious sports injury and sociodemographic factors. For our purposes we are defining serious sports injury as one for which medical (doctor or nurse) attention was sought, and sociodemographic factors would include metropolitan status and gender.

Then one should clearly state a hypothesis. For instance, one research hypothesis might be that rural students would be more likely to suffer serious sports injury than nonrural students.

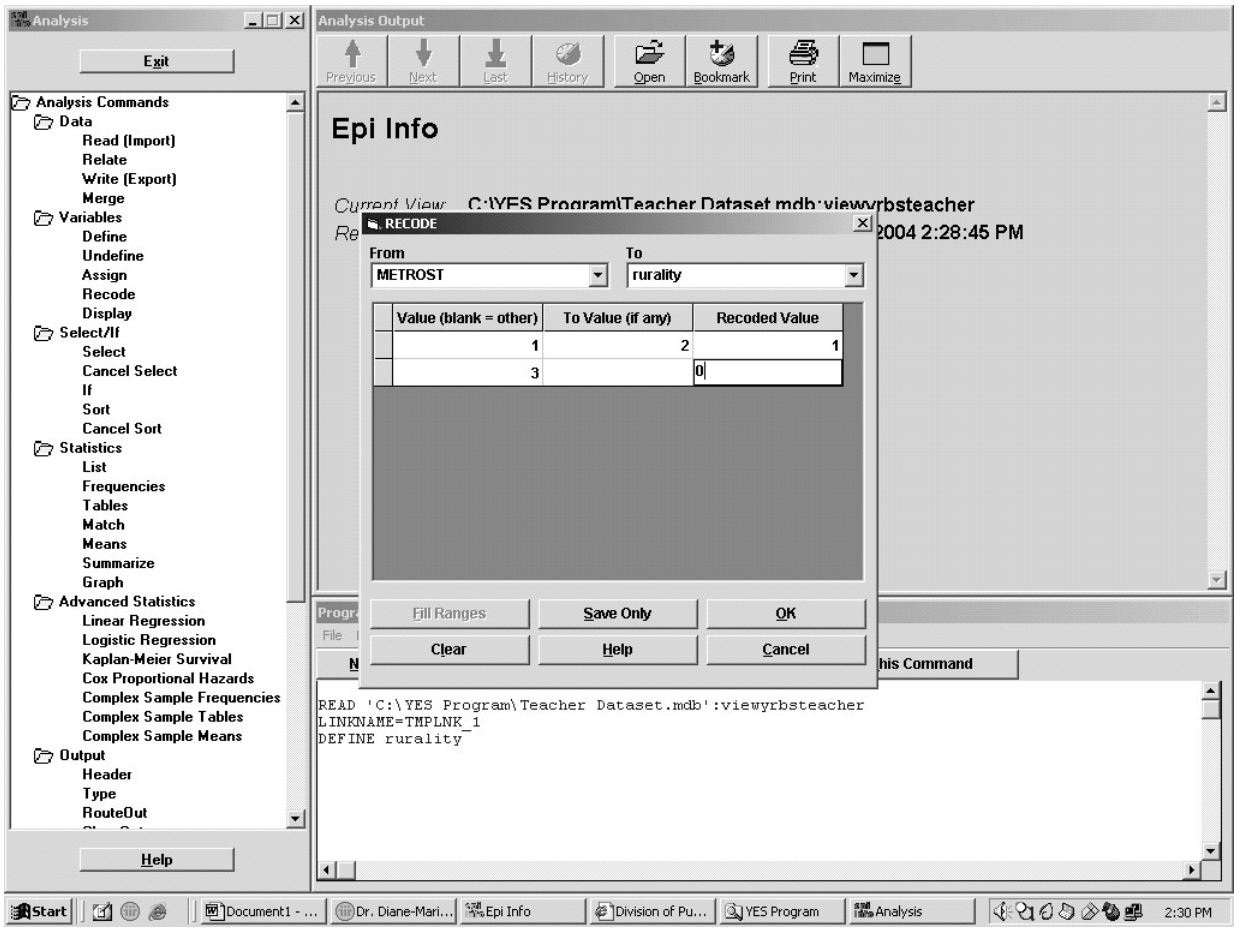
The data analysis strategy can then be devised as follows to test that hypothesis.

1. The metropolitan status variable would need to be collapsed to create two categories (rural versus nonrural), and from the sports injury question (Q92) we would need to create a new variable that excludes those who do not exercise or play sports and as such were not eligible to have had a sports injury. We would then be assessing the relationship between two binary variables. Assuming that all assumptions of the test are met, we could do this using a Chi-square test of the following statistical hypotheses:
    - *Null hypothesis:* The observed distribution of frequencies equals the expected distribution. (There is no relationship between rurality and sports injuries.)
    - *Alternate hypothesis:* The observed distribution of frequencies does not equal the expected distribution. (There is a relationship between rurality and sports injuries.)
2. First, we must open *Epi Info* and access the YRBS data. Click on the *Epi Info* icon to open the dataset and then click on the Analyze Data button. Click on Read from the Analysis-

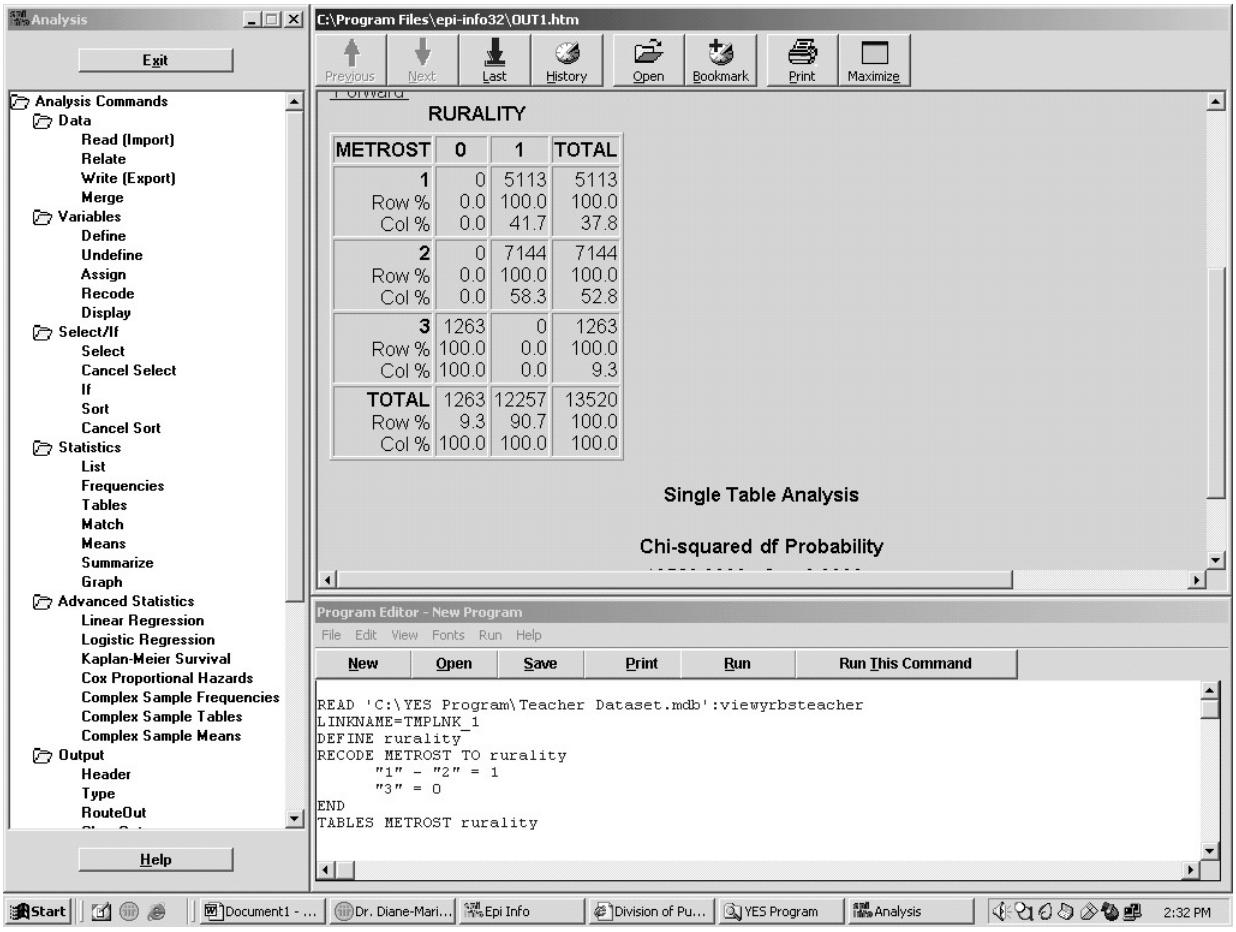
Commands left-hand menu and then a dialog box will appear. Keep the default Data Format (Epi 2000) and then click on the dots to the right of the Data Source box to browse and select the dataset (Teacher Dataset) wherever you have stored it on your computer. Click on viewyrbsteacher. Then click OK. *Epi Info* will tell you that it is creating a temporary link. Click OK.

3. Then create a RURALITY variable from the METROST variable:

   - Click on Define and for variable name, type in RURALITY. Keep Scope (which is the variable type) as standard, the default. Then click OK.
   - Click on Recode and in the From pull-down menu, select METROST. In the To pull-down menu, select RURALITY.
   - Click in the Value box and type 1, click in the To Value box and type 2, and then click in the Recoded Value box and type 1. Next press the Enter key on the keyboard. Then on line 2, click in the Value box and type 3 and click on the Recoded Value box and type 0. Your screen should look as follows:



- Now click OK. You will have your new variable: RURALITY = 0 when METROST = 3 and RURALITY = 1 if METROST = 1 or METROST = 2. To confirm this, click on Tables in the left-hand menu and select METROST from the exposure pull-down menu box and RURALITY from the outcome pull-down menu box and click OK. You should see a table appear in the Output window as follows:

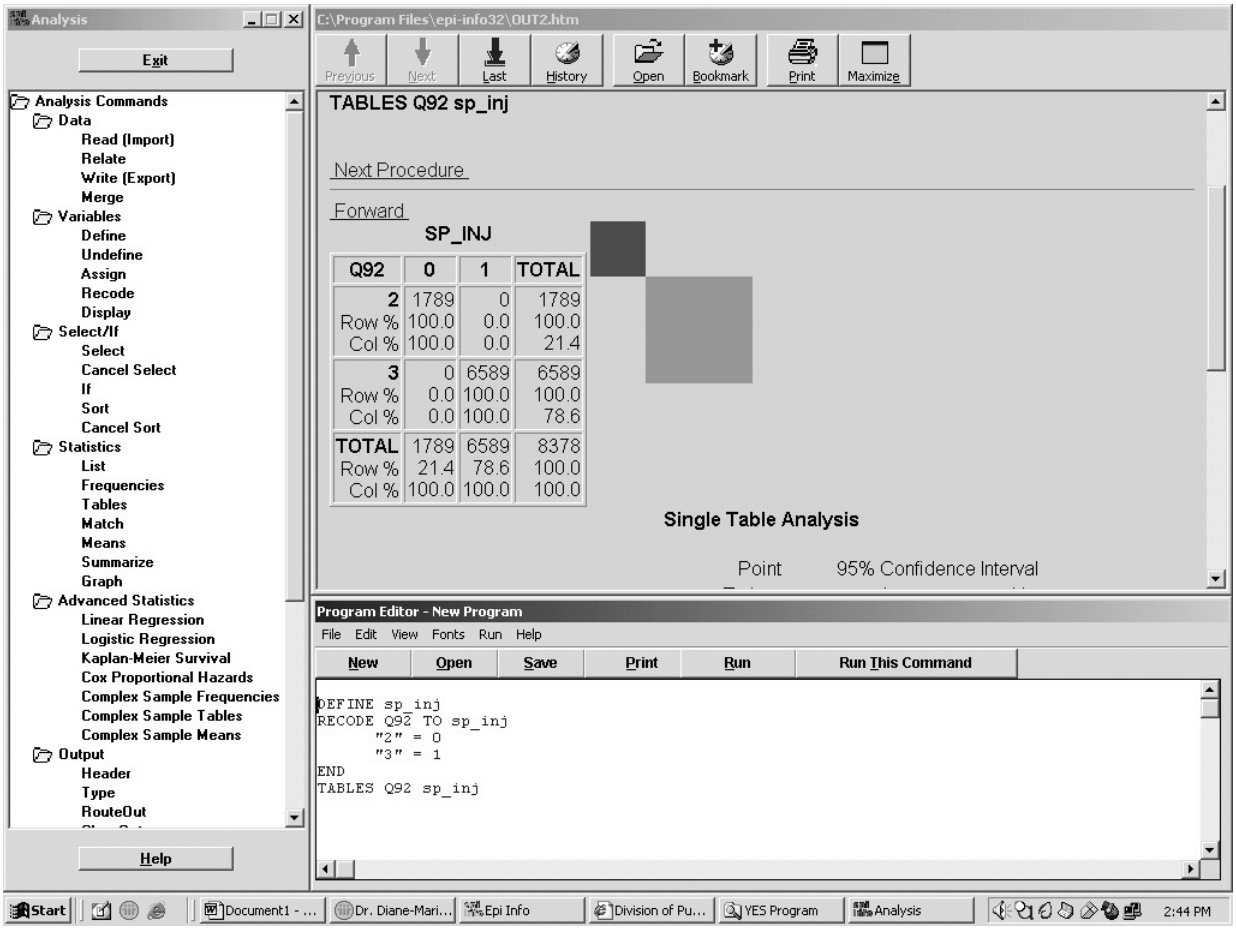


This table confirms that values of 1 and 2 for METROST (urban and suburban) are now coded as 1 for RURALITY (nonrural) and values of 3 for METROST (rural) are now 0 for RURALITY (rural). Note that we ignored METROST=0, which was unknown because one cannot determine whether that was a rural or nonrural respondent, so we treat those as missing values.

4. Then using the same method as above, recode the Q92 variable to create SP-INJ:

   - Click on Define and for variable name, type in sp_inj. Keep Scope (which is the variable type) as standard, the default. Then click OK.

- Click on Recode and in the From pull-down menu, select Q92. In the To pull-down menu, select sp_inj.
- Click in the Value box and type 2 and then click in the Recoded Value box and type 0. Next press the Enter key on the keyboard. Then on line 2, click in the Value box and type 3 and click on the Recoded Value box and type 1.
- To double-check your work and demonstrate that the recode has succeeded, click on Tables in the left-hand menu and select Q92 as your exposure and SP_INJ as your outcome, and you should see the following table appear:



Again, this confirms that the recoding worked as intended.

5. You are now ready to conduct your Chi-square test.

Click on Tables and for the exposure variable, select RURALITY and for the outcome variable, select SP_INJ. The following table should appear:

SP_INJ

rurality	0	1	TOTAL
0	185	676	861
Row %	21.5	78.5	100.0
Col %	10.4	10.3	10.4
1	1595	5858	7453
Row %	21.4	78.6	100.0
Col %	89.6	89.7	89.6
Total	1780	6534	8314
Row %	21.4	78.6	100.0
Col %	100.0	100.0	100.0

Single Table Analysis

	Point Estimate	95% Confidence Interval Lower	Upper
PARAMETERS: Odds-based			
Odds Ratio (cross product)	1.0051	0.8465	1.1935 (T)
Odds Ratio (MLE)	1.0051	0.8449	1.1917 (M)
		0.8417	1.1959 (F)
PARAMETERS: Risk-based			
Risk Ratio (RR)	1.0040	0.8773	1.1490 (T)
Risk Difference (RD%)	0.0859	-2.8114	2.9831 (T)

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

STATISTICAL TESTS	Chi-square	1-tailed p	2-tailed p
Chi square-uncorrected	0.0034		0.9536249154
Chi square-Mantel-Haenszel	0.0034		0.9536277014
Chi square-corrected (Yates)	0.0002		0.9886064019
Mid-p exact		0.4742119341	
Fisher exact		0.4916574598	

The 2 × 2 table tells us that the prevalence of serious sports injury among the rural students was 21.5% and among the nonrural students was 21.4%. From just that information alone, one might say that there is no relationship between rurality and sports injury. However, one can assert that statistically by using the Chi-square test of proportions from the output: $\chi^2 < .01$, $p = 0.95$. Hence we fail to reject the null hypothesis and conclude that there is no relationship between rurality and serious sports injury.

6. We might conduct a similar analysis of the relationship between gender and serious sports injury, based on a hypothesis that females may be more likely to suffer serious sports injury than males.

To do this we set up a similar 2 × 2 table and observe the Chi-square test. Using the Tables command, we identify Q2 as the exposure variable and SP_INJ as the outcome, and the following appears:

SP_INJ

Q2	0	1	TOTAL
1	778	3219	3997
Row %	19.5	80.5	100.0
Col %	43.7	49.0	47.9
2	1001	3354	4355
Row %	23.0	77.0	100.0
Col %	56.3	51.0	52.1
TOTAL	1779	6573	8352
Row %	21.3	78.7	100.0
Col %	100.0	100.0	100.0

Single Table Analysis

	Point Estimate	95% Confidence Interval Lower	Upper
PARAMETERS: Odds-based			
Odds Ratio (cross product)	0.8098	0.7288	0.8999 (T)
Odds Ratio (MLE)	0.8098	0.7287	0.8998 (M)
		0.7277	0.9011 (F)
PARAMETERS: Risk-based			
Risk Ratio (RR)	0.8468	0.7792	0.9204 (T)
Risk Difference (RD%)	-3.5205	-5.2721	-1.7689 (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	15.4091		0.0000877415
Chi square-Mantel-Haenszel	15.4072		0.0000878261
Chi square-corrected (Yates)	15.1998		0.0000978848
Mid-p exact		0.0000426697	
Fisher exact		0.0000474121	

In this example we see that the prevalence of injuries among females is 19.5% compared with 23.0% among males. We again could attempt to make a judgment based only on a comparison of those proportions. However, it would be helpful to look at the Chi-square test and see that $\chi^2(1\ df) = 15.4$ and $p < .01$. So in this case we reject the null hypothesis of no relationship and assert that there is indeed a relationship between gender and sports injuries. Of course the test does not tell us why this relationship holds. Teachers may wish to help the students probe for possible reasons—for example, perhaps there are more males involved in high-impact sports. Students may realize that some of these "why" questions may be answered by other questions in the dataset and some may not. To answer burning questions that may not be addressed using the YRBS data, primary data collection may be necessary.

Assessment

Four options are given, one with suggested answers. Please note that students are not expected to conduct these surveys, but rather they are being asked to demonstrate their skills in survey design.

Young Women and Soda

As is well known, teens drink soda. Some medical experts believe that excessive drinking of soda by teenaged girls may put them at risk. Some studies have demonstrated an association between drinking soda and bone fractures in active girls, though the possible causative biological mechanism is unknown, and scientific study is still at the exploratory stage.

Your task is to design a questionnaire that will be used to gather data to explore the relationship between soda consumption and bone fracture. The variables thought to have some explanatory power are:

1. The activity level of the young women. They may be totally sedentary, or they may engage in physical activity ranging from light to vigorous.
2. The types of activities they engage in. They may have no organized activity; or they may be a member of a high school sports program, a program outside school or both. (If there turns out to be an association between soda and bone fractures, this information will be useful in designing intervention strategies.)
3. Their soda consumption. They may drink no carbonated beverages, or they may drink colas, noncolas or both.

The goal of the study is to find an association—if it exists—between these variables and bone fractures. The bone fractures may be of different types, as there are bones of different sizes subject to different stresses during the normal day. Thus it will be important to get some detailed information about the fractures. It is also possible that other factors may contribute to bone fractures, such as diet. Because diet may play a significant role, your questionnaire should inquire if the respondents are on any sort of special diet, either on the advice of their doctor or by their own choice, such as a vegetarian diet.

Your questionnaire should contain the following elements:

- A short introduction, explaining the purpose of the questionnaire
- Questions to determine the level of the respondent's physical activity

- The sponsorship of the activity
- The types of carbonated beverages consumed, as well as some indication of how much is consumed in a typical period of time
- Dietary habits
- Medical history of bone fractures (and relevant details)

School Violence

Health professionals, educators and parents have recognized that school violence is a major concern and possibly a significant public health problem. Little is known about the nature of adolescent violence in school, and if this health threat is to be effectively countered, the nature of school violence must be studied. One form of school violence of particular concern is fighting. Available evidence is anecdotal, usually based on the experience of school administrators who have questioned individuals after fights. Because the combatants have a vested interest in blaming the other person, this information is suspect at best.

Your task is to design a questionnaire to be given to students in grades 7–10 who have participated in fights in the previous six months. You may assume this questionnaire is confidential and is being given after any punishments have been meted out to the individuals involved in the fight. (On the questionnaire you will need to explain this to the respondents, so they know they are not putting themselves at risk of further punishment.)

Generally your interest should center on the following aspects of the fighting:

1. What were the causes of the fight?
2. What were the genders, ages and grade levels of the combatants and what was the relationship between them?
3. Where was the fight?
4. What was the involvement, if any, of bystanders?
5. What, if any, injuries were sustained and how severe were they?

In previous studies students have been hesitant to be precise in their responses. To counteract this, you should provide some common responses for the respondent to pick from, as well as have a blank line labeled "Other," to be filled in. The common responses will have to be from your own experience in school, and (to provide you with some guidelines) there should be at least five specific common responses if you can identify that many. If respondents pick from responses rather than construct their own, the information tends to be more precise. Therefore, if there are more than five common responses, you should list them if at all possible.

Your questionnaire should contain the following elements:

- A short introduction explaining the purpose of the questionnaire
- Assurances of confidentiality
- Questions that elicit responses about the five general topics above

Teens on the Job

On-the-job injury has become a serious threat to American youth with the increasing numbers of teenagers who hold part-time jobs. Little is known about the working environments of teenage workers, especially their exposure to hazardous equipment and dangerous work environments. It is believed that:

- Common jobs are located at home, retail stores or restaurants.
- Common jobs are as lawn care workers, cashiers and dishwashers.
- Common hazards teens are exposed to on the job are ladders or scaffolding, forklifts, tractors or riding mowers, and working around loud noises.
- Students may be working many hours, evening hours or both.

Your task is to design a questionnaire that would be given to teenagers in this age group (mostly high school students). From your own experience you may have a sense of where and what the jobs are locally, and you should slant the questions to get some detail. For example, if your school is a city school, you need not ask questions about farm labor; if your school is a rural school, there may similar reasonable omissions. Your questionnaire should probe for the types of work teens are doing, the number of hours and the times of day that the students are working. Of special concern are (a) the hazards that teens are exposed to and (b) the *perceived* hazards—that is, the hazards that teens believe are found at work.

Your questionnaire should contain the following elements:

- A short introduction explaining the purpose of the questionnaire
- Questions of a demographic nature: age, gender, etc.
- Questions to determine the location of part-time jobs
- Questions to find out if identified hazards such as those listed above exist in their workplace
- Space for the respondents to list other, unidentified, hazards

You should be particularly concerned with identifying types of workplaces. For example, you should distinguish between a fast-food restaurant and a more formal restaurant. Some teens work construction in the summer. This is a very broad category of jobs, and you should break down such jobs into narrower categories, such as laborer, machine operator, flagger and so on. You should also attempt to obtain descriptions of the types and amount of on-the-job training that students have received.

Health Services for Performing Arts Students

Adolescence is a very important time from a health standpoint. Many behaviors and attitudes relating to health are developed and crystallize during this period. Health problems such as stress, depression and nervousness, as well as social and psychologic concerns, are not uncommon in this age group. Adolescents who are involved in competitive performances are particularly prone to injury. Although it is not commonly realized, students in the performing arts, such as dancers and theater performers, are considered athletes, given the physical demands and training requirements put on them. Classical ballet, for example, results in a high incidence of health problems. Eating disorders, substance abuse and low self-esteem are not uncommon among these performers, who are typically very achievement oriented.

Your task is to design a questionnaire that will be used to gather data about the health risks and concerns of performing arts students, for the purpose of advising health professionals who deal with the health and medical needs of these adolescents. You should specifically ask for the following information:

- Demographic information, such as age, gender, year of school
- What performing arts activities they are engaged in
- Information about any injuries they have had, whether sustained during a performing arts activity or not
- Risk-taking behaviors, such as sexual activity and substance abuse
- Specific health concerns that the respondents may have

Problems that you must address while constructing the questionnaire will be as follows:

1. You must ascertain what performing arts activities now exist at school and in your local area, as well as the level of participation of the respondent in these activities.
2. In your introduction to the questionnaire, you will have to be particularly careful about ensuring confidentiality.
3. You will have to list particular health concerns the respondents can pick from (use your own experience as a guide) as well as categories of concerns they may respond to by listing

specific examples. For example, they may be concerned about sleep patterns. You should list some specific problems, such as little sleep, fitful sleep or nightmares.

It is also possible that you are unfamiliar with performing arts. In that case, prior to writing the questionnaire you will need to find some performing arts persons to help you understand the possible health problems they have.

School Violence Assignment (Teacher's Guide)

There are, of course, several ways in which the questionnaire can be designed. However, it is important that in their questionnaires students have demonstrated that they understand some key issues:

1. Students should have considered the grade level of the respondents (7–10) and selected appropriate vocabulary.
2. They should have included a preamble providing instructions for filling out the questionnaire, including a statement about confidentiality.
3. They should have included questions that solicit information about:
   - The fight (causes, location, bystanders' involvement)
   - Sequelae (injuries, punishments)
   - Combatants' characteristics

Example

You were selected to participate in this research study about school fighting. The information we collect will help us to better understand school fighting and how it can be prevented. We would like to ask you to answer a few questions that should take no more than 10 minutes. Please note that your answers are completely confidential. Your name will not be included in any reports about these results. Your individual answers will not be shared with anyone.

For each question below, please write in the answer or place a check mark in the box.

1. Have you been in a fight at school within the past six months?
No. If no, please stop here. You do not need to answer any more questions. Please fold this survey in half and place it in the sealed box outside the auditorium. Thank you for your time!
Yes. If you answered yes to Question 1, please continue with the questions below.

2. How old are you?

_____years

3. Are you male or female?

☐ Male
☐ Female

4. What grade are you in?

☐ 7th grade
☐ 8th grade
☐ 9th grade
☐ 10th grade

5. In the past six months, how many times have you been in a fight at school?

_____ times in the past six months

The next questions will all be based on the most recent school fight in which you were involved. Please answer based ONLY on the most recent school fight.

6. In what month was your most recent fight?

☐ November
☐ December
☐ January
☐ February
☐ March
☐ April

7. When did your most recent fight occur?

☐ Before school started
☐ During class
☐ During lunchtime
☐ During recess
☐ Between classes (while going from one classroom to another one)
☐ After school was over

8. Where did your most recent school fight occur?

☐ In a classroom
☐ In the school halls
☐ In a bathroom
☐ In the gym

☐ In a teacher's office
☐ Some other place. If you picked this one, please write in where the fight occurred: _____

9. In your most recent fight, did you know the other student?

☐ Yes, I knew the other student and we were friends.
☐ Yes, I knew the other student but we were not friends.
☐ No, I did not know the other student.

10. In your most recent fight, who made the first physical contact? In other words, who started it?

☐ You
☐ Someone else

11. Were there any other students looking at the fight?

☐ Yes
☐ No

12. If there were other students looking at the fight, please describe what they were doing.

☐ They were trying to stop the fight.
☐ They were trying to encourage the fight.
☐ They were doing something else. Please describe: _____
☐ I do not know what they were doing.

13. Who stopped your most recent fight?

☐ I stopped the fight.
☐ The other student stopped the fight.
☐ A teacher stopped the fight.
☐ Someone else stopped the fight. Who? _____

14. What were the reasons for the most recent fight in which you were involved? Please check all the reasons that you feel were important.

☐ I was teasing the other student.
☐ The other student was teasing me.
☐ I got the other student in trouble.
☐ The other student got me in trouble.
☐ I was mad at the other student for something. Please write in what you were mad about: _____
☐ The other student was mad at me for something. Please write in what the other student was mad about: _____

☐ I wanted the other students to know that they shouldn't "mess with me."
☐ I didn't like the other student.
☐ The other student didn't like me.
☐ Another reason. Please write in what the reason was: _____

15. Did you get hurt in your most recent fight?

☐ Yes
☐ No

16. If you got hurt in your most recent fight, please tell us what your injuries were. If you had more than one, please check all.

☐ I did not get hurt.
☐ I had cuts.
☐ I had a black eye (shiner).
☐ I had bruises.
☐ I had scratches.
☐ I had bite marks.
☐ I had a broken bone.
☐ Other. If you had some other injury, please describe it: _____

17. Were you punished for being in the fight?

☐ Yes
☐ No

18. If you were punished, what was the punishment? If there was more than one punishment, please check them all.

☐ I was not punished.
☐ My parents grounded me.
☐ My parents spanked me.
☐ My parents took away my allowance.
☐ I got in-school suspension.
☐ I was suspended from school.
☐ I got detention after school.
☐ I got some other punishment. Please describe it: _____

19. Was the other student who was in the fight punished?

☐ Yes
☐ No
☐ I don't know.

Appendix 1: YRBS 2001 Data Documentation/Codebook*

Q1 How old are you?
1 12 years old or younger
2 13 years old
3 14 years old
4 15 years old
5 16 years old
6 17 years old
7 18 years old or older
†Missing

Q2 What is your sex?
1 Female
2 Male
Missing

Q3 In what grade are you?
1 9th grade
2 10th grade
3 11th grade
4 12th grade
5 Ungraded or other grade
Missing

Q4 How do you describe yourself?
1 American Indian or Alaska Native
2 Asian
3 Black or African American
4 Hispanic or Latino
5 Native Hawaiian or Other Pacific Islander
6 White
7 Multiple—Hispanic
8 Multiple—Non-Hispanic
Missing

*Includes only those items included in the module dataset.

†Missing: Survey respondent did not answer that question.

Q5 How tall are you without your shoes on? (Note: Data are in meters.)

Q6 How much do you weigh without your shoes on? (Note: Data are in kilograms.)

Q7 During the past 12 months, how would you describe your grades in school?
1 Mostly A's
2 Mostly B's
3 Mostly C's
4 Mostly D's
5 Mostly F's
6 None of these grades
7 Not sure
Missing

Q10 How often do you wear a seat belt when riding in a car driven by someone else?
1 Never
2 Rarely
3 Sometimes
4 Most of the time
5 Always
Missing

Q11 During the past 30 days, how many times did you ride in a car or other vehicle driven by someone who had been drinking alcohol?
1 0 times
2 1 time
3 2 or 3 times
4 4 or 5 times
5 6 or more times
Missing

Q12 During the past 30 days, how many times did you drive a car or other vehicle when you had been drinking alcohol?
1 0 times
2 1 time
3 2 or 3 times
4 4 or 5 times
5 6 or more times
Missing

Q16 During the past 30 days, on how many days did you not go to school because you felt you would be unsafe at school or on your way to or from school?
1 0 days
2 1 day
3 2 or 3 days

4 4 or 5 days
5 6 or more days
Missing

Q29 How old were you when you smoked a whole cigarette for the first time?
1 Never smoked a cigarette
2 8 years old or younger
3 9 or 10 years old
4 11 or 12 years old
5 13 or 14 years old
6 15 or 16 years old
7 17 years old or older
Missing

Q30 During the past 30 days, on how many days did you smoke cigarettes?
1 0 days
2 1 or 2 days
3 3 to 5 days
4 6 to 9 days
5 10 to 19 days
6 20 to 29 days
7 All 30 days
Missing

Q32 During the past 30 days, how did you usually get your own cigarettes?
1 Did not smoke cigarettes
2 Store or gas station
3 Vending machine
4 Someone else bought them
5 Borrowed/bummed them
6 A person 18 or older
7 Took them from store/family
8 Some other way
Missing

Q33 When you bought or tried to buy cigarettes in a store during the past 30 days, were you ever asked to show proof of age?
1 Did not buy cigarettes
2 Yes
3 No
Missing

Q41 How old were you when you had your first drink of alcohol other than a few sips?
1 Never other than a few sips

2 8 years old or younger
3 9 or 10 years old
4 11 or 12 years old
5 13 or 14 years old
6 15 or 16 years old
7 17 years old or older
Missing

Q42 During the past 30 days, on how many days did you have at least one drink of alcohol?
1 0 days
2 1 or 2 days
3 3 to 5 days
4 6 to 9 days
5 10 to 19 days
6 20 to 29 days
7 All 30 days
Missing

Q70 During the past 30 days, did you go without eating for 24 hours or more (also called fasting) to lose weight or to keep from gaining weight?
1 Yes
2 No
Missing

Q73 During the past 7 days, how many times did you drink 100% fruit juices such as orange juice, apple juice, or grape juice?
1 Did not drink fruit juice
2 1 to 3 times
3 4 to 6 times
4 1 time per day
5 2 times per day
6 3 times per day
7 4 or more times per day
Missing

Q74 During the past 7 days, how many times did you eat fruit?
1 Did not eat fruit
2 1 to 3 times
3 4 to 6 times
4 1 time per day
5 2 times per day
6 3 times per day
7 4 or more times per day
Missing

Q75 During the past 7 days, how many times did you eat green salad?
1 Did not eat green salad
2 1 to 3 times
3 4 to 6 times
4 1 time per day
5 2 times per day
6 3 times per day
7 4 or more times per day
Missing

Q76 During the past 7 days, how many times did you eat potatoes?
1 Did not eat potatoes
2 1 to 3 times
3 4 to 6 times
4 1 time per day
5 2 times per day
6 3 times per day
7 4 or more times per day
Missing

Q77 During the past 7 days, how many times did you eat carrots?
1 Did not eat carrots
2 1 to 3 times
3 4 to 6 times
4 1 time per day
5 2 times per day
6 3 times per day
7 4 or more times per day
Missing

Q78 During the past 7 days, how many times did you eat other vegetables?
1 Did not eat other vegetables
2 1 to 3 times
3 4 to 6 times
4 1 time per day
5 2 times per day
6 3 times per day
7 4 or more times per day
Missing

Q79 During the past 7 days, how many glasses of milk did you drink?
1 Did not drink milk
2 1 to 3 glasses past 7 days

3 4 to 6 glasses past 7 days
4 1 glass per day
5 2 glasses per day
6 3 glasses per day
7 4 or more glasses per day
Missing

‡Q92 During the past 30 days, did you see a doctor or nurse for an injury that happened while exercising or playing sports?
1 No exercise in past 30 days
2 Yes
3 No
Missing

Q93 When was the last time you saw a doctor or nurse for a check-up or physical exam when you were not sick or injured?
1 During the past 12 months
2 Between 12 and 24 months ago
3 More than 24 months ago
4 Never
5 Not sure
Missing

Q94 When was the last time you saw a dentist for a check-up, exam, teeth cleaning, or other dental work?
1 During the past 12 months
2 Between 12 and 24 months ago
3 More than 24 months ago
4 Never
5 Not sure
Missing

Q95 How often do you wear sunscreen or sunblock with an SPF of 15 or higher when you are outside for more than one hour on a sunny day?
1 Never
2 Rarely
3 Sometimes
4 Most of the time
5 Always
Missing

‡Please note that although most of the above variables are in both datasets, Q92–95 are exclusively in the Teacher Dataset.

GREG Geographic Region
1 Northeast
2 Midwest
3 South
4 West

METROST Metropolitan Status
0 Unknown
1 Urban
2 Suburban
3 Rural

Appendix 2: Interpreting Chi-Square—A Quick Guide for Teachers

For many investigators the excitement of research is a combination of a joy derived from creating new knowledge in their field, from interacting with people when taking surveys, and in the field of epidemiology, from improving the health of the public. However, that excitement is somewhat subdued when it comes to the actual data analysis. Fortunately we now have computers and calculators to do the drudgery of calculation. Unfortunately there still is that part about understanding the computer output—the statistical stuff.

We would like to present a brief guide to understanding the computer output from analyzing surveys, and a lot of assurance that with a little practice, interpretation not only will be less threatening but will become a minor part of any investigation. Interpreting survey data or, for that matter, all data is a mixture of art, science, wisdom and experience. Interpreting the computer output is just a case of knowing what to look for and what to ignore. With this short introduction, we will try to help separate the wheat from the chaff and help you interpret the wheat. It will not be possible to teach you all about the Chi-square statistic—we will give you some Web sites for ready browsing—but we hope to lessen the statistics anxiety a bit.

The very first thing you need to know is that you don't need to know everything! The computer doesn't really know your level of expertise, so it spits out everything, under the tenuous assumption the reader is a professional statistician or epidemiologist. Most of it—trust us—can be safely ignored. Let's consider the *Epi Info* computer output from the sports injury question in the module. Those parts of the computer output that are important for interpreting our 2 $\times$ 2 surveys are printed in bold. (You will be pleasantly surprised to have to search a bit for the bold print.)

Single Table Analysis

	Point Estimate	95% Confidence Interval Lower	Upper
PARAMETERS: Odds-based			
Odds Ratio (cross product)	0.8098	0.7288	0.8999 (T)
Odds Ratio (MLE)	0.8098	0.7287	0.8998 (M)
		0.7277	0.9011 (F)
PARAMETERS: Risk-based			
Risk Ratio (RR)	0.8468	0.7792	0.9204 (T)
Risk Difference (RD%)	-3.5205	-5.2721	-1.7689 (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	15.4091		0.0000877415
Chi square-Mantel-Haenszel	15.4072		0.0000878261
Chi square-corrected (Yates)	15.1998		0.0000978848
Mid-p exact		0.0000426	
Fisher exact		0.0000474	

Just in case you aren't quite sure your eyes are finding the correct bold print, let's pull out the critical information that beginners would need to pay attention to:

STATISTICAL TESTS	**Chi-square**	**1-tailed p**	**2-tailed p**
Chi square-uncorrected	**15.4091**		**0.0000877415**

Analyzing the output from statistical hypothesis testing really breaks down into three considerations:

1. If my null hypothesis is correct, what sort of Chi-square statistic should I see?
2. What sort of evidence counts against my null hypothesis?
3. How much evidence is enough evidence to reject my null hypothesis?

The answer to the first question is that it depends. If the only tables you analyze are 2×2 tables, then the answer is this: If your null hypothesis is correct, you should expect to see Chi-square statistics close to 1.0. The actual number will fluctuate slightly from sample to sample but will not be very far from 1.0. (For tables different from 2×2, your expectation for the

Chi-square value will be different. Before analyzing survey data with responses different from yes or no, consult an elementary statistics book or the Web sites listed below.)

The answer to the second question works for all tables, not just 2×2 ones. Recall that we generally are asking a question about whether two variables are associated. Our null hypothesis is that the variables are not associated. In our example in the module the null hypothesis is that there is no relationship between gender and sports injury. In this statistical test we are looking for any evidence that this null hypothesis is inconsistent with reality. The Chi-square statistic is a measure of this difference between hypothesis and reality (as represented by our data). A Chi-square value of 0.0 would theoretically indicate a perfect match, but this never occurs in real life.

Although it is possible to get values for Chi-square between 0.0 and 1.0, such values are rare. For the most part, numbers larger than 1.0 will count as evidence against the null hypothesis: The larger the number, the more evidence you have against the null hypothesis. This happens because, to repeat, the Chi-square statistic is essentially a measure of mismatch between your actual data and what you would expect to see if your null hypothesis were true. A certain amount of discrepancy between theory and data is tolerated because of the vagaries of sampling. but as the Chi-square statistic gets larger, this is treated as an indication of more and more of a dissonance between what you expect to see when a null hypothesis is true and what you are seeing in the data.

Now for the last question—how much evidence is enough? How big a discrepancy can be tolerated before one is suspicious that the null hypothesis is false? There is no single answer to this question. Some researchers are more tolerant that others. However, researchers and statisticians are in general agreement on how to easily interpret the amount of discrepancy and what levels of tolerance are commonly used. The measure of discrepancy typically used is called a p-value and is reported in the computer output as a 2-tailed p. (The reason for that name will be clear to those who have had some inferential statistics, but it is not necessary to go into that—just remember that the p-values are what you are looking for.) The p-value is actually a probability and is technically defined as follows:

> The p-value is the probability that were a null hypothesis true, one would observe a test statistic value at least as inconsistent with the null hypothesis as what actually resulted.

For our purposes in a 2×2 table, the p-value is the answer to this question: If the two variables I'm interested in (gender and sports injury) are really not associated, what's the probability I'd get a Chi-square statistic this large? A p-value of 0.05 says, “Gee—if my null hypothesis (of no association) were true, I would get this large a value for Chi-square only 5% of the time."

The usual suspects, that is, the levels of suspicion tolerated before rejecting the null hypothesis, are called levels of significance. The commonly accepted levels of significance are

0.10, 0.05 and 0.01, with 0.05 winning most of the time by default. The levels of significance and the Chi-square values associated with them for a 2×2 table are presented below:

Chi-Square Statistics and Their Associated p-Values for a 2×2 Table

Chi-Square Value	p-value
2.70	.10
3.84	.05
6.63	.01

With this in mind, we can interpret the Chi-square as large enough to engender suspicion about the null hypothesis or the p-value as small enough to engender suspicion. Whichever we prefer, we are thereby regarding our data as too unlikely to occur if the null hypothesis is true. So in our example we have a Chi-square value of 15.4 with a 2-tailed p-value of 0.000088. We have a very large Chi-square value and a very small p-value, which tells us that if my null hypothesis were true, i.e., if gender is not related to sports injury, I would get a 15.4 value only 0.008% of the time, which is pretty unlikely indeed. So we feel comfortable rejecting the null hypothesis and claiming that we have evidence for a relationship between gender and sports injury.

We hope this quick guide has been helpful as you wade through the computer output for survey analysis. There are some nice Web sites with information about the Chi-square statistic, presented at an elementary level, so you don't have to be a math major.

Here they are:

Georgetown University Web site. Chi-Square Tutorial page. Available at: http://www.georgetown.edu/faculty/ballc/webtools/web,chi,tut.html

Office for Mathematics, Science and Technology Education, University of Illinois at Urbana-Champaign. Web site. Chi-Square page. Available at: http://www.mste.uiuc.edu/patel/chi-square/intro.html

Hyper Stat Online Web site. Chi-Square page. Available at: http://davidmlane.com/hyperstat/chi-square.html

For those whose preference is for books, we recommend those listed below. They are both well written and generally nonmathematical.

Peck R, Olsen C, Devore JL. *Introduction to Statistics and Data Analysis*. With CD-ROM. Pacific Grove, CA: Duxbury Press; 2001.

Yates D, Moore DS, Starnes DS. *The Practice of Statistics*. 2nd ed. New York: WH Freeman; 2003.