Interpreting Categorical Data

Unit Overview

This unit is an introduction to the analysis of categorical data, that is, data that can be classified into categories. A summary statistic is the frequency (or count) or proportion (or percentage) of the sample or group that falls into each category. For example, you might categorize all voters in the United States as Republicans, Democrats, or other. Then, you could give the number of voters or the proportion of voters who fall into each category.

The basic mathematical tool for the analysis of categorical data is proportional reasoning. Skill with proportional reasoning is crucial for college-bound students and they will receive a lot of practice with proportional reasoning and the use of percentages in this unit. While students have already learned basic computational procedures, such as converting a proportion to a percentage, this unit requires them to decide which procedures are necessary in real-life, non-trivial, situations. They will practice reading reports that present comparisons using proportions, percentages, and other statistics. This unit provides background and practice to help students understand such reports.

Lesson 1 Comparing the Risk

Lessons 1 and 2 involve the comparison of two different groups. For example, two random samples, one of 100 voters under age 30 and another of 100 voters age 30 and older might be categorized according to political party preference. (Statisticians sometimes use the phrase categorized on political party preference.) The age of the voters is called the explanatory variable, while political party preference is called the response variable. The data can be summarized in a two-way frequency table like that below, which would contain the counts of younger and older voters who chose each party preference.

<table>
<thead>
<tr>
<th>Political Party Preference</th>
<th>Under Age 30</th>
<th>Age 30 and Older</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democratic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Republican</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent or Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Students will learn to use such frequency tables and also bar graphs to compare two groups. Many examples in this lesson compare the consequences for those who engage in various risky behaviors and those who do not.

Lesson 2 A Test of Significance

In Lesson 2, students learn the chi-square test of homogeneity. First, they learn what homogeneous groups look like and how to compute the expected frequency in each category under the assumption of homogeneity. For the example above, the statistical inference question will be, “From the observed data, is it plausible that, if we had asked all voters in each age group, the proportions who choose each party preference would be the same? Or, do we have statistically significant evidence that the preferences would be different for the two age groups of voters?”
Lesson 3 The Relationship Between Two Variables

Lesson 3 involves one group or sample, sorted on two different variables. For example, a random sample of 200 voters might be categorized according to both age and political party preference. The two-way frequency table on the following page shows how the data might be summarized. Note that, in the case of a single sample, both row and column totals (called marginal totals) are meaningful. The total number of voters preferring the Democratic party, for example, can be used to compute an estimate of the proportion of all voters who prefer the Democratic party. That is not the case in the table above because the sample sizes do not necessarily reflect the proportion of younger and older voters in the population of all voters. (There are far more voters in the age group 30 and older than in the age group of under 30.) The statistical inference question in Lesson 3 will be, “From the observed data, is it plausible that, if we had asked all voters, the two variables of age and party preference would be independent? Or, do we have statistically significant evidence that the variables would be associated?”

<table>
<thead>
<tr>
<th>Political Party Preference</th>
<th>Under Age 30</th>
<th>Age 30 and Older</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democratic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Republican</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent or Other</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td>200</td>
</tr>
</tbody>
</table>

If it seems like the two situations above are very similar, they are. The mechanics of the chi-square tests used to analyze them are identical. The difference is in how the data are collected: is there one sample or are there two? And the method of data collection affects the conclusions we can draw, even though the mechanics are the same.

Unit 5, Binomial Distributions and Statistical Inference

In the next statistics unit in this course, there will be only one group, sorted on one categorical variable. The variable will have only two categories, such as yes/no, success/failure, or lived/died. The number of people in the group that fall into the first category will be counted. A typical question will be, “In a random sample of 100 voters from a state in which 43% of the voters are Democrats, what is the probability of getting at least 50 Democrats? A question involving statistical inference will be, “If we take a random sample of 100 voters from North Carolina and get 50 Democrats, is it plausible that the percentage of all Democrats amongst North Carolina voters is 43%?”

**Unit Objectives**

- Understand terminology used in medicine to compare the risk of some condition between two groups of people and understand the terminology used to describe the statistical characteristics of a diagnostic or screening test
- Understand the characteristics of a well-designed experiment
- Review conditional probability and the concept of independent events
- Conduct and interpret a chi-square test of homogeneity
- Conduct and interpret a chi-square test of independence
Developing a Collaborative Classroom

*Transition to College Mathematics and Statistics (TCMS)* has been designed to support learning through interactive problem-based investigations. As you help your students get to know each other in the first few days of the school year, you should develop collaborative group behavior guidelines similar to the ones below. (See *Unit Resource Masters* pages 1 and 2 for a sample set of group guidelines and a template for students to self-assess the effectiveness of their collaboration.)

- Each member contributes to the group’s work.
- Each member of the group is responsible for listening carefully when another group member is talking.
- Each member of the group has the responsibility and the right to ask questions.
- Each group member should help others in the group when asked.
- Each member of the group should be considerate and encouraging.
- All group members should work together until everyone understands and can explain the group’s results.

One way to help students work together effectively is to choose a collaborative skill on which to focus for a particular investigation. Introduce the skill to the class by discussing what that skill might look and sound like. Once students have worked in their groups, provide a collaborative processing prompt that leads to a full-class discussion of the particular skill. (See pages T3E and T8 for sample skill and processing prompts.) Suggestions for skills and the corresponding collaboration processing prompts are often provided in the *Teacher’s Guide*. Holding students accountable to established group norms and giving them positive feedback regarding their group’s interaction is particularly important at the early stages of developing effective groups. The *Unit 1 Resource Masters* include a student self-assessment form.

See the front matter of this text for more information on developing student-centered classrooms that embody the CCSS Mathematical Practices.

Access and Equity Matters

Research by Jo Boaler (Boaler, J. 1998, 2000, 2002) and by the Quasar Project (Brown, C. A. Stein, M. K. and Favar, E. A., 1996) has identified particular teaching practices with curricula such as *TCMS* that promote greater access and equity in mathematics classrooms.

These practices include:

- introducing activities through class discussion.
- teaching students to explain and justify.
- making real-world contexts accessible by discussing the contexts.

*TCMS* offers many opportunities for teachers to incorporate these practices into daily routines. The *Think About This Situation* is one built-in opportunity to introduce investigations through discussion. You may wish to begin investigations in a similar way. Since much of the mathematical content is based in real-world contexts, it is important that all students understand the contexts and draw on their own or a classmate’s background knowledge when possible. Opportunities for students to explain and justify their thinking are built into all curriculum features. Encourage the habits of mind of explaining thinking and justifying claims.
The Promoting Mathematical Discourse (PMD) scenarios for this unit are written to incorporate classroom norms and sociomathematical norms. These norms are being developed, so the PMDs show this development by both the teacher and students. These scenarios can be used for teacher discussion purposes and should be interpreted as realistic discussions, not perfectly executed discussions. You may wish to compare the two STM scenarios to see this development.

Common Core State Standards for Mathematics and Mathematical Practices

TCMS is designed so that students engage in the mathematical behaviors identified in the CCSS Mathematical Practices as the primary vehicle for learning the mathematics and statistics elaborated in the CCSS content standards. See the front matter of this Teacher’s Guide for more information related to the CCSS and also the Common Core State Standards for English Language Arts and Literacy.

**TCMS-Tools**

TCMS-Tools is a suite of Java-based mathematical software specifically designed to support student learning and problem solving in each strand of. The software includes four families of programs: *Algebra* (spreadsheet and CAS), *Geometry* (coordinate, synthetic, and vector), *Statistics*, and *Discrete Math*. Each content area includes specific custom apps developed for the TCMS course. Data sets for selected problems are included in the course-specific software.

Features of the statistics software particularly useful for this unit are the stacked bar graphs and the chi-square test. Data sets embedded in TCMS-Tools for this unit are:

**Lesson 2:**
- Poll I and Poll II p. 46
- Bicycle Helmet p. 49
- Fashion p. 50
- Crying p. 53
- SAT Math Scores p. 56
- Artificial Turf p. 57
- Satisfaction with Life p. 58
- Hispanic Origins p. 60
- Poll A and Poll B p. 61

**Lesson 3:**
- Baseball p. 83
- Laptops p. 84
- IBS and Anxiety p. 89
- North Carolina Traffic Accidents p. 89
- Virginia Traffic Crashes p. 93
- Seasons p. 94
- Roofing Invoices p. 97

**Lesson 4:**
- Veterans and PTSD p. 101

**CCSS Mathematical Practice: Use Appropriate Tools Strategically**

By downloading the TCMS-Tools software to school and home computers, students will have many opportunities to develop the mathematical practice of using appropriate tools strategically. This includes selecting and using technology tools to explore and deepen understanding of concepts, to
visualize mathematical ideas and view results of varying assumptions, to model and solve problems, to compare predictions with data, and to become sufficiently familiar with technology tools to make sound decisions about when various features of the software might be helpful.

**Synthesizing and Summarizing Activities**

Students should create a Mathematics Toolkit that organizes important class-generated ideas, selected Summarize the Mathematics (STM) responses, and topic summary templates as they complete the investigations. Unit summary templates should be added to the toolkit when it is completed following the “Looking Back” lesson. Above all, the Math Toolkit should be useful to students. Thus, often offer students the option to include in their toolkit what they consider valuable. Exercising this decision will help them prepare for post-secondary studies. You may wish to discuss this with your class after the STM discussion.

**Review and Practice**

*TCMS* includes review tasks in the homework sets. The purpose of the review tasks is two-fold. Some tasks are just-in-time review of skills needed in the following lesson. These tasks will be designated by a clock icon near the solution. Some tasks provide distributed practice of mathematical skills to maintain procedural fluency. These tasks should be completed outside of class by students. If a few students are identified as needing additional assistance with specific skills, they should be given additional assistance outside of class.

Distributed review tasks that review the basics of trigonometry, logarithms, and linear and exponential functions, although not marked as just-in-time tasks, are preparation for work in Unit 2, *Functions Modeling Change*. If you decide to omit some investigations in this unit, you should still assign these types review tasks from the On Your Own (OYO) set.

**Writing Thorough Responses**

Helping students develop skills in writing complete and concise solutions is one of the goals of this program. However, always writing thorough responses can unnecessarily slow student progress through the investigations. As a guideline, we suggest that during investigations, students should make notes of their thinking and discussion of ideas rather than use complete sentences. Investigation time can be thought of as draft work or getting ideas out for discussion. For investigation problems that ask students to explain reasoning or to compare, you may want to require complete-sentence responses. Student responses to the Summarize the Mathematics and Math Toolkit entries should be more complete. If these responses are written following the class summary of important mathematical ideas, students will be able to write more thorough responses. Homework tasks from the On Your Own sets, particularly Connections tasks should also be thoroughly written. For some OYO tasks, thorough responses should be reviewed following class discussion of the task to ensure correct and secure understanding of the ideas.

**College Readiness Assessment**

Opportunities for additional review and practice are provided in the College Readiness Assessment (CRA) sets included in the *Unit Resource Masters*. Each College Readiness Assessment set presents 10 questions in the form of test items similar to how they often appear on college placement tests. By using these practice sets, students can become familiar with these types of questions and develop effective test-taking strategies for performing well on such tests. More information on this feature is on page xxv of the front matter. The first CRA set for this unit follows Lesson 1 in the *Unit Resource Masters* (pages 15–16).
# Unit 1 Planning Guide

<table>
<thead>
<tr>
<th>Lesson Objectives</th>
<th>On Your Own Assignments*</th>
<th>Suggested Pacing**</th>
<th>Resources</th>
</tr>
</thead>
</table>
| **Lesson 1  Comparing the Risk**  
- Understand terms used to compare risk: absolute risk reduction, relative risk  
- Distinguish between explanatory and response variables  
- Interpret bar graphs displaying categorical data  
- Understand the limitations of anecdotal evidence and the possibility of lurking variables  
- Learn the characteristics of a well-designed experiment, especially the role of randomization  
- Understand how to design an experiment to account for the placebo effect, including single and double blinding  
  
  **After Investigation 1:**  
  A1 or A2, C8, R14, E19, Rv23–Rv25  
  **Note:** Rv23 and Rv24 are just-in-time review for Lesson 1. Assign early.  
  **After Investigation 2:**  
  A3, A4, choose two of C9–C11, R15, R16, E20, Rv26–Rv28  
  **After Investigation 3:**  
  A5–A7, C12, C13, R17 or R18, E21 or E22, Rv29–Rv30  
  
  **9 days** (including assessment)  
  **CCSS NOTE**  
  See the note on page T14 regarding topics in Inv. 3.  
  **• Protractor for Investigation 1, Problem 8 Part c  
  • Unit Resource Masters  
  • CRA Set 1  
  • TCMS-Tools software** |
| **Lesson 2  A Test of Significance**  
- Understand what it means for groups to be homogeneous  
- Distinguish between a population and a sample  
- Compute expected frequencies using marginal totals and proportional reasoning  
- Understand that expected frequencies show what homogeneous samples would have looked like  
- Compute the chi-square statistic and understand that the more different the two samples, the larger it will be  
- Determine if the chi-square statistic is statistically significant and write a conclusion  
  
  **After Investigation 1:**  
  A1, choose one of A2–A4, R16, R17, Rv24–Rv27  
  **After Investigation 2:**  
  A5, A6, C12–C14, R18, Rv28–Rv30  
  **After Investigation 3:**  
  Choose one of A7–A9, A10 or A11, C15, R19, E20, choose one of E21–E23, Rv31, Rv32  
  **10 days** (including assessment)  
  **• Unit Resource Masters  
  • CRA Set 2  
  • TCMS-Tools software** |
| **Lesson 3  The Relationship Between Two Variables**  
- Learn the meaning of the terms false positive and false negative in diagnostic testing  
- Understand the four summary statistics that can be used to describe how well a test performs: sensitivity, specificity, positive predictive value, and negative predictive value  
- Review the concepts of conditional probability and independent events  
- Distinguish between situations calling for a chi-square test of homogeneity and a chi-square test of independence  
- Compute expected frequencies for a chi-square test of independence using the definition of independent events and then compute the chi-square statistic  
- Determine if the chi-square statistic is statistically significant and write a conclusion  
  
  **After Investigation 1:**  
  A1, A2 or A3, C8, R12, Rv20–Rv23  
  **After Investigation 2:**  
  A4, A5, C9, C10, R13, E15, Rv24, Rv25  
  **After Investigation 3:**  
  A6, A7, C11, R14, E16 and E17 or E18 and E19, Rv26, Rv27  
  **8 days** (including assessment)  
  **NOTE Inv. 1 may be omitted without loss of continuity.**  
  **• Unit Resource Masters  
  • CRA Set 3  
  • TCMS-Tools software** |
| **Lesson 4  Looking Back**  
- Review and synthesize the major objectives of the unit  
  
  **Unit Summary Template**  
  **2 days** (including assessment)  
  **• Unit Resource Masters** |

*When choice is indicated, it is important to leave the choice to the student.*  
*Note: It is best if Connections tasks are discussed as a whole class after they have been assigned as homework.*  
**Pacing assumes technology access for homework tasks**
LESSON 1 Comparing the Risk

Teens and Risky Behavior Teens often engage in risky behavior. So much so, that the U.S. Centers for Disease Control and Prevention (CDC) have an ongoing system for monitoring such behavior, called the Youth Risk Behavior Surveillance System. Several states and even individual high schools have instituted their own surveys, based on the one from the CDC. For more information, see www.cdc.gov/healthyyouth/yrbs/.

This lesson gives teens statistical tools useful in thinking about the possible consequences of engaging in risky behavior. While presenting teens with the hard facts does not always, or even usually, change their behavior, it will in some cases and will increasingly as they mature.

Further, your students soon will be 18 years old, if they are not already. At this age, they will be responsible for making their own medical decisions. This unit contains two investigations about medical statistics. In Investigation 1 of Lesson 1, students learn how to quantify the difference in the risk from two behaviors. Then, in Investigation 1 of Lesson 3 they learn that medical tests typically do not diagnose a condition with perfect accuracy and so a patient must be aware of the statistical characteristics of the medical test they are considering taking.

The Content of this Lesson The three investigations in this lesson contain the descriptive statistics used, especially in medical fields, to compare two groups categorized on the same variable. In the first investigation, students learn the term absolute risk (sometimes called prevalence or incidence) and compute it to compare the incidence of some characteristic in two different groups. They compare how well bar graphs and pie charts display categorical data. In the second investigation, students examine how reports in the media use the terms absolute risk reduction and relative risk to compare the incidence of some condition in two different groups. In Investigation 3, students learn how to design an experiment to establish the effect of a particular treatment on some condition. This lesson helps prepare students for statistical inference in Lesson 2. There they will use the chi-square test of homogeneity to test the difference of two groups.

Students learn the following terminology in the three investigations.

I1 “Summarizing and Displaying the Risk”
- categorical data, absolute risk, explanatory and response variables, bar graphs, stacked (segmented) and grouped bar graphs, two-way frequency tables, pie charts

I2 “Comparing Risk”
- prevalence, absolute risk reduction, relative risk, side effects, clinical trial, placebo
I3 “Design of Experiments”

observational study, anecdotal evidence, experiment, subjects,
characteristics of a well-designed experiment, treatments, control group,
placebo, placebo effect, single blind, double blind, lurking variable

Proportions Versus Percentages and the Issue of Rounding
College-bound students need fluidity with both proportions and percentages, so sometimes problems in this unit will ask for the answer as a proportion and sometimes for the answer as a percent. Because they are interchangeable, there seldom is a reason to prefer one over the other except that people typically feel more comfortable with percentages.

Throughout this unit, students are likely to ask how to round proportions and percentages. There is no easy answer to what is best, because the number of digits to report depends on the situation. As usual, no rounding should be done until the final answer. Generally, solutions will be rounded to four decimal places. For example, \(\frac{2}{3}\) will be rounded to 0.6667, or 66.67%. It may be best to ask students to use that same rule, that all proportions and percentages should be rounded to four decimal places. The exceptions are when they are comparing percents, or writing a summary for the general public. In that case, too many decimal places tends to be confusing, and percentages are typically rounded to the nearest whole percent.

Lesson Objectives
- Understand terms used to compare risk: absolute risk reduction, relative risk
- Distinguish between explanatory and response variables
- Interpret bar graphs displaying categorical data
- Understand the limitations of anecdotal evidence and the possibility of lurking variables
- Learn the characteristics of a well-designed experiment, especially the role of randomization
- Understand how to design an experiment to account for the placebo effect, including single and double blinding

Lesson Launch
This Think About This Situation engages students in thinking about how to quantify the possible consequences of engaging in a risky behavior. When discussing Part f, you may wish to ask students how they would respond to a friend who says that his grandfather smoked and lived to be 90. Have students discuss how best to quantify the risk involved with smoking. That is, what statistic do they think would most impress students their age? Note that rounding makes quite a difference here. In Part c, if students rounded the proportions to 0.17 and 0.01, they would get 17 times as likely rather than about 13 times as likely. Hence the direction not to round before the final answer.
Think About This Situation

(a) A smoker is more likely, as \(\frac{283,000}{1,645,000} \approx 0.1720\), or about 17% of smokers eventually will get lung cancer, while \(\frac{88,000}{6,748,000} \approx 0.0130\), or only about 1% of the non-smokers eventually will get lung cancer.

(b) Yes. The chance that a boy who gets lung cancer is a smoker is \(\frac{283,000}{283,000 + 88,000} \approx 0.7628\) and the chance that a boy who gets lung cancer is a non-smoker is \(\frac{88,000}{283,000 + 88,000} \approx 0.2372\). So, about 76% of the cases of lung cancer happen among smokers, while 24% happen among the much larger group of non-smokers.

(c) \(\frac{0.1720}{0.0130} \approx 13.2308\) times as likely (Answers may vary due to rounding.)

(d) If none of the smokers smoke, the estimate of the number who would get lung cancer anyway is \((0.0130)(1,645,000) = 21,385\). So, the number of cases prevented would be \(283,000 – 21,385 = 261,615\).

(e) People mean that smokers are more likely to get lung cancer than non-smokers, all other things (such as genetics, diet, exercise, exposure to toxins) being equal. The evidence here is in favor of that because a smoker has a 0.1720 chance of getting lung cancer while a non-smoker has only a 0.0130 chance. However, we do not know if all other things are equal. For example, the smokers might get lung cancer at a greater rate because they tend to exercise less and drink more alcohol.

(f) There always are going to be exceptions to any general rule about human beings. People who concentrate on the exceptions are likely to ignore the overall high risk of smoking.
Teacher: Please open your math book to the Table of Contents pages. We will take a few minutes for an overview of this course and how it is similar to and different from other math courses you have taken during high school. (The class discussion revolves around the mathematics and statistics content in the course. Then the teacher has students look over the format of Unit 1 Lesson 1, discussing the instructional model that will be used for the course. Students participate in setting some classroom social norms for whole class and group discussions similar to those in the Collaborative Group Guidelines master in the Unit Resource Masters. The teacher collects student recommendations on a poster and organizes students to work in pairs on the first investigation. Each pair will be assigned another pair to confer with while working on the investigation.)

Teacher: Now that we have some initial ideas of our classroom expectations for this course, let’s look at the opening page for the first unit, Interpreting Categorical Data. Sometimes I will ask you to read this material as homework prior to beginning a unit and other times—like today, since it is the first day—we will read this material together. (The teacher asks for volunteers to read the text, including each of the lesson descriptions, on page 1.)

What experiences, either inside or outside of school, have you had thinking about the statistics we will be studying in this unit?

Dominick: I’ve never studied anything like this before. I really don’t know what those lesson descriptions mean. I did get that we will be learning how to decide how risky something is. And have some ideas for making decisions about medical treatments.

Margo: I agree with Dominick. But in my psychology class last year we talked about randomized experiments, so I know a little about that, but not really much.

Krystiana: We had some discussion of medical treatments at our house because my grandfather had skin cancer. He had to decide between two treatments. There were some numbers discussed like: 60% chance of being cured with one treatment and 70% with another. I am not sure if this unit will be about these types of decisions though. It does sound interesting—at least more interesting than solving algebra equations.

Teacher: Thanks for your participation. You might want to know that some of the statistics in this unit were also fairly new for me. I did study some statistics like this in college, but that was a long time ago. When preparing to teach this unit, I spent quite a bit of time working out problems this summer. So, in some respects, we will be learning this material together. Please feel free to ask lots of questions. That is how we will all learn more about how to assess risk and test for statistical significance.

Let’s continue examining the two-way frequency table relating smoking and lung cancer on page 2. Keep your book open to this page so you can see the table. Our discussion questions will come from the Think About This Situation on page 3. Those questions are also displayed on the board. Please get your calculators out so that if you wish, you can use them during our discussion.

Using the data in the table, is a smoker or a non-smoker more likely to get lung cancer? Discuss this with the classmate you are paired with. (The teacher gives a couple of minutes and listens to their thinking and ways to build on their different thinking—in voluntarily not taking questions from students at this time. He is planning to give students time to discuss Parts a and b to seed the whole-class discussion. Some students will discuss other parts.)

Okay, Kelsey, what were you and John discussing about the question in Part a? Is a smoker or a non-smoker more likely to get lung cancer? (This pair is selected to report because they did not use proportions or percents to compare. This choice of students will provide an opportunity for others to comment on their thinking.)
Kelsey: Well, first we were surprised that there were 88,000 boys who eventually got lung cancer who did not smoke. But then we noticed that about three times as many boys got lung cancer who did smoke cigarettes. So, we decided that a smoker is much more likely to get cancer than a non-smoker.

Teacher: Does anyone want to add a comment or ask Kelsey or John a question related to their ideas? Corina?

Corina: Liam and I also thought that a lot of boys got lung cancer who did not smoke. But we noticed that there were many boys who did not smoke—almost 7 million—many more than those who did smoke. So, the 88,000 was really out of 6,748,000. So, the proportion of boys who got lung cancer even though they did not smoke was only about 0.01

Teacher: Does anyone want to add a comment or ask a follow-up question?

Eric: Where did you get 0.01?

Liam: We got that from 88,000 divided by 6,748,000. It came out to be 0.013. This is like saying the proportion of non-smokers who got cancer or as the question asks, how likely is it that a non-smoker gets cancer? Not likely at all.

Teacher: To be clear about our expectations during whole-class discussions, this question by Eric and response by Liam is an example of how you should feel free to ask each other questions. Also, after someone answers your question, you should tell them whether you understand now or you need more explanation. In fact, any time someone says something that you do not understand, please ask for more explanation. We will see that sometimes explanations need refinement or changes to make them correct and clear. When you ask each other questions it should be done respectfully and not be taken as criticism of a person. We are focused on mathematical ideas. This process will help us become better at explaining our thinking or justifying our reasoning throughout the school year. So, let’s try that here. Eric, what is your response to Liam’s explanation? (Eric indicates he understands.) Okay, are there any remaining things unclear about the proportion of non-smokers who got lung cancer? (Students indicate that there are no more questions.) Okay then, let’s go back to the question in Part a. Have we completed our discussion of this question?

Ana: We have not really compared to the proportion of smokers who get cancer yet. We found that about 0.17 of smokers eventually get lung cancer. This number is much larger than 0.01. Dana said that it is easier for her to see the size difference in this number if we use percents. So, we decided that we were comparing about 17% to about 1%. (The teacher waits for comments/questions from other students. He notices that some students are using their calculators.)

Teacher: Well, no one is commenting on Ana’s explanation. Does that mean that you agree with her comments, or have you not processed what she said?

Kelsey: We did not do the proportion before. Remember, we said there were three times as many smokers who got cancer. So, I just checked the proportion that Ana said and also found that 283,000 out of 1,645,000 is about 17%. So, I agree with Ana.

Teacher: Are we ready to move on to Part b then? (Students indicate they are.) Today we are spending extra time on this Think About This Discussion (TATS) in order to set our classroom expectations. Once we understand how our classroom conversations will go, we will not need to have classroom expectation discussions with each TATS. Another of our classroom expectations is that you will explain your reasoning and support your claims so that others do not need to ask you to do so after you make your comments. Often this is mentioned in the student text as with Parts a and b in this TATS. But even if it is not explicitly mentioned in the text, you should justify your reasoning and/or calculations.

Let’s see if we can practice our whole-class discussion expectations while offering thoughts on Part b. Is a boy who gets lung cancer more likely to be a smoker than is a boy who does not get lung cancer? Donell? What are your and Tanya’s thoughts? (This pair has been selected because the teacher observed that they did think about the proportions correctly. He wants to set up the opportunity for other pairs that did not analyze this correctly to ask follow-up questions.)
Donell: We looked at the table and decided that we needed to know the total number of boys who got lung cancer. So, we added 283,000 and 88,000 to get the total with lung cancer, 371,000. Then we found the ratio of the smokers who got lung cancer out of all who got lung cancer: 283,000 out of 371,000. The ratio was about 0.76, or 76%. The ratio for the non-smokers who got lung cancer was 88,000 over 371,000, or about 24%.

Tanya: But after we did the second ratio, we noticed that we could have just done 100% minus 76% to get the 24%.

Serena: But why does that work, Tanya?

Tanya: I’m not sure why. We just noticed it.

Teacher: Was this just a coincidence? Or will it work in other similar cases? Anyone have ideas to put on the table about this? Serena?

Serena: Now that I thought about it, it seems to me that our whole group, or 100% of the boys who got lung cancer, is the sum of those who did and did not smoke. We are only looking at the first row of the table for this question. That is why we can just subtract 76% from 100% to get the non-smokers who got lung cancer—24%. There are no other parts that make up the 100% here. (The teacher notices that many students are nodding in agreement and decides to move on since students will have more opportunities to think about two-way frequency tables and their computations in this lesson.)

Teacher: Some of you did not yet get a chance to discuss Part c. But even if you did not, based on our discussion, what number should go in the blank for Part c? A male smoker is _______ times more likely to get lung cancer as is a male non-smoker?

Jung-Chen: It looks like about three times as many because 76% is a little more than three times 24%.

Teacher: (There seems to be much acceptance of this answer from student’s facial expressions and nodding.) Let’s attach a little more language to these numbers to help us understand what they mean. What does 76% represent?

Robert: Of those who get lung cancer, 76% were smokers.

Teacher: So, is that what we are asked to compare in Part c? Robert?

Robert: No, this is asking the same question as Part a, in a sense. It is asking us to see how many times more likely a smoker is to get cancer than a non-smoker. We found in Part a that the two proportions were about 17% and 1%. A smoker is about 17 times more likely to get cancer than a non-smoker. This question involves using the columns of the table. (The teacher decides not to bring up the issue of rounding during this TATS, but plans to do so when students report out different results due to different rounding in the investigation.)

Jennifer: This is confusing.

Teacher: Yes, we will be looking at many more studies in this unit and working toward understanding what questions are asking and which data to analyze from the tables or information in a report. It is fine to be a little confused at this point. The important thing to understand here is that you need to read tables carefully and discuss what questions are asking, not just pull some numbers out of the table and compute with them. The contexts should help you think about what numbers to use.

So, let’s move on and consider Part d together since almost no pairs have yet discussed this question. How many cases of lung cancer would be prevented if none of these boys smoked? Let’s talk about what this means. Who has an idea to start us off? Maitas, thanks for volunteering.

Maitas: Well, even when they did not smoke, 1% of the boys got cancer anyway. So, we should expect that there still would be 1% of the 1,645,000 smokers—now called non-smokers—who would get cancer. So, about 16,450 rather than the 283,000 in the table would get lung cancer.

Helen: Okay, that makes sense. So, the number with lung cancer would decline by 283,000 minus 16,450 which is … 266,550 fewer lung cancer cases.

Teacher: What exactly do people mean when they say that smoking causes lung cancer? Support your claim by information in the table. Who would like to respond to this question? Thanks for volunteering, Arianna.

Arianna: People mean that smokers are more likely to get lung cancer than non-smokers. The information in the table lead us to say that smokers were about 17 times more likely to get cancer.
**Teacher:** When people say this, they are really making an assumption about the two groups, smokers and non-smokers. What might that assumption be? Leonardo?

**Leonardo:** They are probably thinking that the two groups are only different in terms of smoking or not. But that may not be the case. It may be that smokers also do exercise less than non-smokers and so their lungs are not as healthy. Like we talked about before. The boys in this experiment were not randomly assigned to smoke or not to smoke. So, there may be other things, maybe in addition to smoking, that make the higher rates of lung cancer.

**Teacher:** Are there other thoughts? *(None are offered. Students seem to agree with Leonardo's comments.)* Will someone offer to read Part f of the TATS? Thanks, Chisom.

**Chisom:** *(She reads the question.)*

**Teacher:** Thank you. Who would like to comment? Thanks for offering, Osel.

**Osel:** You really need to know the general trend. There will always be exceptions. Look at the table. There were over a million smokers who did not get lung cancer.

**Chisom:** That is a lot of exceptions.

**Teacher:** You really listened well to each other's ideas and contributed to the discussion by clarifying and asking questions. I think you are already getting a sense of how whole-class discussions will proceed in this class. Now we will begin our first investigation. Please turn to page 3 in your textbook.
Summarizing and Displaying the Risk

In this investigation, students practice computing the absolute risk that a member of some group has a condition. Other words that often are substituted for absolute risk are prevalence and incidence. All give you the proportion or percentage of people in a group who have or are expected to have the condition.

POSSIBLE ROUNGING ERROR In general, students tend to round too much and to be careless about the last digit. For example, it is not unusual for students to round \( \frac{2}{3} \) to 0.7 so that \( \frac{2}{3} \cdot \frac{2}{3} \) becomes 0.49 rather than 0.4444. Encourage them to carry all digits in intermediate steps and to round only at the end. In these solutions, percentages and proportions generally will be rounded to four significant digits. For example, \( \frac{2}{3} \) will be given as 0.6667, or 66.67%.

Launch

To convince students of the importance of learning the terminology in this investigation, have them enter the term “absolute risk” in a news search engine such as Google News or Yahoo! News. Ask students to make a list of the topics that come up and then discuss whether these could be important to them someday. The word prevalence has the same meaning as “absolute risk,” but also other meanings. You will notice this if you do a search for “prevalence.”

1. \( \frac{9}{20} \) or 0.45 or 45%; \( \frac{3}{20} \) or 0.15 or 15%
2. The percentage of women who will get breast cancer is \( \frac{1}{8} = 0.125 \), or 12.5%. Or, the chance that a randomly-selected woman gets breast cancer is 12.5%. Or, the absolute risk that a woman gets breast cancer is 12.5%.

3. Only \( \frac{117}{52,595} \approx 0.002225 \), or 0.2225% of women who take these medications to prevent loss of bone density get this unusual type of fracture. Or, the risk that a woman who takes these medications to prevent loss of bone density gets this unusual type of fracture is only 0.2225%.

**INSTRUCTIONAL NOTE** Absolute risk always has a time element involved, which may be explicit or implied. For example, in Problem 2, the absolute risk of breast cancer is explicitly given as over a lifetime. In Problem 4, however, the time period is unstated but appears to refer to the incidence of adverse effects within a short period after the vaccine is given. Often it is important to know to what time period absolute risk refers.

4. a. After receiving the vaccine, \( \frac{32}{23,000,000} \approx 0.000001391 \), or 0.0001391% of the patients die.

   b. After receiving the vaccine, \( 0.062 \times \frac{54}{100,000} = \frac{3.348}{100,000} \approx 0.00003348 \), or 0.003348% of the patients have a serious adverse event. (If students use the rounded percentage given in the article “fewer than 1 percent” rather than the actual number of “54 out of every 100,000” for the prevalence of adverse events, they will get \((0.062)(0.01) = 0.00062\), or 0.062%.)
There are no teacher’s notes for the corresponding student page.
INSTRUCTIONAL NOTE For Problem 5 Part d, students might incorrectly choose the percent bar graph because of the word “percentage” in the question. Listen to group discussions so that during the STM discussion for Part c, you can have groups who correctly analyzed this question report their thinking. See the Promoting Mathematical Discourse scenario on page T8.

a. Bars are defined by the explanatory variable and segmented according to the response variable.

b. The bar on the left in the stacked bar graph (frequency on the vertical axis) represents the boys who smoke. The bottom segment gives the number of smokers who are expected to get lung cancer. The top segment gives the number who are expected not to get lung cancer.

The bar on the left in the stacked bar graph (percent on the vertical axis) represents the boys who smoke. The bottom segment gives the percentage of smokers who are expected to get lung cancer and the top segment gives the percentage who are expected not to get lung cancer.

The bar on the left in the grouped bar graph (frequency on the vertical axis) shows the number of smokers who are expected to get lung cancer.

c. The stacked bar graph (percent on the vertical axis) shows clearly that the percentage of smokers who get lung cancer is higher than the percentage of non-smokers who get lung cancer.

d. The stacked bar graph (frequency on the vertical axis) clearly shows that there are far more non-smokers than smokers.

e. The grouped bar graph (frequency on the vertical axis) makes it easiest to see that the number of non-smokers who do not get lung cancer is many times more than the number of smokers who do not get lung cancer.

6. a. The groups are the leopard attacks on humans in the three years before the program began and in the three years after the program began. The response variable is the seriousness of the attacks.

b. The absolute risk was \( \frac{2}{12} \approx 0.1667 \) before the program and \( \frac{18}{50} = 0.36 \) after the program, so much higher after the program.
c. Graphs like the following can be made using Excel.

**Stacked Bar Graph (Frequency)**

![Stacked Bar Graph (Frequency)](image)

**Stacked Bar Graph (Percent)**

![Stacked Bar Graph (Percent)](image)

**Grouped Bar Graph (Frequency)**

![Grouped Bar Graph (Frequency)](image)

**INSTRUCTIONAL NOTE**

Students’ stacked bar graph (frequency on the vertical axis) and grouped bar graph (frequency on the vertical axis) may have the same vertical scales.
d. The stacked bar graph (frequency on the vertical axis) shows clearly that there were a much larger number of attacks and a much larger number of lethal attacks after the program than before. Also, the height of the bar shows the total number in each category.

The stacked bar graph (percent on the vertical axis) clearly shows a larger percentage of attacks were lethal after the program than before the program.

The grouped bar graph (frequency on the vertical axis) is the best one to use to compare the numbers of nonlethal attacks (because both bars start at 0). It shows that the number of nonlethal attacks also was much larger after the program than before.

7 a. 7th- and 8th-grade boys and 7th- and 8th-grade girls were compared. The response variable is whether the answer to the question was yes, no, or not sure

<table>
<thead>
<tr>
<th></th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>25</td>
<td>23</td>
</tr>
<tr>
<td>No</td>
<td>24</td>
<td>61</td>
</tr>
<tr>
<td>Not Sure</td>
<td>35</td>
<td>54</td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td>138</td>
</tr>
</tbody>
</table>

c. Boys were more likely to say yes because $\frac{25}{84} \approx 0.2976$ of the boys said yes, but only $\frac{23}{138} \approx 0.1667$ of the girls said yes.

e. A girl is much less likely to say yes than is a boy and much more likely to say no. The two groups are roughly equally likely to say that they are not sure.

INSTRUCTIONAL NOTE
Students may order the stacks with “Yes” on the bottom; other orders are possible but should be the same for all bars. If using TCMS-Tools stacked bar graphs with the table entered as in Part b, the vertical order of the stacks will match the table, but be the reverse of those in the graphs in Part d.

TECHNOLOGY NOTE
You may wish to have students enter these data in TCMS-Tools and make bar graphs during class. If a couple of computers are available, groups could take turns learning how to use this feature, then they could use TCMS-Tools for homework tasks and print their work.
INSTRUCTIONAL NOTE Problem 8 can be used to address the issue of various groups of students being at different places in the investigation. For example, if some groups of students have completed Problem 8 before others, you may wish to discuss their thoughts on Part d, so that they have well-thought-out responses to contribute to the discussion of STM Part d.

Then the students in these groups could be asked to begin the CYU task. Each student could do Part a and then discuss responses with each other. Then follow the same process for Parts b and c. Meanwhile, it is likely that other groups will be working on Problem 8. When some other groups have completed Problem 8 Parts a–c and others have completed Problem 7, you could pull the class together for the STM discussion. See the PMD discussion on pages T8A–T8C.

a. \[ \frac{35}{84} \cdot 360^\circ = 150^\circ \]
LESSON 1  •  COMPARING THE RISK

Summarize the Mathematics

a. Absolute risk, or prevalence, is the proportion or percentage of people who have or will get some condition.

b. The explanatory variable is the variable that "explains" the response or outcome. Typically, the question will be worded something like, "Are men or women more likely to smoke?" The explanatory variable is the group of people men or women and the response would be the behavior of smoking or not smoking.

c. Each bar of a stacked bar graph (frequency on the vertical axis) is segmented by the number of people in the group who fall into each category. Each bar represents one of the groups and its height is the number of people in the group.

   Each bar of a stacked bar graph (percent on the vertical axis) is segmented by the percentage of people in the group who fall into each category. Each bar represents one of the groups. Each bar is the same height, at 100%.

   In a grouped bar graph (frequency on the vertical axis), each group has one bar for each category. The bar shows the number of people in the group who fall into that category. The total number of people in the group is the sum of the frequencies in a category.

d. For each category, multiply the proportion of the group that falls into the category by 360°. Then, measure an angle to represent the first category, continuing from the side of that angle with the next category, and so on, around the circle.
See the Instructional Note with Problem 8. This PMD scenario is based on that approach. A suggested supplemental article is “Orchestrating Discussions” by Smith, Hughes, Engle, & Stein (2009).

Teacher: Excuse me, may I have your attention at this point? I’ve noticed that groups are at various stages in completing the investigation, but I think everyone has completed enough of this investigation so that we will be able to have productive conversations around the Summarize the Mathematics on page 8. So, please turn to that page, even if you are still working on Problem 7. The STMs for each investigation gives you a chance to summarize the mathematical ideas developed in your groups. There may be varying points of view and differing conclusions that we will need to discuss and decide whether they are mathematically acceptable or not. This discussion based on your thinking is crucial to building understanding of mathematical ideas and to be able to effectively apply the mathematics to other problems. For this STM, I will read each question. Then I would like to have you think alone about a response for a moment. Then we will ask for a volunteer to start our discussion of a response.

“In this investigation, you learned to use summary statistics and graphs to compare groups sorted by the same categorical variable. What is absolute risk?” (Teacher waits for students to think. Some students flip back in their textbook to page 3 to review the definition.)

Antonina: It is like the probability of something bad happening. It is the ratio made from the number of times the bad thing occurs out of a bunch of times. That is how I think about it.

Teacher: Okay, remember that without my prompting each time, you should think about whether or not you want to add a comment or ask a question about your classmate’s response. Let’s practice that. Antonina, please describe your thinking about what absolute risk is again for us.

Antonina: It is like the probability of something bad happening. It is the number of times the bad thing occurs out of a bunch of times. So, it is a ratio.

Jackson: I just looked back at the definition on page 3. That seems to fit with Antonina’s explanation. The book says, the proportion or percentage (that is like her “ratio”) for whom an undesirable event occurs. “Undesirable event” is like Antonina’s “bad thing.”

Teacher: One of the resources each of you will develop this year is a Mathematics Toolkit. This toolkit will contain definitions and examples from your work that you would like to have as a reference as you work on the remainder of the unit and other units in the course. Do you think that “absolute risk” is one of the terms you would like to have in your toolkit? (Most students indicate “yes,” so the teacher helps them with the logistics of a math toolkit. See Teacher’s Guide front matter for more information on Math Toolkits. You will find a prompt for this investigation on page T8. The class uses the definition from the textbook and includes some additional information about the ratio such as

\[
\frac{\text{number for whom the bad event occurs}}{\text{number who engage in the risky behavior}}
\]

Teacher: On to Part b. Will someone volunteer to read Part b? Thank you, Charles. Then, again we will wait a moment for you to think about the question before someone offers his or her thoughts.

Charles: “When comparing situations involving risk, how can you tell which is the explanatory variable and which is the response variable?”

Teacher: Who is willing to start by sharing their thinking? Tonya.
Tonya: The explanatory variable is the different groups that are being compared. I think of the response variable as the different outcomes of the risky behavior.

Jori: I agree with Tonya about the explanatory variable, but I think that the response variable is not always an outcome of risky behavior. If you look at Problem 7, there the response variable was boys’ and girls’ answers to a question. But it does fit with the idea of “responses.”

Charles: That sounds good. But I also think of the explanatory variable as what “explains” the response. So, when I was reading Problem 7, I really thought about what was the response was first. This helped me think about the explanation for the response or the two groups being compared.

Juanita: Hey, I like that idea, Charles! I’m going to try that with some problems. Thinking about what goes in the rows of the table first (the responses) may help me make my table correctly.

Teacher: Anything else we should add to this discussion? Okay, then Part c: “Describe three types of bar graphs and the characteristics of the data that are best shown by each one.” Let’s display the three types of bar graphs from page 5 so that we can refer to them for our discussion. Juan, what are the characteristics of the data best shown in the first stacked bar graph?

Juan: Well, each bar represents one of the groups and its total height is the number of people in the group. The lines in the bars represent the different categories. Like how many people are in the “Get Lung Cancer” and “Do Not Get Lung Cancer” categories. This graph makes it easy to compare the numbers for each of these categories. You can also easily see if one group had more people in it by which bar is higher.

Monica: You can also compare the groups by noticing how the categories in one group make up the whole. Like for the non-smokers, say the categories “Get Lung Cancer” and “Do Not Get Lung Cancer” are split into about 10% and 90% of the whole bar, respectively. Then you can see if the smokers had a similar split for the categories.

Teacher: (The teacher waits.) I am assuming that no one wishes to add anything or ask a question at this point. One way to talk about the divisions in the bars is to say that the bars are “segmented” into different categories. So, Melissa, what are the characteristics of the data best shown in the second stacked bar graph?

Melissa: This graph shows percent on the vertical axis. So, both bars are the same height. Each bar shows the percentage of people in each category.

Martin: I’d like to add something. This graph is good when the two groups have quite different numbers. It is often better to compare percents rather than numbers of people.

Teacher: Thank you. What about the characteristics of the third bar graph on the page? Tonya.

Tonya: This graph is just the first graph with the sections of the bars unstacked. It shows the numbers of people in each of the categories for each group.

George: In some ways, this is an easier way to see the numbers for example of “those who did not get lung cancer” out of the group who smoked. In the first graph, you need to estimate those who did get lung cancer and subtract that from the total bar height. Here you can just read the frequency.

Sam: George, I did not understand what you said.

Teacher: George, would you come up to the display and show us what you mean? (George comes to the display and re-explains using the graphs.)

Sam: Now I get it.

Teacher: I realize that some of you did not finish Problem 8, but others did. Who would like to explain how to make a pie chart?

June: You make a pie chart for each group. Then, the pie is divided up by the proportions of the categories. For the girls’-answers pie, we took the proportion of “yes” answers and multiplied by 360°. That was 60°, so we drew a radius of the circle and then measured to make a 60° angle. We did the same for the proportion of “no” answers. Timesed it by 360°. That is how you find the pieces of the pie.
Teacher: Thanks, June. Mark, your group had some discussion around Problem 8 Part d. Would you please read this part for us and then someone else from your group can report what you discussed? (The teacher is being more directive at this point because only a few students actually discussed Part d.)

Mark: “Most statisticians recommend including stacked bar graphs rather than pie charts in reports, even if you have the technology that makes both easily. Give a reason why you think stacked bar graphs are better than pie charts.”

Alfonso: We talked about the fact that it is easier to compare categories, like the “yes” category for the two different groups in the stacked bar graphs. You can read the percents or frequencies on the vertical axis and see the relative heights pretty easy. In pie charts, it is more difficult to see whether or not a category is a little more in one group than in the other group.

Suki: We also said that even without reading the frequencies on the vertical axis in stacked bar graphs, it is easy to compare the size of the segments of the bars in the two groups. I find it hard to estimate angle sizes.

Teacher: Thanks, Suki. I noticed that in our discussion of this STM, you were taking the initiative to respond to each other’s thoughts without my prompting. This was a helpful discussion. Now let’s consider whether or not you would like to add other ideas in addition to that of absolute risk from this discussion to your Math Toolkit.
Check Your Understanding

a. |                      | Adult      | Child      |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Food Secure</td>
<td>194,579,000</td>
<td>57,010,000</td>
</tr>
<tr>
<td>Low Food Security</td>
<td>20,741,000</td>
<td>16,209,000</td>
</tr>
<tr>
<td>Very Low Food Security</td>
<td>12,223,000</td>
<td>988,000</td>
</tr>
<tr>
<td>Total</td>
<td>227,543,000</td>
<td>74,207,000</td>
</tr>
</tbody>
</table>

b. An adult is much more likely to live in a household with very low food security because the proportion of adults who live in such households is \( \frac{12,223,000}{227,543,000} \approx 0.0537 \) while the proportion of children who live in such households is \( \frac{988,000}{74,207,000} \approx 0.0133 \).

c. The stacked bar graph (percent on the vertical axis) below shows that while a smaller percentage of children than adults live in households with very low food security, a higher percentage of children than adults live in households with either low or very low food security. In other words, a lower percentage of children than adults live in food secure households.

A second bar graph representation is:
In this investigation, students use two different methods to compare the risk that two groups have of getting some condition:

absolute risk reduction = absolute risk for one group – absolute risk for another group

relative risk = \frac{absolute risk for one group}{absolute risk for another group}

Usually the larger risk comes first in absolute risk reduction, resulting in a positive number of percentage points. Typically the larger risk is in the numerator of the relative risk fraction, so that we can say, for example, that if you persist in this behavior, your chance of dying is 2 times what it would be if you stop. In some cases, however, we might want to say something like, if you stop this behavior, your chances of dying are half of what they would be if you persist. Then, the larger risk would go into the denominator.

**Launch**

You may wish to do Problems 1 and 2 and perhaps Problem 3 as a class discussion. If so, you may wish to use a think-pair-share strategy to allow students to individually think about the problem and then discuss with a classmate for a few minutes prior to the class discussion. The methods of Alma and Bill in Problem 1 illustrate absolute risk reduction and relative risk, respectively, although students will not learn those terms until Problem 3. Problem 2 is important so that students become aware of the possible ambiguity in the language “is increased by 75%.”

1. Alma computed the difference in the percentages: 62% – 55% = 7%.
   The percentage of boys who drank sugar-sweetened soda the previous day was 7% higher than the percentage of girls who drank this type of soda. Bill computed a ratio involving the percentages: \frac{62}{55} \approx 113\%.

2. a. Let x be the absolute risk of melanoma for those under age 30 who do tan regularly.
   Alma: x – 0.002 = 0.75, so x is 0.752, or 75.2%.
   Bill: \frac{x}{0.002} = 1.75, so x is 0.0035, or 0.35%.

   b. Obviously, it is not the case that 75% of women who use tanning beds get melanoma. That would mean an incredible number of deaths. The IARC statement must be using reasoning like Bill’s and the absolute risk (when tanning regularly begins before age 30) is 0.35%.
**TERMINOLOGY NOTE** Sometimes the term “prevalence ratio” is used instead of “relative risk.” The statistic relative risk reduction (RRR) also is widely used in reports. See, for example, evidence.ahc.umn.edu/arr-s5.htm or en.wikipedia.org/wiki/Relative_risk_reduction/. RRR is defined this way:

\[
RRR = \frac{\text{larger absolute risk} - \text{smaller absolute risk}}{\text{larger absolute risk}}
\]

3. Alma computed absolute risk reduction, \(62\% - 55\% = 7\%\).
   Bill computed relative risk: \(\frac{62}{55} \approx 113\%\).

b. Statement II describes absolute risk reduction most clearly.
   Statement III describes relative risk most clearly.

**TERMINOLOGY NOTE** Stress that the units for absolute risk reduction (or, the difference of any two percentages) should be labeled **percentage points**. If the media would use this terminology consistently, there would be a lot less confusion. For example, in Problem 1, saying that there is a 7 percentage point difference in the rates for boys and girls makes the meaning clear.

4. a. The difference in the percentage of women who tan regularly before age 30 and the percentage of women who do not tan regularly who get melanoma is \(0.35\% - 0.2\%\), or 0.15 percentage points.

b. A woman who tans regularly before age 30 is \(\frac{0.0035}{0.002} = 1.75\) times more likely to get melanoma than a woman who does not tan regularly.

c. The units are either percent or no units. For example, for Bill’s computation in Problem 1, the units were percent—one rate was 113% of the other. But, Bill could just as well have said that the boys’ rate was 1.13 times the girls’ rate. So, more typically, there are no units, as in the example in Part b.

d. If 2.3 million teens use tanning beds, we expect that \(0.0035 \times 2,300,000 = 8,050\) teens will get melanoma. If none of them used tanning beds, we would expect \(0.002 \times 2,300,000 = 4,600\) teens to get melanoma, a reduction of \(8,050 - 4,600 = 3,450\) cases of melanoma.
LESSON 1  •  COMPARING THE RISK

a. The proportions are $\frac{4}{26}$, or about 15.38%, for those who pitched at least 100 innings in one year and $\frac{6}{128}$, or about 4.688%, for those who did not. The difference of 10.692 percentage points is called absolute risk reduction.

A small study of baseball pitchers ages 9 to 14 found that those who pitched at least 100 innings in one year were more likely to have had elbow surgery, shoulder surgery, or retirement due to a throwing injury than pitchers who did not pitch that much in any year. The increase in risk was 10.7 percentage points, rising from about 4.7% for those who did not throw at least 100 innings a year to 15.4% for those who did throw at least 100 innings a year.

(Note that in descriptions for the general public, percentages generally should be carried to not more than one decimal place, if that.)

b. $\frac{4}{6} \approx 3.282$. This is called relative risk.

If a pitcher pitches at least 100 innings, the risk of having one of these problems is about 3.3 times higher than for a pitcher who does not pitch this much.

POSSIBLE MISCONCEPTION
Some of your students may have been taught that in mathematics, “difference” always refers to subtraction. In Problem 5 Part a, “difference” does imply subtraction but it does not in Part b. Sometimes “difference” refers to how things are “different.” Watch for other instances in this unit. See the Terminology Note on page T49.

a. $\frac{35}{1,653}$ means that of the 1,653 patients who received salmeterol, 35 were hospitalized; $\frac{16}{1,622}$ means that of the 1,622 patients who received the placebo, 16 were hospitalized.

b. $\frac{35}{1,653} \approx 0.0212$, or about 2% and $\frac{16}{1,622} \approx 0.009864$, or less than 1% (but not much less than 1%).

c. $\frac{35}{1,653} \approx 2.1465$; The relative risk is approximately 2.1.

d. The absolute risk reduction for not taking salmeterol is $\frac{35}{1,653} - \frac{16}{1,622} \approx 0.01131$, or about 1.131 percentage points.

e. Salmeterol is one of the components of Advair, a medication for asthma. A study randomly divided patients aged 12 to 18 into two groups. One group took salmeterol for 28 weeks. The other group took a placebo (with no salmeterol or other active ingredient in it). After the 28 weeks, about 2% of the 1,653 patients who took salmeterol had been hospitalized, while about 1% of the 1,622 patients who took the placebo were hospitalized. This counted all hospitalizations for any reason, not necessarily because of the salmeterol. So, the chance of being hospitalized, while still small, is double that for asthma patients this age who take salmeterol rather than a placebo.
a. A: 1 percentage point  
    B: 34 percentage points  
    C: 1 percentage point

b. Diseases A and C have the same absolute risk reduction,  
   1 percentage point. Suppose you choose the vaccine for Disease A.  
   For every 100 people who get vaccinated, you expect that 1 person  
   will get Disease A rather than 2, saving 1 person from disease. If  
   you choose to vaccinate for Disease C, for every 100 people who  
   get vaccinated, you expect that 98 rather than 99 people will get  
   Disease C, also saving 1 person from disease. So, it does not matter  
   which disease you choose.

c. A: 2; B: 2; C: 1.0102

d. Diseases A and B have the same relative risk: an unvaccinated  
   person has twice the risk of getting the disease as a vaccinated  
   person.

   Suppose you choose the vaccine for Disease A. For every  
   100 people who get vaccinated, you expect that 1 person will get  
   Disease A rather than 2, saving 1 person from disease. If you choose  
   to vaccinate for Disease B, for every 100 people who get vaccinated,  
   you expect that 34 people will get Disease B rather than 68, saving  
   34 people from disease. You should provide vaccinations for Disease B.

**Summary**

Ask students to provide an example in Part a of the STM. Ask them for one  
that is different than those in the investigation. As students discuss Part c,  
they will likely focus on relative risk because it sounds more alarming. If not,  
you might insert the question, “Which method of comparing risk—absolute  
risk reduction or relative risk—sounds more alarming?”

**Summarize the Mathematics**

| a | Absolute risk reduction is the difference in the percentage of people in two groups who have or will get some condition. For example, if 24% of unvaccinated people get some disease and 18% of vaccinated people get the disease, then the absolute risk reduction as a result of being vaccinated is 6 percentage points. |
| b | Relative risk is the ratio of the absolute risks for two groups. For the situation above, an unvaccinated person is 1.5 times more likely to get the disease than is a vaccinated person. |
| c | As computed in Problem 2, if a woman does not tan, the absolute risk reduction is only 0.15 percentage points. This does not sound impressive at all. But saying that a woman who tans regularly before age 30 is 1.75 times more likely to get melanoma than a woman who does not tan regularly sounds much more alarming. In this case, if you want to get a woman’s attention, give her the relative risk. |
Teacher: Please turn to page 12 to discuss the STM for this investigation. The questions are also displayed on the white board. Remember how we will be conducting this discussion. When someone provides their thinking, if you do not understand, please formulate a question to ask your classmate. If you have a question or something to add, please raise your hand. That way we can have contributions from a variety of people today. For this STM, I would like you to discuss the three parts in the STM in your groups before we have our class discussion. Make some brief notes of your thinking so you can share from those notes. I will give you 8 minutes. You are asked in Parts a and b for an example. Think of an example that is different from the ones you have done in this investigation. (The teacher listens to groups discussing the questions to inform the discussion.)

Dominick. Please read Part a and get the conversation started.

Dominick: “What is absolute risk reduction? Give an example of how it is interpreted.” It is the difference between the two risks.

Margo: I noticed that we usually subtract the smaller percentage from the larger percentage. This means we do not have negative numbers. And we are finding which group had the reduced risk.

Teacher: Will someone take Dominick’s and Margo’s ideas to give a more formal explanation of absolute risk reduction?

Krystiana: I’ll try. Absolute risk reduction is the difference between the absolute risk for one group and the absolute risk for the other group. Can I show how I wrote it in my notes? (The teacher approves.) It goes like this: $\text{Abs risk reduct (2 groups)} = \text{Abs risk (larger)} - \text{abs risk (smaller)}$.

Corina: That makes sense. You really just need to remember that the reduction is a subtraction and is not like the relative risk, which is a ratio.

Teacher: Since you have mentioned “relative risk,” Corina, please explain what it means. That may clarify your statement.

Corina: “Relative risk” is the ratio of the absolute risks for the two groups. If I wrote it like Krystiana did, it would look like this. (She writes on the white board.)

$$\text{Rel Risk} = \frac{\text{abs risk (larger)}}{\text{abs risk (smaller)}}.$$ This tells you how many times more likely the risker behavior is.

Derek: But you have to be careful when you say that “relative risk” is a ratio. It is really a ratio of two ratios. This is because “absolute risk” is also a ratio. I like the short-hand way that we wrote these two, but it would be even more helpful to me to think about the “absolute risk” as a subtraction of two ratios or percents and the “relative risk” as the ratio of two ratios or percents.

Teacher: Okay since there are no other comments, I am assuming that you each have thought through this enough so that we can make Math Toolkit entries for these two ways to compare risks. But let’s first have an example of “absolute risk reduction” and “relative risk.” Think of a risky behavior that is different from the ones you thought about in this investigation.

Kelsey: My brother just got a motorcycle. That worries me. You might have, say, two groups of 20-year old guys. One group might have motorcycles and the other group cars. Then if you find from some data that about 20% of guys who ride motorcycles have accidents and about 10% of those with cars have accidents, then the absolute risk reduction is $20\% - 10\% = 10\%$ and the relative risk is $20\%/10\% = 2$.

Teacher: Thanks for the example, Kelsey. I’d like you to read and discuss Part c with your group before we hear from someone. (The teacher gives them one minute to discuss. Then calls on Eric to reply for his group.)
Eric: We used Kelsey’s example even though the risk was not extremely rare to make our point. Saying to her brother that the risk of having an accident is double that of others in his class who drive cars might make a bigger impression than just saying his risk is 10% more.

Liam: We thought about the risk for concussions for guys who play high school football and guys who do not play football. We used a couple of very small absolute risks: do not play football: 0.03% and do play football: 1.2%. So, the relative risk is 1.2%/0.03% = 40. That means football players 40 times more likely to have a concussion than those who do not play football. This sounds much more serious than saying that the absolute risk reduction is 1.17% for those who do not play football.

Teacher: Other comments? No, then let’s craft a Math Toolkit entries for absolute risk reduction and relative risk.
**Check Your Understanding**

a. Explanatory variable: whether the child was vaccinated against measles or not
   Response variable: whether the child got measles or not.

b. |            | Vaccinated | Not Vaccinated |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Got Measles</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Did Not Get Measles</td>
<td>599</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>609</td>
<td>16</td>
</tr>
</tbody>
</table>

c. Vaccinated: $\frac{10}{609} \approx 0.01642$
   Not vaccinated: $\frac{7}{16} = 0.4375$

d. $\frac{7}{16} - \frac{10}{609} \approx 0.4211$

e. $\frac{7}{10} \approx 24.6438$

f. Measles can be very serious, including causing death. If your child is not immunized, he or she is 25 times more likely to get measles than a child who was vaccinated.
   There was an outbreak of measles in Colorado in 1994. Of the 16 children who were not vaccinated, 44% got measles. Of the 609 children who were vaccinated, less than 2% got measles, and those cases occurred among those who had not gotten both recommended doses. So, to put it another way, if your child is not vaccinated, it may increase his or her chance of getting measles in an outbreak by 42 percentage points.
1. **Applications**

a. The groups are Korean and American Midwest college students. The response variable is whether the email included apologies or not.

b. 

<table>
<thead>
<tr>
<th></th>
<th>Koreans</th>
<th>Americans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apologized</td>
<td>105</td>
<td>51</td>
</tr>
<tr>
<td>Did Not Apologize</td>
<td>22</td>
<td>46</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>127</td>
<td>97</td>
</tr>
</tbody>
</table>

d. Koreans are much more likely to include an apology than are Americans; about 83% compared to only 53%.

c. A stacked bar graph (percent on the vertical axis) is the best to use to compare proportions.
a. The groups being compared are injuries in games played on FieldTurf and injuries in games played on grass. The response is the number of players with minor, substantial, and severe injuries in each game.

b. The first bar graph shows that there were more injuries on grass and the second shows us that they also tended to be worse. There are a higher proportion of injuries that are severe on grass than on FieldTurf and a higher proportion of injuries that are substantial.

c. Artificial turf in athletic fields was first introduced in the 1960s. Its safety has been controversial since then. One issue that has been investigated is whether injuries of football players tend to be more serious on an artificial turf than on grass. A study followed 24 NCAA division 1A college football teams over three seasons. (Source: Michael C. Meyers, Thaddeus C. Mechanisms, and Severity of Game-Related College Football Injuries on FieldTurf Versus Natural Grass: A 3-Year Prospective Study, American Journal of Sports Medicine, Vol. 38, 2010, pp. 687-697)

d. Encouraging students to use TCMS-Tools at home or school for this task will help them gain experience to be able to use TCMS-Tools' built-in data sets for some problems in Lessons 2 and 3. See URM 6 for a Technology Tip.
a. For example, the central angle for girls who meet this recommendation is \(0.206(360) \approx 74.16^\circ\).

b. The following bar graph shows the distribution of meeting the recommendation for boys and girls.

INSTRUCTIONAL NOTE In Connections Task 10 and Task 11 and similar problems, students must make a table of frequencies by multiplying the group size by the percentage that falls into one of the categories. These tables are approximations because there actually could be many tables that satisfy the conditions given in the problem. For example, in Connections Task 10, we are told that 89% of the 143,473 newborns whose mothers lived, survived to the age of 10. The number of such newborns can be estimated by multiplying 143,473 by 0.89, getting about 127,690.97 and rounding to 127,691, as was done in Part a below. But note that many other whole numbers, when divided by 143,473 give a proportion that rounds to 0.89. For example, the larger number 128,408, when divided by 143,473, is about 0.894998, which rounds to 89%. So, the frequencies in such tables are approximations.
10. a. | Mother Died | Mother Lived | Total |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Child Died</td>
<td>1,053</td>
<td>15,782</td>
</tr>
<tr>
<td>Child Lived</td>
<td>332</td>
<td>127,691</td>
</tr>
<tr>
<td>Total</td>
<td>1,385</td>
<td>143,473</td>
</tr>
</tbody>
</table>

b. \( \frac{1,053 + 15,782}{144,858} \approx 0.1162 \), or about 11.62%

c. \( \frac{1,053}{1,053 + 15,782} \approx 0.06255 \), or about 6.255%

d. In this rural area of Bangladesh, about \( \frac{1,053}{1,385} \approx 0.76 \), or 76% of all babies whose mothers die within their first 10 years of life have themselves died by age 10. Of the babies whose mothers have not died, about \( \frac{15,782}{143,473} \approx 0.11 \), or 11% died by age 10, for a reduction of 65 percentage points.

e. If the mother of a newborn dies within 10 years of the birth, the chance her baby dies before age 10 is about \( \frac{76}{11} \approx 6.9 \) times as high as that for a newborn whose mother remains alive.

### Reflections

Yes, this is possible. For example, the table below actually was used to make the bar graph. Here, a larger number of people in Group B than in Group A have the condition, but a smaller percentage of people in Group B than in Group A have the condition.

<table>
<thead>
<tr>
<th>Have Condition</th>
<th>Group A</th>
<th>Group B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do Not Have Condition</td>
<td>700</td>
<td>5,400</td>
</tr>
<tr>
<td>Total</td>
<td>1,000</td>
<td>6,000</td>
</tr>
</tbody>
</table>

Let \( x \) be the number of people in Fulton County. Then, \( \frac{16.83}{10,000} = \frac{25}{x} \), so 14,854 people live in Fulton County.

### Extensions

a. \( \frac{16.7}{100,000} = 0.000167 \)

b. Solving \( \frac{x}{100,000} = 0.000097 \), \( x \) is 9.7. The risk in New Jersey is 9.7 per 100,000.
c. Solve the equation \( \frac{x}{100,000} = \text{absolute risk} \).

d. Solving \( \frac{x}{100,000} = 0.000596 \), \( x \) is 59.6. The risk in Wyoming is 59.6 per 100,000.

e. Solving \( \frac{x}{10,000} = 0.000596 \), \( x \) is 5.96. The risk in Wyoming is 5.96 per 10,000.

f. Over the years from 1990 to 2008 in the United States, 17-year-olds had higher rates of involvement in a fatal crash than did 16-year-olds, and the difference has stayed fairly consistent over the years, about 10 more fatal crashes per 100,000 17-year-olds than 16-year-olds. Overall, the rates for both ages are dropping, from a combined rate of about 38 per 100,000 to about 16 per 100,000 in 2008.

**Review**

**Just in Time**

23 a. 23,719(0.217) \( \approx \) 5,147.023. So, approximately 5,147 people in Caswell County, NC lived in poverty in 2010.

b. \( \frac{1.60}{17.60 + 1.60} = \frac{1.60}{19.20} \approx 0.083 \). The average price of milk decreased by 8.3% between April 2011 and April 2012.

c. \( \frac{1}{6} (315,000,000) = 52,500,000 \). So, approximately 53,000,000 people in the United States were expected to suffer from food poisoning during 2012.

d. \( \frac{12 \text{ million}}{313 \text{ million}} = \frac{1}{x} \) \( x \approx 26 \). So, approximately 1 out of every 26 people in the United States in 2010 was a carrier of the defective gene for CF without having the disease.

28 a. i. \( \log 1 = 0 \)  ii. \( \log 100 = 2 \)

   iii. \( \log 100,000 = 5 \)  iv. \( \log 0.01 = -2 \)

b. i. \( 10^1 = 10 \)  ii. \( 10^3 = 1,000 \)

   iii. \( 10^{-1} = 0.1 \)  iv. \( 10^{-3} = 0.001 \)

c. \( \log 485 \) is the power of 10 that results in 485. Since \( 10^2 = 100 < 485 < 10^3 = 1,000 \), \( 2 < \log 485 < 3 \).
d. Student estimates will vary. Calculations are provided below.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>log 16</td>
<td>between 1 and 2</td>
<td>1.204</td>
</tr>
<tr>
<td>log 1.6</td>
<td>between 0 and 1</td>
<td>0.204</td>
</tr>
<tr>
<td>log 1,600</td>
<td>between 3 and 4</td>
<td>3.204</td>
</tr>
<tr>
<td>log 3</td>
<td>between 0 and 1</td>
<td>0.477</td>
</tr>
<tr>
<td>log 0.3</td>
<td>between –1 and 0</td>
<td>–0.523</td>
</tr>
<tr>
<td>log 30</td>
<td>between 1 and 2</td>
<td>1.477</td>
</tr>
</tbody>
</table>

e. \( \log 100x = 2 + \log x \)

Let \( \log x = a \). Then \( 10^a = x \). Since \( 100 = 10^2 \), \( \log 100x = \log 10^2x = \log 10^210^a = \log 10^2 + a \).

But \( \log 10^2 + a = 2 + a = 2 + \log x \). So, \( \log 100x = \log 10^210^a = \log 10^2 + a = 2 + \log x \).

**DIFFERENTIATION NOTE** The rigor of the reasoning that you should expect for part e depends on your students and their ability to write general arguments as well as their prior experience with logarithms. For some classes, you may want to just notice the pattern and omit the general proof. If students know the rules of logarithms, that is that \( \log ab = \log a + \log b \), then this general argument is simpler than that shown here. This part of the problem may also be an opportunity for you to learn what your students remember from their previous work with logarithms.

60 a. The monthly interest will be \( 0.005(750) = 3.75 \).

b. | Number of Months | Loan Balance | Number of Months | Loan Balance |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$753.75</td>
<td>7</td>
<td>$776.25</td>
</tr>
<tr>
<td>2</td>
<td>$757.50</td>
<td>8</td>
<td>$780.00</td>
</tr>
<tr>
<td>3</td>
<td>$761.25</td>
<td>9</td>
<td>$783.75</td>
</tr>
<tr>
<td>4</td>
<td>$765.00</td>
<td>10</td>
<td>$787.50</td>
</tr>
<tr>
<td>5</td>
<td>$768.75</td>
<td>11</td>
<td>$791.25</td>
</tr>
<tr>
<td>6</td>
<td>$772.50</td>
<td>12</td>
<td>$795.00</td>
</tr>
</tbody>
</table>

c. The balance is increasing by $3.75 each month.

d. The only correct recursive formula is iv.

e. \( B = 750 + 3.75t \)

**CRA Set 1**

1. (d) 2. (c) 3. (b) 4. (a) 5. (e) 6. (c) 7. (d) 8. (c) 9. (d) 10. (a)