Workshop Statistics: Discovery with Data, A Bayesian Approach

James H. Albert and Allan J. Rossman

May 23, 2009

ACKNOWLEDGMENT

Portions of this book have been reproduced from *Workshop Statistics* by Alan J. Rossman, published by Springer-Verlag New York, in 1996. This material was reprinted by permission of Springer-Verlag New York. Any further reproduction is strictly prohibited.

PREFACE

This book is a collection of classroom and homework activities designed to introduce the student to concepts in data analysis, probability, and statistical inference. Students work toward learning these concepts through the analysis of genuine data and hands-on probability experiments and through interaction with one another, with their instructor, and with technology. Providing a one-semester introduction to fundamental ideas in statistics for college and advanced high school students, **Activities for Statistics: Data, Probability, and Learning from Data** is designed for courses that employ an interactive learning environment by replacing lectures with hands-on activities.

This book is distinctive with respect to two aspects — its emphasis on active learning and its use of the Bayesian viewpoint to introduce the basic notions of statistical inference. Below we first describe the workshop pedagogical approach. We then outline the traditional and Bayesian approaches for introducing statistical inference, and explain why a Bayesian viewpoint may be advantageous in communicating basic concepts of inference in a first statistics course.

Active Learning

This text is written for use with the workshop pedagogical approach, which fosters active learning by minimizing lectures and eliminating the conventional distinction between laboratory and lecture sessions. The book's activities require students to collect data and perform random experiments, make predictions, read about studies, analyze data, discuss findings, and write explanations. The instructor's responsibilities in this setting are to check students' progress, ask and answer questions, lead class discussions, and deliver "mini-lectures" where appropriate. The essential point is that every student is actively engaged with learning the material through reading, thinking, discussing, computing, interpreting, writing, and reflecting. In this manner students construct their own knowledge of probability and statistical ideas as they work through the activities.

The activities also lend themselves to collaborative learning. Students can work together through the book's activities, helping each other to think through the material. Some activities specifically call for collaborative effort through the pooling of class data.

The text also stresses the importance of students' communications skills. As students work through the activities, they constantly read, write, and talk with one another. Students should be encouraged to write their explanations and conclusions in full, grammatically correct sentences, as if to an educated layperson.

Traditional Method of Introducing Statistical Inference

Statistical inference is traditionally taught using the frequentist approach. To illustrate this approach, suppose one is interested in learning about the proportion of all undergraduate students at a particular college who drink coffee regularly. One learns about this unknown proportion by taking a random sample of 100 students, asking each student if they drink coffee regularly, and computing the proportion of students who drink coffee in the sample. Suppose that this sample proportion is $21/100 = .21$. Based on this data, what have we learned about the proportion of all undergraduates who drink coffee?

The frequentist approach to inference is based on the concept of a sampling distribution. Suppose that one is able to take samples of size 100 repeatedly from the population of undergraduates and compute the sample proportion for each sample selected. The collection of proportions from all of the samples is called the sampling distribution of the proportion. The knowledge of the shape, mean, and standard deviation of this sampling distribution is used to construct confidence intervals for the proportion of interest, and to make decisions about the location of the proportion.

To correctly interpret traditional inferential procedures, students need to understand the notion of a sampling distribution. The students will analyze only one sample in their data analysis. But he or she has to think what could happen if we took a large number of random samples (like the one just selected) from the population. Indeed, a statement such as "95% confidence" for an interval estimate refers to the behavior of the statistical procedure when samples are repeatedly taken from the population.

The Bayesian Viewpoint

The Bayesian viewpoint toward inference is based on the subjective interpretation of probability. In our example, the proportion of undergraduates drinking coffee is an unknown quantity, and a probability distribution is used to represent a person's belief about the location of this proportion. This probability distribution, called the prior, reflects a person's knowledge about the proportion before any data is collected. After the sample survey is taken and data are observed, then one's opinions about the proportion will change. Bayes' rule is the recipe for computing the new probability distribution for the proportion, called the posterior, based on knowledge of the prior probability distribution and the sample survey data. All inferences about the proportion are made by computing appropriate summaries of the posterior probability distribution.

To use Bayes' rule in an introductory statistics class, the student needs to learn some

basic probability concepts. Topics 11, 12, 13 discuss the interpretation of probabilities and methods of computing and interpreting probability tables. Conditional probability is introduced in Topic 14 by means of a two-way probability table, and this two-way table is used in Topic 15 to introduce Bayes' rule. The basic methodology used in Topic 15 is extended to problems of inference for one proportion, one mean, and two proportions in Topics 16, 17, 18, 19, and 21.

Why Not the Traditional Approach?

Although the traditional approach to teaching statistical inference leads to the familiar methods used in practice, it can be difficult to learn in a first course. The notion of a sampling distribution is perhaps the most difficult concept, since the student is asked to think about the variation in samples other than the one that he or she observed. If the student does not fully understand the repeated sampling idea that is inherent in a sampling distribution, then he or she will not be able to correctly interpret traditional inferential conclusions like "I am 95% confident that the proportion is contained in my interval." Since the traditional inferential concepts are hard to learn, the instructor may focus on teaching the mechanics of statistical inference instead of the concepts. These mechanics include the use of a variety of statistical recipes and the correct programming of these recipes using a statistics computer package. This type of "cookbook" class is counter to the modern movement in statistics instruction which encourages more thinking in a first class and fewer recipes.

Why Bayes?

The Bayesian viewpoint has several features that make it especially attractive in the teaching of an introductory statistics class.

- **Conditional inference.** All Bayesian inferential conclusions are made conditional on the observed data. Unlike the traditional approach, one need not be concerned with datasets other than the one that is observed. There is no need to discuss sampling distributions using the Bayesian approach.

- **Inferential conclusions are understandable.** From a Bayesian viewpoint, it is legitimate to talk about the probability that the proportion falls in a specific interval, say (.1, .3), or the probability that a hypothesis is true. These inferential conclusions are generally the ones that seem intuitive for the students. In contrast, traditional inferential conclusions are frequently misstated. For example, if a computed traditional

confidence interval is (.1, .3), it is common for the student to incorrectly state that the proportion falls in the interval (.1, .3) with probability .90. The students forget that the probability (in a traditional viewpoint) refers to the behavior of the interval estimate under repeated sampling.

- **One recipe.** Bayes' rule is really the only inferential method that needs to be taught. Once Bayes' rule is used to compute the posterior distribution, the student just needs to summarize this probability distribution to make inferences.

- **Illustrates the use of the scientific method.** One goal of an introductory statistics class is to learn how statisticians use the scientific method to answer questions. In the scientific method, one begins with a hypothesis or theory about some event, data is collected relevant to the problem, and then the theory is revised according to the data results. The Bayesian viewpoint provides a convenient paradigm for implementing the scientific method. The prior probability distribution can be used to state initial beliefs about the population of interest, relevant sample data is collected, and the posterior probability distribution reflects one's new beliefs about the population in light of the data that were collected.

Contents

Topic 1: Data and Variables

Introduction

Statistics is the science of collecting, organizing, and interpreting data. **Data** is the general term for information that we will collect. There are many different types of data. Data can be the temperatures of different cities across the world on a particular day, the political affiliations of 20 voters surveyed in a small community, the heights of football players on your favorite professional team, or the times it takes a pizza restaurant to deliver pizzas to the homes of fifteen customers.

There are different reasons for collecting data. One reason is to gain an understanding of the data by organizing and graphing the individual values. One can learn about the heights of professional football players by plotting the heights in some reasonable manner. This plot is helpful for answering some questions about these heights. What is an average size of a football player on this team? Are there many players that are under six feet? Am I tall enough to potentially play for this football team? This chapter will describe a number of methods that are useful for organizing, presenting, and summarizing data.

A second, perhaps more important reason for collecting data is to draw conclusions about a larger group of information. The pizza restaurant is interested in the times it takes to deliver pizza. They may wish to make a guarantee to the customers that a pizza will be delivered in 25 minutes or less. If the pizza is not delivered in this time interval, then the pizza will be free. Is this a reasonable guarantee? Will most pizzas be delivered within the 25 minute time interval? To answer these questions, the restaurant is interested in the distribution of *all* times it will take to deliver pizzas in the next year. These times are not measurable. It is certainly impossible to measure the time that it takes to deliver a pizza six months in the future. However, the restaurant can measure the time to deliver pizza for fifteen deliveries during the past week. They can graph and summarize these fifteen time measurements. Suppose that these fifteen measurements can be viewed as representative of delivery times for the next year. Then the information that is gained about this data can be used to learn about the larger group of delivery times in the next year.

PRELIMINARIES

1. Write a sentence describing what the word *statistics* means to you. (Here and throughout the course, please write in complete, well-constructed, grammatically correct sentences.)

2. Record in the table below the responses of students in this class to the following questions:

   - What is your gender?
   - Which of the following terms best describes your political views: liberal, moderate, or conservative?
   - Do you agree with the statement that "Activities of married women are best confined to home and family?"
   - Do you think that the United States should retain or abolish the penny as a coin of currency?
   - Rank your opinion of the value of statistics in society on a numerical scale of 1 (completely useless) to 9 (incredibly important).

3. Take a wild guess as to the number of different states that have been visited (or lived in) by a typical student at this college. Also guess what the fewest and most states visited by the students in this class will be. Put the three numbers in the space below.

4. Take a guess as to the proportion of students at this college who have been to Europe.

5. Place a check beside each state that you have visited (or lived in or even just driven through), and count how many states you have visited.

stu	gender	polit view	family?	abol penny?	value of stat	stu	gender	polit view	family?	abol penny?	value of stat
1						16					
2						17					
3						18					
4						19					
5						20					
6						21					
7						22					
8						23					
9						24					
10						25					
11						26					
12						27					
13						28					
14						29					
15						30					

state	visited?	state	visited?	state	visited?
Alabama		Louisiana		Ohio	
Alaska		Maine		Oklahoma	
Arizona		Maryland		Oregon	
Arkansas		Massachusetts		Pennsylvania	
California		Michigan		Rhode Island	
Colorado		Minnesota		South Carolina	
Connecticut		Mississippi		South Dakota	
Delaware		Missouri		Tennessee	
Florida		Montana		Texas	
Georgia		Nebraska		Utah	
Hawaii		Nevada		Vermont	
Idaho		New Hampshire		Virginia	
Illinois		New Jersey		Washington	
Indiana		New Mexico		West Virginia	
Iowa		New York		Wisconsin	
Kansas		North Carolina		Wyoming	
Kentucky		North Dakota			

6. Record in the table below the following data concerning each student in this class:

   - the number of states in the U.S. that he/she has visited
   - the number of countries that he/she has visited
   - whether or not he/she has been to Walt Disney World (WDW) in Florida

- whether or not he/she has been to Europe

student	states	nations	WDW	Europe	student	states	nations	WDW	Europe
1					16				
2					17				
3					18				
4					19				
5					20				
6					21				
7					22				
8					23				
9					24				
10					25				
11					26				
12					27				
13					28				
14					29				
15					30				

7. How many words are in the sentence that you wrote in response to question 1?

8. Count the number of letters in each word that you wrote in response to question 1. Record these below:

9. Among the occupations listed below, label with a "W" an occupation where you would expect to find a high proportion of women. Label with an "M" an occupation where you would expect to find a high proportion of men.

   Architect Nurse Elementary School Teacher Social Worker
   Physician Musician Photographer Lawyer

10. Take a guess as to the percentage of lawyers who are women.

11. For each of the following pair of sports, identify the one that you consider more hazardous to its participants:

    - bicycle riding and football:
    - soccer and ice hockey:
    - swimming and skateboarding:

Categorical and measurement variables

Data is typically collected from a number of people or things. We define a **variable** to be some characteristic of an individual person or thing that can be assigned a number or a category. The person or thing which is assigned the number or category is called the **case** or **observational unit**. In the "states visited" example above, the students in this class are the cases of interest. If we were analyzing the number of residents in each of the 50 states, the states themselves would be the cases.

In this book, we will distinguish between two different **types** of variables. A **categorical variable** is a characteristic of an individual which can be broken down into different classes or categories. Simple examples of a categorical variable are the eye color of a student, the political affiliation of a voter, the manufacturer of your current car, and the letter grade in a particular class. Typically, a categorical variable is nonnumerical, although numbers are occasionally used in classification. The social security number of a person is an example of a categorical variable, since its main purpose is to identify or classify individuals. **Binary variables** are categorical variables for which only two possible categories exist.

A **measurement variable** is a number associated with an individual that is obtained by means of some measurement. Examples of a measurement variable include your age, your height, the weight of your car, and the distance that you traveled during your Thanksgiving vacation. A measurement variable will have a range of possible numerical values. A person's age, for example, ranges from 0 to approximately 100.

The techniques for graphing and summarizing data depend on the type of the variable that is collected. The way eye colors of students are described and summarized differs significantly from the way a group of students' heights is organized.

IN-CLASS ACTIVITIES

Activity 1-1: Types of Variables

(a) For each of the variables listed below, indicate whether it is a measurement or a categorical variable. If it is a categorical variable, indicate whether or not it is a binary variable.

gender:

political identification:

penny question:

value of statistics:

# of states:

# of countries:

Europe?:

WDW?:

letters per word:

(b) Suppose that instead of recording the number of letters in each word of your sentence, you had been asked to classify each word according to the following criteria:

1-3 letters:	small word
4-6 letters:	medium word
7-9 letters:	big word
10 or more letters:	very big word

In this case, what type of variable is size of word?

(c) Suppose that instead of recording whether or not you have been to Walt Disney World, you had been asked to report the number of times that you have been to Walt Disney World. What type of variable would this have been?

As the term variable suggests, the values assumed by a variable differ from case to case. Certainly not every student has visited the same number of states or is of the same gender! In other words, data display **variability**. The pattern of this variability is called the **distribution** of the variable. Much of the practice of statistics concerns distributions of variables, from displaying them visually to summarizing them numerically to describing them verbally.

Count tables and bar graphs

To illustrate methods for displaying and summarizing a batch of categorical data, consider the following example. I am shopping for a car. I decide that new cars are too expensive and so I decide to shop for a used car. I open the classified section of my local newspaper and look at the used car ads. I record the year and car manufacturer for the first twenty cars that are listed. The observations are recorded in the table below

1992 Honda	1994 Oldsmobile	1967 Chevrolet	1985 Nissan
1994 Toyota	1987 Ford	1968 Chevrolet	1980 Ford
1989 Pontiac	1993 Toyota	1991 Chevrolet	1988 Oldsmobile
1990 Volkswagon	1984 Mercury	1994 Ford	1989 Ford
1992 Ford	1986 Volkswagon	1995 Buick	1977 Jeep

Here there are 20 observations or cars in the table. For each car, two categorical variables have been recorded: the year the car was manufactured and the name of the manufacturer. (Actually, the year can be viewed as a measurement variable. From the manufacturer year, one can compute the age of the car which is a measurement.)

Suppose that I am interested in learning about the different car manufacturers that are listed in the car ad. I can organize this data by means of a **count table**. First I list all of the car manufacturers in one column and make a **tally column** to the right of the first column. As I read through each data value, I make a mark (|) in the tally column corresponding to the particular manufacturer. After I go through all 20 cars, the following table is produced.

Manufacturer	Tally
Honda	|
Oldsmobile	||
Chevrolet	|||
Nissan	|
Toyota	|
Ford	|||||
Pontiac	|
Volkswagon	||
Mercury	|
Buick	|
Jeep	|

From this tally table, I can construct a **count table**. For each car manufacturer, I record the number of tallies. This number is the count of cars which are manufactured by that particular manufacturer. I place these numbers in a column labeled 'Count'. So we see that five of the cars are Fords, three are Chevrolet, and only one car was a Honda.

It is also useful to compute **proportions** in this table. The proportion of Chevrolet, say, is the proportion of all of the cars that were manufactured by Chevrolet. By summing the count column of the table, we see that the total number of car ads is 20. To find the proportion of Chevrolet, we divide the count 3 by the total number 20 — the proportion of Chevrolet is 3/20 = .15. We make this computation for each car manufacturer and place the proportions in a separate column. The final table with counts and proportions is shown in the table below. The total row at the bottom of the table gives the sums of the count and proportion columns. The sum of the proportion column should always be equal to one.

Manufacturer	Count	Proportion
Honda	1	.05
Oldsmobile	2	.10
Chevrolet	3	.15
Nissan	1	.05
Toyota	2	.10
Ford	5	.25
Pontiac	1	.05
Volkswagon	2	.10
Mercury	1	.05
Buick	1	.05
Jeep	1	.05
TOTAL	20	1

What have we learned from this count table? In the car ads that I read, the most popular car brands were Ford and Chevrolet. In fact, 40% (25% + 15%) of the cars were one of the two types. What proportion of the cars were American made? The American brands correspond to Oldsmobile, Chevrolet, Ford, Pontiac, Mercury, Buick and Jeep. Summing the counts for these brands, we see that 14 cars are American which corresponds to a proportion of 14/20 = .7 or 70%. From this relatively small sample, it appears that there are a high proportion of American cars on the used car market.

A **bar graph** is a graphical display of a count table for a categorical variable. To construct this graph for the car manufacturer data, we list the different brands along the horizontal axis. The vertical axis corresponds to the count. We place tic marks along this axis starting at 0 to the largest count in the table. Then for each car manufacturer, we draw a vertical bar with height equal to the corresponding count of the brand. The resulting graph is displayed below.

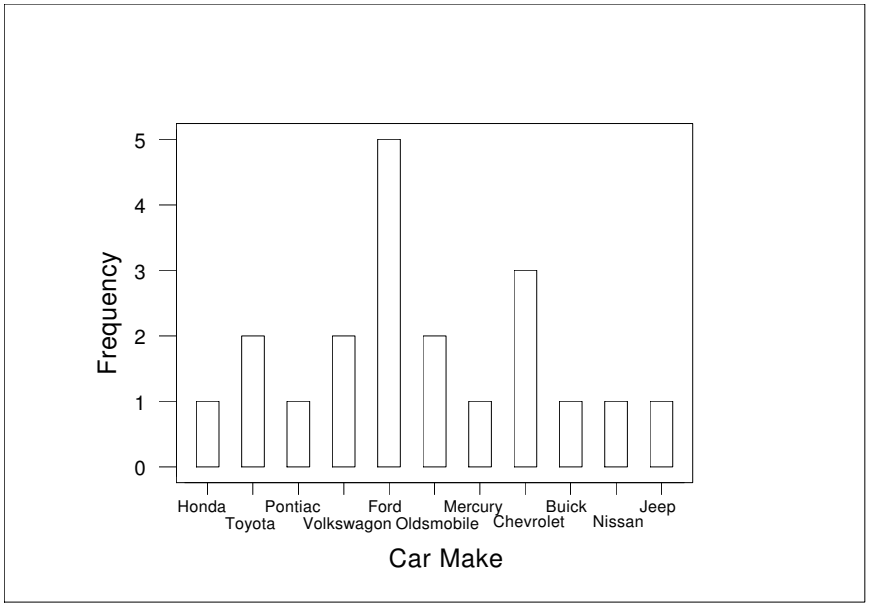
This bar graph provides a visual summary of the car brand data. It is easy to see from the graph that Chevrolet and Ford were the more popular brands in the newspaper ad.

Activity 1-2: Penny Thoughts

(a) How many students responded to the question about whether the United States should retain or abolish the penny? How many of these voted to retain the penny? What proportion of the respondents is this?

(b) How many and what proportion of the respondents voted to abolish the penny?

(c) Create a visual display of this distribution of responses using a bar graph.



Bar graph of car manufacturers.

(d) Write a sentence or two describing what your analysis reveals about the attitudes of students in this class toward the penny.

Activity 1-3: Value of Statistics

Consider the question of students' rating the value of statistics in society on a numerical score of 1 to 9. Tally the responses by counting how many students answered 1, how many answered 2, and so on.

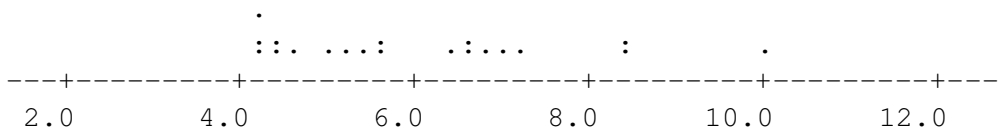
rating	1	2	3	4	5	6	7	8	9
tally (count)									

Dotplots

Remember that a measurement variable is a number associated with a person or thing that is the result of a measurement. As an example, consider the data below which gives the July 1994 unemployment rate for the twenty largest cities in the United States. The unemployment rate for a particular city, say Chicago, is computed by dividing the number of people unemployed who are seeking work in Chicago by the total number of eligible workers. We're interested in learning about the extent of unemployment in the major cities. Do the major cities have similar levels of unemployment? What is an average or typical unemployment rate? Are there particular cities that have unusually high or low levels of unemployment?

City	Rate	City	Rate	City	Rate
New York City	8.3	Phoenix	4.9	Jacksonville	5.1
Los Angeles	10.0	Detroit	6.8	Columbus	4.1
Chicago	5.6	San Antonio	5.6	Milwaukee	4.5
Houston	6.9	San Jose	7.1	Memphis	4.4
Philadelphia	6.5	Indianapolis	4.4	Washington, D. C.	4.2
San Diego	8.3	San Francisco	6.5	Boston	5.3
Dallas	5.6	Baltimore	6.3		

The first step in understanding the variation in the unemployment rates is to graph the data. One basic graphical display is the **dotplot**. To make this display, we first make a number line with sufficient range to cover all of the data values. In this example note that all of the unemployment rates fall between 4.1 and 10, so the number line can go from 4 to 10. Then we represent each data item by a large dot at the corresponding location. So we place a dot at 8.3 for New York City, a dot at 10 for Los Angeles, and so on. It may happen that we wish to put a dot at a location where a dot already exists. Then we put the new dot above the location of the previous dot. The final dotplot for the employment rate data looks like the Minitab computer display below.

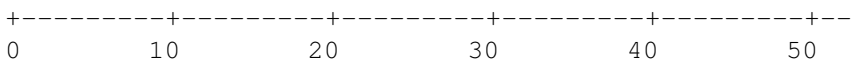


What have we learned from this picture of the data? There is a cluster of values between 4 and 7 which indicates that most of the employment rates of the major cities in the United States fall in the 4-7 percent range. There are three large unemployment rate values that appear to be separated

from the main cluster. Looking back at the table on page 10, we see that the largest unemployment rate is Los Angeles (10.0) and New York City and San Diego have rates of 8.3.

Activity 1-4: Students' Travels

(a) Create a dotplot of distribution of the numbers of states visited. A horizontal scale has been drawn below; you are to place a dot for each student above the appropriate number of states visited For repeated values, just stack the dots on top of each other.



(b) Circle your own value on the display. Where do you seem to fall in relation to your peers with regard to number of states visited?

(c) Based on this display, comment on the accuracy of your guesses in the Preliminaries section.

(d) Write a paragraph of at least four sentences describing various features of the distribution of states visited. Imagine that you are trying to explain what this distribution looks like to someone who cannot see the display and has absolutely no idea about how many states people visit. Here and throughout the course, please relate your comments to the context; remember that these are states visited and not just arbitrary numbers!

Activity 1-5: Women Employed

The table below lists the number of men and women employed in the United States in 1995 for different occupations. The numbers listed are in thousands, so the number 163 in the Architect row means that there were 163,000 architects employed in the U.S. in 1995.

Occupation	Male	Female	Total	Percentage
Architect	131	32	163	
Engineer	1772	162	1934	
Math/Computer Science	813	382	1195	
Natural Science	377	142	519	
Physician	524	169	693	
Registered Nurse	136	1841	1977	
Teacher - PK, K	9	489	498	
Teacher - Elem.	276	1462	1738	
Teacher - Second.	530	702	1232	
Lawyer	658	236	894	
Musician	101	60	161	
Photographer	99	37	136	
Barber	73	14	87	
Hairdresser	60	690	750	
Social Worker	233	494	727	
Librarian	31	164	195	

(a) Compute the percentage of women employed in each occupation. Put your percentages in the table above.

(b) Identify the three occupations with the *highest* percentages of women employed and the three occupations with the *lowest* percentages of women employed.

(c) Construct a dotplot of the distribution of percentage women. Based on examining this dotplot, write a brief paragraph describing the key features of this distribution of percentages.

(d) One can compute that the "average" percentage of women employed among these 16 occupations is 50%. Looking at the dotplot, would it be reasonable to say that 50% is a typical percentage of women employed among these occupations? Why or why not?

HOMEWORK ACTIVITIES

Activity 1-6: Types of Variables (cont.)

Suppose that each of the following was a variable that you were to measure for each student in this class. Indicate whether it is a measurement variable or a categorical variable; if it is categorical indicate whether it is also binary.

(a) height

(b) armspan

(c) ratio of height to armspan

(d) time spent sleeping last night

(e) whether or not the individual went to sleep before midnight last night

(f) month of birth

(g) numerical score (out of a possible 100 points) on the first exam in this course

(h) whether or not the individual scores at least 70 points on the first exam in this course

(i) distance from home

(j) whether the individual owns a credit card

Activity 1-7: Types of Variables (cont.)

For each of the following variables, indicate whether it is a measurement variable or a categorical (possibly binary) variable. Also identify the case (observational unit) involved. (You will encounter each of these variables as the course progresses.)

(a) whether a spun penny lands *heads* or *tails*

(b) the color of a Reese's Pieces candy

(c) the number of calories in a fast food sandwich

(d) the life expectancy of a nation

(e) whether an American household owns a cat or does not own a cat

(f) the number of years that a faculty member has been at a college

(g) the comprehensive fee charged by a college

(h) for whom an American voted for in the 1992 Presidential election

(i) whether or not a newborn baby tests HIV-positive

(j) the running time of an Alfred Hitchcock movie

(k) the age of an American penny

(l) the weight of an automobile

(m) whether an automobile is foreign or domestic to the United States

(n) the classification of an automobile as small, midsize, or large

(o) whether or not an applicant for graduate school is accepted

(p) the occupational background of a Civil War general

(q) whether or not an American child lives with both parents

(r) whether a college student has abstained from the use of alcohol for the past month

(s) whether or not a participant in a sport suffers an injury in a given year

(t) a sport's injury rate per 1000 participants

(u) a state's rate of automobile thefts per 1000 residents

(v) the airfare to a selected city from Harrisburg, Pennsylvania

(w) the average low temperature in January for a city

(x) the age of a bride on her wedding day

(y) whether the bride is older, younger, or the same as the groom in a wedding couple

(z) the difference in ages (groom's age minus bride's age) of a wedding couple

Activity 1-8: Students' Political Views

Consider the students' self-descriptions of political inclination as liberal, moderate, or conservative.

(a) Calculate the proportion of students who identified themselves as liberal, the proportion who regard themselves as moderate, and the proportion who lean toward the conservative.

(b) Create a bar graph to display this distribution of political inclinations.

(c) Comment in a sentence or two on what your calculations and bar graph reveal about the distribution of political inclinations among these students.

Activity 1-9: Value of Statistics (cont.)

Return again to the data collected above concerning students' ratings of the value of statistics in society.

(a) Create a dotplot of the distribution of students' responses.

(b) How many and what proportion of these students rated the value of statistics as a 5 on a 1-9 scale?

(c) How many and what proportion of these students rated the value of statistics as higher than 5?

(d) How many and what proportion of these students rated the value of statistics as less than 5?

(e) Summarize in a few sentences what the dotplot reveals about the distribution of students' ratings of the value of statistics. Specifically comment on the degree to which these students seem to be in agreement. Also address whether students seem to be generally optimistic, pessimistic, or undecided about the value of statistics.

Activity 1-10: Students' Travels (cont.)

Referring to the data collected above, construct a dotplot of the numbers of countries visited. Write a paragraph describing this distribution.

Activity 1-11: Word Lengths

(a) Create a dotplot of the lengths (number of letters) of words that you recorded in the Preliminaries section.

(b) Is there one particular length that occurs more often than any other? If so, what is it?

(c) Try to identify a length such that about half of your words are longer than that length and about half are shorter.

(d) Write a few sentences describing this distribution of word lengths.

Activity 1-12: Hazardousness of Sports

The following table lists the number (in thousands) of sports-related injuries treated in U.S. hospital emergency rooms in 1991, along with an estimate of the number of participants in the sports:

sport	injuries	participants	sport	injuries	participants
Basketball	647	26,200	Fishing	84	47,000
Bicycle riding	601	54,000	Horseback riding	71	10,100
Baseball, softball	460	36,100	Skateboarding	56	8,000
Football	454	13,300	Ice hockey	55	1,800
Soccer	150	10,000	Golf	39	24,700
Swimming	130	66,200	Tennis	30	16,700
Volleyball	130	22,600	Ice skating	29	7,900
Roller skating	113	26,500	Water skiing	27	9,000
Weightlifting	86	39,200	Bowling	25	40,400

(a) If one uses the number of injuries as a measure of the hazardousness of a sport, which sport is more hazardous between bicycle riding and football? between soccer and ice hockey? between swimming and skateboarding?

(b) Calculate the *rate* of injuries per thousand participants for the sports listed below. (Find this rate by dividing the *number of injuries* by the *number of participants* and then multiplying by one thousand. The first rate is calculated for you.)

Sport	Rate of injuries per 1000 particpants
Bicycle riding	601/54,000 $\times$ 1000 = 11.1
Football	
Soccer	
Ice Hockey	
Swimming	
Skateboarding	

(c) In terms of the injury rate per thousand participants, which sport is more hazardous between bicycle riding and football? between soccer and ice hockey? between swimming and skateboarding?

(d) How do the answers to (a) and (c) compare to each other? How do they compare to your intuitive perceptions from the Preliminaries section?

(e) Find the most and least hazardous sport according to the injury rate per thousand participants.

(f) Identify some other factors that are related to the hazardousness of a sport. In other words, what information might you use to produce a better measure of a sport's hazardousness?

Activity 1-13: Super Bowls and Oscar Winners

Select either National Football League Super Bowls or movies which have won the Academy Award for Best Picture as the cases of interest in a study. List two measurement variables and two binary categorical variables that one might study about those cases.

Activity 1-14: Variables of Personal Interest

Please list three variables that you would be interested in studying. These can be related to anything at all and not need be things that are feasible for us to study in class. Be sure, however, that these correspond to the definition of a variable given above. Also indicate in each instance what the case is. Please be very specific.

WRAP-UP

Since statistics is the science of **data**, this topic has tried to give you a sense of what data are and a glimpse of what data analysis entails. Data are not mere numbers: data are collected for some purpose and have meaning in some context. The guessing exercises in these activities have not been simply for your amusement; they have tried to instill in you the inclination to think about data in their context and to anticipate reasonable values for the data to be collected and analyzed.

You have encountered two very important concepts in this topic that will be central to the entire course: **variability** and **distribution**. You have also learned to distinguish between **measurement** and **categorical** variables. These activities have also hinted at a fundamental principle of data analysis: One should always begin analyzing data by looking at a visual display (i.e., a "picture") of the data. You have discovered two simple techniques for producing such displays: **bar graphs** for categorical variables and **dotplots** for measurement variables.

The next topic will introduce you to some more graphical displays for displaying distributions and will also help you to develop a checklist of features to look for when describing a distribution of data.

Topic 2: Displaying and Describing Distributions

Introduction

In the first topic you discovered the notion of the distribution of a set of data. You created visual displays (bar graphs and dotplots) and wrote verbal descriptions of some distributions. In this topic you will discover some general guidelines to follow when describing the key features of a distribution and also encounter two new types of visual displays— stemplots and histograms.

PRELIMINARIES

1. Take a guess as to length of the longest reign by a British ruler since William the Conqueror. What ruler do you think reigned the longest?

2. What do you think is a typical age for an American man to marry? How about for a woman?

3. Among all of the states, guess at a typical divorce rate (number of divorces for each 1000 people) in 1994.

4. Which would you guess is longer for most people: height or armspan (distance from fingertip to fingertip when arms are extended as far as possible)?

5. Take a guess concerning a typical foot length in centimeters for a student in this class.

6. Record below the gender, foot length, height, and armspan for the students in this class. Record the measurement variables in centimeters.

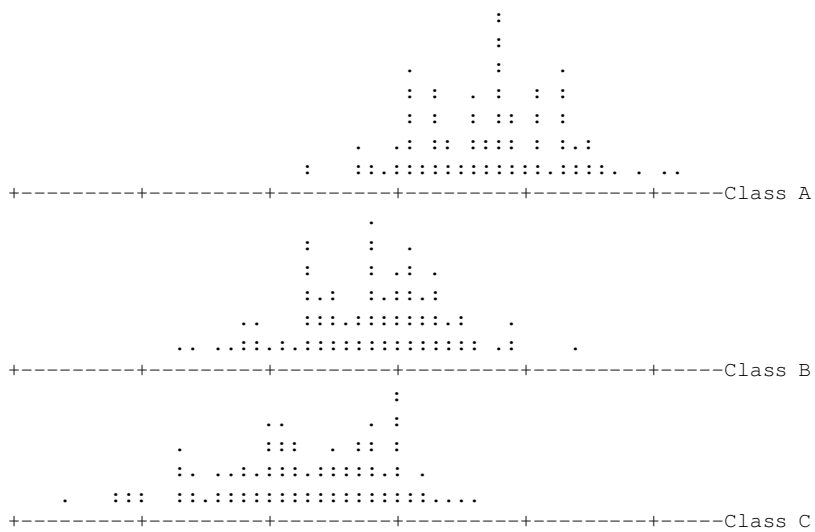
stu	gender	f length	height	armspan	stu	gender	f length	height	armspan
1					16				
2					17				
3					18				
4					19				
5					20				
6					21				
7					22				
8					23				
9					24				
10					25				
11					26				
12					27				
13					28				
14					29				
15					30				

IN-CLASS ACTIVITIES

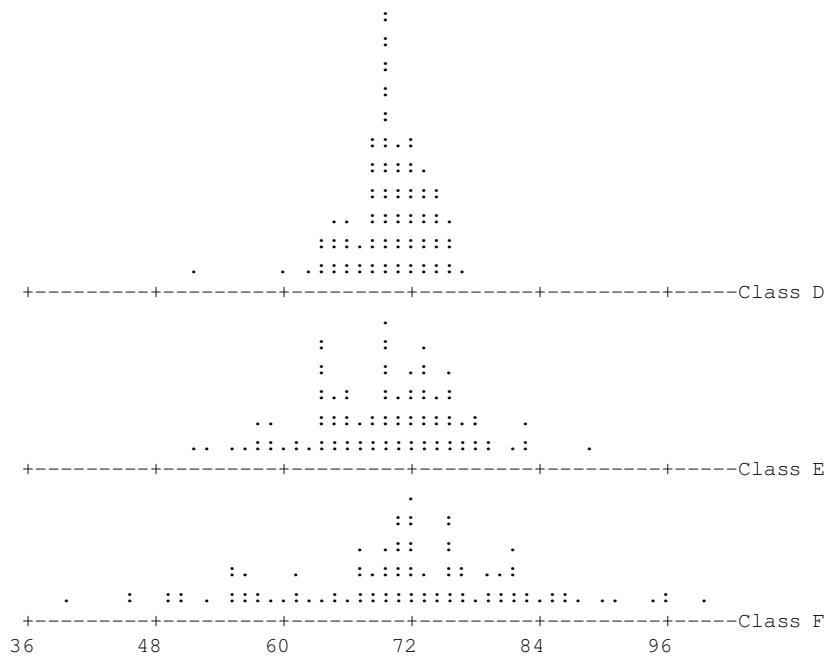
Activity 2-1: Hypothetical Exam Scores

Presented below are dotplots of distributions of (hypothetical) exam scores for twelve different classes. The questions following them will lead you to compile a ”checklist” of important features to consider when describing distributions of data.

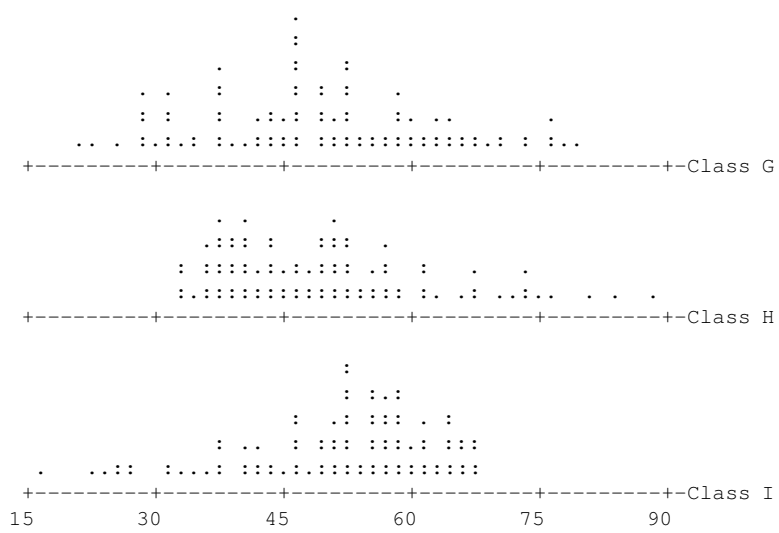
(a) Looking closely at the dotplots below, what do you think is the most distinctive difference among the distributions of exam scores in classes A, B, and C?



(b) What is most distinctive difference among the distributions of scores in classes D, E, and F?



(c) What is the most distinctive difference among the distributions of scores in classes G, H, and I?



Basic features of a data distribution

These hypothetical exam scores illustrate three basic features that are often of interest when analyzing a distribution of data:

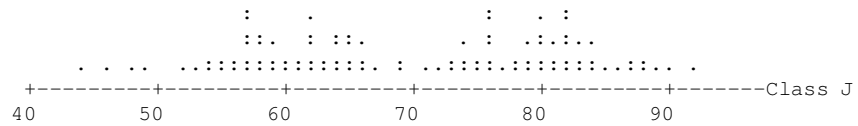
- The **center** of a distribution is usually the most important aspect to notice and describe. Where are the data?
- A distribution's **variability** is a second important feature. How spread out are the data?
- The **shape** of a distribution can reveal much information. By shape, we are referring to a general statement how the data values are distributed. While data distributions come in a limitless variety of shapes, certain shapes arise often enough to have their own names.
  - A distribution is **symmetric** if one half is roughly a mirror image of the other.
    Mound-shaped data has a symmetric distribution. For this data, most of the values fall in the middle and the values at the left and to the right trail off in about the same way.
  - A distribution is **skewed to the right** if there is a cluster of values at the left and the values trail off much farther to the right than to left.
  - A distribution that is **skewed to the left** is a mirror image of right-skewed.
    There is a cluster of values on the right and the values trail off much farther to the left than to the right.

The figure below represents the three basic data shapes using smooth curves.

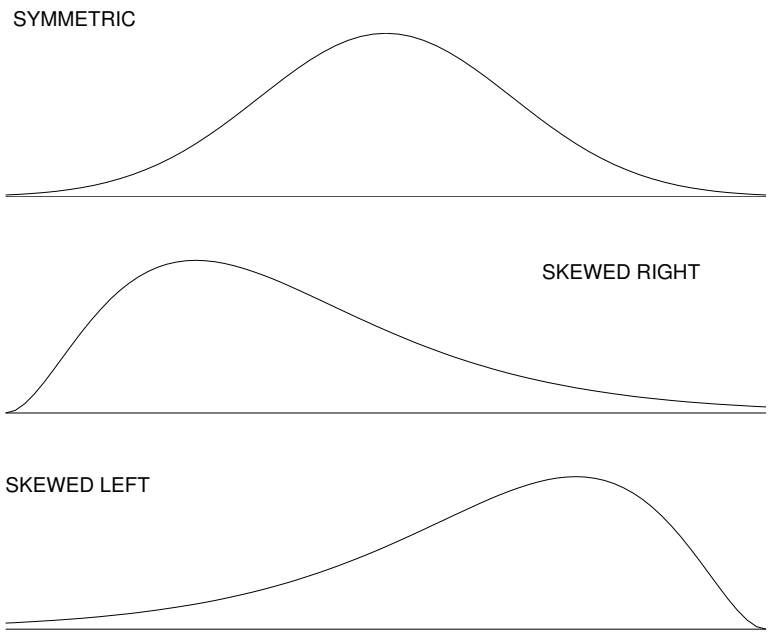
Activity 2-2: Hypothetical Exam Scores (continued)

This activity illustrates other features of data distributions that can be important.

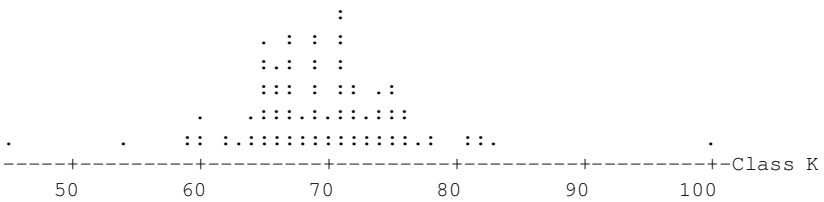
(a) What is a distinctive feature of the test scores in class J pictured below?



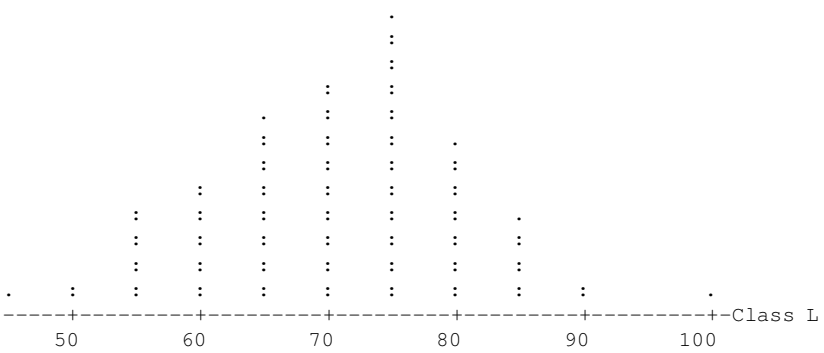
(b) What is a distinctive feature of the test scores in class K?



Three shapes of measurement data.



(c) What is a distinctive feature of the test scores in class L?



Other features of data distributions

This activity illustrates three additional features of data distributions:

- A distribution may have **peaks** or **clusters** which indicate that the data fall into natural subgroups.
- **Outliers**, observations which differ markedly from the pattern established by the vast majority, often arise and warrant close examination.
- A distribution may display **granularity** if its values occur only at fixed intervals (such as multiples of 5 or 10).

Please keep in mind that although we have formed a checklist of important features of distributions, you should not regard these as definitive rules to be applied rigidly to every data set that you consider. Rather, this list should only serve to remind you of some of the features that are typically of interest. Every data set is unique and has its own interesting features, some of which may not be covered by the items listed on our checklist.

Stemplots

A second basic graph for measurement data that is easy to construct by hand is the **stemplot**. Again we illustrate this graph on the set of unemployment data. To make this display, we first break each data value into two parts. In this example, we divide all of the values of unemployment between the ones place and the tenth place. For example, the first unemployment rate 8.3 is divided as follows:

$$8|3$$

We call the number to the left of the break the **stem** and the digit to the right of the break the **leaf**. Next we list all possible values of the stem as a single column. The smallest rate 4.1 has a stem of 4 and the largest rate 10.0 has a stem of 10, so we list stem values from 4 to 10.

```
 4|
 5|
 6|
 7|
 8|
 9|
10|
```

We next tally the data items on a stemplot by writing down the value of the leaf next to the corresponding stem. The first unemployment rate is 8.3; we record the value 3 on the 8 line. We record 4.9 by writing a 9 next to the 4 line, record 5.1 by writing 1 next to the 5 line, and so on. When we get to the value 4.1, we record a 1 next to the 9 on the 4 line. After passing through the data table by rows, we obtain the following stemplot.

```
 4| 915442
 5| 16636
 6| 89553
 7| 1
 8| 33
 9|
10| 0
```

This stemplot and the dotplot give similar information about the unemployment rates. We see from the stemplot that most of the rates fall in the 4, 5, and 6 lines. Los Angeles stands out in the stemplot with a large unemployment rate. One advantage of the stemplot is that we see the actual values of the data in the display. One can see quickly in this display how Columbus' unemployment rate (4.1) compares to the rates of other major cities.

Activity 2-3: British Rulers' Reigns

The table below records the lengths of reign (rounded to the nearest year) for the rulers of England and Great Britain beginning with William-the-Conqueror in 1066.

ruler	reign	ruler	reign	ruler	reign	ruler	reign
William I	21	Edward III	50	Edward VI	6	George I	13
William II	13	Richard II	22	Mary I	5	George II	33
Henry I	35	Henry IV	13	Elizabeth I	44	George III	59
Stephen	19	Henry V	9	James I	22	George IV	10
Henry II	35	Henry VI	39	Charles I	24	William IV	7
Richard I	10	Edward IV	22	Charles II	25	Victoria	63
John	17	Edward V	0	James II	3	Edward VII	9
Henry III	56	Richard III	2	William III	13	George V	25
Edward I	35	Henry VII	24	Mary II	6	Edward VIII	1
Edward II	20	Henry VIII	38	Anne	12	George VI	15

(a) How long was the longest reign? Who ruled the longest?

(b) What is the shortest reign? Who ruled the shortest time? What do you think this value really means?

(c) Fill in the stemplot below by putting each leaf on the row with the corresponding stem. I have gotten you started by filling in the reign lengths of William I (21 years), William II (13 years), Henry I (35 years), and Stephen (19 years).

```
0 |
1 | 3 9
2 | 1
3 | 5
4 |
5 |
6 |
```

(d) The final step to complete the stemplot is to order the leaves (from smallest to largest) on each row. Reproduce the stemplot below with the leaves ordered.

(e) Write a short paragraph describing the distribution of lengths of reign of British rulers. (Keep in mind the checklist of features that we derived above. Also please remember to relate your comments to the context.)

The stemplot is a simple but useful visual display. Its principal virtues are that it is easy to construct by hand (for relatively small sets of data) and that it retains the actual values of the observations. Of particular convenience is the fact that the stemplot also sorts the values from smallest to largest.

(f) Find a value such that one-half of these 40 British rulers enjoyed reigns longer than that value and one-half of them ruled for fewer years than that value. (Make use of the fact that the stemplot has arranged the values in order.)

(g) Find a value such that one-quarter of these 40 rulers enjoyed longer reigns and three-quarters of them had shorter reigns than that value.

(h) Find a value such that three-quarters of these 40 British rulers ruled for more years and one-quarter of them ruled for fewer years than that value.

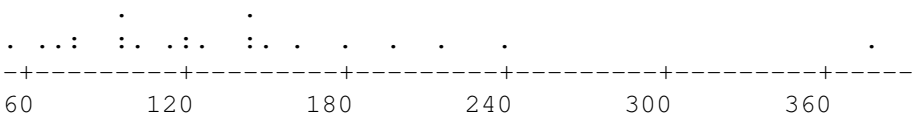
Variations of stemplots

To illustrate some alternative methods of constructing stemplots, suppose that I am thinking about buying a house. I read the real estate section of the paper and note that many of the houses will be open for inspection over the weekend. How expensive are these homes? I jot down the selling prices of 23 homes that will be open. The prices, in thousands of dollars, are displayed in the table below.

94	98	104	115	239	159	142	179
54	145	119	199	149	69	375	215
145	78	64	126	121	94	79	

Prices in thousands of dollars of 23 homes that are open for inspection.

Let us construct both a dotplot and a stemplot for these house prices. A dotplot is made by drawing a number line from 50 to 400 and recording the values on the line.



To draw a stemplot, we have to decide how to break the values. If we broke the price 104 between the tens and ones places,

$$10|4$$

we would have too many possible stem values and our display would be too long. So we break the data between the hundreds and tens places.

$$1|04$$

For this particular data value, the value of the stem is 1 and the leaf value is 0 (we ignore the remaining digit 4). For a house price of 94, we rewrite it as 094, so the stem value is 0 and the leaf is 9. Using this choice of stem and leaf, the stemplot is displayed below.

```
0| 56677999
1| 011224444579
2| 13
3| 7
```

This is not an effective display for this data since all of the house prices are put on four lines of the stemplot. The display appears to be a bit too squeezed. In cases such as this, one can see more pattern in the data if the display is stretched out. We can spread out the basic stemplot by writing down each stem value twice. For example, the stem value 0 is written as two values, '4*' and '4.'. Then the leaf values that are 0, 1, 2, 3, 4 are placed on the first line (indicated by an asterisk) and the leaf values from 5 to 9 are placed on the second line (indicated by a period). Since there are five possible leaf values on each stem line, we call this display a **stemplot with five leaves per stem**.

```
0*|
0.|56677999
1*|011224444
1.|579
2*|13
2.|
3*|
3.|7
```

We see some more detail in this display. For example, the one large house price appears to stand out more than it did in the basic stemplot display.

We can stretch out a stemplot a second way. Suppose that that each possible stem value is written down five times. For example, instead of just writing down the stem value 1 once, we write it down five times.

```
1*|
1t|
1f|
1s|
1.|
```

The leaf value for a particular stem value is written on one of the five stem lines. A leaf value of 0 or 1 is placed on the '*' line, a leaf of 2 or 3 is placed on the 't' line, a leaf of 4 or 5 is placed on the 'f' line, a leaf of 6 or 7 is put on the 's' line and leafs of 8 or 9 is placed on the '.' line. (The reason why the letters t, f, and s are used is that the words two, three start with the letter t, four and five start with the letter f, and six and seven start with the letter s.) If the house prices are placed on this stretched stemplot, we get the display below. Comparing with the basic stemplot, note that the data values in the first line $0|56677999$ have been placed on three lines in the expanded display.

```
0f|5
0s|6677
0.|999
1*|011
1t|22
1f|44445
1s|7
1.|9
2*|1
2t|3
2f|
2s|
2.|
3*|
3t|
3f|
3s|7
```

We call this graph a **stemplot with two leaves per stem** since there are two possible leaf values on each line of the display.

The dotplot and the stretched out stemplot are both helpful in understanding the variation in house prices. There is a bump of values in the 60 thousand dollar range, another cluster of values at 140 thousand dollars, and there is a trail of large numbers which corresponds to expensive homes. Since the house prices trail off toward large values instead of small values, I would describe the shape of this data as right skewed. Also, there is a definite large extreme value — the house selling for 370 thousand dollars.

Activity 2-4: Weights of Newborns

The weights (in ounces) and the gender of 27 newborn babies were taken from the hospital page in a local newspaper. The data is recorded in the table below.

weight	gender	weight	gender
147	g	110	g
102	g	125	g
107	b	129	b
90	g	114	b
126	g	124	b
105	g	102	g
143	g	97	g
123	g	123	g
117	b	126	g
126	b	118	g
132	b	136	g
110	b	121	g
121	b	133	b
87	g		

(a) In constructing a stemplot for the babies' weights, one has to first decide how to break the data into the stem and the leaf. Here there are two possibilities: one could break a weight, say 147, between the hundredths place and the tens place (so the stem is 1 and the leaf would be 4), or one could break between the tens and units places (so the stem would be 14 and the leaf would be 7). Which would be a better choice for this data? Why?

(b) Construct the basic stemplot .

(c) Construct the stemplot with five leaves per stem.

(d) Construct the stemplot with two leaves per stem.

(e) Compare the three stemplots that you constructed. Which display seems to give the best picture of the data distribution?

(f) Describe the distribution of babies' weights using at least two sentences.

Histograms

The dotplot and stemplot are helpful in organizing and displaying a relatively small amount of measurement data. When we have a large amount of data, say over 50 values, a **histogram** is a useful display. This graph is made by first grouping the data into a number of equally spaced intervals and then graphing the proportions of data values in the intervals by means of a bar chart.

To demonstrate the construction of a histogram, we will work with a larger set of house selling prices. From the newspaper, I see an ad for one of the larger realty companies in the city. The prices (in thousands of dollars) for all of the houses listed for sale for this company are listed in the below table.

148	164	307	121	161	59	139	109	127	205	154
149	61	113	34	97	72	225	289	115	79	36
119	79	168	122	118	129	189	118	217	179	65
179	67	117	104	73	225	209	134	264	126	298
177	99	116	79	62	449	144	143	129	76	595
126	215	86	89	68	119	139	57	169	119	115
134	54	110	77	124	74	206	58	89	125	254
121	91	79	89	135	168	64	109	123	125	158
155	64	53	137	237	109	52	229	94	39	184
139	154	175	119	75						

Prices in thousands of dollars of 104 homes for sale by a particular realtor.

We can organize this set of 104 prices by means of a **grouped count table**. To make this table, one first breaks the range of the data into a number of intervals of equal width. The house prices range from a low of 34 thousand dollars to a high of 595 thousand dollars, and these intervals should cover these extreme values. Suppose that we use the intervals (1, 50), (51, 100), (101, 150), (151, 200), (201, 250), (251, 300), (301, 350), (351, 400), (401, 450), (451, 500), (501, 550), (551, 600). Here we are using twelve intervals and each interval contains 50 possible house prices. For example, the first interval (1, 50) will contain the 50 possible prices 1, 2, ..., 50.

To construct a grouped count table, we list the twelve intervals, and make a tally for each data value next to the interval in which the value falls. For example, the first house price 148 falls in the (101, 150) interval and I would make a tally mark next to this interval. After tallying the complete set of data and converting the tally marks to counts, we obtain the following table

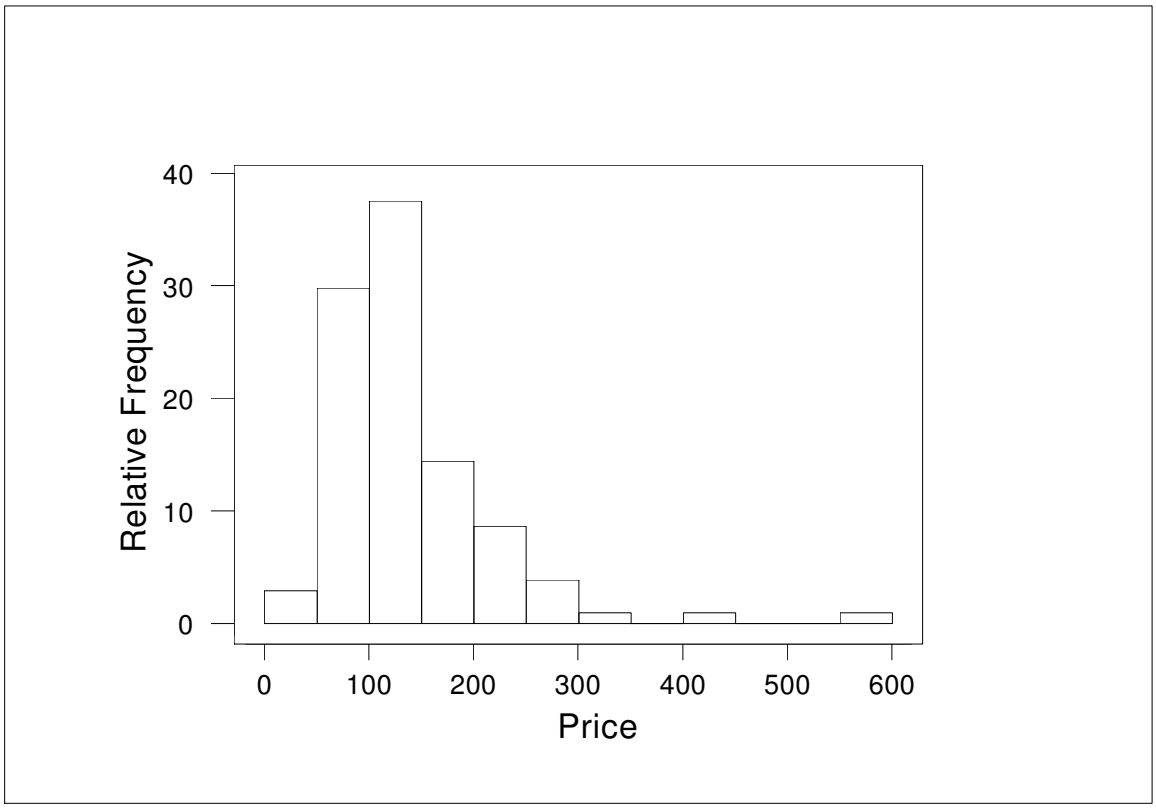
Interval	Count	Proportion
(1, 50)	3	.03
(51, 100)	31	.30
(101, 150)	39	.38
(151, 200)	15	.14
(201, 250)	9	.09
(251, 300)	4	.04
(301, 350)	1	.01
(351, 400)	0	0
(401, 450)	1	.01
(451, 500)	0	0
(501, 550)	0	0
(551, 600)	1	.01
TOTAL	104	1

Group count table for house prices data.

Next to the count column of the table, we add a proportion column. To find the proportions, we add the count column to get the total number of data values (104) and then divide each count by this total number. To illustrate, the count in the first interval (1, 50) is 3; the corresponding proportion is 3/104 = .03.

This grouped count table is helpful in understanding the distribution of prices of homes sold by this particular realtor. Note that 33% of the homes are selling for 100 thousand or less. This statement may be reassuring to a family that is looking for a relatively inexpensive house. What percentage of homes sell for more than 200 thousand? We are interested in the proportion of prices of all of the intervals from (201, 250) through (551, 600). Adding the counts in these intervals, we see that the number of houses in this range is 16 = $9 + 4 + 1 + 0 + 1 + 0 + 0 + 1$. So the proportion of prices greater than 200 thousand dollars is 16/104 = .15 or 15%.

A histogram is a graph of a grouped count table. To construct this graph we place a number line of possible house prices on the horizontal axis. The number line should start at the smallest value in the first interval and continue until the largest value in the last interval. On the vertical scale, one marks off count values or proportion values starting with zero. Then, for each interval, one makes a bar with width covering the interval and height equal to the corresponding count or proportion. A histogram of the grouped count table for the 104 house prices is shown in the below figure.



Histogram of prices of 104 houses.

Note that this histogram shows proportions in percentages. For example, we see that the proportion of houses selling between 150 and 200 thousand is approximately 15%. Also observe that the bars of the histogram are touching. This distinguishes this graph from the bar graph that we constructed earlier for categorical data.

Like a dotplot or stemplot, a histogram tells us about the shape of the distribution of house prices. We see that most of the prices fall in the 50-200 thousand dollar range. There are only a few homes that sell for under 50 thousand dollars and the prices trail off slowly at the high end. We would call this distribution of data right skewed. Also we notice a few expensive homes that are separated from the main body of data. The home price in the 550-600 interval stands out as an usually large value.

Activity 2-5: Tuitions of the Largest Colleges

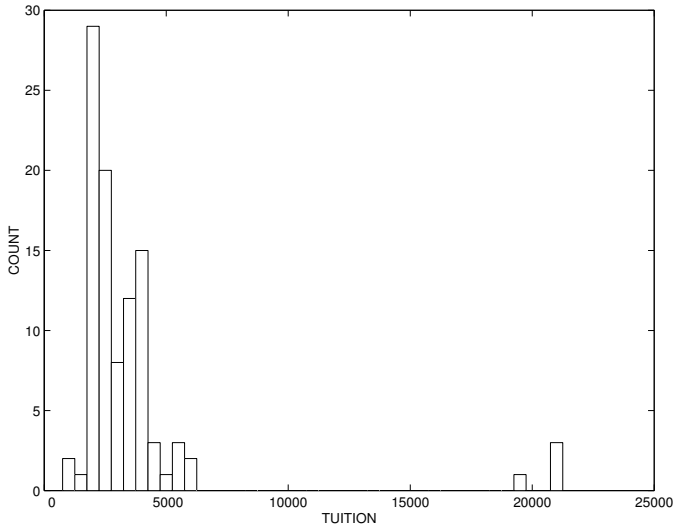
The Wall Street Journal Almanac 1998 listed the 1996-97 undergraduate tuitions and fees for the 100 largest colleges ranked by the total undergraduate and graduate enrollment. A count table for the tuitions is presented below. Some intervals for which no counts are observed are not listed. A histogram of these counts is also presented below.

Interval	Count	Interval	Count
(750, 1250)	2	(4750, 5250)	1
(1250, 1750)	1	(5250, 5750)	3
(1750, 2250)	29	(5750, 6250)	2
(2250, 2750)	20	(6250, 6750)	0
(2750, 3250)	8	(19250, 19750)	1
(3250, 3750)	12	(19750, 20250)	0
(3750, 4250)	15	(20250, 20750)	0
(4250, 4750)	3	(20750, 21250)	3

(a) How many of the colleges had tuitions between \$1750 and \$2750?

(b) How many colleges had tuitions greater than \$3750? What proportion of the colleges listed had tuitions greater than \$3750?

(c) From looking at the histogram, make some general comments about the interesting features of the distribution.



Histogram of tuitions of largest colleges.

(d) How many distinct clusters (or peaks) can you identify in the distribution? Roughly where do they fall? Can you come up with a reasonable explanation of which kinds of colleges tend to fall in which clusters?

Activity 2-6: Students' Measurements

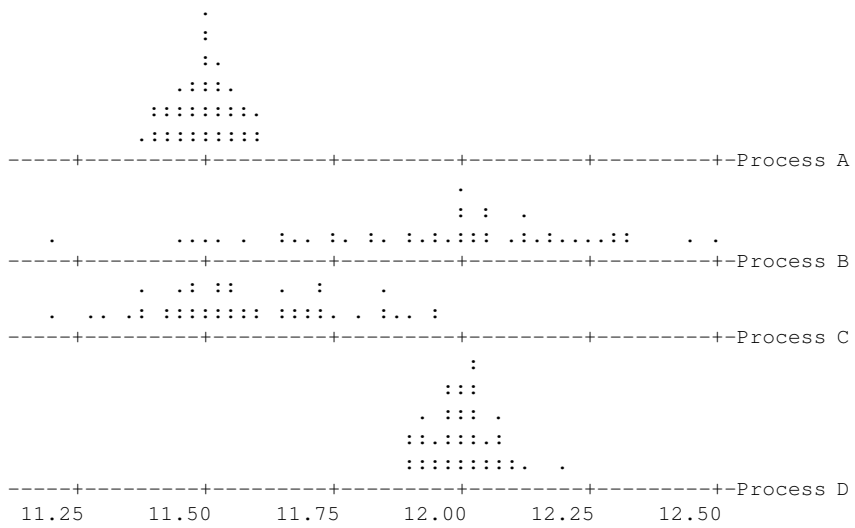
Consider the data on students' physical measurements that you gathered above. Construct a dotplot and histogram of the foot lengths. Write a paragraph commenting on key features of this distribution. (Please remember always to relate such comments to the context. Imagine that you are describing this distribution of foot lengths to someone who has absolutely no idea how long college students' feet are. By all means think about the checklist of features that we have developed as you write this paragraph.)

HOMEWORK ACTIVITIES

Activity 2-7: Hypothetical Manufacturing Processes

Suppose that a manufacturing process strives to make steel rods with a diameter of 12 centimeters, but the actual diameters vary slightly from rod to rod. Suppose further that rods with diameters

within ± 2 centimeters of the target value are considered to be within specifications (i.e., acceptable). Suppose that 50 rods are collected for inspection from each of four processes and that the dotplots of their diameters are as follows:



Write a paragraph describing each of these distributions, concentrating on the center and variability of the distributions. Also address each of the following questions in your paragraph:

- Which process is the best as is?
- Which process is the most stable; i.e., has the least variability in rod diameters?
- Which process is the least stable?
- Which process produces rods whose diameters are generally farthest from the target value?

Activity 2-8: Marriage Ages

Listed below are the ages of a sample of 24 couples taken from marriage licenses filed in Cumberland County, Pennsylvania in June and July of 1993.

couple	husband	wife	couple	husband	wife	couple	husband	wife
1	25	22	9	31	30	17	26	27
2	25	32	10	54	44	18	31	36
3	51	50	11	23	23	19	26	24
4	25	25	12	34	39	20	62	60
5	38	33	13	25	24	21	29	26
6	30	27	14	23	22	22	31	23
7	60	45	15	19	16	23	29	28
8	54	47	16	71	73	24	35	36

(a) Select either the husbands' ages or the wives' ages, and construct a stemplot of their distribution. (Indicate which spouse you are analyzing.)

(b) Write a short paragraph describing the distribution of marriage ages for whichever spouse you chose.

Activity 2-9: Hitchcock Films

The following table lists the running times (in minutes) of the videotape versions of 22 movies directed by Alfred Hitchcock:

film	time	film	time
The Birds	119	Psycho	108
Dial M for Murder	105	Rear Window	113
Family Plot	120	Rebecca	132
Foreign Correspondent	120	Rope	81
Frenzy	116	Shadow of a Doubt	108
I Confess	108	Spellbound	111
The Man Who Knew Too Much	120	Strangers on a Train	101
Marnie	130	To Catch a Thief	103
North by Northwest	136	Topaz	126
Notorious	103	Under Capricorn	117
The Paradine Case	116	Vertigo	128

(a) Construct a stemplot of this distribution.

(b) Comment on key features of this distribution.

(c) One of these movies is particularly unusual in that all of the action takes place in one room and Hitchcock filmed it without editing. Explain how you might be able to identify this unusual film based on the distribution of the films' running times.

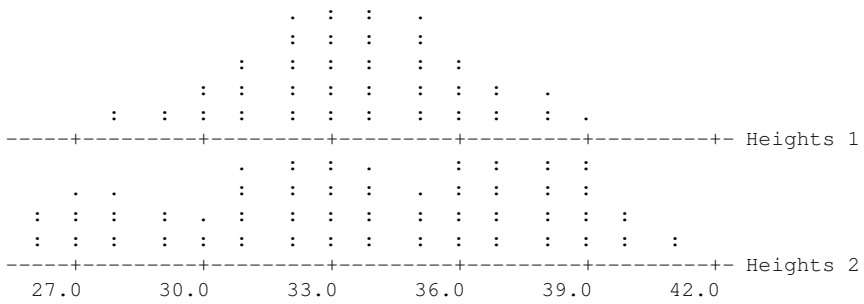
Activity 2-10: Jurassic Park Dinosaur Heights

In the blockbuster movie Jurassic Park, dinosaur clones run amok on a tropical island intended to become mankind's greatest theme park. In Michael Crichton's novel on which the movie was based, the examination of dotplots of dinosaur heights provides the first clue that the dinosaurs are not as controlled as the park's creator would like to believe. Here are reproductions of two dotplots presented in the novel:

(a) Comment briefly on the most glaring difference in these two distributions of dinosaur heights.

(b) The cynical mathematician Ian Malcolm (a character in the novel) argues that one of these distributions is characteristic of a normal biological population, while the other is what one would have expected from a controlled population which had been introduced in three separate batches (as these dinosaurs had). Identify which distribution corresponds to which type of population.

(c) Take a closer look at the first distribution. There is something suspicious about it that suggests that it does not come from real data but rather from the mind of an author. Can you identify its suspicious quality?

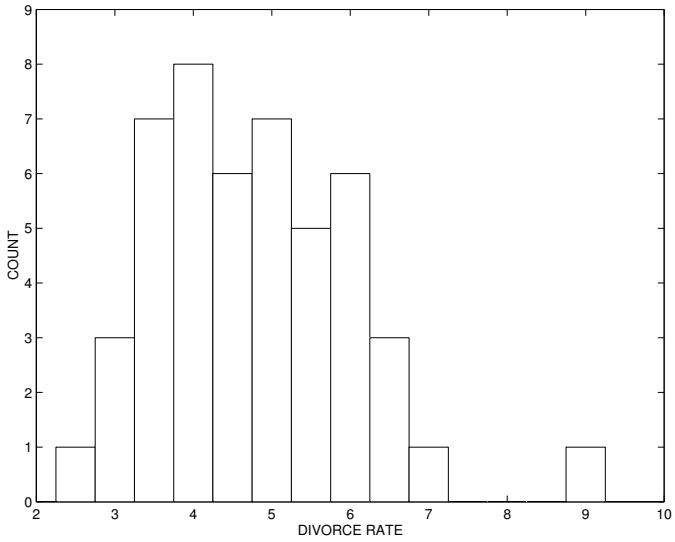


Activity 2-11: Divorce Rates

state	rate	state	rate	state	rate
Alabama	6.2	Kentucky	5.8	North Dakota	3.4
Alaska	5.5	Louisiana	*	Ohio	4.5
Arizona	5.8	Maine	4.4	Oklahoma	6.7
Arkansas	7.1	Maryland	3.5	Oregon	5.3
California	*	Massachusetts	2.4	Pennsylvania	3.3
Colorado	5.1	Michigan	4.1	Rhode Island	3.2
Connecticut	2.8	Minnesota	3.6	South Carolina	4.2
Delaware	4.8	Mississippi	5.7	South Dakota	4.2
Dist. of Col.	3.9	Missouri	5.0	Tennessee	6.6
Florida	5.9	Montana	4.9	Texas	5.4
Georgia	5.2	Nebraska	4.0	Utah	4.7
Hawaii	4.2	Nevada	9.0	Vermont	4.0
Idaho	6.2	New Hampshire	4.4	Virginia	4.6
Illinois	3.7	New Jersey	3.0	Washington	5.6
Indiana	*	New Mexico	6.0	West Virginia	5.0
Iowa	3.9	New York	3.3	Wisconsin	3.4
Kansas	4.7	North Carolina	5.1	Wyoming	6.5

The table above lists the 1994 divorce rate (per 1,000 population) for all of the states of the United States and the District of Columbia. An * in the table indicates that the divorce rate is not

available for that state. The distribution of the divorce rates is displayed using the histogram below. The intervals that are used in constructing the histogram are 2.25-2.75, 2.75-3.25, 3.25-3.75, and so on. Use the histogram to answer the following questions.



Histogram of 1994 divorce rates

(a) How many states had divorce rates that were between 3.75 and 6.25?

(b) What *proportion* of states had divorce rates smaller than 3.75?

(c) Describe the general shape of the data.

(d) What is a typical divorce rates for the states?

(e) Are there any outliers in this dataset? Looking back at the table, what states correspond to these outliers? Can you explain why these states have unusual divorce rates?

Activity 2-12: Turnpike Distances

The Pennsylvania Turnpike extends from Ohio in the west to New Jersey in the east. The distances (in miles) between its exits as one travels west to east are listed below:

exit	name	miles	exit	name	miles
1	Ohio Gateway	*	16	Carlisle	25
1	New Castle	8	17	Gettysburg Pike	9.8
2	Beaver Valley	3.4	18	Harrisburg West Shore	5.9
3	Cranberry	15.6	19	Harrisburg East	5.4
4	Butler Valley	10.7	20	Lebanon-Lancaster	19
5	Allegheny Valley	8.6	21	Reading	19.1
6	Pittsburgh	8.9	22	Morgantown	12.8
7	Irwin	10.8	23	Downingtown	13.7
8	New Stanton	8.1	24	Valley Forge	14.3
9	Donegal	15.2	25	Norristown	6.8
10	Somerset	19.2	26	Fort Washington	5.4
11	Bedford	35.6	27	Willow Grove	4.4
12	Breezewood	15.9	28	Philadelphia	8.4
13	Fort Littleton	18.1	29	Delaware Valley	6.4
14	Willow Hill	9.1	30	Delaware River Bridge	1.3
15	Blue Mountain	12.7			

(a) In preparation for constructing a histogram to display this distribution of distances, count how many values fall into each of the sub-intervals listed in the table below:

miles	0.1-5.0	5.1-10.0	10.1-15.0	15.1-20.0
tally (count)				
miles	20.1-25.0	25.1-30.0	30.1-35.0	35.1-40.0
tally (count)				

(b) Construct (by hand) a histogram of this distribution.

(c) Comment in a few sentences on key features of this distribution.

(d) Find a value such that half of the exits are more than this value apart and half are less than this value. Also explain why such a value is not unique.

(e) If a person has to drive between consecutive exits and only has enough gasoline to drive 20 miles, is she very likely to make it? Assume that you do not know which exits she is driving between, and explain your answer.

(f) Repeat (e) if she only has enough gasoline to drive 10 miles.

Activity 2-13 - Sales of Restaurant Chains

The Wall Street Almanac 1998 gives the sales (in 1996) of the 25 largest restaurant chains. For each restaurant chain, the table below gives the total sales (in millions of dollars) and the number of units or restaurants. By dividing the total sales by the number of units, one obtains the sales (in

millions of dollars) per restaurant unit that is displayed in the third column. This sales per unit can be considered an average earnings of a restaurant from a particular chain. For example, an "average" McDonald's restaurant earned \$1.4 milllion during the year 1996.

chain	sales	units	sales per unit	chain	sales	units	sales per unit
McDonald's	16369	12094	1.4	Red Lobster	1776	683	2.6
Burger King	7484	7057	1.1	Dunkin' Donuts	1593	3200	0.5
Pizza Hut	4917	8701	0.6	Applebee's	1464	823	1.8
Taco Bell	4419	6645	0.7	Jack in the Box	1248	1251	1.0
Wendy's	4284	4369	1.0	Olive Garden	1239	475	2.6
KFC	3900	5079	0.8	Shoney's	1209	844	1.4
Hardee's	2988	3225	0.9	Boston Market	1166	1087	1.1
Subway	2700	10848	0.2	Chili's	1055	487	2.2
Dairy Queen	2602	5035	0.5	Cracker Barrel	1029	283	3.6
Domino's	2300	4300	0.5	Outback Steakhouse	1021	372	2.7
Arby's	1915	2859	0.7	Sonic Drive-In	1012	1587	0.6
Denny's	1883	1571	1.2	TGI Friday's	935	319	2.9
Little Caesar	1800	4810	0.4				

(a) Using the number line below, construct a dotplot for the sales per unit.

```
+------+------+------+------+------+------+------+---
0.0   0.5    1.0    1.5    2.0    2.5    3.0    3.5  SALES PER UNIT
```

(b) Describe in a few sentences the basic features of this dataset.

(c) From looking at the dotplot, how many restaurant chains had an average earning exceeding \$2 million? What proportion of chains had an average exceeding \$2 million?

(d) From the dotplot, it should appear that there are two distinctive clusters of observations. Looking at the list of restaurants, can you describe the types of restaurants that have average sales that fall in each cluster?

Activity 2-14: ATM Withdrawals

The following table lists both the number and total amount of cash withdrawals for a particular individual from an automatic teller machine (ATM) during each month of 1994. (For example, January saw nine withdrawals which totaled \$1020.)

month	#	total	month	#	total	month	#	total
January	9	$1020	May	8	$980	September	10	$850
February	8	$890	June	13	$1240	October	10	$1110
March	10	$970	July	4	$750	November	7	$860
April	9	$800	August	9	$1130	December	14	$1680

(a) Create a dotplot of the distribution of the number of withdrawals in each month, and comment briefly on the distribution.

(b) Create a dotplot of the distribution of the total amount withdrawn in each month, and comment briefly on the distribution.

(c) Which month had the most withdrawals and by far the most cash withdrawn? Suggest an explanation for this.

(d) This individual took two extended trips in one of these months. Can you guess which month it was based on these data? Explain.

Activity 2-15: Word Lengths (cont.)

Reconsider the data that you collected in Topic 1 concerning the number of letters in your words. Comment on features of this distribution with regard to the six features enumerated above.

WRAP-UP

With this topic you have progressed your study of distributions of data in many ways. You have created a checklist of various features to consider when describing a distribution verbally: center, spread, shape, clusters/peaks, outliers, granularity. You have encountered three shapes that distributions often follow: symmetry, skewness to the left, and skewness to the right. You have also discovered two new techniques for displaying a distribution- stemplots and histograms.

Even though we have made progress by forming our checklist of features, we have still been somewhat vague to this point about describing distributions. The next two topics will remedy this ambiguity somewhat by introducing you to specific numerical measures of certain features (namely, center and spread) of a distribution. It will also lead you to studying yet another visual display- the boxplot.

Topic 3: Measures of Center

Introduction

You have been exploring distributions of data, representing them graphically and describing their key features verbally. Graphs such as dotplots, stemplots and histograms give us a general understanding of the distribution of a variable. In particular, from a display such as a histogram, we see where the data is concentrated and see the variation in the data. In the histogram of the house prices in the last section, we saw that many of the homes were priced near 100 thousand dollars, and the prices ranged from approximately 25 to 575 thousand dollars.

For convenience, it is often desirable to have a single numerical measure to summarize a certain aspect of a distribution. Although a graphical display is the first step in understanding data, we are next interested in describing the data by the use of one or two numbers. These numbers are helpful in communication. It is difficult to describe in words the shape of a particular distribution of data and it much easier to state numbers which summarize this distribution. We'll call these descriptive numbers **summaries**.

In this topic you will encounter some of the more common summaries of the center of a distribution, investigate their properties, apply them to some genuine data, and expose some of their limitations.

PRELIMINARIES

1. Take a guess as to how long a typical member of the U.S. Supreme Court has served.

2. Take a guess concerning the distance from the Sun for a typical planet in our solar system.

3. Take a guess as to the average salary of a Major League Baseball player in 1998.

4. Make guesses about the closest and farthest a student in this class is from "home". Also guess the distance from home of a typical student in the class.

5. Record in the table below the distances from home (estimated in miles) for each student in the class.

student	distance	student	distance	student	distance
1		11		21	
2		12		22	
3		13		23	
4		14		24	
5		15		25	
6		16		26	
7		17		27	
8		18		28	
9		19		29	
10		20		30	

Two averages – the mean and the median

What numbers should be used to describe measurement data? First, we should state a number which is an **"average"** value of the measurements. An average value is a number which is a typical measurement or a measurement that is in the center of the distribution of data.

There are two averages that are commonly used in summarizing data — the **mean** and the **median**. The mean is the well-known arithmetic average that is obtained by finding the sum of the measurements and dividing the sum by the number of data items. The median is the number which divides the data in half; half of the measurements are smaller than the median and half are larger.

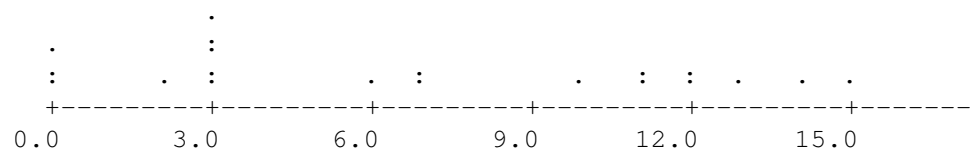
To illustrate the computation of these two measures of average, let's consider some nutrition data. One of the most popular American foods is the breakfast cereal. There are over one hundred brands of breakfast cereal that are sold in grocery stores and the cereals produced by some companies are heavily advertised on television. One general concern about breakfast cereals is their nutritional content. In particular, people think that many of the cereals popular among children contain too much sugar.

One can learn about the quantity of sugar in a breakfast cereal by reading its nutritional label on the side of the box. The following table lists the amount of sugar in grams in a single serving for twenty different cereals. What is an average amount of sugar in this collection of cereals?

Cereal	Sugar (gm)	Cereal	Sugar (gm)
100% Bran	6	Cocoa Puffs	13
Cracklin' Oat Bran	7	Crispix	3
Crispy Wheat & Raisins	10	Frosted Flakes	11
Frosted Mini-Wheats	7	Fruitful Bran	12
Honey Graham Ohs	11	Kix	3
Lucky Charms	12	Maypo	3
Nutri-grain Wheat	2	Rice Krispies	3
Shredded Wheat	0	Shredded Wheat 'n'Bran	0
Shredded Wheat spoon size	0	Smacks	15
Total Raisin Bran	14	Wheat Chex	3

Grams of sugar in one serving of twenty breakfast cereals.

First, we graph the data below using a dotplot. We see that the sugar contents of the twenty cereals appear to be distributed in four clumps. Three cereals have zero grams of sugar, a second group has approximately three grams, a third group is centered about 7, and a large group of cereals is centered about 12 grams of sugar.



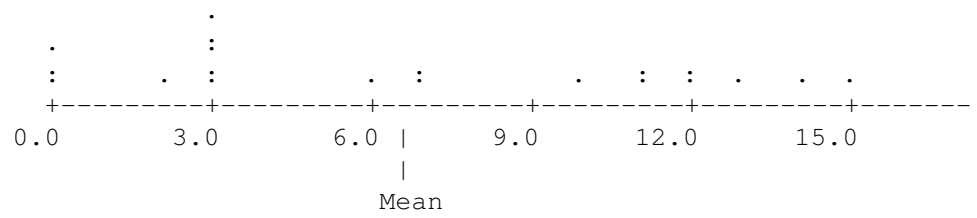
The Mean

One average sugar content of the cereals is given by the **mean** which we denote by the symbol $\bar{x}$. The mean is found by adding up all of the sugar contents and dividing by the number of data items (20).

$$\bar{x} = \frac{6+7+10+7+11+12+2+0+0+14+13+3+11+12+3+3+3+0+15+3}{20}$$

$$= 6.75$$

What is the interpretation of the mean? Suppose that the above number line is a stiff board and each dot (data item) is a weight. Then the mean is the fulcrum or point of support of the board such that the two sides are balanced. We indicate the position of the mean on the dotplot below.



The Median

A second measure of center for measurement data is the **median** which we will denote by M. This is the middle measurement when the data has been arranged in ascending order. To find the median for the cereal data, we first sort the numbers from smallest to largest:

$$0,\ 0,\ 0,\ 2,\ 3,\ 3,\ 3,\ 3,\ 3,\ 6,\ 7,\ 7,\ 10,\ 11,\ 11,\ 12,\ 12,\ 13,\ 14,\ 15$$

If we have an odd number of measurements, then the median M will be the middle value. When the number of measurements is even, then M will be the mean of the middle two numbers. Here the number of measurements is 20 which is even. So the median will be the mean of the two middle numbers, which correspond to the 10th and 11th smallest measurements. The 10th smallest number is 6 and the 11th smallest is 7; therefore the median is

$$M = \frac{6+7}{2} = 6.5$$

What is the interpretation of the median? The median M is the value where roughly half of the data is smaller than M and half of the data is larger than M. In this example, approximately one half of the sugar contents are smaller than $M = 6.5$.

Can the mean and median be different?

In this example, the values of the mean (6.75) and the median (6.5) are relatively close. Can these two measurements of "average" be different? The general answer is yes. One situation where the two measures of center can be different is when extreme data values are observed. To illustrate this situation, suppose that we make an error in recording the sugar content of the cereal Smacks. Instead of the correct value of 15 grams, I enter the sugar content incorrectly as 51 grams. How does this change affect the values of $\bar{x}$ and M? The new value of the mean is

$$\bar{x} = \frac{6+7+10+7+11+12+2+0+0+14+13+3+11+12+3+3+3+0+51+3}{20}$$

$$= 8.55$$

However, the value of the median M remains the same since the two middle measurements are still 6 and 7.

Why are the two measures of center different in this case? The mean is the arithmetic average of the entire collection of measurements. Since it uses all of the measurements in its computation, it can be affected by a single extreme data value, such as the 51 above. In contrast, the value of the median is dependent only on the order of the observations. Since the sugar content of Smacks was already the largest value, the order of the observations is not changed by increasing this one value to 51. Since the order of the observations is the same, the value of the median will not change.

One difference between the two measures of center is how they are affected by extreme data values. The median M is said to be a **resistant** measure since it is resistant or not affected by extreme observations. The incorrect large sugar measurement of Smacks had no impact on the value of the median. In contrast, the mean $\bar{x}$ can be affected by a single extreme observation. The mean is an example of a summary measure which is sensitive or **nonresistant** to extreme data values.

IN-CLASS ACTIVITIES

Activity 3-1: Supreme Court Service

The table below lists the justices comprising the Supreme Court of the United States as of October 1994. Also listed is the year of appointment and the tenure (years of service) for each.

Supreme Court Justice	year	tenure
William Rehnquist	1972	22
John Paul Stevens	1975	19
Sandra Day O'Connor	1981	13
Antonin Scalia	1986	8
Anthony Kennedy	1988	6
David Souter	1990	4
Clarence Thomas	1991	3
Ruth Bader Ginsburg	1993	1
Stephen Breyer	1994	0

(a) Create a dotplot of the distribution of these years of service.

(b) What number might you choose if you were asked to select a single number to represent the center of this distribution? Briefly explain how you arrive at this choice.

(c) Calculate the mean of these years of service. Mark this value on the dotplot above with an $\times$.

(d) How many of the nine people have served more than the mean number of years? How many have served less than the mean number of years?

(e) Calculate the median of these years of service. Mark this value on the dotplot above with an $\bigcirc$.

(f) How many of the nine people have served more than the median number of years? How many have served less than the median number of years?

It is easy enough to pick out the median (the middle observation) in a small set of data, but we will try to come up with a general rule for finding the location of the median. The first step, of course, is to arrange the observations in order from smallest to largest. Let n denote the sample size, the number of observations in the data set.

(g) With the data you analyzed above (where $n = 9$), the median turned out to be which ordered observation (the second, the third, the fourth, ...)?

(h) Suppose that there had been $n = 5$ observations; the median would have been which (ordered) one? What if there had been $n = 7$ observations? How about if $n = 11$? What about $n = 13$?

(i) Try to discover the pattern in the question above to determine a general formula (in terms of the sample size) for finding the location of the median of an odd number of ordered observations.

Activity 3-2: Faculty Years of Service

The following table presents the years of service of eight college professors:

professor	yrs	professor	yrs	professor	yrs	professor	yrs
Baric	31	Hastings	7	Reed	1	Stodghill	28
Baxter	15	Prevost	3	Rossman	6	Tesman	6

(a) Rewrite these values in order from smallest to largest.

(b) Calculate the mean of these years of service.

(c) Determine the mode of these years of service. (The **mode** is the most common value; that is, the value which occurs with the highest frequency.)

(d) Explain why finding the median is less straightforward in this case than in the case of nine Supreme Court justices.

If there are an even number of observations, the median is defined to be the average (mean) of the middle two observations.

(e) With this definition in mind, calculate the median of these years of service.

These activities should have led you to discover that the median of an odd number of observations can be found in position $\frac{n+1}{2}$ (once the values are arranged in order), while the median of an even number of observations is the mean of those occupying positions $\frac{n}{2}$ and $\frac{n}{2}+1$.

More important than the ability to calculate these values is understanding their properties and interpreting them correctly.

Activity 3-3: Properties of Averages

Reconsider the hypothetical exam scores whose distribution you analyzed in Activity 2-1.

(a) Based on the distributions as revealed in the dotplots in Activity 2-1, how would you expect the mean exam score to compare among classes A, B, and C? What about the median exam scores in these three classes?

(b) Your instructor will tell you the mean exam scores in classes A, B, and C. Record the results below. Do the calculations confirm your intuitions expressed in (a)?

	class A	class B	class C
mean			
median			

(c) Consider class G. Do you expect the mean and median of this distribution to be close together, do you expect the mean to be noticeably higher than the median, or do you expect the median to be noticeably higher than the mean?

(d) Repeat (c) in reference to class H.

(e) Repeat (c) in reference to class I.

(f) Your instructor will tell you the mean and median exam scores in classes G, H, and I. Record the results below, and indicate whether the calculations confirm your intuitions expressed in (c), (d), and (e).

	class G	class H	class I
mean			
median			

(g) Summarize what classes G, H, and I indicate about how the shape of the distribution (symmetric or skewed in one direction or the other) relates to the relative location of the mean and median.

To investigate the effect of outliers on these measures of center, reconsider the data from Activity 3-1 on Supreme Court justices' years of tenure.

(h) In Activity 3-1, you computed the mean and median of the years of tenure. Put these values in the first row of the table below.

(i) Now imagine that Justice Rehnquist has served for 42 years rather than 22. Would you expect the mean and the median to change? If so, which would change more? After answering these questions, ask your instructor to give you the new mean and medians values for this new data. Record the values in the second row of the table below.

(j) Finally, suppose that Justice Rehnquist's 22 years of service had been mistakenly recorded as 222 years. Your instructor will give you the mean and median for the Justices' data with this really big outlier. Record the values in the third row of the table.

	mean	median
Justices		
Justices with "big" outlier		
Justices with "huge" outlier		

(k) A measure whose value is relatively unaffected by the presence of outliers in a distribution is said to be **resistant.** Based on these calculations, would you say that the mean or the median is resistant? Based on the definition of each, explain briefly why it is or is not resistant.

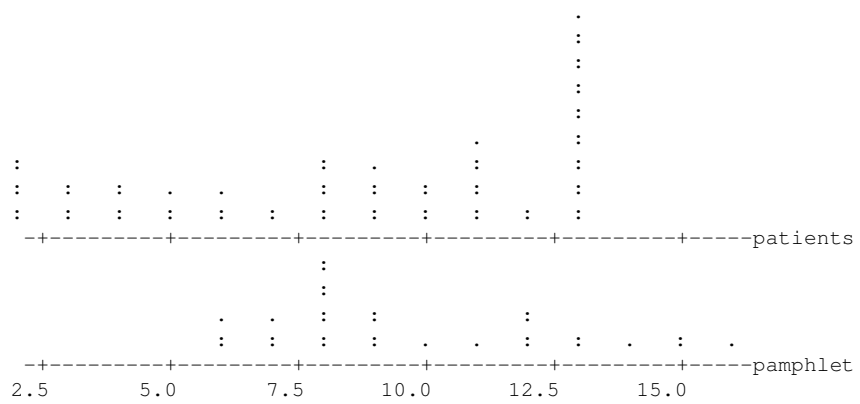
One can calculate the mean only with measurement variables. The median can be found with measurement variables and with categorical variables for which a clear ordering exists among the categories. The mode applies to all categorical variables but is only useful with some measurement variables.

Activity 3-4: Readability of Cancer Pamphlets

Researchers in Philadelphia investigated whether pamphlets containing information for cancer patients are written at a level that the cancer patients can comprehend. They applied tests to measure the reading levels of 63 cancer patients and also the readability levels of 30 cancer pamphlets (based on such factors as the lengths of sentences and number of polysyllabic words). These numbers correspond to grade levels, but patient reading levels of under grade 3 and above grade 12 are not determined exactly.

The tallies in the following table indicate the number of patients at each reading level and the number of pamphlets at each readability level. The dotplots below display these distributions with under 3 displayed as 2 and above 12 displayed as 13 for convenience.

Patients' reading level	Tally		Pamphlets' readability	Tally
Under 3	6		6	3
3	4		7	3
4	4		8	8
5	3		9	4
6	3		10	1
7	2		11	1
8	6		12	4
9	5		13	2
10	4		14	1
11	7		15	2
12	2		16	1
Above 12	17			



(a) Explain why the form of the data do not allow one to calculate the mean reading skill level of a patient.

(b) Determine the median reading level of a patient.

(c) Determine the median readability level of a pamphlet.

(d) How do these medians compare? Are they fairly close?

(e) Does the closeness of these medians indicate that the pamphlets are well-matched to the patients' reading levels? Compare the dotplots above to guide your thinking.

(f) What proportion of the patients do not have the reading skill level necessary to read even the simplest pamphlet in the study? (Examine the dotplots above to address this question.)

This activity illustrates that while measures of center are often important, they do not summarize all aspects of a distribution.

Activity 3-5: Students' Distances from Home

Consider the data collected above concerning students' distances from home.

(a) Construct a stemplot of the distribution.

(b) Would you say that the distribution of distances from home is basically symmetric or skewed in one direction or the other?

(c) Based on the shape of the distribution, do you expect the mean distance from home to be close to the median distance, smaller than the median, or greater than the median?

(d) Have the computer calculate the mean and median distances from home, recording them below. Comment on the accuracy of your expectation in (c).

(e) What is your own distance from home? Is it above or below the mean? How does it compare to the median? About where does your distance from home fall in the distribution?

(f) If you could not look at the dotplot and histogram of these distances but were only told the mean and median distance, would you have a thorough understanding of the distribution of distances from home? Explain.

HOMEWORK ACTIVITIES

Activity 3-6: Planetary Measurements

The following table lists the average distance from the sun (in millions of miles), diameter (in miles), and period of revolution around the sun (in Earth days) for the nine planets of our solar system.

planet	distance (mill. miles)	diameter (miles)	revolution (days)
Mercury	36	3,030	88
Venus	67	7,520	225
Earth	93	7,926	365
Mars	142	4,217	687
Jupiter	484	88,838	4,332
Saturn	887	74,896	10,760
Uranus	1,765	31,762	30,684
Neptune	2,791	30,774	60,188
Pluto	3,654	1,428	90,467

(a) Calculate (by hand) the median value of each of these variables.

(b) If a classmate uses the formula and obtains a median diameter of 88,838 miles, what do you think would be the most likely cause of his/her mistake? (Notice that this is Jupiter's diameter.)

Activity 3-7: Supreme Court Service (cont.)

Use the computer to display the distribution of years of service for all Supreme Court justices who preceded the ones listed in Activity 3-1. These data are listed below:

5	31	32	23	21	21	19	22	6	6	23
1	4	14	6	9	20	3	5	34	13	31
20	34	32	8	14	15	4	8	19	7	3
8	30	28	23	34	10	5	15	23	18	4
6	16	4	18	6	2	26	16	36	7	24
9	18	28	28	7	16	5	7	9	16	17
1	33	15	14	20	13	10	16	5	16	24
13	22	19	34	11	26	10	11	1	33	15
15	20	27	8	5	29	26	15	12	5	14
4	2	5	10							

(a) Describe the general shape of the distribution.

(b) Based on this shape, do you expect the mean to be larger than the median, smaller than the median, or about the same as the median?

(c) Have the computer calculate the mean and median of these years of service. Report these values and comment on your expectation in (b).

(d) How do the mean and median of these years of service compare with those you found in Activity 3-1 for active Supreme Court justices. Offer an explanation for any differences that you find.

Activity 3-8: Food prices

Recently one of us found a newspaper from the year 1979. This paper had an ad from a foodstore listing the sale prices for a number of grocery items. The 1979 price for each of the 18 products is listed below, together with the prices of the identical items in 1995 (16 years later). For each item, we can compute the *percentage increase*. For example, in 1979 an 8.2 oz. tube of Aqua Fresh toothpaste cost \$1.18 and in 1995 it cost \$2.04. The difference in price is $\$2.04 - \$1.18 = \$0.86$, and if we divide this difference by the orginal price $\$0.86/\$1.18 = .73$, we see that this tube of toothpaste increased in price by 73%. We put all of these percentage increases in the third column of the table.

Product	1979 price	1995 price	% inc
Aqua Fresh toothpaste - 8.2 oz	1.18	2.04	73
Head and Shoulders shampoo - 7 oz	1.28	2.69	110
Hill Bros coffee - 1 lb	3.19	3.39	6
Handi-Wrap - 250 sq feet	0.99	2.99	202
Musselman's applesauce - 50 oz	1.09	2.09	92
Hunts Tomato Paste - 6 oz	0.30	0.43	43
Kraft singles - 8 oz	0.99	1.99	101
carrots - 1 lb	0.20	0.50	150
Del Monte Beets - 16 oz	0.50	0.55	10
On Cor Lasagne - 32 oz	1.99	3.00	51
Cool Whip - 8 oz	0.69	1.59	130
Ore Ida French Fries - 2 lb	0.85	2.45	188
Purex Bleach - 1 gallon	0.69	1.69	145
Cremette Spaghetti - 1 lb	0.53	0.75	42
Cheez-It Crackers - 16 oz	0.89	2.50	181
Pillsbury Cakemix - 1 lb	0.65	1.39	114
Vlasic pickles - 1 qt	0.99	3.39	242
Wishbone Italian Dressing - 16 oz	0.99	2.79	182

(a) Draw a stemplot of the price increases for the 18 items.

(b) Describe the basic shape of the distribution. Are there any items which had unusually low or high price increases over the 16 year period? Can you give a possible explanation for these extreme values?

(c) Compute the mean and median for the group of price increases.

(d) Do the mean and median agree? If so, what does that say about the shape of the dataset?

Activity 3-9: Consumer Price Index

Activity 3-8 considered the change in various grocery items in a 16 year period. More generally, how have prices of various types of expenses, such as food, housing, and alcoholic beverages changed over time? To help answer this question, the table below gives the percentage change in the consumer price index (cpi) for various expenditure categories for cities in the United States in the eight-year period between 1988 and 1996. To help read this table, note that *food at home* had a percentage increase of 40, which means that food at home generally cost 40% more in 1996 than in 1988.

Expenditure category	Percent increase
food at home	40
food away from home	30
alchoholic beverages	39
housing	33
fuel and other utlilities	27
house furnishings	7
apparel products	14
new vehicles	25
motor fuel	32
airline fares	65
intracity public transportation	46
prescription drugs	69
entertainment products	28
tobacco and smoking products	71
personal care	29
college tuition	98

(a) Construct a stemplot of the percent increases for the 16 expenditure categories.

(b) From looking at the stemplot, what seems to be a typical increase in the cpi index among the 16 categories?

(c) Are there any unusually small or large cpi increases? What categories do these extreme values correspond to?

(d) Compute the mean and median percentage increase. Do these averages agree with your guess in the typical increase in part (b)?

Activity 3-10: ATM Withdrawals (cont.)

Reconsider the data presented in Activity 2-14 concerning ATM withdrawals. There were a total of 111 withdrawals during the year.

(a) Do the data as presented in that activity enable you to identify the mode value of those 111 withdrawal amounts? If so, identify the mode. If not, explain.

(b) Do the data as presented in that activity enable you to determine the median value of those 111 withdrawal amounts? If so, determine the median. If not, explain.

(c) Do the data as presented in that activity enable you to calculate the mean value of those 111 withdrawal amounts? If so, calculate the mean. If not, explain.

(d) The following table summarizes the individual amounts of the 111 withdrawals. For example, 17 withdrawals were for the amount of $20 and 3 were for the amount of $50; no withdrawals were made of $40. Use this new information to calculate whichever of the mode, median, and mean you could not calculate before.

amount	tally	amount	tally
$20	17	$150	16
$50	3	$160	10
$60	7	$200	8
$100	37	$240	1
$120	3	$250	1
$140	8		

(e) Create (by hand) a dotplot of the distribution of these 111 withdrawals. Comment on key features of the distribution, including its unusual granularity.

Activity 3-11: Professional Baseball Salaries

Consider the salaries (in millions of dollars) of 1998 Major League Baseball players from four teams as reported in the table below.

Atlanta Braves	9.6, 8, 7.8, 7, 4.5, 4.1, 3.8, 3.5, 3, 2.5, 1.2, .7, .6, .4, .3, .2, .2, .2, .2, .2, .2, .2, .2, .2, .2, .2
Florida Marlins	10, 7, 5.9, 3.3, 1.4, 1.1, .6, .5. ,3, .3, .3, .2, .2, .2, .2, .2, .2, .2, .2, .2, .2, .2, .2, .2, .2, .2, .2
New York Yankees	8.2, 6.7, 6, 5.4, 4.3, 4.3, 4.2, 3.8, 2.9, 2.8, 2.5, 1.9, 1.8, 1.1, .9, .9, .9, .8, .8, .8, .8, .8, .3, .3, .2, .2
Cleveland Indians	7.5, 6.5, 5.4, 4.8, 3.8, 3.4, 3, 2.8, 2.8, 2.8, 2.7, 2.2, 2, 1.8, 1.7, 1.3, 1, 1, .6, .6, .5, .3, .2, .2, .2, .2, .2, .2

(a) Before you do any calculations, do you expect a mean team salary to exceed the median team salary or vice versa? Explain.

(b) Select your favorite baseball team (among the four given), and calculate the mean and median salary for that team; record them below. (Indicate the team that you are examining.)

(c) If you are the owner of the Atlanta Braves and are concerned about the costs of owning a baseball team, would you be more interested in the mean salary or the median salary? Explain.

Activity 3-12: Wrongful Conclusions

For each of the following arguments, explain why the conclusion drawn is not valid. Also include a simple hypothetical example which illustrates that the conclusion drawn need not follow from the information.

(a) A real estate agent notes that the mean housing price for an area is $125,780 and concludes that half of the houses in the area cost more than that.

(b) A businesswoman calculates that the median cost of the five business trips that she took in a month is $600 and concludes that the total cost must have been $3000.

(c) A company executive concludes that an accountant must have made a mistake because she prepared a report stating that 90% of the company's employees earn less than the mean salary.

(d) A restaurant owner decides that more than half of her customers prefer chocolate ice cream because chocolate is the mode when customers are offered their choice of chocolate, vanilla, and strawberry.

WRAP-UP

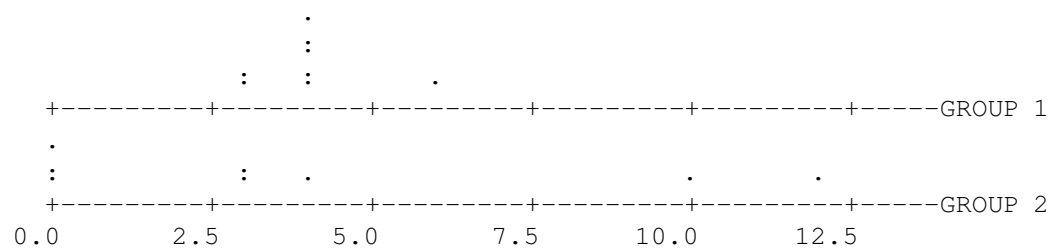
You have explored in this topic how to calculate a number of measures of the center of a distribution. You have discovered many properties of the mean, median, and mode (such as the important concept of resistance) and discovered that these statistics can produce very different values with certain data sets. Most importantly, you have learned that these statistics measure only one aspect of a distribution and that you must combine these numerical measures with what you already know about displaying distributions visually and describing them verbally.

In the next topic, you will discover similar measures of another aspect of a distribution of data- its variability.

Topic 4: Measures of Spread

Introduction

The mean or median is a summary measure which gives one an idea about the "average" measurement. But there is more to a distribution of measurements than just this average. To illustrate, suppose that we are comparing the sugar contents of two groups of cereals. The first group of cereals (called group 1) has sugar contents (in grams) of 4, 3, 4, 4, 6, 3, 4, 4. The second group (called group 2) has sugar contents of 0, 0, 12, 4, 10, 3, 3, 0. The dotplots of the two groups of measurements are displayed below using the same scale.



To summarize the two groups of data, one can use the mean. The mean of the sugar contents of group 1, call it $\bar{x}_1$, is given by

$$\bar{x}_1 = \frac{4+3+4+4+6+3+4+4}{8} = 4$$

For the second group of sugar contents, the mean $\bar{x}_2$ is

$$\bar{x}_2 = \frac{0+0+12+4+10+3+3+0}{8} = 4$$

Although the two groups of measurements have the same mean of 4 grams, it is clear from the two dotplots that the two distributions are very different. The sugar contents of group 1 are tightly clustered about the mean value of 4. In contrast, the sugar measurements of group 2 are widely scattered; two cereals have 0 grams of sugar and one has 12 grams of sugar. It is clear that a single "average" measure such as the mean is not enough to distinguish these two groups of data. We need a second summary measure which tells us something about the **spread** of the measurements. This topic will discuss different measures of spread for a batch of measurement data.

PRELIMINARIES

1. Take a guess concerning the average high temperature in January in Chicago. Do the same for the average high temperature in San Diego in January.

2. Take a guess concerning the average high temperature in July in Chicago. Do the same for the average high temperature in San Diego in July.

3. Think about the average high temperatures in January for selected cities from across the United States. Would you expect this distribution to be more or less variable than the average high temperatures in July for those same cities?

The Quartiles and the Quartile Spread

Recall the interpretation of the median M. It was the value which divided the data into two halves – half of the measurements were smaller than M and half were larger than M. Suppose we take this division idea one step further. We divide the smaller half of the measurements into two halves, and likewise divide the larger half of the measurements into two halves. The two new division values are called the **quartiles**. The distance between the two quartiles is the **quartile spread**, which is a measure of spread of the dataset.

Let's compute the quartiles and quartile spread for the sugar contents of the twenty breakfast cereals that we considered in Topic 3. Recall that we computed the median M by first writing the measurements in ascending order

$$0,\ 0,\ 0,\ 2,\ 3,\ 3,\ 3,\ 3,\ 3,\ 6,\ 7,\ 7,\ 10,\ 11,\ 11,\ 12,\ 12,\ 13,\ 14,\ 15$$

and then finding the middle value. To find the quartiles, we break the data into two halves. If the number of measurements is even, then the two halves will each contain exactly half of the measurements. In this example, there are 20 values, and one half will contain the smallest 10 values, and the second half will contain the largest 10 values. If the number of measurements is odd, then we remove the median M and divide the remaining observations in half. For example, if there were 15 observations, the median is the 8th smallest observation. If the median is removed, there are 14 remaining data values, and these would be divided into the 7 smallest and the 7 largest.

To find the lower quartile, we find the median of the half containing the smallest observations. Here the 10 smallest sugar contents are

$$0,\ 0,\ 0,\ 2,\ 3,\ 3,\ 3,\ 3,\ 3,\ 6$$

The median of these measurements is the average of the two middle values which is $(3+3)/2 = 3$. So the lower quartile, denoted by Q_L, is 3. Likewise, the upper quartile is found by finding the

median of the largest 10 sugar contents:

$$7,\ 7,\ 10,\ 11,\ 11,\ 12,\ 12,\ 13,\ 14,\ 15$$

The upper quartile, denoted by Q_U, is given by $(11 + 12)/2 = 11.5$.

We measure the spread or variation of these sugar contents by computing the quartile spread QS which is the difference between the upper and lower quartiles. For this dataset

$$QS = Q_U - Q_L = 11.5 - 3 = 8.5$$

What is the meaning of the quartiles and the quartile spread QS? Approximately one quarter of the measurements are smaller than the lower quartile Q_L and approximately one quarter of the data values are larger than the upper quartile Q_U. In our example, five sugar contents are larger than the upper quartile Q_U which represents exactly one quarter of the 20 measurements. Four of the 20 measurements are smaller than Q_L which is approximately one quarter of the data.

In addition, approximately half of the measurements fall between the lower and upper quartiles Q_L and Q_U. The quartile spread QS is the distance of the middle half of the data. In our example, $QS = 8.5$. This interpretation is that the spread of the middle half of the sugar contents of the cereals is 8.5 grams.

IN-CLASS ACTIVITIES

Activity 4-1: Supreme Court Service (cont.)

Reconsider the data from Activity 3-1 concerning the length of service of Supreme Court Justices.

(a) Create again a dotplot of the distribution of justices' years of service.

(b) Recall from Topic 3 the mean and median of the justices' years of service; record them below.

(c) The **range** is simply the difference between the largest and smallest values in the distribution. Calculate the range of the distribution of justices' years of service.

(d) To find the lower quartile Q_L of the distribution, first list the observations which fall below the location of the median. (Since there are an odd number of observations, do not include the median itself.) Then determine the median of this list.

(e) Similarly, to find the upper quartile Q_U of the distribution, list the observations which fall above the location of the median. Then determine the median of this list.

(f) Calculate the quartiles spread QS of the distribution by determining the difference between the quartiles.

The Five-number Summary

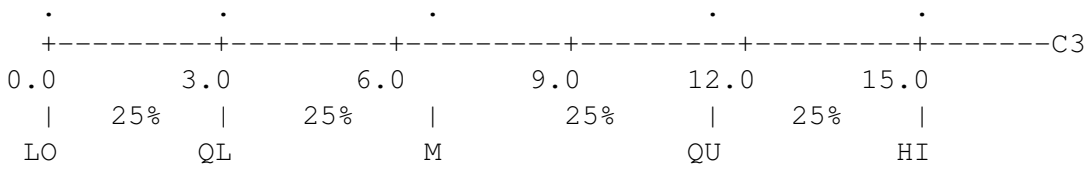
A useful set of summary numbers for a group of measurements is the median M, the lower and upper quartiles Q_L and Q_U, and the smallest and largest measurements, which we denote by LO and HI. The five-numbers written in ascending order

$$(LO,\ Q_L,\ M,\ Q_U,\ HI)$$

is called a five-number summary. For the sugar contents of the twenty cereals, the five-number summary is given by

$$(0, 3, 6.5, 11.5, 15)$$

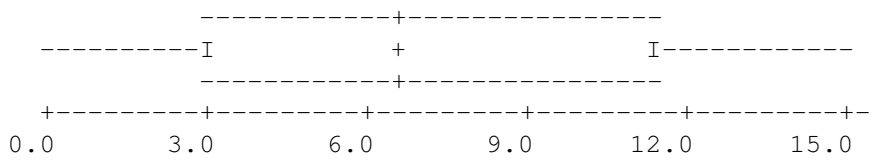
This tells us that all of the sugar contents range between 0 and 15 grams. The three middle summary values divide the data into four equal parts. On the dotplot below, we plot all of the summary values.



The numbers below the dotplot indicate that, roughly, 25% of the sugar amounts fall between 0 and 3, 25% fall between 3 and 6.5, 25% fall between 6.5 and 11.5, and 25% fall between 11.5 and 15.

One graphical display of a distribution known as the **boxplot** is based on this five-number summary. To construct a boxplot, one draws a box between the quartiles, thus indicating where the middle 50% of the data fall. Horizontal lines called whiskers are then extended from the middle of the sides of the box to the minimum and to the maximum. The median is then marked with a vertical line inside the box.

The boxplot of the sugar amounts of the cereals is graphed below.



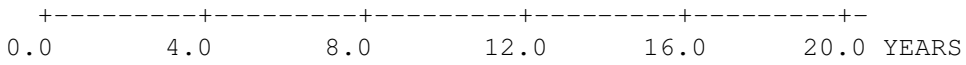
Activity 4-2: Supreme Court Service (cont.)

(a) Write down the 5-number summary of the justices' lengths of service.

(b) Half of the lengths of service are smaller than what number?

(c) Twenty five percent of the lengths of service are larger than what number?

(d) Draw a boxplot of the lengths of service on the number line below.



The Standard Deviation

To introduce a second measure of spread, consider again the sugar contents of the "group 1" cereals described in the introduction. Suppose that we look at the deviations of each data observation from the mean. In our example, the mean $\bar{x} = 4$ and the deviation of the first observation 4 from the mean is $4 - 4 = 0$, the deviation of the second observation 3 from the mean is $3 - 4 = -1$, and so on. What if we used as our measure of spread the arithmetic average of these deviations from the mean? The table below gives the calculations for our dataset. In the "data" column we list the data items and in the "deviation from mean" column, we give the deviation of each data value from the mean $\bar{x} = 4$. The arithmetic average of the deviations is given by the sum of the deviations divided by the number of observations. But note that the sum of the deviations is 0 (in fact, this is always the case), and so the arithmetic average of the deviations is also 0. So this doesn't appear to give a good measure of spread.

The average deviation from the mean is not a good measure of spread since the deviations from the mean that are positive in sign cancel out the deviations that are negative in sign. But what if we consider the squares of the deviations? In the table below, we take the square of each deviation that we just computed and place the answer in a new column labeled "deviation from mean squared".

The squared values of the deviations provide a more useful notion of spread. The measurements that are close to the mean will have a small squared deviation and the measurements, such as 4,

data	deviation from mean
4	$4-4=0$
3	$3-4=-1$
4	$4-4=0$
4	$4-4=0$
6	$6-4=2$
3	$3-4=-1$
4	$4-4=0$
4	$4-4=0$
SUM	0

Deviations from the mean $\bar{x}=4$ for sugar contents of eight cereals.

data	deviation from mean	deviation from mean squared
4	$4-4=0$	$0^2=0$
3	$3-4=-1$	$(-1)^2=1$
4	$4-4=0$	$0^2=0$
4	$4-4=0$	$0^2=0$
6	$6-4=2$	$2^2=4$
3	$3-4=-1$	$(-1)^2=1$
4	$4-4=0$	$0^2=0$
4	$4-4=0$	$0^2=0$
SUM		6

Computations for the standard deviation for sugar contents of eight cereals.

that are far from the mean will have a large positive squared deviation. To find our first measure of spread, the standard deviation, we divide the sum of squared deviations by one less than the number of measurements, and then take the square root of the result. We call the result s.

$$s=\sqrt{\frac{\text{sum of squared deviations from the mean}}{\text{number of measurements}-1}}=\sqrt{\frac{6}{8-1}}=.926.$$

Why do we take the square root at the end? We would like a measure of spread that has the same units as the data. In our example, all of the sugar contents are in grams. When we take the square of each deviation from the mean, the units of the squared deviations will be in grams^2. By taking the final square root, the unit of the standard deviation s will be in grams.

Activity 4-3: Supreme Court Service (cont.)

(a) To calculate the standard deviation of the distribution of years of service, begin by filling in the missing entries in the following table: (The mean here is $\bar{x}=8.44$.)

	original data	deviation from mean	deviation from mean squared
	22	13.56	183.87
	19	10.56	111.51
	13		
	8	-0.44	0.19
	6		
	4	-4.44	19.71
	3	-5.44	29.59
	1	-7.44	55.35
	0	-8.44	71.23
column sum	76		

(b) Divide the sum of the last column (the sum of squared deviations) by 8 (one less than the sample size). Put your answer below.

(c) Take the square root of the result in part (b) to find the standard deviation of the distribution.

Interpreting the standard deviation

What's the meaning of the standard deviation s? First, it is useful in comparing the spreads of two sets of measurements. Let's return to the comparison of the sugar amounts for the two sets of cereals in the introduction. We've already computed the standard deviation for the first set of cereal measurements. We'll compute the standard deviation of the second set of sugar amounts (0, 0, 12, 4, 10, 3, 3, 0) in one step. Recall that the mean for the second dataset was $\bar{x} = 4$. The standard deviation for this group of measurements is

$$\begin{aligned} s &= \sqrt{\frac{(0-4)^2+(0-4)^2+(12-4)^2+(4-4)^2+(10-4)^2+(3-4)^2+(3-4)^2+(0-4)^2}{8-1}} \\ &= \sqrt{\frac{16+16+64+0+36+1+1+16}{8-1}} \\ &= \sqrt{\frac{150}{8-1}} \\ &= 4.63 \end{aligned}$$

The standard deviation of the first group of cereal sugar contents is .926 and the standard deviation of the second group is 4.63. The two values of s tell us that the second set of measurements is more spread out than the first set.

What are possible values of the standard deviation s? Suppose that we look at a third group of six cereals, all with high sugar content. The numbers of grams of sugar of these cereals are 12, 12, 12, 12, 12, 12. Note that this dataset does not show any variability; all of the values are equal to the mean $\bar{x} = 12$. In this case, all of the deviations from the mean will be equal to zero and the standard deviation $s = 0$. In general, the computed value of a standard deviation can be any nonnegative number, and a value of 0 indicates that there is no spread in the measurements.

Activity 4-4: Properties of Measures of Spread

Reconsider the hypothetical exam scores that you first analyzed in Activity 2-1.

(a) Based on the distributions as revealed in the dotplots in Activity 2-1, how would you expect the standard deviations of exam scores to compare among classes D, E, and F? What about the quartile spreads among these three classes?

(b) Your instructor will give you standard deviations and quartile spreads for exam scores in classes D, E, and F. Do the calculations support your expectations from (a)?

(c) Suppose there are three classes of students, each class of size 5, that take the same exam. In the space below, make up test scores for the three groups of students, where the scores for the first class have very small variation, the scores for the second class have moderate variation, and the scores from the third class have high variation.

The 68 - 95 - 99.7 rule

The standard deviation is best understood when the group of measurements has a particular shape. Suppose that the distribution of measurements is symmetric and approximately **mound-shaped**. That is, most of the measurements are clustered in the center of the distribution and the values trail

off to the left and to the right in about the same rate. (See the top graph on page 23.) If the dataset is mound-shaped, then approximately,

- **68%** of the observations will fall within **one** standard deviation of the mean
- **95%** of the observations will fall within **two** standard deviations of the mean
- **99.7%** of the observations will fall within **three** standard deviations of the mean

We call these statements collectively the 68 - 95 - 99.7 rule.

To illustrate this rule, let's consider a measurement variable which has an approximate mound-shaped distribution. In baseball, one measure of the quality of a hitter is the batting average which is defined to be the number of base hits divided by the number of official at-bats. Simply, a batting average is the proportion of times the player gets a base hit. For the 1995 baseball season, we collect the batting averages for all of the regular players in the American League who had at least 200 at-bats. Here are a few of the measurements: Rafael Palmeiro 310, Bret Barberie 241, Manny Alexander 236, Cal Ripken 262. Note that we are ignoring the decimal point in recording these numbers: Ripken's 262 batting average is actually a proportion value of .262.

A stemplot of the batting averages of the 148 regular players in the American League is given below.

```
16| 8
17|
18|
19| 5
20| 3
21| 67999
22| 344569
23| 2566778
24| 1333344457788
25| 0111346677778
26| 00111122223333456688999
27| 0011345778888888
28| 234557788899
29| 012335557789
30| 0000011236667777888
31| 0001344478
32| 01134
33| 033
34|
35| 6
```

Note that this collection of batting averages is approximately mound-shaped. The measurements are centered about 270 and the shape of the distribution for small batting averages is approximately a mirror image of the shape for large averages.

Using the computer, one can compute the mean and standard deviation of this dataset: $\bar{x} = 273.4$ and $s = 31.8$.

Using the 68 - 95 - 99.7 rule, we expect that, approximately,

- 68% of the batting averages will fall between $\bar{x} - s$ and $\bar{x} + s$ or 273.4 - 31.8 = 241.6 and 273.4 + 31.8 = 305.2.
- 95% of the batting averages will fall between $\bar{x} - 2s$ and $\bar{x} + 2s$ or 273.4 - 2(31.8) = 209.8 and 273.4 + 2(31.8) = 337.0.
- 99.7% of the batting averages will fall between $\bar{x} - 3s$ and $\bar{x} + 3s$ or 273.4 - 3(31.8) = 178.0 and 273.4 + 3(31.8) = 368.8.

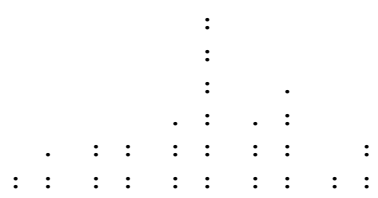
What proportion of the batting averages actually fall in the above ranges? Looking at the stem-plot above, we can count that 98 of the batting averages fall in the interval (241.6, 305.2), so the proportion that fall in this interval is 98/148 = .66 or 66%. All but 4 observations fall in the interval (209.8, 337), so the proportion in this interval is 144/148 = .97 or 97%. Only one observation is outside of the interval (178, 368.8) and the proportion inside this interval is 147/148 = .99 or 99%. Note that the observed percentages 66, 97 and 99 are very close to the percentages given in the above rule.

Activity 4-5: Placement Exam Scores

The Department of Mathematics and Computer Science of Dickinson College gives an exam each fall to freshman who intend to take calculus; scores on the exam are used to determine into which level of calculus a student should be placed. The exam consists of 20 multiple-choice questions. Scores for the 213 students who took the exam in 1992 are tallied in the following table:

Score	1	2	3	4	5	6	7	8	9	10
Count	1	1	5	7	12	13	16	15	17	32
Score	11	12	13	14	15	16	17	18	19	
Count	17	21	12	16	8	4	7	5	4	

The mean score on this exam is $\bar{x} = 10.221$. The standard deviation of the exam scores is $s = 3.859$. A dotplot of this distribution follows:



```
       .  :   :  :   :  :   :  :   :  :   :  :   :     :  .
 .  :  :   :  :   :  :   :  :   :  :   :  :   :  :   :  :   :
 -----+---------+---------+---------+---------+- SCORE
     4.0       8.0      12.0      16.0      20.0
```

(a) Does this distribution appear to be roughly symmetric and mound-shaped?

(b) Consider the question of how many scores fall within one standard deviation of the mean (denoted by $\bar{x}$). Determine the upper endpoint of this interval by adding the value of the standard deviation to that of the mean. Then determine the interval's lower endpoint by subtracting the value of the standard deviation from that of the mean.

(c) Look back at the table of tallied scores to determine how many of the 213 scores fall within one standard deviation of the mean. What proportion of the 213 scores is this?

(d) Determine how many of the 213 scores fall within two standard deviations of the mean, which turns out to be between 2.503 and 17.939. What proportion is this?

(e) Determine how many of the 213 scores fall within three standard deviations of the mean, which turns out to be between -1.356 and 21.798. What proportion is this?

Measure of Relative Standing

Suppose that a student takes a math test and receives a score of 80. She is probably interested in the letter grade – was this a high enough score to get a "B"? She may also be interested in how her grade compared to other grades in the class. In other words, what was the **relative standing** of her score in comparison with the other scores of students in her class?

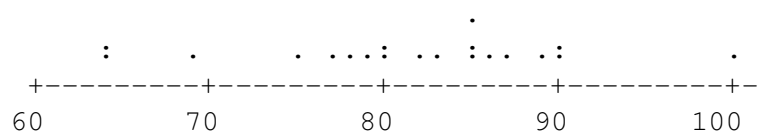
Suppose that we are given some measurement data. One way of measuring the relative standing of a particular measurement is based on the mean and standard deviation of the dataset. Suppose that we call the single measurement X and, as usual, denote the mean and standard deviation of the group of data by $\bar{x}$ and s, respectively. The **z-score** of the measurement X, denoted by z, is found by subtracting the mean from the measurement and then dividing by the standard deviation:

$$z = \frac{X - \bar{x}}{s}$$

When we subtract the mean from the measurement, we are seeing how far the measurement is from the mean. By dividing this quantity by the standard deviation, we are converting this distance from the mean to **standard deviation units**.

To illustrate the computation of a z-score, consider the collection of 20 test scores given in the table below. A dotplot of these scores is also shown.

77	69	85	86	90
64	78	80	90	100
89	80	83	85	82
85	79	64	87	75



Note that most of the test scores are in the 80-90 range. There is one perfect score (100) that stands out at the high end and three low scores under 70 that are separated from the cluster in the middle.

For this dataset, the mean $\bar{x} = 81.8$ and $s = 8.8$. Consider the particular score of 64. The corresponding z-score is

$$z = \frac{64 - 81.8}{8.8} = -2.02.$$

Since this z-score is negative, this indicates that this student's score was below the mean. The magnitude of this score refers to units of standard deviation. So the student's score was 2 standard deviations under the mean.

A second student scored 90. What was his relative standing in the class? The z-score of 90 is

$$z = \frac{90 - 81.8}{8.8} = .93.$$

The interpretation of this z-score is that this student scored .93 (approximately 1) standard deviations above the mean.

To summarize, a z-score is a way of understanding the relative standing of a measurement. It can be thought as the number of standard deviations above or below the mean. Specifically,

- if a z-score is equal to 0, then the measurement is at the mean
- if a z-score is positive, then the measurement is above the mean
- if a z-score is negative, then the measurement is below the mean
- the magnitude or size of a z-score is the number of standard deviations away from the mean

Z-scores are useful in comparing the relative position of scores in different datasets. To illustrate, suppose that I take two English tests. On both tests I scored a 80. Although the two scores

were the same, my relative standing in the class may be different for the two tests. In particular, suppose that the mean and standard deviation of the test 1 scores were 90 and 10 respectively and the mean and standard deviation of the test 2 scores were 70 and 20. The z-scores of my two grades, z_1 and z_2, are

$$z_1 = \frac{80 - 90}{10} = -1$$

$$z_2 = \frac{80 - 70}{20} = .5$$

So actually my relative standing was higher on the second test. I scored one standard deviation below the mean on test 1 and 1/2 standard deviation above the class mean on test 2. If the grades on the tests are assigned on a curve (based on relative standing), then it is possible that my grade on test 2 would be higher than my grade on test 1.

Activity 4-6: SAT's and ACT's

Suppose that a college admissions office needs to compare scores of students who take the Scholastic Aptitude Test (SAT) with those who take the American College Test (ACT). Suppose that among the college's applicants who take the SAT, scores have a mean of 896 and a standard deviation of 174. Further suppose that among the college's applicants who take the ACT, scores have a mean of 20.6 and a standard deviation of 5.2.

(a) If applicant Bobby scored 1080 on the SAT, how many points above the SAT mean did he score?

(b) If applicant Kathy scored 28 on the ACT, how many points above the ACT mean did she score?

(c) Is it sensible to conclude that since your answer to (a) is greater than your answer to (b), Bobby outperformed Kathy on the admissions test? Explain.

(d) Determine how many standard deviations above the mean Bobby scored by dividing your answer to (a) by the standard deviation of the SAT scores.

(e) Determine how many standard deviations above the mean Kathy scored by dividing your answer to (b) by the standard deviation of the ACT scores.

(f) Which applicant has the higher z-score for his/her admissions test score?

(g) Explain in your own words which applicant performed better on his/her admissions test.

(h) Calculate the z-score for applicant Mike who scored 740 on the SAT and for applicant Karen who scored 19 on the ACT.

(i) Which of Mike and Karen has the higher z-score?

(j) Under what conditions does a z-score turn out to be negative?

HOMEWORK ACTIVITIES

Activity 4-7: Hypothetical Manufacturing Processes (cont.)

Look back at the dotplots from Activity 2-7 of data from hypothetical manufacturing processes. The following table lists the means and standard deviations of these processes. Match each dotplot (process A, B, C, and D) with its numerical statistics (process 1, 2, 3, or 4).

	process 1	process 2	process 3	process 4
mean	12.008	12.004	11.493	11.723
std. dev.	0.274	0.089	0.041	0.18

Activity 4-8: Comparing Baseball Hitters

In the history of major league baseball, Mike Schmidt and Mickey Mantle were two of the greatest homerun hitters. In the table below, the number of homeruns that Schmidt and Mantle hit during each season of their careers is listed. Note that all of the baseball seasons for the two players are not shown — only the seasons where the players played full-time and had at least 300 opportunities to hit (called official at-bats) are listed.

Mike Schmidt		Mickey Mantle	
Season	Homeruns	Season	Homeruns
1973	18	1951	13
1974	36	1952	23
1975	38	1953	21
1976	38	1954	27
1977	38	1955	37
1978	21	1956	52
1979	45	1957	34
1980	48	1958	42
1981	31	1959	31
1982	35	1960	40
1983	40	1961	54
1984	36	1962	30
1985	33	1964	35
1986	37	1965	19
1987	35	1966	23
1988	12	1967	22
		1968	18

Back-to-back stemplots of the numbers of season homeruns for the two players are shown below.

```
  SCHMIDT      MANTLE

       2 | 1 | 3
       8 | 1 | 89
       1 | 2 | 1233
         | 2 | 7
      31 | 3 | 014
88876655 | 3 | 57
       0 | 4 | 02
      85 | 4 |
         | 5 | 24
```

(a) For each player, compute the five-number-summary of the numbers of season homeruns.

(b) On the average, who was the better homerun hitter during a season? Explain the reason for your choice?

(c) Compute the quartile spread QS for each player.

(d) Draw parallel boxplots of the season homeruns on the scale below.

```
--------+---------+---------+---------+---------+--------HOMERUNS
      16.0      24.0      32.0      40.0      48.0
```

(e) In the *Historical Baseball Abstract*, Bill James describes one of these two players as one of the most consistent power hitters — he had only one relatively poor season in a fifteen year stretch. Based on the above graphs and the summaries that you calculated, which player was James talking about? Why?

Activity 4-9: Climatic Conditions

The following table lists average high temperatures in January and in July for selected cities from across the U.S.

city	Jan hi	July hi	city	Jan hi	July hi
Atlanta	50.4	88	Nashville	45.9	89.5
Baltimore	40.2	87.2	New Orleans	60.8	90.6
Boston	35.7	81.8	New York	37.6	85.2
Chicago	29	83.7	Philadelphia	37.9	82.6
Cleveland	31.9	82.4	Phoenix	65.9	105.9
Dallas	54.1	96.5	Pittsburgh	33.7	82.6
Denver	43.2	88.2	St. Louis	37.7	89.3
Detroit	30.3	83.3	Salt Lake City	36.4	92.2
Houston	61	92.7	San Diego	65.9	76.2
Kansas City	34.7	88.7	San Francisco	55.6	71.6
Los Angeles	65.7	75.3	Seattle	45	75.2
Miami	75.2	89	Washington	42.3	88.5
Minneapolis	20.7	84			

(a) Calculate (by hand or with the computer) the quartile spread for the January high temperatures and for the July high temperatures.

(b) Use the computer to calculate the standard deviations for both variables.

(c) Which variable has greater variability in its distribution? Comment on the accuracy of your guess from the Preliminaries section.

(d) Which generally has higher temperatures: January or July high temperatures?

(e) Do you think that if one variable tends to cover larger values than another, then the former variable must have more variability in its values as well? Explain.

Activity 4-10: Planetary Measurements (cont.)

Refer back to the data presented in Activity 3-6 concerning planetary measurements.

(a) Calculate (by hand) the five-number summary of distance from the sun.

(b) Draw (by hand) a boxplot of the distribution of distances from the sun.

(c) Would you classify this distribution as roughly symmetric, skewed left, or skewed right?

Activity 4-11: Students' Travels (cont.)

Reconsider the data that was collected in Topic 1 concerning the number of states visited by students.

(a) Calculate (by hand) the five-number summary of this distribution.

(b) Draw (by hand) a boxplot of this distribution.

(c) Between what two values does the middle 50% of these data fall?

Activity 4-12: Word Lengths (cont.)

Reconsider the data collected in Topic 1 concerning the lengths of your words.. Calculate (by hand) the five-number summary of this distribution and draw (also by hand) a boxplot. Comment on what the boxplot reveals about the distribution of your word lengths.

Activity 4-13: Students' Distances from Home (cont.)

Refer back to the data collected in Topic 3 on students' distances from home.

(a) Look at a dotplot of the distribution of distances from home (either hand-drawn or computer-produced). Based on the shape of this dotplot, would you expect the 68-95-99.7 rule to hold in this case? Explain.

(b) Use the computer to calculate the mean and standard deviation of the distances from home. Report their values.

(c) Determine what proportion of the observations fall within one standard deviation of the mean. How close does this proportion match what the 68-95-99.7 rule would predict?

(d) Determine what proportion of the observations fall within two standard deviations of the mean. How close does this proportion match what the 68-95-99.7 rule would predict?

Activity 4-14: SAT's and ACT's (cont.)

Refer back to Activity 4-6 in which you calculated z-scores to compare test scores of college applicants.

(a) Suppose that applicant Tom scored 820 on the SAT and applicant Mary scored 19 on the ACT. Calculate the z-score for Tom and Mary and comment on which of them has the higher test score.

(b) Suppose that scores in a certain state on the Math portion of the SAT have a mean of 474 and a standard deviation of 136, while scores on the Verbal section of the SAT have a mean of 422 and a standard deviation of 122. If Charlie scores 660 on the Math portion and 620 on the Verbal portion, on which portion has he done better in the context of the other test takers in the state?

Activity 4-15: SAT's and ACT's (cont.)

Refer back to the information about SAT and ACT test scores in Activity 4-6. To answer the following questions, assume that the test scores follow a mound-shaped distribution.

(a) According to the 68-95-99.7 rule, about 95% of SAT takers score between what two values?

(b) What does the 68-95-99.7 rule say about the proportion of students who score between 722 and 1070 on the SAT?

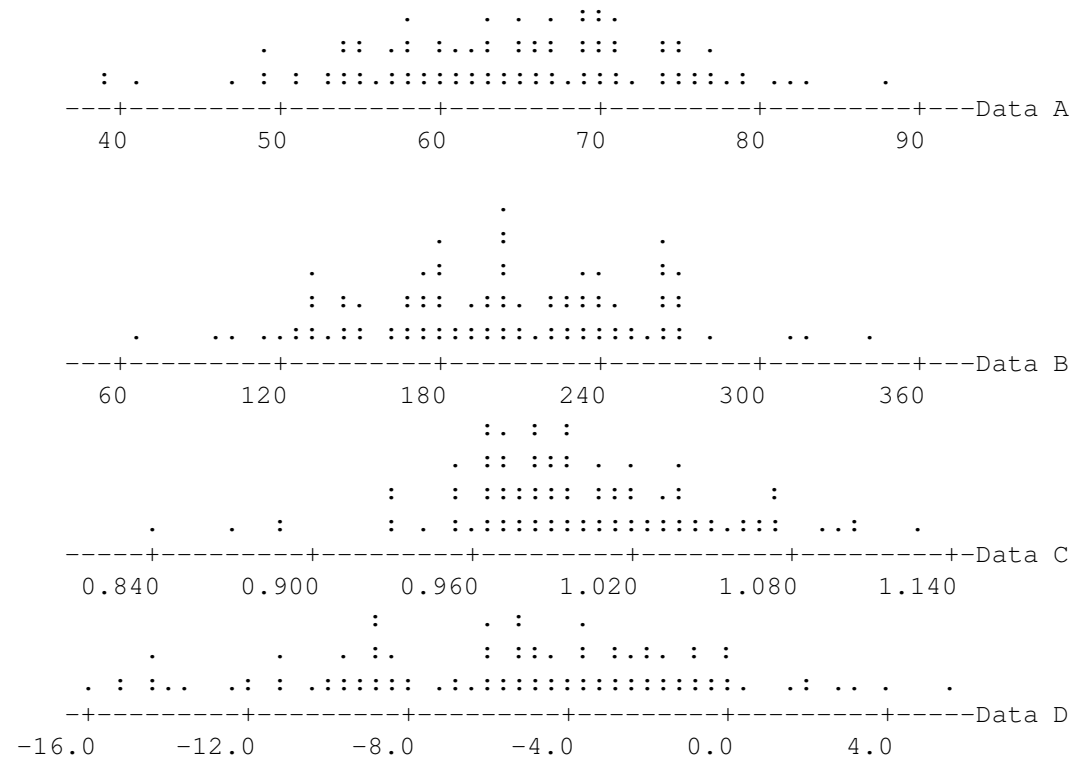
(c) What does the 68-95-99.7 rule say about the proportion of students who score between 10.2 and 31.0 on the ACT?

(d) According to the 68-95-99.7 rule, about 68% of ACT takers score between what two values?

Activity 4-16: Guessing Standard Deviations

Notice that each of the following hypothetical distributions is roughly symmetric and mound-shaped.

(a) Use the 68-95-99.7 rule to make an educated guess about the mean and standard deviation of each distribution. (Hint: Remember that approximately 95% of the data fall within two standard deviations of the mean. In other words, the width of an interval which contains practically all of the data is $(\bar{x} + 2 \times s) - (\bar{x} - 2 \times s) = 4 \times s$.)

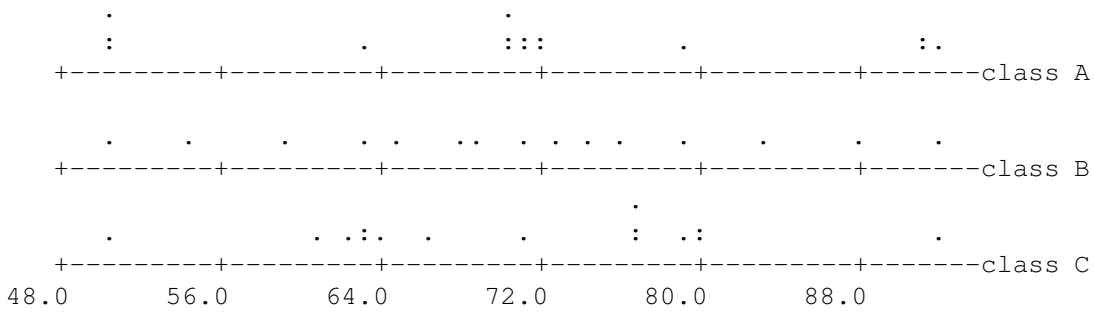


(b) Your instructor will give you the actual means and standard deviations. Comment on the accuracy of your guess.

Activity 4-17: Limitations of Boxplots

Consider the hypothetical exam scores presented below for three classes of students. Dotplots of the distributions are also presented.

class A	50	50	50	63	70	70	70	71	71	72	72	79	91	91	92
class B	50	54	59	63	65	68	69	71	73	74	76	79	83	88	92
class C	50	61	62	63	63	64	66	71	77	77	77	79	80	80	92



(a) Do these dotplots reveal differences among the three distributions of exam scores? Explain briefly.

(b) Find (by hand) five-number summaries of the three distributions. (Notice that the scores are already arranged in order.)

(c) Create boxplots (on the same scale) of these three distributions.

(d) If you had not seen the actual data and had only been shown the boxplots, would you have been able to detect the differences in the three distributions?

Activity 4-18: Creating Examples

Suppose five students have test scores that are all integers between 0 and 100, inclusive.

(a) Write down the five test scores so that the standard deviation is as large as possible.

(b) Write down five scores that have a standard deviation as small as possible.

(c) The interquartile range is zero, but the standard deviation is greater than 0.

(d) The mean score is much smaller than the median score.

(e) All of the scores fall within one standard deviation of the mean.

WRAP-UP

In this topic you have learned to calculate and studied properties of the range, quartile spread, and standard deviation as measures of the variability of a distribution. You have also discovered a new visual display, the boxplot, and studied the five-number summary on which it is based. In addition, you have explored the 68-95-99.7 rule and z-scores as applications of the standard deviation.

To this point you have dealt primarily with one distribution at a time. Often it is more interesting to compare distributions between/among two or more groups. In the next topic, you will discover some new techniques and also apply what you have learned so far to that task.

Topic 5: Comparing Distributions

Introduction

You have been analyzing distributions of data by constructing various graphical displays (dotplot, histogram, stemplot, boxplot), by calculating numerical measures of various aspects of that distribution (mean, median, and mode for center; range, quartile spread, and standard deviation for spread), and by commenting verbally on features of the distribution revealed in those displays and statistics. Thus far you have concentrated on one distribution at a time. With this topic you will apply these techniques in the more interesting case of analyzing, comparing, and contrasting distributions from two or more groups simultaneously.

PRELIMINARIES

1. Which state would you guess was the fastest growing (in terms of population) in the U.S. between 1990 and 1993?

2. Would you expect to find much of a difference in motor vehicle theft rates between eastern and western states? If so, which region would you expect to have higher theft rates?

3. Consider the assertion that men are taller than women. Does this mean that every man is taller than every woman? If not, write one sentence indicating what you think the assertion does mean.

4. What would you guess for the average lifetime of a notable scientist of the past? How about for a notable writer of the past?

5. Take a guess as to the most money won by a female golfer in 1990. Do the same for a male golfer.

6. Which of the seven seasons of the television show Star Trek: The Next Generation would you rate as the best? If you have no opinion, guess which season a panel of reviewers would rate

as the best.

7. Take a guess as to how long a typical penny has been in circulation.
8. Of the four common American coins (penny, nickel, dime, quarter), which do you think tends to have the longest lifetime?
9. Pool the spare change of the class and record the years of the first 25 pennies, nickels, dimes, and quarters that we come across:

pennies	nickels	dimes	quarters

Comparing Groups Using Boxplots

A graph of a five-number summary is called a **boxplot**. To illustrate the construction of a boxplot and to show how it can be used to compare **groups of measurements**, let's consider some data connected with most famous popular group in rock music history.

The Beatles were a rock-and-roll band that achieved stardom in the 1960's. They recorded many albums that remain popular to the current day. Among Beatles fans, it is interesting to examine the musical development of the group. The style of the Beatles' music changed significantly over their career. One can describe this change in musical style in a variety of ways. Here we focus on the lengths of the songs that the Beatles recorded. In particular, we look into the lengths of the Beatles' songs for three of their most popular albums recorded during the years 1966-1968.

We first look at the Beatles album "Rubber Soul". There are 14 songs on this album; the lengths in seconds of all the songs are listed in the following table.

150	126	202	165	139
164	162	171	154	147
148	137	144	139	

Times in seconds of the tracks of the Beatles album "Rubber Soul".

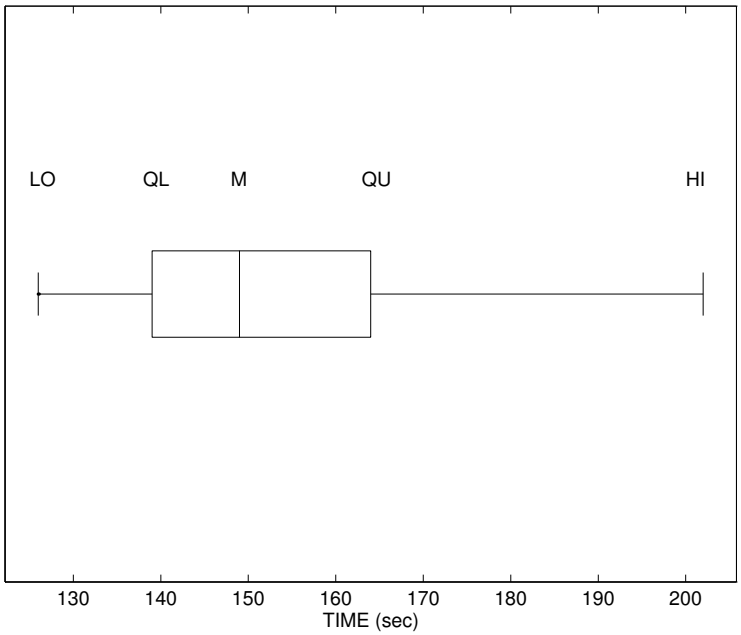
First we graph the times using a stemplot:

```
12| 6
13| 799
14| 478
15| 04
16| 245
17| 1
18|
19|
20| 2
```

Note that all of the lengths of the songs on this particular album fall between 126 and 202 seconds, or roughly between 2 to 3 1/2 minutes. To summarize these numbers, we'll find the median M and the quartiles Q_L and Q_U. There are 14 data items and the median will be the average of the 7th and 8th smallest observations:

$$M = \frac{148 + 150}{2} = 149$$

The lower quartile Q_L is the median of the smallest 7 times. From the stemplot, we see that $Q_L = 139$. Similarly, we see that the upper quartile Q_U, the median of the largest 7 times, is 164. So the



Boxplots of song titles from "Rubber Soul".

five-number summary of the times of the songs on "Rubber Soul" is

$$(126, 139, 149, 164, 202)$$

We display these five numbers using a graph called a **boxplot**. First, we draw a number line with sufficient range to cover all of the possible data values. Then we draw a box above the number line. The placement of the sides of the box corresponds to the locations of the quartiles Q_L and Q_U. The vertical line inside the box is placed at the location of the median M. The boxplot is completed by drawing horizontal lines from the box to the locations of the two extreme values LO and HI. The completed boxplot of the "Rubber Soul" times is shown below.

The boxplot tells us something about the shape of the data. Note that the division line in the box is closer to the left end than the right end. This means that the distance from the lower quartile to the median is less than the distance between the median and the upper quartile. This indicates some right skewness in the middle half of the song times. In addition, the line to the left of the box is shorter than the line to the right of the box. This reflects some right skewness in the outside or **tail** portion of the data. Actually, this long line to the right is due to the single large value of 202 seconds.

A boxplot provides a quick graphical summary of a single group of measurement data. Boxplots are most useful when we are interested in **comparing groups** of measurements. Suppose that we are

interested in comparing the lengths of songs on the Beatles' three albums "Rubber Soul", "Sergeant Pepper", and "The White Album" that were released in the years 1966, 1967 and 1968, respectively. The following table gives the times (in seconds) of the music tracks of "Sergeant Pepper" and "The White Album".

Sergeant Pepper					The White Album				
119	166	205	167	153	165	240	130	190	62
204	156	303	158	163	185	286	167	148	121
155	80	303			140	124	213	232	102
					106	177	160	241	166
					145	195	270	188	253
					162	175	191	495	194

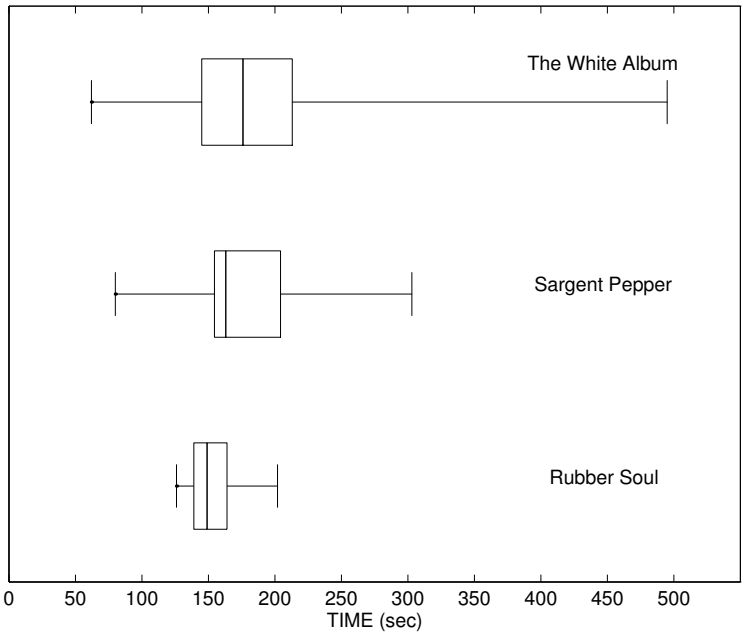
Times in seconds of the tracks of the Beatles albums "Sergeant Pepper" and "The White Album".

Suppose that we wish to compare these three groups of times. For each batch of measurements, we can organize the data by use of a stemplot. Suppose that the three stemplots corresponding to the times of the songs of the three albums are displayed in **parallel form** below. We write down the stem values that are possible in all three albums on three parallel lines. Then the times for the first album ("Rubber Soul") are placed on the first line and the times for the second and third albums are placed on the second and third lines, respectively. In the "White Album" batch, there is one song ("Revolution No. 9") which is unusually long (495 seconds or approximately 8 minutes). If we included this value on the stemplot, the display would take too many lines. So we place this extreme value on a separate 'HI' line. (This convention is helpful when one is constructing a stemplot with unusually large values. Unusually low values can be replaced on a separate 'LO' line.)

```
RUBBER     SERGEANT      THE WHITE
 SOUL       PEPPER         ALBUM

 6|           6|           6| 2
 7|           7|           7|
 8|           8| 0         8|
 9|           9|           9|
10|          10|          10| 26
11|          11| 9        11|
12| 6        12|          12| 14
13| 799      13|          13| 0
14| 478      14|          14| 058
15| 04       15| 3568     15|
16| 245      16| 367      16| 02567
17| 1        17|          17| 57
18|          18|          18| 58
19|          19|          19| 0145
20| 2        20| 45       20|
21|          21|          21| 3
22|          22|          22|
23|          23|          23| 2
24|          24|          24| 01
25|          25|          25| 3
26|          26|          26|
27|          27|          27| 0
28|          28|          28| 6
29|          29|          29|
30|          30| 33       30|
                           HI| 495,
```

This parallel stemplots display is very helpful in understanding the differences in the song lengths between the three albums. The songs on “Rubber Soul” were very homogenous in length – practically all of the songs had lengths between 2 and 3 minutes. About half of the “Sergeant Pepper” tracks have lengths between 150-160 seconds. But two songs are under two minutes, two songs are in the 3-4 minute range and two songs are long (approximately five minutes). The Beatles were experimenting with unusual styles of songs on this album and this may account for the unusual song lengths. The songs on “The White Album” have a large spread. The song times are scattered somewhat uniformly from 1 to 4 minutes and there is one experimental song that is very long.

One can display the differences between these three groups of measurements by the use of **parallel boxplots**. For each album, we compute the five-number summary of the times of the songs. For “Sergeant Pepper”, the five-number summary is (80, 155, 163, 205, 303). This means that the length of an average song on this album was 163 seconds (2 2/3 minutes) and half of the songs lasted between 153 and 205 seconds. Likewise, the five number summary (62, 145, 176, 213, 495) for “The White Album” is computed. Based on these summaries, a boxplot can be constructed



Parallel boxplots of song times from three Beatles albums.

for each of the three groups. In the figure below, the three boxplots are plotted using the same scale. The height of the box is the same for each boxplot and a label is printed to the right of the boxplot indicating the name of the group.

These parallel boxplots display effectively compares the lengths of Beatles' songs on the three albums. The small spread of the "Rubber Soul" tracks and the relative large variability of the "The White Album" tracks is very clear. As the Beatles mature from 1966 to 1968, their songs move away from the 2 1/2 minute pattern to longer songs that exhibit more variability.

IN-CLASS ACTIVITIES

Activity 5-1: Shifting Populations

The following table lists the percentage change in population between 1990 and 1993 for each of the 50 states.

(a) Indicate in the table whether the state lies mostly to the east (E) or to the west (W) of the Mississippi River.

state	% change	region	state	% change	region
Alabama	3.6		Montana	5.1	
Alaska	8.9		Nebraska	1.8	
Arizona	7.4		Nevada	15.6	
Arkansas	3.1		New Hampshire	1.4	
California	3.7		New Jersey	1.9	
Colorado	8.2		New Mexico	6.7	
Connecticut	-0.3		New York	1.1	
Delaware	5.1		North Carolina	4.8	
Florida	5.7		North Dakota	-0.6	
Georgia	6.8		Ohio	2.3	
Hawaii	5.7		Oklahoma	2.7	
Idaho	9.2		Oregon	6.7	
Illinois	2.3		Pennsylvania	1.4	
Indiana	3.0		Rhode Island	-0.3	
Iowa	1.3		South Carolina	4.5	
Kansas	2.1		South Dakota	2.8	
Kentucky	2.8		Tennessee	4.5	
Louisiana	1.9		Texas	6.2	
Maine	0.9		Utah	7.9	
Maryland	3.8		Vermont	2.3	
Massachusetts	-0.1		Virginia	4.9	
Michigan	2.0		Washington	8.0	
Minnesota	3.3		West Virginia	1.5	
Mississippi	2.7		Wisconsin	3.0	
Missouri	2.3		Wyoming	3.7	

In a side-by-side stemplot, a common set of stems is used in the middle of the display with leaves for each category branching out in either direction, one to the left and one to the right. The convention is to order the leaves from the middle out toward either side.

(b) Construct a side-by-side stemplot of these population shift percentages according to the state's region; use the stems listed below.

```
West                          East
                | -0  |
                |  0  |
                |  1  |
                |  2  |
                |  3  |
                |  4  |
                |  5  |
                |  6  |
                |  7  |
                |  8  |
                |  9  |
                | 10  |
                | 11  |
                | 12  |
                | 13  |
                | 14  |
                | 15  |
```

Remember to arrange the leaves in order from the inside out:

```
West                          East
                | -0  |
                |  0  |
                |  1  |
                |  2  |
                |  3  |
                |  4  |
                |  5  |
                |  6  |
                |  7  |
                |  8  |
                |  9  |
                | 10  |
                | 11  |
                | 12  |
                | 13  |
                | 14  |
                | 15  |
```

(c) Calculate the median value of the percentage change in population for each region.

(d) Identify your home state and comment on where it fits into the distribution.

(e) Does one region (east or west) tend to have higher percentage changes than the other? Explain.

(f) Is it the case that every state from one region has a higher percentage change than every state from the other? If not, identify a pair such that the eastern state has a higher percentage change than the western state.

(g) If you were to randomly pick one state from each region, which would you expect to have the higher percentage change? Explain.

Statistical Tendency

You have discovered an important (if somewhat obvious) concept in this activity- that of statistical tendency. You found that western states tend to have higher percentage changes in population than do eastern states. It is certainly not the case, however, that every western state has a higher percentage change than every eastern state.

Similarly, men tend to be taller than women, but there are certainly some women who are taller than most men. Statistical tendencies pertain to average or typical cases but not necessarily to individual cases. Just as Geena Davis and Danny DeVito do not disprove the assertion that men are taller than women, the cases of California and Georgia do not contradict the finding that western states tend to have higher percentage changes in population than eastern states.

Activity 5-2: Professional Golfers' Winnings

The following table presents the winnings (in thousands of dollars) of the 30 highest money winners on each of the three professional golf tours (PGA for males, LPGA for females, and Seniors for

males over 50 years of age) in 1990. (Even if you're not a golf fan, perhaps you're interested either in making money or in studying earning differences between the genders.)

	PGA	winn-	LPGA	winn-	Seniors	winn-
rank	golfer	ings	golfer	ings	golfer	ings
1	Norman	1165	Daniel	863	Trevino	1190
2	Levi	1024	Sheehan	732	Hill	895
3	Stewart	976	King	543	Coody	762
4	Azinger	944	Gerring	487	Archer	749
5	Mudd	911	Bradley	480	Rodriguez	729
6	Irwin	838	Jones	353	Dent	693
7	Calcavecchia	834	Okamoto	302	Charles	584
8	Simpson	809	Lopez	301	Douglass	568
9	Couples	757	Ammacapane	300	Player	507
10	O'Meara	707	Rarick	259	McBee	480
11	Morgan	702	Coe	240	Crampton	464
12	Mayfair	693	Mochrie	231	Henning	409
13	Wadkins	673	Walker	225	Geiberger	373
14	Mize	668	Johnson	187	Hill	354
15	Kite	658	Richard	186	Nicklaus	340
16	Baker-Finch	611	Geddes	181	Beard	327
17	Beck	571	Keggi	180	Mowry	314
18	Elkington	548	Crosby	169	Thompson	308
19	Jacobsen	547	Massey	166	Dill	278
20	Love	537	Figg-Currier	157	Zembriski	276
21	Grady	527	Johnston	156	Barber	274
22	Price	520	Green	155	Moody	273
23	Tway	495	Mucha	149	Bies	265
24	Roberts	478	Eggeling	147	Kelley	263
25	Gallagher	476	Rizzo	145	Jimenez	246
26	Pavin	468	Brown	140	Shaw	235
27	Gamez	461	Mallon	129	Massengale	229
28	Cook	448	Hammel	128	January	216
29	Tennyson	443	Benz	128	Cain	208
30	Huston	435	Turner	122	Powell	208

(a) Notice that one cannot construct side-by-side stemplots to compare these distributions since there are three groups and not just two to be compared. One can, however, construct comparative boxplots of the distributions. Start by calculating (by hand) five-number summaries for each of the three groups, recording them below. Notice that the observations have already been arranged in order and that they are numbered, which should greatly simplify your calculations.

tour	minimum	lower quartile	median	upper quartile	maximum
PGA					
LPGA					
Senior					

(b) Construct (by hand) boxplots of these three distributions on the same scale; I have drawn the axis for you below.

```
+----------+----------+----------+----------+-------WINNINGS
0         300        600        900       1200
```

(c) The boxplot for the Senior golfers reveals one of the weaknesses of boxplots as visual displays. Clearly Lee Trevino, the top money winner among the Seniors, was an outlier, earning almost $300,000 more than his nearest competitor. Would the boxplot look any different, however, if his nearest competitor had won only $5 less than him?

Modified Boxplots

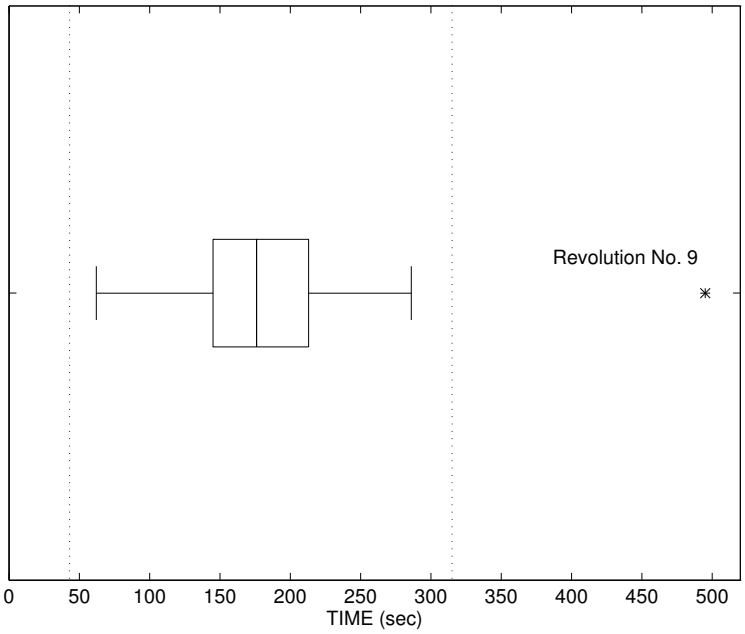
One way to address this problem with boxplots is to construct what are sometimes called modified boxplots. These treat outliers differently by marking them with a special symbol (*) and then only extending the boxplot's "whiskers" to the most extreme non-outlying value. To do this requires an explicit rule for identifying outliers. The rule that we will use regards outliers as observations lying more than 1.5 times the quartile spread away from the nearer quartile.

Let us illustrate the construction of a modified boxplot for the times of songs on the Beatles "White Album". For this group of songs, recall that the lower quartile was $Q_L = 145$, the upper quartile was $Q_U = 213$, and the quartile spread was $QS = 213 - 145 = 68$. To look for possible outliers, we subtract 1.5 times the quartile spread from the lower quartile, and add this same quantity (1.5 times QS) to the upper quartile:

$$145 - 1.5 \times 68, \;\; 213 + 1.5 \times 68$$

or

$$43, \;\; 315.$$



Modified boxplot of song titles from "The White Album".

Our rule says that a Beatles song on the White Album is unusually short or unsually long if it is shorter than 43 seconds or longer than 315 seconds. Looking at the data, we see that one song ("Revolution No. 9") is a outlier on the high end.

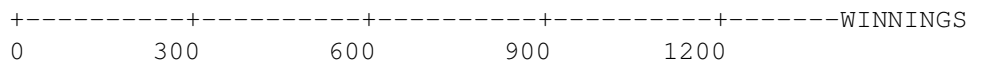
The figure below displays a modified version of the basic boxplot which shows outlying observations. In the figure, the limits for determining outliers (the quartiles plus and minus 1.5 times the quartile spread) are shown using dashed lines. The whiskers of the boxplots are drawn to the data values closest to but not exceeding these outlier limits. Any data value which exceeds the limits are plotted separately using an asterisk (*). This modified boxplot clearly shows that the Beatles' "Revolution No. 9" was much longer than any of the other songs on the album.

Activity 5-3: Professional Golf Earnings (cont.)

With the Seniors' winnings the lower quartile is Q_L = 265 and the upper quartile is Q_U = 568, so the quartile spread is QS = 568 - 265 = 303. Thus, 1.5 x QS = 1.5 (303) = 454.5, so any observation more than 454.5 away from its nearer quartile will be considered an outlier. To look for such observations, we add 454.5 to Q_U, obtaining 568 + 454.5 = 1022.5. Since Trevino's 1190 exceeds this, it is an outlier. On the other end, we subtract 454.5 from Q_L; since the result is negative, clearly no observations fall below it, so there are no outliers on the low end.

(a) Use this rule to check for and identify outliers on the PGA and LPGA lists.

(b) Modified boxplots are constructed by marking outliers with an * and extending whiskers only to the most extreme (i.e., largest on the high end, smallest on the low end) non-outlier. Construct modified boxplots of the three distributions below.



(c) Do the modified boxplots provide more visual information than the "unmodified" ones? Explain.

(d) Write a paragraph comparing and contrasting key features of the distributions of the earnings of the top money winners on these three golf tours.

HOMEWORK ACTIVITIES

Activity 5-4: Students' Measurements (cont.)

Refer back to the data collected in Topic 2 and select one of the three variables: foot length, height, or armspan. Compare men's and women's distributions of this variable by answering the following questions, supplying displays and/or calculations to support your answers:

(a) Does one gender tend to have larger values of this variable than the other gender? If so, by about how much do the genders differ on the average with respect to this variable? Does

every member of one gender have a larger value of this variable than every member of the other gender?

(b) Does one gender have more variability in the distribution of these measurements than the other gender?

(c) Are the shapes of both gender's distributions fairly similar?

(d) Does either gender have outliers in their distribution of this variable?

Activity 5-5: Students' Travels (cont.)

Reconsider the data collected in Topic 1 concerning the number of states visited by college students. Construct (by hand) a side-by-side stemplot which compares the distributions of men and women. Then write a paragraph comparing and contrasting the distributions.

Activity 5-6: Sugar Contents of Ready-to-Eat Cereals

In Topic 3, we looked at the quantity of sugar in a single serving for twenty ready-to-eat cereals. In the cereal aisle of a typical grocery store, the cereals are placed on different shelves, and the cereals placed on some shelves are more visible to children riding in shopping carts than the cereals placed on other shelves. That raises the interesting question: Are there significant differences in the nutritional contents of cereals placed on different shelves?

Let's focus on the sugar contents per serving for 76 cereals. The stemplots below show the sugar contents for the cereals placed on the bottom shelf, the cereals place on the middle shelf, and the cereals on the top shelf. To construct each stemplot, the division point between the stem and the leaf is at the decimal point — the sugar content of 3.0 is recorded as a "0" on the "3" line.

```
BOTTOM SHELF   MIDDLE SHELF    TOP SHELF

 0|000           0|0            0|000
 1|0             1|             1|
 2|00            2|             2|0
 3|00000         3|00           3|000000
 4|              4|             4|0
 5|              5|0            5|0000
 6|00            6|0            6|0000
 7|              7|0            7|000
 8|0             8|             8|0000
 9|              9|000          9|0
10|00           10|            10|000
11|00           11|0           11|00
12|             12|000000      12|0
13|             13|000         13|0
14|             14|0           14|00
15|0            15|0           15|
```

(a) Look at the stemplot of the sugar contents of cereals placed on the bottom shelf. What is a typical sugar content for these cereals? Describe any clusters in this data set.

(b) Now look at the sugar contents of the cereals in the middle shelf. Again give a typical sugar content and describe any clusters that you see in these middle shelf cereals.

(c) Repeat (b) for the sugar contents in the top shelf.

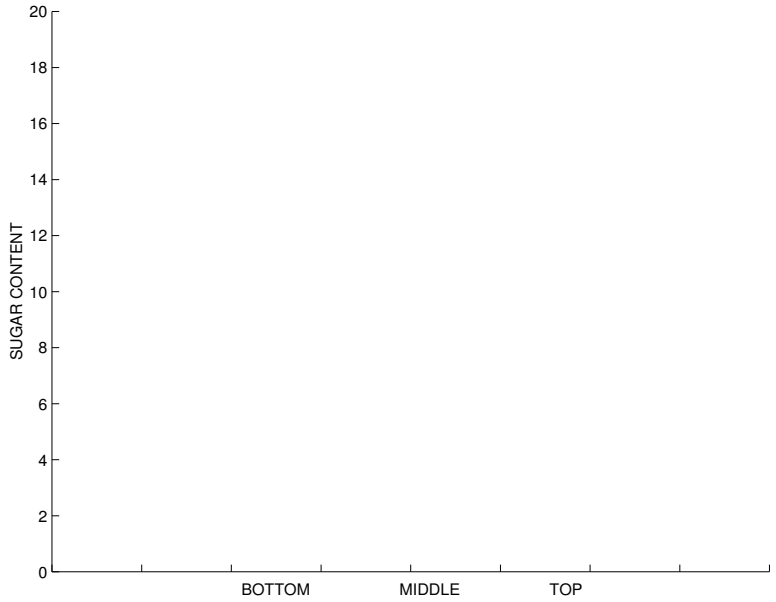
(d) The table below gives the five number summaries of each set of sugar contents. Use this information to draw three parallel boxplots for the sugar contents for cereals on the three shelves. Draw each boxplot in vertical fashion where the extreme lines of the boxplot go from top to bottom.

	LO	Q_L	M	Q_H	HI
Bottom Shelf	0	2	3	10	15
Middle Shelf	0	6.5	12	12.5	15
Top Shelf	0	3	6	9.5	14

(e) Answer the original question: Do the cereals on the different shelves tend to have different sugar amounts? If so, can you provide an explanation for the differences that you see?

Activity 5-7: Automobile Theft Rates

Investigate whether states in the eastern or western part of the U.S. tend to have higher rates of motor vehicle thefts. The following table divides states according to whether they lie east or west of the Mississippi River and lists their 1990 rate of automobile thefts per 100,000 residents.



Eastern state	theft rate	Western state	theft rate
Alabama	348	Alaska	565
Connecticut	731	Arizona	863
Delaware	444	Arkansas	289
Florida	826	California	1016
Georgia	674	Colorado	428
Illinois	643	Hawaii	381
Indiana	439	Idaho	165
Kentucky	199	Iowa	170
Maine	177	Kansas	335
Maryland	709	Louisiana	602
Massachusetts	924	Minnesota	366
Michigan	714	Missouri	539
Mississippi	208	Montana	243
New Hampshire	244	Nebraska	178
New Jersey	940	Nevada	593
New York	1043	New Mexico	337
North Carolina	284	North Dakota	133
Ohio	491	Oklahoma	602
Pennsylvania	506	Oregon	459
Rhode Island	954	South Dakota	110
South Carolina	386	Texas	909
Tennessee	572	Utah	238
Vermont	208	Washington	447
Virginia	327	Wyoming	149
West Virginia	154		
Wisconsin	416		

(a) Create a side-by-side stemplot to compare these distributions. Ignore the last digit of each states rate; use the hundreds digit as the stem and the tens digit as the leaf.

(b) Calculate (by hand) the five-number summary for each distribution of automobile theft rates.

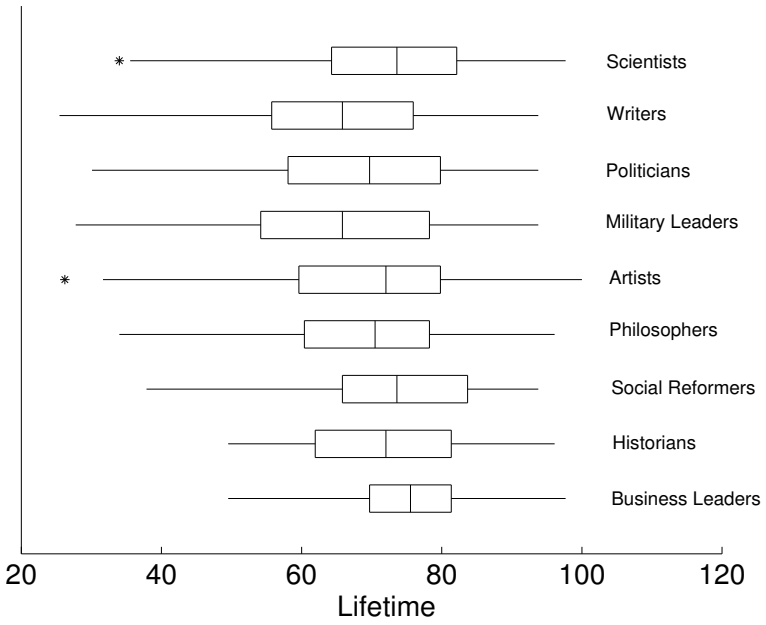
(c) Conduct (by hand) the outlier test for each distribution. If you find any outliers, identify them (by state).

(d) Construct (by hand) modified boxplots to compare the two distributions.

(e) Write a paragraph describing your findings about whether motor vehicle theft rates tend to differ between eastern and western states.

Activity 5-8: Lifetimes of Notables

The 1991 World Almanac and Book of Facts contains a section in which it lists "noted personalities." These are arranged according to a number of categories, such as "noted writers of the past" and "noted scientists of the past." One can calculate (approximately, anyway) the lifetimes of these people by subtracting their year of birth from their year of death. Distributions of the lifetimes of the people listed in nine different categories have been displayed in boxplots below:



Use the information contained in these boxplots to answer the following questions. (In cases where the boxplots are not clear enough to make a definitive determination, you will have to make educated guesses.)

(a) Which group has the individual with the longest lifetime of anyone listed; about how many years did he/she live?

(b) Which group has the largest median lifetime; what (approximately) is that median value?

(c) Which group has the largest range of lifetimes; what (approximately) is the value of that range?

(d) Which group has the largest quartile spread of lifetimes; what (approximately) is the value of that QS?

(e) Which group has the smallest quartile spread of lifetimes; what (approximately) is the value of that QS?

(f) Which group has the smallest median lifetime; what (approximately) is the value of that median?

(g) Describe the general shape of the distributions of lifetimes.

(h) Suggest an explanation for writers tending to live shorter lives than scientists.

Activity 5-9: Hitchcock Films (cont.)

Reconsider the data from Activity 2-9 concerning running times of movies directed by Alfred Hitchcock. Perform (by hand) the outlier test to determine if any of the films constitute outliers in terms of their running times. Comment on your findings in light of your analysis in Activity 2-9.

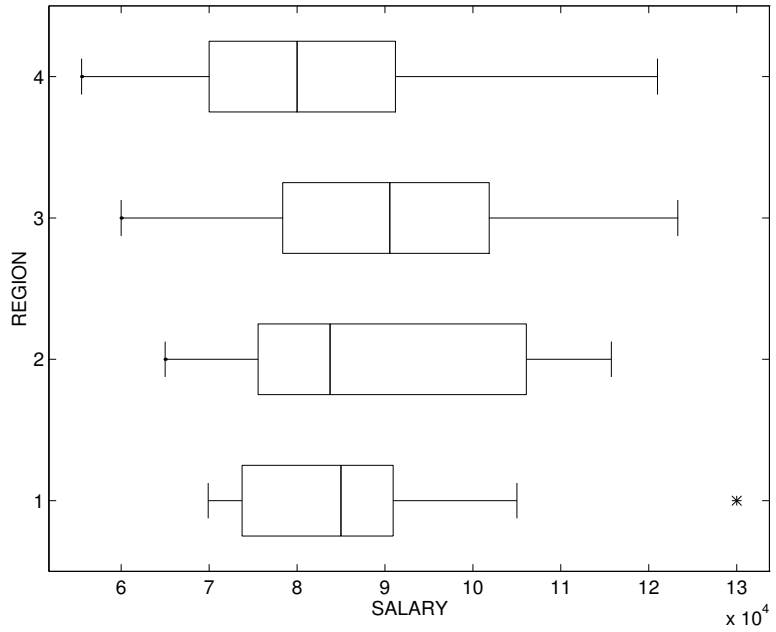
Activity 5-10: Value of Statistics (cont.)

Compare the responses of men and women to the question from Topic 1 about rating the value of statistics in society. Write a paragraph or two describing and explaining your findings; include whatever displays or calculations you care to.

Activity 5-11: Governor Salaries

The boxplots on the next page display the distributions of the 1993 governor's salaries according to the state's geographic region of the country. Region 1 is the Northeast, 2 the Midwest, 3 the South, and 4 the West.

Write a paragraph comparing and contrasting key features of these distributions.



Boxplots of governor's salaries by region

Activity 5-12: Lengths of Number One Songs

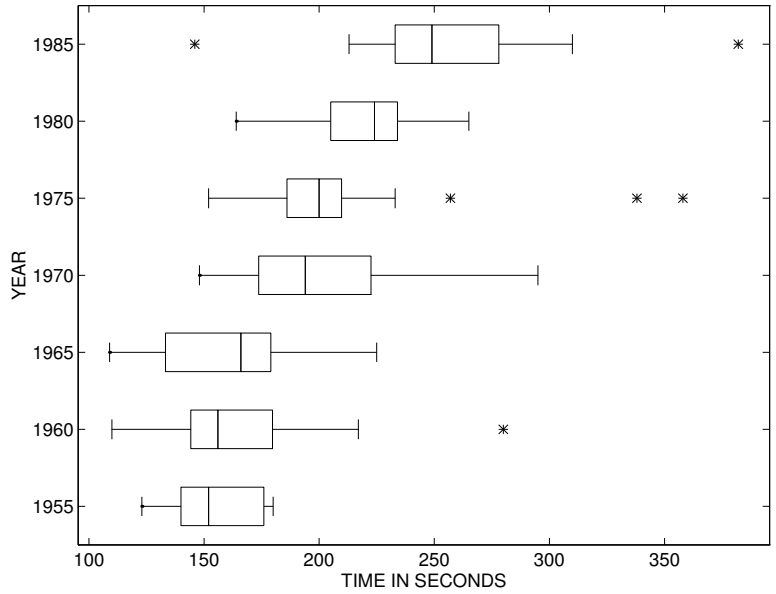
Earlier in this topic, we explored the length of Beatles' songs from different stages of their career. We found that the lengths of the Beatles' songs tended to get longer as the group matured from 1963 to 1988. One may wonder if this pattern is true for other singing groups. Specifically, have lengths of popular songs changed over the years? To help answer this question, the times of all of the songs that reached number 1 on the Billboard Top 100 pop chart were found for the years 1955, 1960, 1965, ..., 1985. The times of all the songs were recorded in seconds and grouped by the year that the songs reached number one. The graph below shows parallel boxplots of the lengths of the number one songs for the seven years.

(a) For the number one songs in 1955, what was a typical length in minutes? What was the shortest song and longest song during this year?

(b) Compare the lengths of the number one songs in 1955 and 1960. Do the average song lengths differ for the two years? Is the spread of the 1955 song lengths different than the spread of the 1960 song lengths? Which year had a greater variation in song lengths?

(c) Which year had some unusually long songs? How long (in minutes) were these songs?

(d) How did the average length of the number one song change from 1955 to 1985? Describe in a few sentences how the song lengths changed in this 30 year period. Is this pattern consistent



Boxplots of lengths (in seconds) of number one songs from different years.

with the pattern in the Beatles' songs that was found earlier?

Activity 5-13: Cars' Fuel Efficiency

Consumer Reports magazine puts out a special issue with information about new cars. The following boxplots display the distributions of cars' miles per gallon ratings. The category designations are those of the magazine: 1=small, 2=median, 3=minivan, 4=coupe, 5=luxury, 6=sports-utility.

(a) Which category of car tends to get the best fuel efficiency? second best? worst?

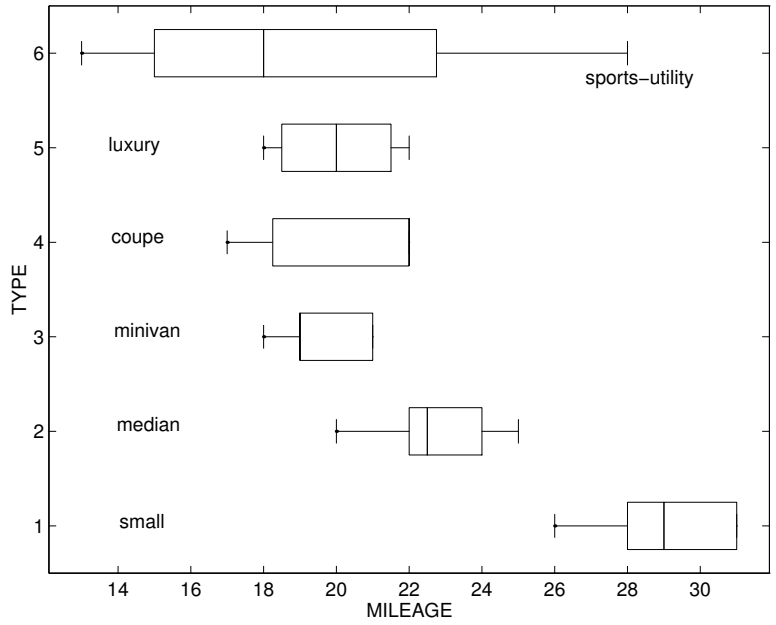
(b) Does every car in the best fuel efficiency group have a higher miles per gallon rating than every car in the second best group?

(c) Does every car in the best fuel efficiency group have a higher miles per gallon rating than every car in the worst fuel efficiency group?

(d) Which group has the most variability in miles per gallon ratings?

Activity 5-14: Sentence Lengths

In *The Cambridge Encylopedia of the English Language*, there is a discussion of the changes in the linguistic style of newspaper articles over the past 90 years. As a case in point, this book reprints the front page of London's *Daily Sketch* from April 11, 1914. This front page of this newspaper contains



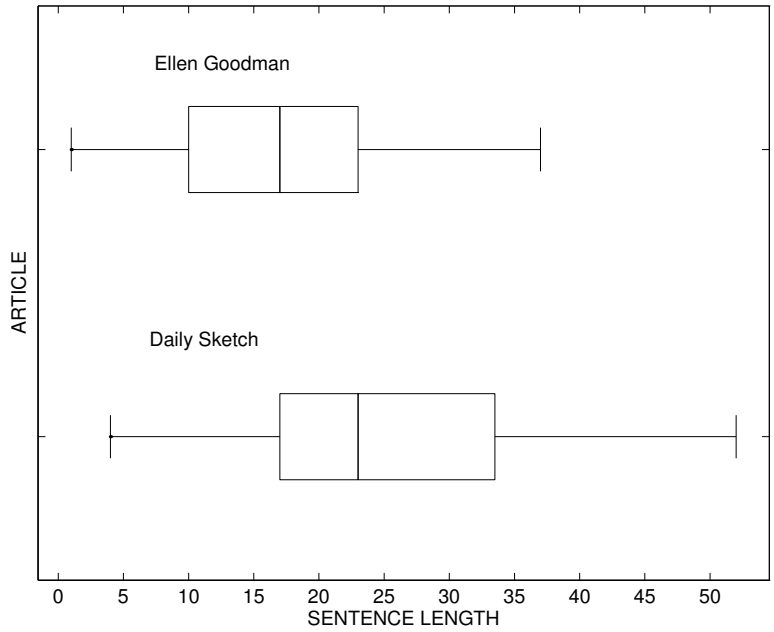
Boxplots of fuel efficiencies of cars of different categories.

a preview of the famous play "Pygmalion" (later to become famous as the musical "My Fair Lady"). We will define the length of a sentence to be the number of words it contains. The table below gives the lengths of 32 sentences taken from this 1914 preview article. To see how sentence lengths have possibly changed over the years, the lengths of 46 sentences were recorded by an opinion piece by Ellen Goodman that appeared in the *Findlay Courier* on April 9, 1998 (almost 84 years later). These sentence lengths are also recorded in the table. The figure below displays parallel boxplots comparing the lengths of sentences from the 1914 article with the lengths in the 1998 article.

Daily Sketch								Ellen Goodman							
4	23	17	18	25	41	29	52	28	22	37	26	23	16	5	16
17	12	14	23	37	29	14	40	37	4	29	17	10	22	22	1
10	22	17	30	21	22	24	45	14	1	26	17	23	29	26	19
44	40	16	18	49	29	7	24	10	22	17	9	19	8	25	17
								10	9	25	5	15	11	19	8
								26	10						

(a) From looking at the graph, find the five-number summaries of the sentence lengths of the 1914 and 1998 articles.

(b) Which article had the larger average sentence length? How much longer were the sentences for the one article, on average?



Lengths of sentences from two newspaper articles.

(c) Did the articles differ with respect to the spread of the sentence lengths? What numbers can you use to compare the spreads of the lengths?

(d) Do you think that it was fair to compare the sentence lengths of the preview of a play with the sentence lengths of a opinion piece? What type of articles in a modern-day newspaper might have short sentences, and what type of articles would have long sentences?

Activity 5-15: Mutual Funds' Returns

Mutual funds are collections of stocks and bonds which have become a very popular investment option in recent years. Some funds require the buyer to pay a fee called a "load" which is similar to a commission charge. A natural question to ask is whether funds that require a load charge perform better than funds which require no load. The table below lists the 1993 percentage return for the mutual funds of the Fidelity Investment group:

No load					Load	
13.9	17.8	21.9	12.4	22.1	24.5	19.5
15.4	13.9	12.5	9	9.1	33.4	20.2
23.3	36.7	13.1	13.1	5.2	21.4	24.7
26.3	81.8	13.9	6.7	13.1	26.4	24.7
19.3	21.3	11.2	13.2	14	26.8	8.3
13.8	18.9	35.1	12.9	19.1	19.9	40.1
13.6	9.8	16.2	12.8	10.2	27.2	63.9
25.5	6.5	20.5	12.6	15.6	16.2	21.1
25.9	18.4	12.2	21.4	22.9		
8.1	5.5	13.8	12.5	36.5		

Analyze these data to address the question of whether funds that charge load fees tend to perform better than those without loads. Write a brief summary of your findings, including whatever visual displays or summary statistics you deem appropriate. Also be sure to comment on what it means to say that one type of fund tends to outperform another. (You may use the computer.)

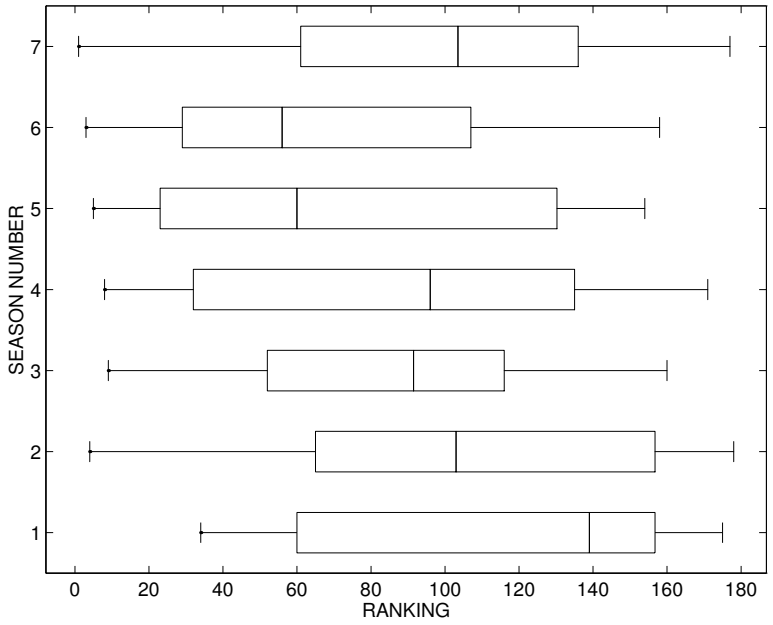
Activity 5-16: Star Trek Episodes

Editors of an Entertainment Weekly publication ranked every episode of Star Trek: The Next Generation from best (rank 1) to worst (rank 178). These appear below, separated according to the season of the shows seven-year run in which the episode aired.

season 1	season 2	season 3	season 4	season 5	season 6	season 7
50	103	113	8	5	23	146
51	90	143	153	16	68	157
101	104	151	25	109	57	78
164	70	91	77	140	127	121
173	105	92	120	76	81	97
148	62	95	108	6	18	61
163	48	160	38	20	3	168
149	161	67	69	42	13	80
175	31	100	112	135	122	128
40	4	87	167	60	64	83
84	172	33	118	39	36	14
139	94	123	82	73	116	58
125	176	12	166	53	117	141
34	75	35	115	147	59	177
162	132	37	144	47	43	124
145	156	88	129	26	49	110
86	52	98	65	7	29	56
155	66	133	137	71	107	134
85	165	72	46	138	32	54
63	99	119	171	111	158	136
150	106	19	21	152	55	74
169	45	89	30	24	11	93
44	178	102	28	79		174
41	126	9	96	15		114
142	159		170	130		1
			17	131		2
			22	27		
				154		
				10		

Five-number summaries for these rankings by season appear below along with comparative boxplots:

	no.	minimum	lower quartile	median	upper quartile	maximum
season 1	25	34	57	139	158.5	175
season 2	25	4	64	103	157.5	178
season 3	24	9	44.5	91.5	117.5	160
season 4	27	8	30	96	137	171
season 5	29	5	22	60	130.5	154
season 6	22	3	27.5	56	109.25	158
season 7	26	1	60.25	103.5	137.25	177



Boxplots of rankings of Star Trek episodes by season.

(a) In which season did the highest ranked episode appear?

(b) In which season did the lowest ranked episode appear?

(c) Do any seasons not have a single episode in the top ten? If so, identify it/them.

(d) Do any seasons not have a single episode in the bottom ten? If so, identify it/them.

(e) Which season seemed to have the best episodes according to these reviewers? Explain your choice.

(f) Which season seemed to have the worst episodes according to these reviewers? Explain your choice.

(g) Comment on whether the five-number summaries and boxplots reveal any tendencies toward episodes getting better or worse as the seasons progressed (again, of course, accepting the ranking judgments of these reviewers).

Activity 5-17: Shifting Populations (cont.)

Refer back to the data in Activity 5-1 concerning states' population shifts.

(a) Identify the eastern states with the biggest and the smallest percentage increases in population. Do the same for the western states. Also record the values of those percentage increases.

(b) Calculate (by hand) five-number summaries of the percentage changes in population for eastern states and for western states.

(c) Perform (by hand) the outlier test for both regions, identifying any outliers that you find.

(d) Construct (by hand) comparative modified boxplots of these distributions.

(e) Comment on what these boxplots reveal about the distributions of percentage changes in population between eastern and western states.

WRAP-UP

You have been introduced in this topic to methods of comparing distributions between/among two or more groups. This task has led you to the very important concept of statistical tendencies. You have also expanded your knowledge of visual displays of distributions by encountering side-by-side stemplots and modified boxplots. Another object of your study has been a formal test for determining whether an unusual observation is an outlier.

In the next unit you will begin to study the important issue of exploring relationships between variables.

Topic 6: Graphical Displays of Association

Introduction

We have described a number of methods for displaying and summarizing a single measurement variable. In many situations, more than one measurement is taken from each person or object. For example, one may measure the height and weight for each male student in a math class, or record the engine size and mileage for ten cars that are interested in purchasing. In some cases, the definition of the "person or object" and the measurements may not be as clear as the two examples above. For example, suppose that we record the high temperature for two specific cities, say Raleigh and Detroit, for twenty days. In this case, a day can be viewed as the unit on which measurements are taken and the temperatures in the cities Raleigh and Detroit correspond to the measurements.

In this section, we focus on bivariate data where two measurements are taken on each subject. We are interested in using graphs and summary statistics to learn about how the two measurements are related. For example, if we measure the height and weight for twenty male students, we might be interested in understanding the relationship between a student's height and weight. Certainly, there is some relationship — taller men are generally heavier. But can we make a more precise statement about the relationship? If we meet a student who is six feet tall, can we accurately guess at his weight? In other cases, the relationship between two variables may not be as well known. In the car example, we may be unsure about how a car's engine size is related to its mileage. The goal of this section is to provide some tools that are helpful for learning about such relationships.

We'll start talking about bivariate measurement data in the context of the following example. My family lives in Findlay, Ohio and we were given the opportunity to take new jobs in Raleigh, North Carolina for a ten month period. Our friends were encouraging us to take these temporary jobs. One advantage of the move was the warmer weather in the south, especially during the winter months. After we moved, we wondered about the relationships between the temperatures in Ohio and North Carolina. If we knew the temperature in Ohio for a particular day, could we accurately

predict the temperature in North Carolina? To answer this question, we needed some data. For a number of days throughout the year, we recorded the high temperatures in Findlay, Ohio and Raleigh, North Carolina. Part of the data is given in the table below. Ten days are listed and, for each day, the maximum temperatures for the two cities are listed.

Day	Ohio Temperature	North Carolina Temperature
January 18	31	50
February 3	20	51
October 23	65	79
December 8	38	54
December 26	36	57
January 19	29	46
February 23	53	70
November 7	57	66
December 10	35	55
December 28	39	65

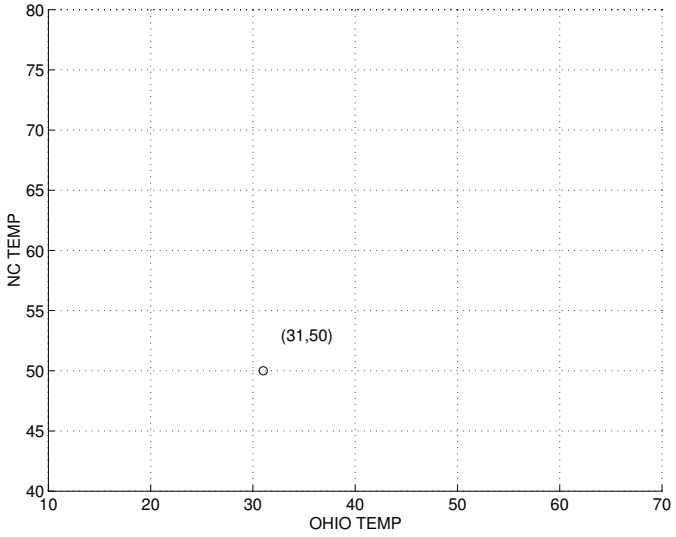
The maximum temperatures in Findlay, Ohio and Raleigh, North Carolina for ten days.

Looking at the Data - The Scatterplot

Our first step in understanding this data is to make a graph called a **scatterplot**. This graph is related to the dotplot that we constructed for a single measurement variable. In a dotplot, we placed a dot on a number line which corresponded to the location of a particular data item. In a scatterplot, we place dots on a rectangular grid. The location of a dot on the rectangular grid gives the location for a single bivariate measurement.

The figure on the next page sets up the construction of the scatterplot for our temperature data. On the horizontal axis, we construct a number line corresponding to the possible Ohio temperatures. We mark the number line off in equal spaces between 10 and 70 degrees (Fahrenheit); the range 10 to 70 covers all of the Ohio temperatures in our dataset. The vertical axis corresponds to the North Carolina temperatures and we construct a number line from 40 to 80 degrees. For the major divisions on the two axes, we draw a series of horizontal and vertical lines. The resulting grid is helpful in placing the dots on the scatterplot.

The two observations (Ohio temperature, NC temperature) for a single day are plotted as a dot on the graph. The first observation is the ordered pair (31, 50) — this corresponds to a temperature of 31 degrees in Ohio and a temperature of 50 degrees in North Carolina. We plot this point by first finding the value 31 along the horizontal axis and then moving up to the value 50 along the vertical axis. We place a dot at this location which is shown on the figure.



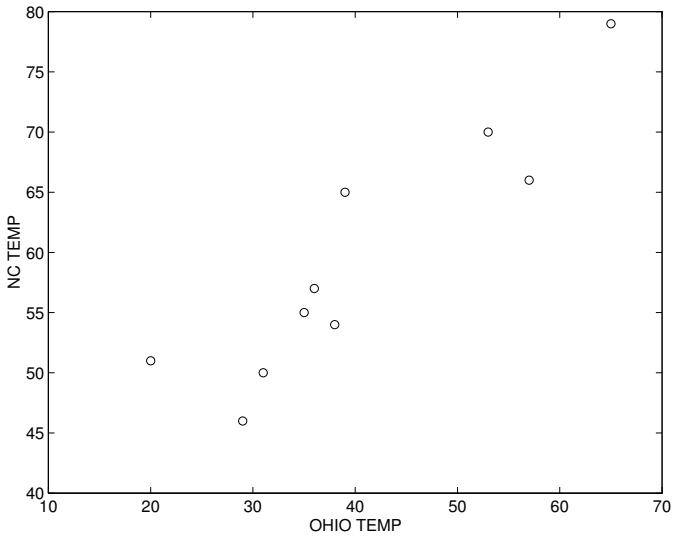
Initial construction of a scatterplot for temperature data.

We repeat this procedure for the remaining nine days in the table. The resulting scatterplot is shown in the figure on the next page. Note that the grid of horizontal and vertical lines is not shown. This grid is useful in the construction process but covers up the points that we have plotted.

A scatterplot is a picture of the relationship between the Ohio and North Carolina temperatures. What do we see in this graph? Generally, the points seem to have a clear pattern; the points appear to go from the lower left to the upper right sections of the display. This means that low Ohio temperatures are generally associated with low North Carolina temperatures and high Ohio temperatures with high North Carolina temperatures. This is an understandable relationship. Both Ohio and North Carolina are affected by the same weather systems that move across the United States. Days that are relatively warm in Ohio will also be relatively warm in North Carolina. Likewise, days that are "cold" in Ohio will be "cold" in North Carolina.

We can describe this scatterplot pattern as **positive**. As the Ohio temperatures increase from small to large values (left to right), the North Carolina temperatures generally increase from small to large values (bottom to top).

A scatterplot gets us started in our understanding of the relationship between the two variables. Later in this section, we'll study this relationship further. What is the strength of the relationship? We'll use a number, the correlation coefficient, to measure the association pattern that we see in the scatterplot. Also, if there is a strong association between the Ohio and North Carolina temperatures, we wish to explain it. We'll describe the association by means of a line that goes through most of the points on the scatterplot. This line will give us a simple formula that relates the two temperatures and will allow us to use a Ohio temperature to predict the North Carolina temperature on the same



Scatterplot of temperature data.

day.

A second example

As a second example, suppose that I am interested in purchasing a car with good mileage. I am also interested in a car with good acceleration, so the size of the engine is important. But the size of a car's engine might have something to do with its mileage. Is it possible to purchase a car with good mileage and a relatively large engine?

To explore the connection between mileage and engine size, I look at a consumer's magazine which profiles the 1996 cars. I focus on the class of economy cars, since that is the type of car that I'm interested in purchasing. For each of 19 cars, the magazine gives the engine displacement in cubic centimeters and its estimated mileage (in miles per gallon) in city driving. The values of these two variables "displacement" and "mileage" are given in the below table. In addition, the "Car" column gives the name of each automobile.

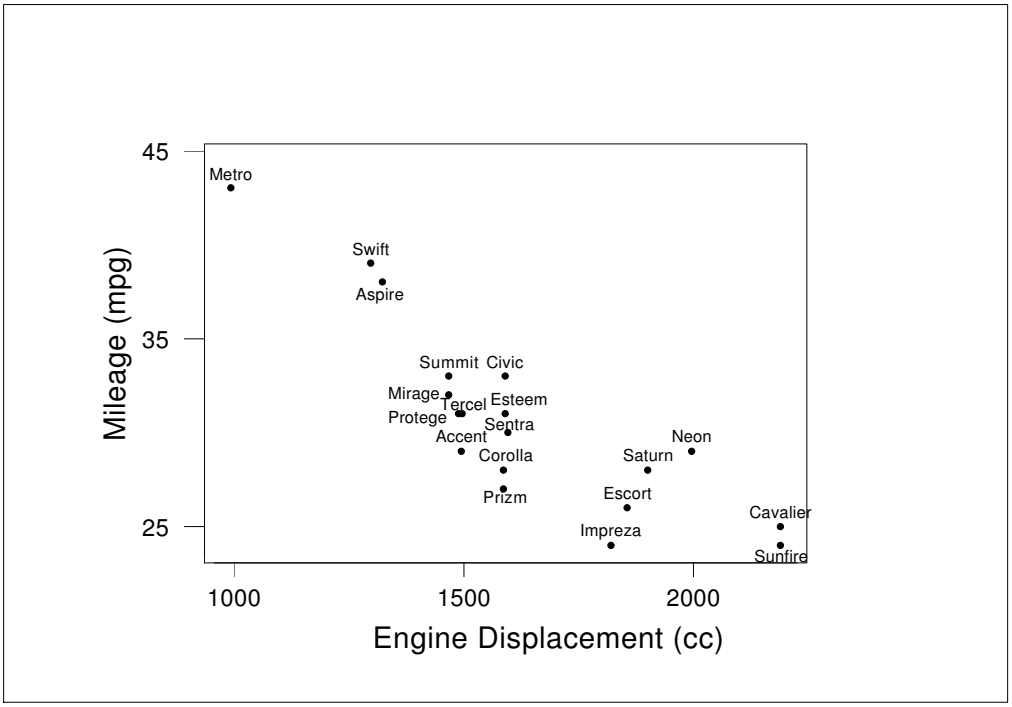
Car	Displacement	Mileage
Cavalier	2189	25
Neon	1996	29
Summit	1468	33
Aspire	1324	38
Escort	1856	26
Metro	993	43
Prizm	1587	27
Civic	1590	33
Accent	1495	29
Protege	1489	31
Mirage	1468	32
Sentra	1597	30
Sunfire	2189	24
Saturn	1901	28
Impreza	1820	24
Esteem	1590	31
Swift	1298	39
Corolla	1587	28
Tercel	1497	31

Engine displacement (cc) and mileage (mpg) for 19 economy cars.

To explore the relationship between engine size and mileage, I draw a scatterplot. I place the displacement variable on the horizontal axis and the mileage on the vertical axis. The first car in the table, Cavalier, has a displacement of 2189 cc and a mileage of 25 miles per gallon. To start, I plot the ordered pair (2189, 25) on the graph. Since I am interested in particular car models, I will label each point with its name.

The figure on the next page shows the completed scatterplot with the points labeled with the car names. There appears to be a relatively strong relationship between engine size and mileage. Specifically, the points appear to drift from the upper left to the lower right regions of the graph. The Metro car has a small engine (approximately 1000 cc) and a high mileage close to 45. In contrast, the two cars, Cavalier and Sunbird, have relatively large engines and low mileages. We describe this scatterplot pattern as **negative**. As the engine sizes increase from small to large (left to right), the car mileages decrease from large values to small values.

As in the temperature example, this scatterplot motivates some questions. How strong is the relationship between engine size and mileage? We'll describe the strength of this relationship by means of a special number called the correlation coefficient. Also, we may be interested in explaining the negative pattern that we see in the scatterplot. Soon we will talk about the use of a straight line to describe the association pattern we see. This straight line can be used to predict the mileage of an economy car with a specified engine size.



Scatterplot of car displacement/mileage data.

PRELIMINARIES

1. Do you expect that there is a tendency for heavier cars to get worse fuel efficiency (as measured in miles per gallon) than lighter cars?

2. Do you think that if one car is heavier than another that it must always be the case that its gets a worse fuel efficiency?

3. Take a guess as to a typical temperature for a space shuttle launch in Florida.

4. Do you think that people from large families tend to marry people from large families, people from small families, or do you think that there is no relationship between the family sizes of married couples?

5. Do you think that people from large families tend to have large families of their own, small families of their own, or do you think that there's no relationship there?

6. Record in the table below the number of siblings (brothers and sisters) of each student in the class, the number of siblings of each student's father, and the number of siblings of each student's mother. So that everyone counts the same way, decide as a class whether or not to count step-siblings and half-siblings.

student	own	father	mother	student	own	father	mother
1				16			
2				17			
3				18			
4				19			
5				20			
6				21			
7				22			
8				23			
9				24			
10				25			
11				26			
12				27			
13				28			
14				29			
15				30			

IN-CLASS ACTIVITIES

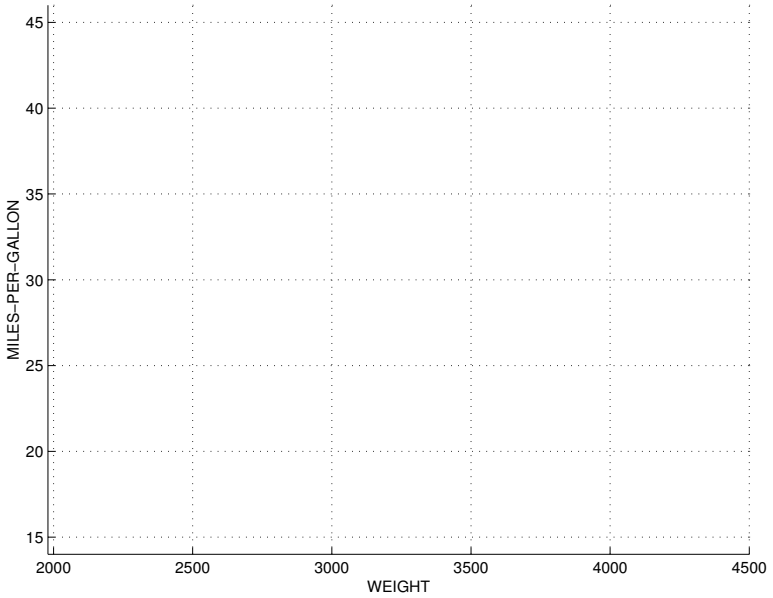
Activity 6-1: Cars' Fuel Efficiency (cont.)

Refer back to Activity 5-13, which dealt with a Consumer Reports issue examining 1995 cars. For a small group of 16 car models, the following table lists the weight of the car (in pounds) and the fuel efficiency (in miles of gallon) achieved in a 150-mile test drive.

model	weight	mpg	model	weight	mpg
BMW 3-Series	3250	28	Ford Probe	2900	28
BMW 5-Series	3675	23	Ford Taurus	3345	25
Cadillac Eldorado	3840	19	Ford Taurus SHO	3545	24
Cadillac Seville	3935	20	Honda Accord	3050	31
Ford Aspire	2140	43	Honda Civic	2540	34
Ford Crown Victoria	4010	22	Honda Civic del Sol	2410	36
Ford Escort	2565	34	Honda Prelude	2865	30
Ford Mustang	3450	22	Lincoln Mark VIII	3810	22

(a) Use the axes on the next page to construct a scatterplot of miles per gallon vs. weight.

(b) Does the scatterplot reveal any relationship between a car's weight and its fuel efficiency? In other words, does knowing the weight reveal any information at all about the fuel efficiency? Write a few sentences about the relationship between the two variables.



Two variables are said to be **positively associated** if larger values of one variable tend to occur with larger values of the other variable; they are said to be **negatively associated** if larger values of one variable tend to occur with smaller values of the other. The strength of the association depends on how closely the observations follow that relationship. In other words, the strength of the association reflects how accurately one could predict the value of one variable based on the value of the other variable.

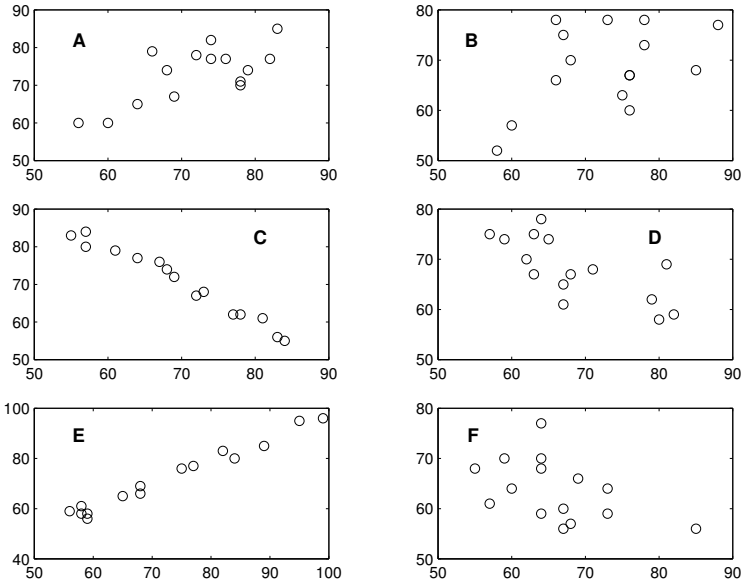
(c) Is fuel efficiency positively or negatively associated with weight?

(d) Can you find an example where one car weighs more than another and still manages to have a better fuel efficiency than that other car? If so, identify such a pair and circle them on the scatterplot above.

Clearly the concept of association is an example of a statistical tendency. It is not always the case that a heavier car is less fuel efficient, but heavier cars certainly do tend to be.

Activity 6-2: Guess the Association

The figure above contains a collection of 6 scatterplots, labeled A through F. Look at scatterplot A of hypothetical scores on the first and second exams of a course. The first exam scores are plotted along the horizontal axis and the second exam scores along the vertical axis.



Scatterplots of some hypothetical exam scores.

(a) Describe briefly (in words related to the context) what the scatterplot reveals about the relationship between scores on the first exam and scores on the second exam. In other words, does knowing a student's score on the first exam give you any information about his/her score on the second exam? Explain.

(b) In the figure there are five more scatterplots of hypothetical exam scores (labelled B, C, D, E, F). Your task is to evaluate the direction and the strength of the association between scores on

the first exam and scores on the second exam for each of the hypothetical classes A through F. Do so by filling in the table below with the letter (A through F) of the class whose exam scores have the indicated direction and strength of association. (You are to use each class letter once and only once, so you might want to look through all six before assigning letters.)

	most strong	moderate	least strong
negative			
positive			

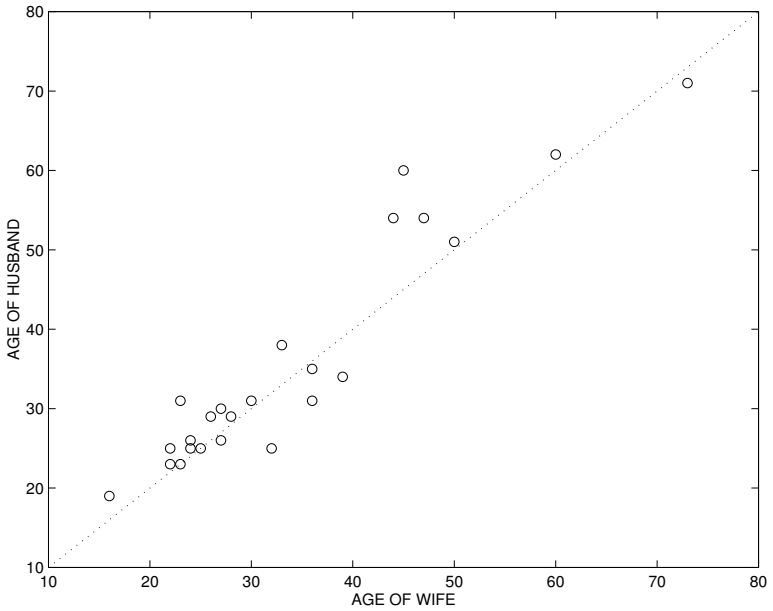
(c) Indicate what you would expect for the direction (positive, negative, or none at all) and strength (none, weak, moderate, or strong) of the association between the pairs of variables listed below.

pair of variables	direction of association	strength of association
height and armspan		
height and shoe size		
height and G.P.A.		
SAT score and college G.P.A.		
latitude and avg. January temp. of American cities		
lifetime and weekly cigarette consumption		
serving size and calories of fast food sandwiches		
air fare and distance to destination		
cost and quality rating of peanut butter brands		
governor's salary and avg. pay in the state		
course enrollment and average student evaluation		

Activity 6-3: Marriage Ages (cont.)

Refer to the data from Activity 2-8 concerning the ages of 24 couples applying for marriage licenses. The following scatterplot displays the relationship between husband's age and wife's age. The line drawn on the scatterplot is a 45° line where the husband's age would equal the wife's age.

(a) Does there seem to be an association between husband's age and wife's age? If so, is it positive or negative? Would you characterize it as strong, moderate, or weak? Explain.



Scatterplot of husbands' ages against wifes' ages.

(b) Look back at the original listing of the data to determine how many of the 24 couples' ages fall exactly on the line. In other words, how many couples listed the same age for both the man and the woman on their marriage license?

(c) Again looking back at the data, for how many couples is the husband younger than the wife? Do these couples fall above or below the line drawn in the scatterplot?

(d) For how many couples is the husband older than the wife? Do these couples fall above or below the line drawn in the scatterplot?

(e) Summarize what one can learn about the ages of marrying couples by noting that the majority of couples produce points which fall above the 45° line.

This activity illustrates that when working with paired data, including a 45° line on a scatterplot can provide valuable information about whether one member tends to have a larger value of the variable than the other member of the pair.

A categorical variable can be incorporated into a scatterplot by constructing a labeled scatterplot, which assigns different labels to the dots in a scatterplot. For example, one might use an 'x' to denote a man's exam score and an 'o' to denote a woman's.

Activity 6-4: Fast Food Sandwiches

The following table lists some nutritional information about sandwiches offered by the fast food chain Arby's. Serving sizes are in ounces.

sandwich	meat	serving	calories
Regular Roast Beef	roast beef	5.5	383
Beef and Cheddar	roast beef	6.9	508
Junior Roast Beef	roast beef	3.1	233
Super Roast Beef	roast beef	9	552
Giant Roast Beef	roast beef	8.5	544
Chicken Breast Fillet	chicken	7.2	445
Grilled Chicken Deluxe	chicken	8.1	430
French Dip	roast beef	6.9	467
Italian Sub	roast beef	10.1	660
Roast Beef Sub	roast beef	10.8	672
Turkey Sub	turkey	9.7	533
Light Roast Beef Deluxe	roast beef	6.4	294
Light Roast Turkey Deluxe	turkey	6.8	260
Light Roast Chicken Deluxe	chicken	6.8	276

(a) Create (by hand) a labeled scatterplot of calories vs. serving ounces, using different labels for the three types of meat (roast beef, chicken, turkey). For example, you might use 'r' for roast beef, 'c' for chicken, and 't' for turkey.

(b) Disregarding for the moment the distinction between the types of meat, does the scatterplot reveal an association between a sandwich's serving size and its calories? Explain.

(c) Now comment on tendencies you might observe with regard to the type of meat. Does one type of meat tend to have more or fewer calories than other types of similar serving size? Elaborate.

Activity 6-5: Space Shuttle O-Ring Failures

The following data were obtained from 23 shuttle launches prior to Challenger's fatal launch. Each of four joints in the shuttle's solid rocket motor is sealed by two O-ring joints. After each launch, the reusable rocket motors were recovered from the ocean. This table lists the number of O-ring seals showing evidence of thermal distress and the launch temperature for each of the 23 flights.

flight date	O-ring failures	temperature	flight date	O-ring failures	temperature
4/12/81	0	66	11/8/84	0	67
11/12/81	1	70	1/24/85	3	53
3/22/82	0	69	4/12/85	0	67
11/11/82	0	68	4/29/85	0	75
4/4/83	0	67	6/17/85	0	70
6/18/83	0	72	7/29/85	0	81
8/30/83	0	73	8/27/85	0	76
11/28/83	0	70	10/3/85	0	79
2/3/84	1	57	10/30/85	2	75
4/6/84	1	63	11/26/85	0	76
8/30/84	1	70	1/12/86	1	58
10/5/84	0	78			

(a) Use the computer to construct a scatterplot of O-ring failures vs. temperature. Write a few sentences commenting on whether the scatterplot reveals any association between O-ring failures and temperature.

(b) The forecasted low temperature for the morning of the fateful launch was 31° F. What does the scatterplot reveal about the likeliness of O-ring failure at such a temperature?

(c) Eliminate for the moment those flights which had no O-ring failures. Considering just the remaining seven flights, ask the computer to construct a new scatterplot of O-ring failures vs. temperature on the axes provided below. Does this scatterplot reveal association between O-ring failures and temperature? If so, does it make the case as strongly as the previous scatterplot did?

(d) NASA officials argued that flights on which no failures occurred provided no information about the issue. They therefore examined only the second scatterplot and concluded that there was little evidence of an association between temperature and O-ring failures. Comment on the wisdom of this approach and specifically on the claim that flights on which no failures occurred provided no information.

HOMEWORK ACTIVITIES

Activity 6-6: Students' Family Sizes

Enter the data that you collected above concerning family sizes into the computer.

(a) First examine just the distribution of students' siblings by itself. Comment briefly on the distribution, and also report the mean, standard deviation, and five-number summary.

(b) Now examine the relationships between students' siblings and fathers' siblings, between students' siblings and mothers' siblings, and between fathers' siblings and mothers' siblings. Have the computer produce the relevant scatterplots. Write a paragraph summarizing your findings concerning whether an association exists between any pairs of these variables.

Activity 6-7: Air Fares

The table below lists distances and cheapest airline fares to certain destinations for passengers flying out of Baltimore, Maryland (as of January 8, 1995):

destination	distance	air fare	destination	distance	air fare
Atlanta	576	178	Miami	946	198
Boston	370	138	New Orleans	998	188
Chicago	612	94	New York	189	98
Dallas/Fort Worth	1216	278	Orlando	787	179
Detroit	409	158	Pittsburgh	210	138
Denver	1502	258	St. Louis	737	98

(a) Create (by hand) a scatterplot of air fare vs. distance.

(b) Is air fare associated with distance? If so, is the association positive or negative? Would you characterize it as strong, moderate, or weak?

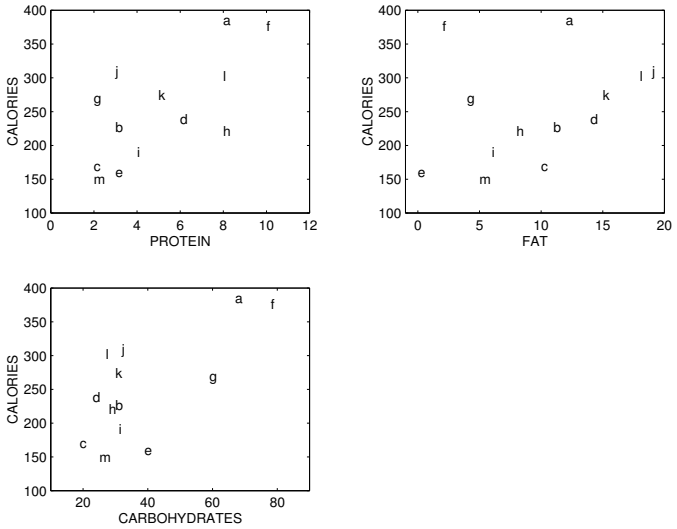
(c) Find at least two examples of pairs of cities that go against the association. In other words, if the association is positive, find two pairs of cities where the more distant city is cheaper to fly to. Circle these pairs of cities on your scatterplot.

(d) Explain in your own words how (b) and (c) address the issue of association as a statistical tendency.

Activity 6-8: Nutritional Content of Desserts

The table below gives the nutritional content of a number of popular desserts as given by the *www.homearts.com* website. For a single serving of 13 desserts, the table lists the number of calories, the protein content. the amount of fat, and the carbohydrate content (all measured in grams).

The scatterplots on the next page plot the number of calories (vertical axis) against the protein content, the fat amount, and the carbohydrate content. The plotting point in each graph is the dessert label that is shown in the table. For example, the point labeled “a” corresponds to the chocolate pudding.



dessert	dessert label	calories	protein	fat	carbohydrates
Chocolate pudding	a	385	8	12	67
Unglazed cake doughnut	b	227	3	11	30
Glazed cake doughnut	c	170	2	10	19
Eclair	d	239	6	14	23
Fruit gelatin	e	161	3	0	39
Ice cream cone	f	377	10	2	78
Orange sherbet	g	270	2	4	59
Tapioca pudding	h	221	8	8	28
Ice cream sandwich	i	191	4	6	31
Cherry turnover	j	310	3	19	32
Danish pastry	k	274	5	15	30
Cream puff	l	303	8	18	27
Raspberry granola bar	m	150	2	5	25

(a) First look at the scatterplot which plots calories against protein.

(1) Which desserts have high number of calories?

(2) List some desserts that are high in protein.

(3) Describe the relationship between calories and protein.

(b) Now look at the scatterplot of calories against fat.

(1) Name one dessert that is low in fat and low in calories.

(2) Name one dessert that is low in fat and high in calories.

(3) Again describe the relationship between calories and fat. Is the relationship as strong as the one between calories and protein?

(c) The scatterplot of calories against carbohydrates appears to form two clusters of points. Name the three desserts that are relatively high with respect to these two variables. Discuss how these three desserts are similar.

Activity 6-9: Students' Measurements (cont.)

Reconsider the data collected in Topic 2 concerning students' heights, and armspans.

(a) Use the computer to create a scatterplot of height vs. armspan. Draw a 45° line on the scatterplot.

(b) Comment on the association between height and armspan as revealed in the scatterplot.

(c) Can you find examples of pairs of students where the shorter student has the longer armspan?

(d) What can you say about students who fall above the 45°

(e) What is true about the ratio of height to armspan for those students who fall below the 45° line?

Activity 6-10: Students' Measurements (cont.)

Reconsider the data collected in Topic 2 concerning students' heights and foot lengths.

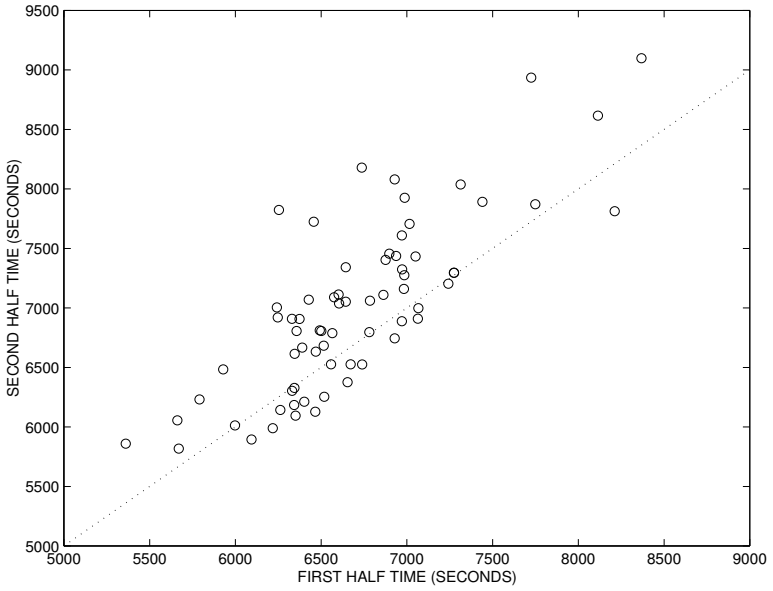
(a) Use the computer to create a labeled scatterplot of height vs. foot length, using different labels for each gender.

(b) Disregarding gender for the moment, write a sentence or two to describe the association between height and foot length.

(c) Comment on any gender differences that the scatterplot reveals.

Activity 6-11: Boston Marathon Times

A marathon is a cross-country footrace of 26 miles, 385 yards. The best-known marathon that is run in the United States is the Boston Marathon that is held each April. Thousands of men and women run this particular marathon. We will focus on the completion times for the group of women runners who reside in Ohio. For each runner, the time (in seconds) to complete the first half of the race and



the time (also in seconds) to complete the second half of the race are recorded. The scatterplot above plots the first half completion time on the horizontal axis against the second half completion time on the vertical axis. The line on the scatterplot is a 45° line where the second half time is equal to the first half time.

(a) Circle one point on the graph (and label it "F") which corresponds to a runner who completed both the first half and the second half of the race very fast. What were her first and second half completion times in hours? (One hour is equivalent to 3600 seconds.)

(b) Circle a point on the graph (and label it "S") which corresponds to a slow runner. What were her first half and second half times?

(c) Circle a point (and label it "B") which corresponds to a runner who ran the second half of the race faster than the first half.

(d) Are the majority of the points above or below the 45° line? What does this tell us about the relationship between the first half time and the second half time for most runners? Can you explain this relationship?

Activity 6-12: Peanut Butter

The September 1990 issue of Consumer Reports rated thirty-seven varieties of peanut butter. Each variety was given an overall sensory quality rating, based on taste tests by a trained sensory panel.

Also listed was the cost (per three tablespoons, based on the average price paid by CU shoppers) and sodium content (per three tablespoons, in milligrams) of each product. Finally, each variety was classified as creamy (cr) or chunky (ch), natural (n) or regular (r), and salted (s) or unsalted (u). The results are below:

brand	cost	sodium	quality	cr/ch	r/n	s/u
Jif	22	220	76	cr	r	s
Smucker's Natural	27	15	71	cr	n	u
Deaf Smith Arrowhead Mills	32	0	69	cr	n	u
Adams 100% Natural	26	0	60	cr	n	u
Adams	26	168	60	cr	n	s
Skippy	19	225	60	cr	r	s
Laura Scudder's All Natural	26	165	57	cr	n	s
Kroger	14	240	54	cr	r	s
Country Pure Brand	21	225	52	cr	n	s
NuMade	20	187	43	cr	r	s
Peter Pan	21	225	40	cr	r	s
Peter Pan	22	3	35	cr	r	u
A&P	12	225	34	cr	r	s
Hollywood Natural	32	15	34	cr	n	u
Food Club	17	225	33	cr	r	s
Pathmark	9	255	31	cr	r	s
Lady Lee	16	225	23	cr	r	s
Albertsons	17	225	23	cr	r	s
Shur Fine	16	225	11	cr	r	s
Smucker's Natural	27	15	89	ch	n	u
Jif	23	162	83	ch	r	s
Skippy	21	211	83	ch	r	s
Adams 100% Natural	26	0	69	ch	n	u
Deaf Smith Arrowhead Mills	32	0	69	ch	n	u
Country Pure Brand	21	195	67	ch	n	s
Laura Scudder's All Natural	24	165	63	ch	n	s
Smucker's Natural	26	188	57	ch	n	s
Food Club	17	195	54	ch	r	s
Kroger	14	255	49	ch	r	s
A&P	11	225	46	ch	r	s
Peter Pan	22	180	45	ch	r	s
NuMade	21	208	40	ch	r	s
Health Valley 100% Natural	34	3	40	ch	n	u
Lady Lee	16	225	34	ch	r	s
Albertsons	17	225	31	ch	r	s
Pathmark	9	210	29	ch	r	s
Shur Fine	16	195	26	ch	r	s

(a) Select any pair of measurement variables that you would like to examine. Use the computer to produce a scatterplot of these variables, and write a paragraph of a few sentences commenting on the relationship between them.

(b) Now select one of the three categorical variables, and use the computer to produce a labeled scatterplot of your two variables from (a) using this new variable for the labels. Comment on any effect of this categorical variable as well as on any other features of interest in the plot.

Activity 6-13: States' SAT Averages

The August 31, 1992 issue of The Harrisburg Evening-News lists the average SAT score for each of the fifty states and the percentage of high school seniors in the state who take the SAT test; these are reproduced below.

state	avg SAT	% taking	state	avg SAT	% taking
Alabama	996	8	Montana	988	24
Alaska	908	42	Nebraska	1018	11
Arizona	933	27	Nevada	922	27
Arkansas	990	6	New Hampshire	923	76
California	900	46	New Jersey	891	75
Colorado	960	29	New Mexico	996	12
Connecticut	900	79	New York	882	75
Delaware	895	66	North Carolina	855	57
Florida	884	50	North Dakota	1068	6
Georgia	842	65	Ohio	951	23
Hawaii	878	56	Oklahoma	1007	9
Idaho	963	17	Oregon	925	55
Illinois	1010	15	Pennsylvania	877	68
Indiana	868	58	Rhode Island	881	70
Iowa	1096	5	South Carolina	831	59
Kansas	1033	10	South Dakota	1040	6
Kentucky	988	11	Tennessee	1013	13
Louisiana	991	9	Texas	876	44
Maine	882	66	Utah	1041	5
Maryland	907	66	Vermont	897	69
Massachusetts	902	80	Virginia	893	63
Michigan	987	11	Washington	916	50
Minnesota	1053	10	West Virginia	924	17
Mississippi	1004	4	Wisconsin	1029	11
Missouri	1004	11	Wyoming	978	13

(a) Use the computer to look at a scatterplot of these data. Write a paragraph describing the

relationship between average SAT score and percentage of students taking the test. Include a reasonable explanation for the type of association that is apparent.

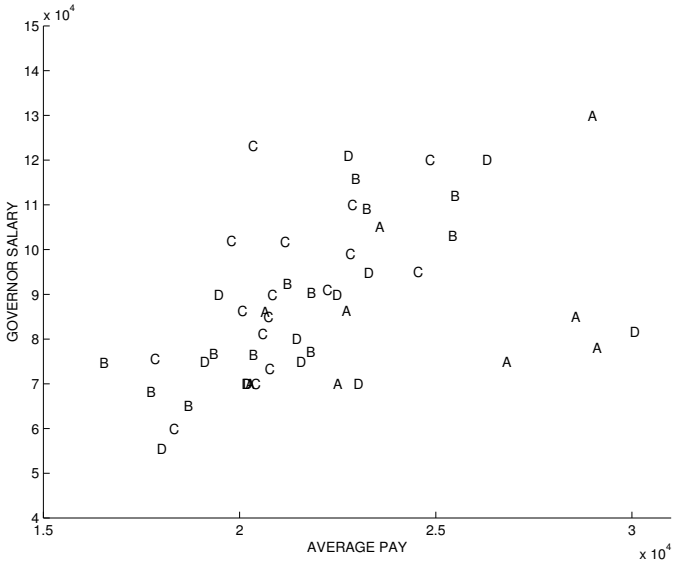
(b) Which state has the highest SAT average? Would you conclude that this state does the best job of educating its students? Which state has the lowest SAT average? Would you conclude that this state does the worst job of educating its students? Explain.

(c) How does your home state compare to the rest in terms of SAT average and percentage of students taking the test? (Identify the state also.)

Activity 6-14: Governor Salaries (cont.)

Recall from Activity 5-11 the data concerning governor salaries. The following table lists each state's average pay and governor's salary (as of 1993) and also indicates the state's region of the country (Northeast, Midwest, South, or West).

state	region	avg pay	gov sal	state	region	avg pay	gov sal
Alabama	S	20,468	81,151	Montana	W	17,895	55,502
Alaska	W	29,946	81,648	Nebraska	MW	18,577	65,000
Arizona	W	21,443	75,000	Nevada	W	22,358	90,000
Arkansas	S	18,204	60,000	New Hampshire	NE	22,609	86,235
California	W	26,180	120,000	New Jersey	NE	28,449	85,000
Colorado	W	22,908	70,000	New Mexico	W	19,347	90,000
Connecticut	NE	28,995	78,000	New York	NE	28,873	130,000
Delaware	S	24,423	95,000	North Carolina	S	20,220	123,300
Florida	S	21,032	101,764	North Dakota	MW	17,626	68,280
Georgia	S	22,114	91,092	Ohio	MW	22,843	115,752
Hawaii	W	23,167	94,780	Oklahoma	S	20,288	70,000
Idaho	W	18,991	75,000	Oregon	W	21,332	80,000
Illinois	MW	25,312	103,097	Pennsylvania	NE	23,457	105,000
Indiana	MW	21,699	77,200	Rhode Island	NE	22,388	69,900
Iowa	MW	19,224	76,700	South Carolina	S	19,669	101,959
Kansas	MW	20,238	76,476	South Dakota	MW	16,430	74,649
Kentucky	S	19,947	86,352	Tennessee	S	20,611	85,000
Louisiana	S	20,646	73,440	Texas	S	22,700	99,122
Maine	NE	20,154	70,000	Utah	W	20,074	70,000
Maryland	S	24,730	120,000	Vermont	NE	20,532	85,977
Massachusetts	NE	26,689	75,000	Virginia	S	22,750	110,000
Michigan	MW	25,376	112,025	Washington	W	22,646	121,000
Minnesota	MW	23,126	109,053	West Virginia	S	20,715	90,000
Mississippi	S	17,718	75,600	Wisconsin	MW	21,101	92,283
Missouri	MW	21,716	90,312	Wyoming	W	20,049	70,000



Labeled scatterplot of governor salary against average wage.

This labeled scatterplot on the next page displays governor salary vs. average wage, using different letters for the regions (A=Northeast, B=Midwest, C=South, D=West).

(a) Comment on the relationship between these two variables.

(b) List three pairs of states for which the state with the higher average pay has the lower governor salary.

(c) Name a state which appears to have a governor salary much higher than would be expected for a state with its average pay.

(d) Identify a cluster of four states which seem to have much lower governor salaries than would be expected for states with their average pay.

Activity 6-15: Teaching Evaluations

Investigate whether there seems to be an association between the number of students in a class and the students' average rating of the instructor on the course evaluation. The following table lists these variables for 25 courses taught by an instructor over a six-year period. The students' ratings of the instructor are on a scale of 1 to 9.

course	students	avg rating	course	students	avg rating	course	students	avg rating
1	11	6.7	10	24	5.3	18	20	6.9
2	12	5.9	11	20	6.7	19	10	8.5
3	21	6.8	12	24	7.8	20	24	7.8
4	32	5.3	13	20	5.7	21	5	7.8
5	23	5.2	14	17	6.5	22	23	7.3
6	13	7.2	15	23	6.4	23	12	7.3
7	20	5	16	17	6.4	24	21	7
8	20	5.5	17	13	7.6	25	8	7.9
9	8	6.5						

Use the computer to examine these data for evidence of an association between these two variables. If you detect an association, comment on its direction and strength.

Activity 6-16: Variables of Personal Interest (cont.)

Think of two pairs of measurement variables whose relationship you might be interested in studying. (Remember what a measurement variable is any characteristic of a person or object that can assume a range of numbers.) Be very specific in describing these variables; identify the cases as well as the variables.

WRAP-UP

This topic has introduced you to an area that occupies a prominent role in statistical practice: exploring relationships between variables. You have discovered the concept of association and learned how to construct and to interpret scatterplots and labeled scatterplots as visual displays of association.

Just as we moved from graphical displays to numerical summaries when describing distributions of data, we will consider in the next topic a numerical measure of the degree of association between two variables- the correlation coefficient.

Topic 7: Correlation Coefficient

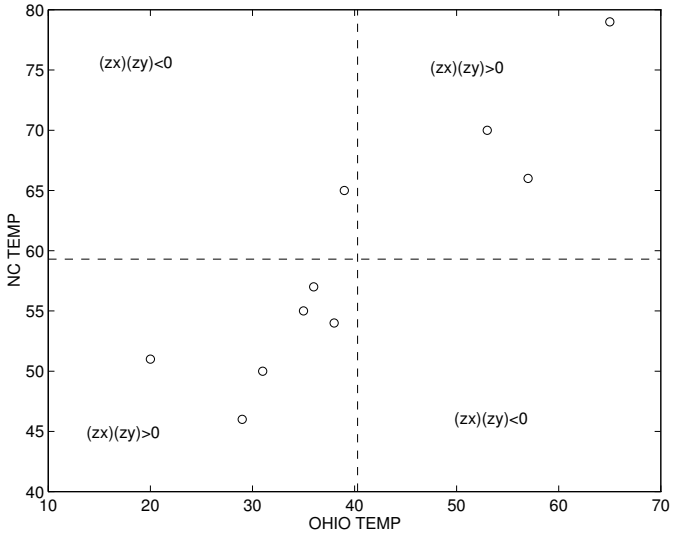
Introduction

You have seen how scatterplots provide useful visual information about the relationship between two variables. Just as you made use of numerical summaries of various aspects of the distribution of a single variable, it would be also handy to have a numerical measure of the association between two variables. This topic introduces you to just such a measure and asks you to investigate some of its properties. This measure is one of the most famous in statistics- the correlation coefficient.

Recall that we used a scatterplot to see the relationship between the two measurement variables. In the temperature example described in Topic 6, we saw that there was a positive relationship between the temperature in North Carolina and the temperature in Ohio. In the car example of Topic 6, we observed a negative relationship between the engine size of an automobile and its mileage. We would like to summarize the association pattern that we see in a scatterplot using a single number. This number should indicate if there is a positive or negative relationship between the two variables. In addition, the number should reflect the strength of the relationship that is seen in the scatterplot.

To motivate one measure of association, let us return to the temperature example. To get started, we compute the mean for the Ohio temperatures and for the North Carolina temperatures. Suppose that we let X denote the variable that is plotted along the horizontal axis and Y the variable along the vertical axis. We denote the mean of the X data as $\bar{x}$ and the mean of the Y variable as $\bar{y}$. Here the mean of the Ohio temperatures is $\bar{x} = 40.3$ and the mean of the North Carolina temperatures is $\bar{y} = 59.3$.

Consider the scatterplot of the $\{$Ohio temperature, North Carolina temperature$\}$ data that was constructed in the last section. In the figure on the next page, we divide the plot using the two means. First we draw a vertical line at the mean X value $\bar{x} = 40.3$. Then we draw a horizontal line at the mean Y value $\bar{y} = 59.3$. Note that we have divided the scatterplot square into four regions. The lower left region, which we will call Region 1, consists of the points where both the Ohio temperature and the North Carolina temperature are smaller than their respective means. The upper



Scatterplot of temperature data with points divided into four quadrants.

left region, Region 2, is where the Ohio temperature is smaller than its mean and the North Carolina temperature is larger than its mean. The upper right region, Region 3, is where both temperatures are larger than their means, and the lower right region, Region 4, is where the Ohio temperature is larger than its mean and the North Carolina is smaller than its mean. Note that most of the points in the scatterplot are in Regions 1 and 3. This means that small Ohio temperatures are associated with small North Carolina temperatures (Region 1) and large Ohio temperatures are associated with large North Carolina temperatures (Region 3).

Our measure of association is based on the computed z-scores for each of the two variables. Let s_X and s_Y denote the standard deviation of the X data and the Y data, respectively. For each Ohio temperature X, we replace its value by its z-score

$$z_X = \frac{X - \bar{x}}{s_X}.$$

Likewise, we replace a North Carolina temperature Y by its z-score

$$z_Y = \frac{Y - \bar{y}}{s_Y}.$$

For this data set, recall that the means of the two variables are $\bar{x} = 40.3$ and $\bar{y} = 59.3$. The corresponding standard deviations are given by $s_X = 13.86$ and $s_Y = 10.35$. So the z-scores of the temperatures 31 (Ohio) and 50 (North Carolina) are given by

$$z_X = \frac{31 - 40.3}{13.86} = -.67$$

and

$$z_Y = \frac{50 - 59.3}{10.35} = -.90.$$

Remember that a z-score tells us if a data value is smaller than its mean. So in Region 1 where both variables are smaller than their means, both z_X and z_Y will be negative. In Region 3 where both variables are larger than their means, both z-scores will be positive. In both of these two regions, note that the product of the z-scores

$$z_X \times z_Y$$

is positive.

In the two remaining regions, Regions 2 and 4, one of the z-scores will be positive and one will be negative. For example, in Region 2, we have a "small" Ohio temperature and a "large" North Carolina temperature, so z_X is negative and z_Y is positive. Here the product $z_X \times z_Y$ is negative. Likewise, in Region 4, z_X is positive, z_Y is negative and the product of the z-scores will be negative.

Our measure of association, the **correlation coefficient**, which we will denote by r, is the average of the products of the z-scores of the two variables. That is,

$$r = \frac{\text{sum of the products } z_X \times z_Y}{\text{number of points} - 1}$$

The table below illustrates the computation of the correlation coefficient for the temperature example. The "Ohio Temp." and "NC Temp." columns list the values of the temperatures. The column "Ohio z-score" gives the z-scores for the Ohio temperatures and "NC z-score" lists the z-scores for the North Carolina temperatures. The final column "(Ohio z-score)(NC z-score)" contains the products of the z-scores. The sum of this column is placed at the bottom in the "SUM" row. Since there are 10 points or 10 pairs of temperatures, we can compute the correlation coefficient:

$$r = \frac{8.21}{10 - 1} = .91$$

Ohio Temp.	Ohio z-score	NC Temp.	NC z-score	(Ohio z-score)(NC z-score)
31	-0.67	50	-0.90	0.60
20	-1.46	51	-0.80	1.17
65	1.78	79	1.90	3.38
38	-0.17	54	-0.51	0.09
36	-0.31	57	-0.22	0.07
29	-0.82	46	-1.29	1.06
53	0.92	70	1.03	0.95
57	1.20	66	0.65	0.78
35	-0.38	55	-0.42	0.16
39	-0.09	65	0.55	-0.05
SUM				8.21

Calculation of correlation coefficient for temperature example.

How do we interpret the correlation coefficient r? Since r is the average of the product of the z-scores $z_X \times z_Y$, it reflects the general pattern of these products. In this example, most of the points were located in Regions 1 and 3 where the product of the z-scores is positive. Thus the average of these products of z-scores will also be positive. In general, a positive value of r indicates that points cluster in Regions 1 and 3. This happens when there is a positive relationship in the scatterplot.

Suppose, instead that there was a negative relationship in the scatterplot. In this case, small values of X will be associated with large values of Y and visa versa. In this case, the points of the scatterplot will cluster in Regions 2 and 4. In these regions, the product of the z-scores is negative and so the correlation coefficient r (the average of the z-scores) will also be negative.

So the sign of the correlation coefficient r, that is, whether it is positive or negative, tells us if it is a positive or negative relationship. Also the size of the value of r tells us about the strength of the relationship. One fact about r is that it always falls between -1 and $+1$. The correlation coefficient will be equal to its extreme values, -1 or $+1$, only when all of the points fall on a straight line. In contrast, a value of $r = 0$ indicates a very weak relationship between the two variables.

For our temperature data, we computed $r = .91$. This indicates that there is a strong positive relationship between the Ohio temperatures and the North Carolina temperatures.

PRELIMINARIES

1. Take a guess as to the number of people per television set in the United States in 1990; do the same for China and for Haiti.

2. Do you expect that countries with few people per television tend to have longer life expectancies, shorter life expectancies, or do you suspect no relationship between televisions and life expectancy?

3. Taken from *An Altogether New Book of Top Ten Lists*, here are David Letterman's "Top Ten Good Things About Being a Really, Really Dumb Guy":

   A. Get to have own talk show with Canadian bandleader
   B. G.E. executive dining room has great clam chowder
   C. Already know the answer when people ask, What are you- an idiot?
   D. Can feel superior to really, really, really dumb guys
   E. May get to be Vice-President of the United States
   F. Pleasant sense of relief when Roadrunner gets away from Coyote
   G. Fun bumper sticker: I'd Rather Be Drooling
   H. Never have to sit through long, boring Nobel Prize banquet

I. Stallone might play you in the movie

J. Seldom interrupted by annoying request to "Put that in layman's terms"

These have been presented in a randomly determined order. Rank these ten jokes from funniest (1) to least funny (10). Record your ranking of each joke next to its letter designation:

joke letter	A	B	C	D	E	F	G	H	I	J
your ranking										

4. Take a guess concerning the average SAT score in your home state in 1991.

5. Would you expect states in which few students take the SAT to tend to have higher or lower average SAT scores than states in which most students take the SAT?

IN-CLASS ACTIVITIES

The correlation coefficient, denoted by r, is a measure of the degree to which two variables are associated. The calculation of r is very tedious to do by hand, so you will begin by letting the computer calculate correlations while you explore their properties.

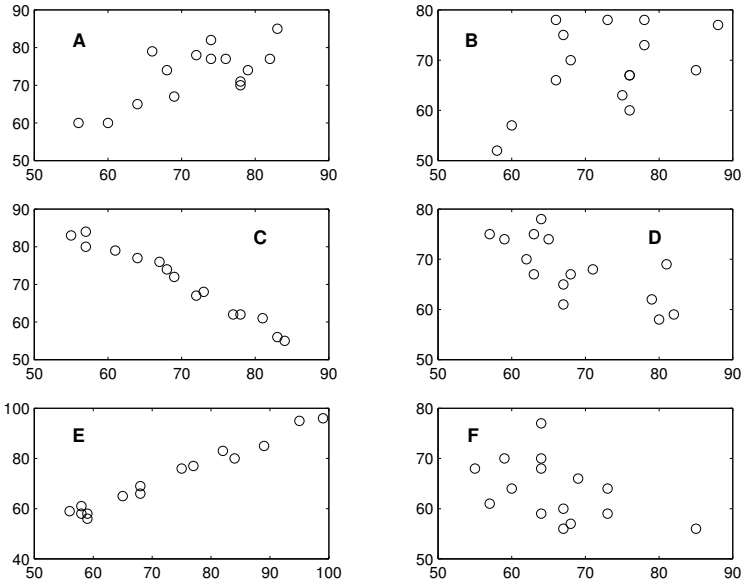
Activity 7-1: Properties of Correlation

(a) Look above at the scatterplots of the six classes (A-F) of hypothetical exam scores that you examined in Topic 6. The table below indicates the direction and strength of the association in each case. Your instructor will give you the correlation coefficient in each class and record its value in the table beside the appropriate letter designation.

	Strong	Moderate	Least Strong
Negative	C	D	F
Positive	E	A	B

(b) Based on these results, what do you suspect is the largest value that a correlation coefficient can assume? What do you suspect is the smallest value?

(c) Under what circumstances do you think the correlation assumes its largest or smallest value; i.e., what would have to be true of the observations in that case?



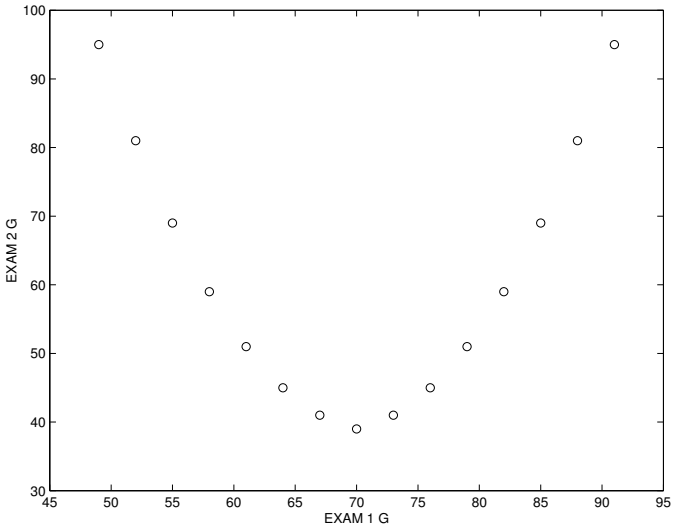
Scatterplots of some hypothetical exam scores.

(d) How does the value of the correlation relate to the direction of the association?

(e) How does the value of the correlation relate to the strength of the association?

Consider the scatterplot (shown above) of another hypothetical data set — exam scores from class G:

(f) Does there seem to be any relationship between the exam scores in class G? If so, describe the relationship.



Scatterplot of exam scores from class G

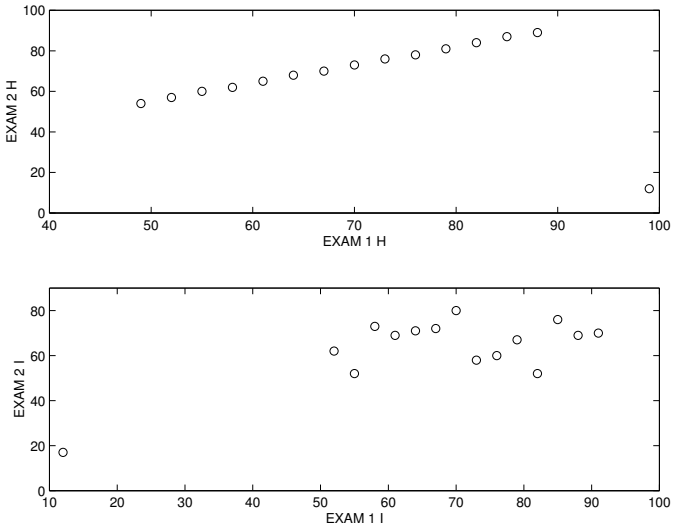
(g) Have the computer compute the correlation coefficient for class G; record it below. Does its value surprise you? Explain.

The example above illustrates that the correlation coefficient measures only linear (straight-line) relationships between two variables. More complicated types of relationships (such as curvilinear ones) can go undetected by r. Thus, there might be a relationship even if the correlation is close to zero. One should be aware of such possibilities and examine a scatterplot as well as the value of r.

Consider the scatterplots of two more hypothetical data sets from classes H and I that are shown above.

(h) In class H, do most of the observations seem to follow a linear pattern? Are there any exceptions?

(i) In class I, do most of the observations seem to be scattered haphazardly with no apparent pattern? Are there any exceptions?



Scatterplot of scores from class H (top) and class I (bottom).

(j) Have your instructor compute the correlation coefficient for each of these classes; record them below. Are you surprised at either of the values? Explain.

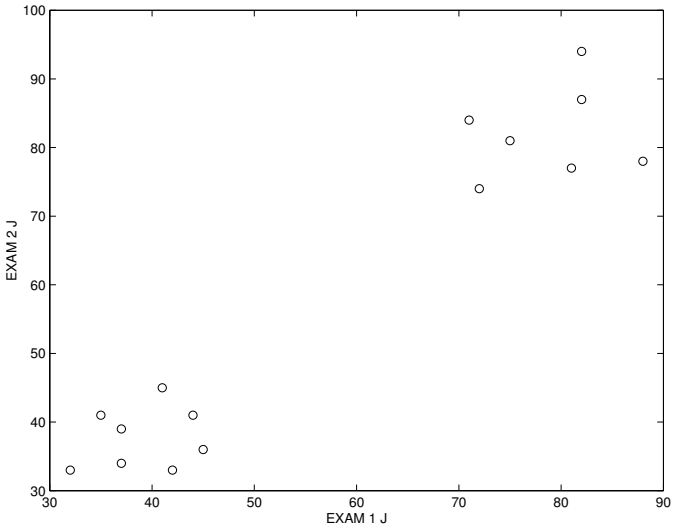
Consider the scatterplot above of one final hypothetical data set:

(k) Describe what the scatterplot reveals about the relationship between exam scores in class J.

(l) Your instructor will tell you the correlation coefficient between exam scores in class J. Is its value higher than you expected?

Activity 7-2: Televisions and Life Expectancy

The following table provides information on life expectancies for a sample of 22 countries. It also lists the number of people per television set in each country.



Scatterplot of exam scores from class J

country	life exp	per TV	country	life exp	per TV
Angola	44	200	Mexico	72	6.6
Australia	76.5	2	Morocco	64.5	21
Cambodia	49.5	177	Pakistan	56.5	73
Canada	76.5	1.7	Russia	69	3.2
China	70	8	South Africa	64	11
Egypt	60.5	15	Sri Lanka	71.5	28
France	78	2.6	Uganda	51	191
Haiti	53.5	234	United Kingdom	76	3
Iraq	67	18	United States	75.5	1.3
Japan	79	1.8	Vietnam	65	29
Madagascar	52.5	92	Yemen	50	38

(a) Which of the countries listed has the fewest people per television set? Which has the most? What are those numbers?

(b) Use the computer to produce a scatterplot of life expectancy vs. people per television set. Does there appear to be an association between the two variables? Elaborate briefly.

(c) Have the computer calculate the value of the correlation coefficient between life expectancy and people per television.

(d) Since the association is so strongly negative, one might conclude that simply sending television sets to the countries with lower life expectancies would cause their inhabitants to live longer. Comment on this argument.

(e) If two variables have a correlation close to +1 or to -1, indicating a strong linear association between them, does it follow that there must be a cause-and-effect relationship between them?

This example illustrates the very important distinction between association and causation. Two variables may be strongly associated (as measured by the correlation coefficient) without a cause-and-effect relationship existing between them. Often the explanation is that both variables are related to a third variable not being measured; this variable is often called a lurking or confounding variable.

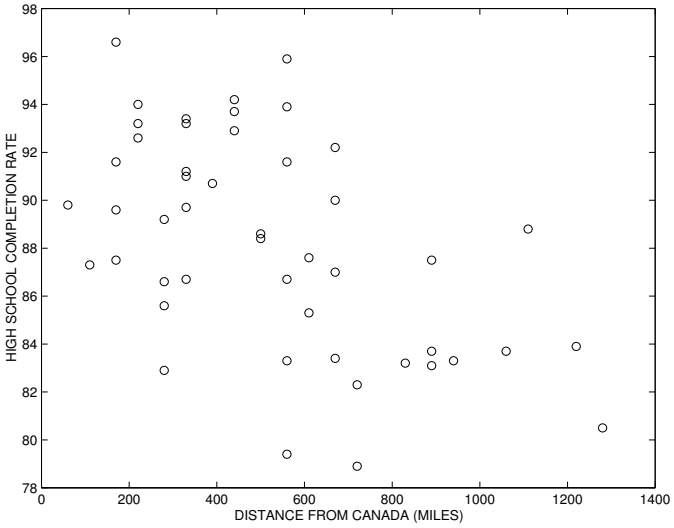
(f) In the case of life expectancy and television sets, suggest a confounding variable that is associated both with a country's life expectancy and with the prevalence of televisions in the country.

Activity 7-3: High School Completion Rates

State	Completion Rate	Distance	State	Completion Rate	Distance
Alabama	83.3	940	Nebraska	95.9	560
Arizona	83.7	1060	Nevada	83.4	670
Arkansas	87.5	890	New Hampshire	86.6	280
California	78.9	720	New Jersey	91.0	330
Colorado	87.6	610	New Mexico	83.7	890
Connecticut	92.6	220	New York	87.5	170
Delaware	93.7	440	North Carolina	85.3	610
Florida	83.2	830	North Dakota	96.6	170
Georgia	79.4	560	Ohio	89.6	170
Idaho	86.7	330	Oklahoma	83.1	890
Illinois	86.7	560	Oregon	82.9	280
Indiana	88.4	500	Pennsylvania	89.7	330
Iowa	94.2	440	Rhode Island	90.7	390
Kansas	92.2	670	South Carolina	87.0	670
Kentucky	83.3	560	South Dakota	93.2	330
Louisiana	83.9	1220	Tennessee	82.3	720
Maine	94.0	220	Texas	80.5	1280
Maryland	92.9	440	Utah	93.9	560
Massachusetts	91.2	330	Vermont	89.8	60
Michigan	89.2	280	Virginia	88.6	500
Minnesota	93.2	220	Washington	87.3	110
Mississippi	88.8	1110	West Virginia	85.6	280
Missouri	90.0	670	Wisconsin	93.4	330
Montana	91.6	170	Wyoming	91.6	560

One measure of the educational level of the people who live in a particular state is the percent of adults who have received a high school diploma. The table above gives the adult high school completion rate (as a percentage) for each of the 48 states. (These data were obtained from the 1998 Wall Street Journal Almanac.) Scanning this table, one notes considerable variability in these rates. For example, Nebraska (a northern state) has a high school completion rate of 95.9 %, while Georgia (a southern state) has a rate of only 79%. That raises an interesting question. Is there a relationship between the state's high school completion rate and its geographic location? To help answer this question, I got out our family's map of the United States and measured the distance from each state's capital to the Canadian border. These distances (in miles) are also recorded in the table.

The figure above plots the distance from Canada (horizontal axis) against the completion rate (vertical axis).



Scatterplot of high school completion rate against distance from the Canadian border for all states.

(a) By looking at the scatterplot, does there appear to be a relationship between distance and completion rate? What direction is the relationship? Is it a strong relationship?

(b) Make an intelligent guess at the value of the correlation r.

(c) Circle one point on the graph which corresponds to a state which is close to Canada and has a relatively small completion rate. Label this point 'A'. Looking at the data, which state does this correspond to?

(d) Circle a second point on the graph which corresponds to a state which is far from Canada and has a relatively large completion rate value. Label this point 'B'. Which state does this point correspond to?

(e) Is there a lurking variable present here? In other words, can you think of another variable which is closely related to both completion rate and distance that might help explain the relationship that we observe in the scatterplot?

(f) Suppose that a southern state is concerned about its relatively low high school completion rate. A state representative comments that maybe the solution to this problem is to move the residents of the state to a new location closer to Canada. Do you agree? Why or why not?

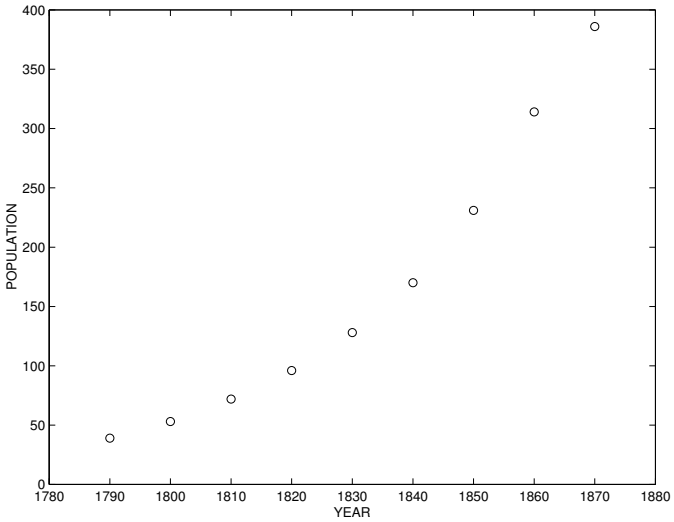
Activity 7-4: Cars' Fuel Efficiency (cont.)

For the fuel efficiency data that you analyzed in Activity 6-1, the weights have a mean of 3208 and a standard deviation of 590. The mean of the miles per gallon ratings is 27.56, and their standard deviation is 6.70. The table below begins the process of calculating the correlation between weight and miles per gallon by calculating the z-scores for the weights and miles per gallons and then multiplying the results.

model	weight	weight z-score	mpg	mpg z-score	product
BMW 3-Series	3250	0.07	28	0.07	0.00
BMW 5-Series	3675	0.79	23	-0.68	-0.54
Cadillac Eldorado	3840	1.07	19	-1.28	-1.37
Cadillac Seville	3935	1.23	20	-1.13	-1.39
Ford Aspire	2140	-1.81	43	2.30	-4.17
Ford Crown Victoria	4010	1.36	22	-0.83	-1.13
Ford Escort	2565	-1.09	34	0.96	-1.05
Ford Mustang	3450	0.41	22	-0.83	-0.34
Ford Probe	2900	-0.52	28	0.07	-0.03
Ford Taurus	3345	0.23	25	-0.38	-0.09
Ford Taurus SHO	3545	0.57	24	-0.53	-0.30
Honda Accord	3050	-0.27	31	0.51	-0.14
Honda Civic	2540	-1.13	34	0.96	-1.09
Honda Civic del Sol	2410	-1.35	36	1.26	-1.70
Honda Prelude	2865		30	0.36	
Lincoln Mark VIII	3810	1.02	22		

(a) Calculate the z-score for the weight of a Honda Prelude and for the miles per gallon rating of a Lincoln Mark VIII. Show your calculations below and record the results in the table.

(b) Add the results of the "product" column and then divide the result by 15 (one less than the sample size of 16 cars) to determine the value of the correlation coefficient between weight and mileage.



Population of the United States plotted against year number.

(c) What do you notice about the miles per gallon z-score of most of the cars with negative weight z-scores? Explain how this results from the strong negative association between weight and miles per gallon.

Activity 7-5: Population of the United States

The table below gives the population of the United States (in one hundred thousands) in the early years of its existence from 1790 to 1870. The figure on the next page plots the population against the year number.

Year	1790	1800	1810	1820	1830	1840	1850	1860	1870
Population	39	53	72	96	128	170	231	314	386

(a) From looking at the scatterplot, does there appear to be a relationship between the population and the year? If so, is it increasing or decreasing?

(b) Comment on the strength of the relationship (weak, moderate, strong).

(c) Make a guess at the correlation r for this data.

(d) Your instructor will give you the value of r found using a computer. Are you surprised at this value? Is it lower or higher than you what you expected?

This last activity emphasizes the fact that r measures the "straight-line" association between the two variables. The relationship between the U.S. population and time for the early years of its growth is strong, but this relationship is not described well by a straight line.

HOMEWORK ACTIVITIES

Activity 7-6: Properties of Correlation (cont.)

Suppose that every student in a class scores ten points higher on the second exam then on the first exam.

(a) Make up some data which satisfies this condition. (Grades for two students have already been put in the table.)

Student	Exam 1	Exam 2
1	65	75
2	80	90
3		
4		
5		

(b) Produce (by hand) a rough sketch of what the scatterplot would look like.

(c) Calculate the value of the correlation coefficient between the two exam scores.

(d) Repeat (a), (b), (c) supposing that every student scored twenty points lower on the second exam than on the first.

(e) Repeat (a), (b), (c) supposing that every student scored twice as many points on the second exam as on the first.

(f) Repeat (a), (b), (c) supposing that every student scored one-half as many points on the second exam as on the first.

(g) Based on your investigation of these questions, does the size of the slope that you see in a scatterplot (the tilt in the graph) have anything to do with the correlation between the two variables?

Activity 7-7: States' SAT Averages (cont.)

Reconsider the data from Activity 6-13 concerning the percentage of high school seniors in a state who take the SAT test and the state's average SAT score.

(a) Use the computer to calculate the correlation coefficient between the percentage of high school seniors in a state who take the SAT and the average SAT score in that state.

(b) Would you conclude that a cause-and-effect relationship exists between these two variables? Explain.

Activity 7-8: Ice Cream, Drownings, and Fire Damage

(a) Many communities find a strong positive correlation between the amount of ice cream sold in a given month and the number of drownings that occur in that month. Does this mean that ice cream causes drowning? If not, can you think of an alternative explanation for the strong association? Write a few sentences addressing these questions.

(b) Explain why one would expect to find a positive correlation between the number of fire engines that respond to a fire and the amount of damage done in the fire. Does this mean that the damage would be less extensive if only fewer fire engines were dispatched? Explain.

Activity 7-9: Evaluation of Course Effectiveness

Suppose that a college professor has developed a new freshman course that he hopes will instill students with strong general learning skills. As a means of assessing the success of the course, he waits until a class of freshmen that have taken the course proceed to graduate from college. The professor then looks at two variables: score on the final exam for his freshman course and cumulative college grade point average. Suppose that he finds a very strong positive association between the two variables (e.g., $r = 0.92$). Suppose further that he concludes that his course must

have had a positive effect on students' learning skills, for those who did well in his course proceeded to do well in college; those who did poorly in his course went on to do poorly in college. Comment on the validity of the professor's conclusion.

Activity 7-10: Space Shuttle O-Ring Failures (cont.)

Reconsider the data from Activity 6-5 concerning space shuttle missions. Use the computer to determine the value of the correlation between temperature and number of O-ring failures. Then exclude the flights in which no O-rings failed and recalculate the correlation. Explain why these correlations turn out to be so different.

Activity 7-11: Climatic Conditions

The following table lists a number of climatic variables for a sample of 25 American cities. These variables measure long-term averages of:

- January high temperature (in degrees Fahrenheit)
- January low temperature
- July high temperature
- July low temperature
- annual precipitation (in inches)
- days of measurable precipitation per year
- annual snow accumulation
- percentage of sunshine

city	Jan hi	Jan lo	July hi	July lo	precip	precday	snow	sun
Atlanta	50.4	31.5	88	69.5	50.77	115	2	61
Baltimore	40.2	23.4	87.2	66.8	40.76	113	21.3	57
Boston	35.7	21.6	81.8	65.1	41.51	126	40.7	58
Chicago	29	12.9	83.7	62.6	35.82	126	38.7	55
Cleveland	31.9	17.6	82.4	61.4	36.63	156	54.3	49
Dallas	54.1	32.7	96.5	74.1	33.7	78	2.9	64
Denver	43.2	16.1	88.2	58.6	15.4	89	59.8	70
Detroit	30.3	15.6	83.3	61.3	32.62	135	41.5	53
Houston	61	39.7	92.7	72.4	46.07	104	0.4	56
Kansas City	34.7	16.7	88.7	68.2	37.62	104	20	62
Los Angeles	65.7	47.8	75.3	62.8	12.01	35	0	73
Miami	75.2	59.2	89	76.2	55.91	129	0	73
Minneapolis	20.7	2.8	84	63.1	28.32	114	49.2	58
Nashville	45.9	26.5	89.5	68.9	47.3	119	10.6	56
New Orleans	60.8	41.8	90.6	73.1	61.88	114	0.2	60
New York	37.6	25.3	85.2	68.4	47.25	121	28.4	58
Philadelphia	37.9	22.8	82.6	67.2	41.41	117	21.3	56
Phoenix	65.9	41.2	105.9	81	7.66	36	0	86
Pittsburgh	33.7	18.5	82.6	61.6	36.85	154	42.8	46
St. Louis	37.7	20.8	89.3	70.4	37.51	111	19.9	57
Salt Lake City	36.4	19.3	92.2	63.7	16.18	90	57.8	66
San Diego	65.9	48.9	76.2	65.7	9.9	42	0	68
San Francisco	55.6	41.8	71.6	65.7	19.7	62	0	66
Seattle	45	35.2	75.2	55.2	37.19	156	12.3	46
Washington	42.3	26.8	88.5	71.4	38.63	112	17.1	56

(a) Use the computer to calculate the correlation coefficient between all pairs of these eight variables, recording your results in a table like the one below. (You need not record each value twice.)

	Jan hi	Jan lo	July hi	July lo	precip	precday	snow	sun
Jan hi	xxxx							
Jan lo		xxxx						
July hi			xxxx					
July lo				xxxx				
precip					xxxx			
precday						xxxx		
snow							xxxx	
sun								xxxx

(b) Which pair of variables has the strongest association? What is the correlation between them?

(c) Which pair of variables has the weakest association? What is the correlation between them?

(d) Suppose that you want to predict the annual snowfall for an American city and that you are allowed to look at that city's averages for these other variables. Which would be most useful to you? Which would be least useful?

(e) Suppose that you want to predict the average July high temperature for an American city and that you are allowed to look at that city's averages for these other variables. Which would be most useful to you? Which would be least useful?

(f) Use the computer to explore the relationship between annual snowfall and annual precipitation more closely. Look at and comment on a scatterplot of these variables.

Activity 7-12: Guess the Correlation

This activity will give you practice at judging the value of a correlation coefficient by examining the scatterplot.

(a) Have the computer generate some "pseudo-random" data and look at a scatterplot. Based solely on the scatterplot, take a guess at the value of the correlation coefficient r, recording your guess in the table below. (Make it a guessing contest with a partner!) Then have the computer compute the actual value of r and record it in the table also. Repeat this for a total of ten "pseudo-random" data sets.

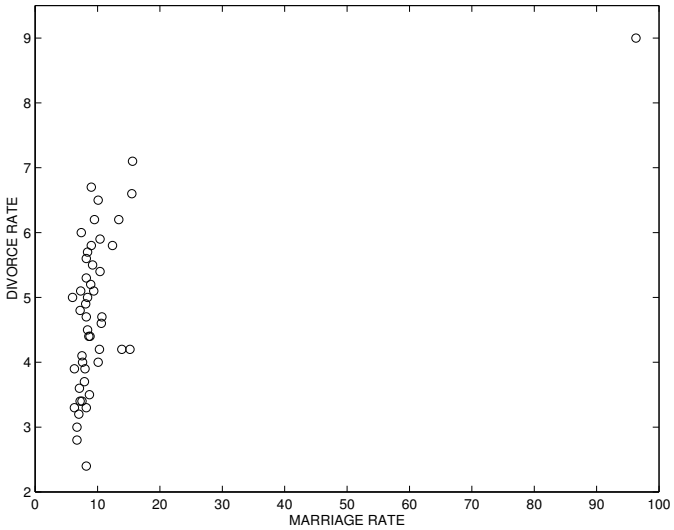
repetition	1	2	3	4	5	6	7	8	9	10
guess										
actual										

(b) Make a guess as to what the value of the correlation coefficient between your guesses for r and the actual values of r would be.

(c) Enter your guesses for r and the actual values of r into the computer and have the computer produce a scatterplot of your guesses vs. the actual correlations. Then ask it to compute the correlation between them; record this value below.

Activity 7-13: Marriage and Divorce Rates

In Activity 2-11, we considered the 1994 divorce rates of all of the states of the United States and the District of Columbia. The almanac also listed the marriage rates for all of the states. To investigate a relationship between the two rates, we construct a scatterplot pictured above, where the divorce rate is plotted on the vertical axis and the marriage rate on the horizontal axis.



Scatterplot of divorce rates against marriage rates for U.S. states.

(a) Describe in a few sentences any pattern or distinctive aspects of this scatterplot. Does a relationship appear to exist between the divorce rate and the marriage rate?

(b) The correlation coefficient for this dataset is calculated to be $r = .556$. Explain in words what this value means.

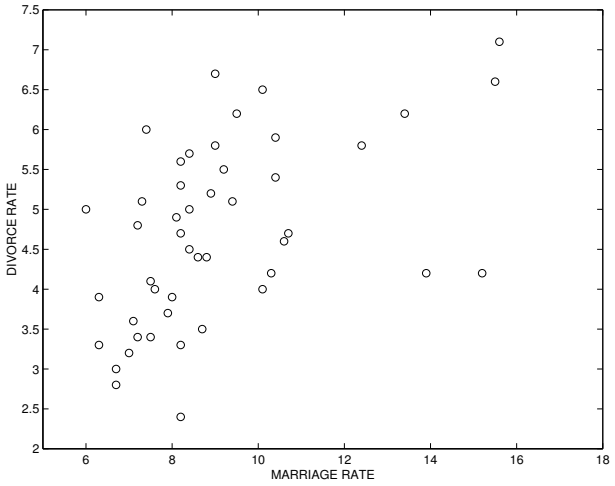
(c) Suppose that the marriage rate for Nevada (the unusual point in the scatterplot) is recorded incorrectly. Instead of 96.3 (the correct value), it is incorrectly recorded as 9.6. How would this data entry mistake affect the value of the correlation? Would the value of r go up, go down, or stay about the same?

(d) Your instructor will tell you the new value of r using the incorrect divorce rate for Nevada. Why is this new value so different from the old value?

(e) Suppose that we redraw the scatterplot with the Nevada point removed — the new scatterplot is shown above. Does this change have an affect on the appearance of the scatterplot? Which scatterplot do you prefer — the first one or the second one? Why?

Activity 7-14: Students' Family Sizes (cont.)

Reconsider the data on students' family sizes collected in Topic 6. Use the computer to calculate the correlation coefficients for each of the three pairs of variables. Which pair has the highest (in absolute value) correlation?



Scatterplot of divorce rates against marriage rates with Nevada excluded.

Activity 7-15: Students' Travels (cont.)

Reconsider the data on students' travels collected in Topic 1. Use the computer to look at a scatterplot of the number of states visited vs. number of countries visited and to calculate the correlation coefficient between these variables. Comment on what the correlation indicates about a relationship between the variables.

Activity 7-16: Students' Measurements (cont.)

Reconsider the data on students' measurements collected in Topic 2. Use the computer to calculate the correlation coefficients for each of the three pairs of variables. Which pair has the highest (in absolute value) correlation?

Activity 7-17: "Top Ten" Rankings

Recall your ranking of the "Top Ten" list from the Preliminaries section. David Letterman's ranking appears in the table below.

joke letter	A	B	C	D	E	F	G	H	I	J
your ranking										
Letterman's ranking	6	8	2	4	3	9	1	10	5	7

(a) Construct (by hand) a scatterplot of Letterman's ranking vs. your ranking of these jokes.

(b) Enter these rankings into the computer and calculate the value of the correlation between your ranking and Letterman's ranking. Record its value, and comment on how closely your ranking matches Dave's.

(c) Enter the rankings of your partner or another classmate into the computer and calculate the correlation between your rankings. Are your rankings more strongly correlated with Letterman's or with your classmate's?

Activity 7-18: Star Trek Episodes (cont.)

Refer back to the data concerning ranking Star Trek episodes in Activity 5-16.

(a) Based on the summary statistics and boxplots presented in Activity 5-16, would you expect an episode's ranking to be positively or negatively correlated with its chronological number?

(b) Use the computer to examine a scatterplot of an episode's ranking vs. its chronological number. Does the scatterplot reveal any obvious association between the two?

(c) Have the computer calculate the correlation coefficient between an episode's ranking and its chronological number. Report the value of this correlation; does its sign agree with your answer to (a)?

Activity 7-19: Variables of Personal Interest (cont.)

Think of a situation in which you would expect two variables to be strongly correlated even though no cause-and-effect relationship exists between them. Describe these variables in a paragraph; also include an explanation for their strong association.

WRAP-UP

In this topic you have discovered the very important correlation coefficient as a measure of the linear relationship between two variables. You have derived some of the properties of this measure, such as the values it can assume, how its sign and value relate to the direction and strength of the association, and its lack of resistance to outliers. You have also practiced judging the direction and strength of a relationship from looking at a scatterplot. In addition, you have discovered the distinction between correlation and causation and learned that one needs to be very careful about inferring causal relationships between variables based solely on a strong correlation.

The next topic will expand your understanding of relationships between variables by introducing you to least squares regression, a formal mathematical model which is often useful for describing such relationships.

Topic 8: Least Squares Regression

Introduction

When we have bivariate measurement data, we first use a scatterplot to see the relationship between the two variables. The correlation coefficient r is a number which describes the type of relationship (positive or negative) and the strength of the relationship. If there is a strong relationship between the variables, indicated by a large value of r, then we typically want to say something more about the association. We will describe this association by means of a straight line. We will find a line which goes through most of the points of the scatterplot. This line will tell us more about how the X variable is related to the Y variable. In particular, this line will let us make an intelligent prediction about the value of the Y variable if we know a value of the X variable.

In this section, we will introduce the use of a straight line in understanding relationships. First, we introduce the notion of a **straight line**. We will talk about the equation of a line and how one can plot a line on a graph. Then we will illustrate the computation of one particular line, the **least squares line**, which gives a good fit to the points of a scatterplot. We will talk about the interpretation of the least squares line and how it can be used to make **predictions**.

Lines

We review the concept of a line using an example. Suppose that there are two variables, X and Y, and each variable can take on values between 0 and 5. We represent possible values for the two measurements by means of the rectangular grid that is shown in the figure on the next page.

Consider the collection of points (X, Y) that line on a single straight line. For example, consider the points (0, 2), (1, 2.5), (2, 3), (3, 3.5), (4, 4), (5, 4.5) and all of the points that fall between these points. We can describe the collection of points of a straight line by the equation of the general form

$$Y = a + bX$$

The equation of a line is described by two numbers, a and b. The number a is called the **y-intercept**. This is the value of the variable Y when the value of the X variable is equal to 0. The

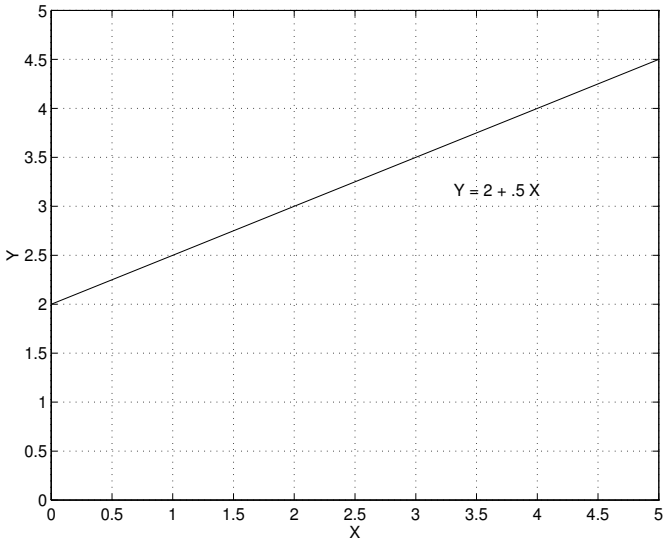


Figure 1: A graph of a line.

second number b is called the **slope**. The slope describes the tilt or steepness of the line. If (x_1, y_1) and (x_2, y_2) are two points on the line, then the slope is defined as the difference in the Y values divided by the difference in the X values:

$$b = \frac{y_2 - y_1}{x_2 - x_1}$$

A slope can be viewed as the change of the variable Y when the variable X is increased by one unit. For example, if $b = 2$, the variable Y increases by two units when X increases by a single unit. Since Y increases as X increases, a positive value of b corresponds to a line which is increasing. In contrast, if $b = -3$, the variable Y decreases by three units for a unit increase in X. Here the slope b is negative and the corresponding line is decreasing.

On the following figure, a line has been plotted. What is its equation? First we look for the y-intercept. When $X = 0$, we note that $Y = 2$. So the y-intercept $a = 2$. Next, we find the slope. We can find the slope by finding two points on the graph and then using the slope formula. Since the y-intercept is 2, one point on the line is (0, 2). We also see that the point (2, 3) is on the graph. So the slope of the line is

$$b = \frac{3 - 2}{2 - 0} = .5$$

The value $b = .5$ means that as X increases by one unit, Y will increase by .5 units. Its equation is given by

$$Y = 2 + .5X$$

Above we found the equation of the line from looking at its graph. Later in this chapter we will be given the equation of a line and wish to plot the line on the graph. Suppose that we are given a line with the equation

$$Y = 2.3X + 10$$

and wish to plot the line on the grid on the figure above.

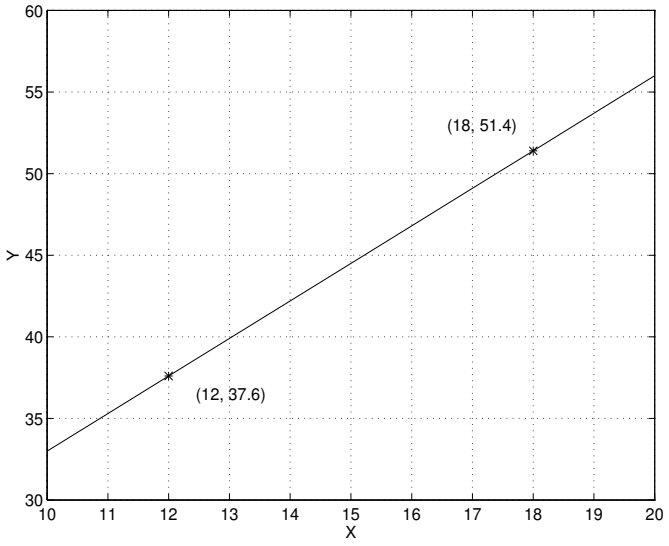


Figure 2: A graph of a line through two points.

An easy way to plot a line on a grid is to find two points on the line and connect the points by use of a straight edge. Note that the values of the variable X on the graph range from 10 to 20. Suppose I choose the value $X = 12$; the value of Y on the line is

$$Y = 2.3(12) + 10 = 37.6$$

Similarly, if I choose the second value $X = 18$, the corresponding value of Y is

$$Y = 2.3(18) + 10 = 51.4$$

The two points on the line are (12, 37.6) and (18, 51.4). On the figure, I plot the two points using thick dots and connect the points with a line.

PRELIMINARIES

1. If one city is farther away than another, do you expect that it generally costs more to fly to the farther city?

2. Take a guess as to how much (on average) each additional mile adds to the cost of air fare.

3. Would you guess that distance explains about 50% of the variability in air fares, about 65% of this variability, or about 80% of this variability?

4. Take a guess concerning the highest yearly salary awarded to a rookie professional basketball player in 1991.

IN-CLASS ACTIVITIES

Activity 8-1: Feeding Fido

An article in the February 1998 issue of Consumer Reports evaluates different brands of dog food with respect to their nutritional content. We will focus on two variables that are measured for each brand of canned dog food.

- FAT - the fat content (in grams) of a daily feeding of the dog food
- PRICE - the price (in cents) of this daily feeding

We are interested in the relationship between FAT and PRICE. A low fat diet is desirable for a dog, and so we would hope that higher priced dog foods will have lower fat contents. If Y represents the fat content of a particular brand of dog food and X represents its cost, then we will see that a simple *line model* for the relationship is given by

$$Y = 100 - 0.3 \times X.$$

If we replace the variables Y and X by the words FAT and PRICE, then this relationship can be expressed as

$$\text{FAT} = 100 - 0.3 \times \text{PRICE}.$$

(a) If a dog food costs 200 cents ($2), then use the line $\text{FAT} = 100 - 0.3 \times \text{PRICE}$ to guess at the fat content.

(b) If a dog food costs 120 cents, use the line equation to guess the fat content.

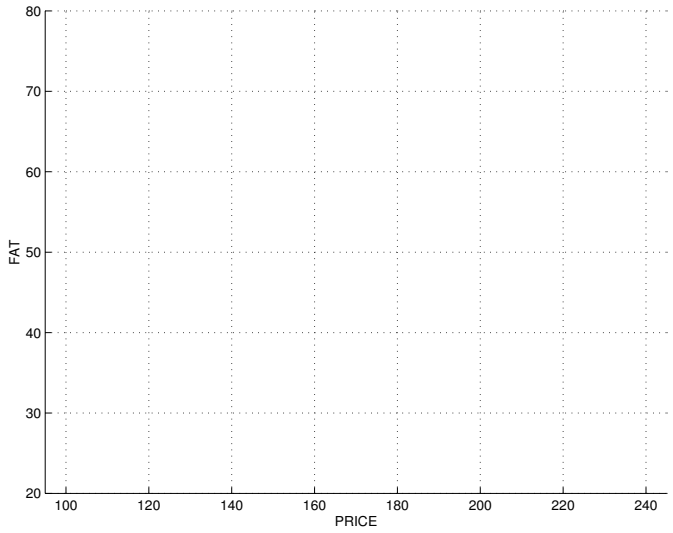


Figure 3: Grid for plotting relationship between FAT and PRICE.

(c) On the graph above, plot the ordered pairs (PRICE, FAT) that you found in parts (a) and (b). Connect the two points with a line. This is the graph of the equation of the line displayed above.

(d) Recall that the slope of the line is the change in the Y variable for each unit increase in the X variable. Here the slope is -0.3. Since it is negative, one expects FAT to decrease by 0.3 grams for each penny increase in PRICE. Suppose that the price increases by 10 cents. Do you expect the FAT content of the dog food to increase or decrease? How much do you expect the FAT content to change?

Finding a Best Line Using a String

A line provides a simple description of the pattern that we see in a scatterplot. Let's return to our temperature example. We know that there is a positive relationship between the Ohio temperature and the North Carolina temperature. This was clear from the pattern of the scatterplot and the computed value of the correlation coefficient r. Can we find a straight line which summarizes the positive pattern that we see in the plot?

We are interested in finding a line which is a good fit to the points. One way of finding this line is by experimentation. Suppose that we get a string and stretch it tight over the cluster of points. After some trial and error, it should be possible to find a line which appears to pass close to a large number of the points.

How do we find the equation of the line using this string method? When the string is stretched at its "best" location, use a pen to mark two points on the line. From the graph, find the X and Y values for these two points. The slope of the line can be found using the formula of the previous section.

We illustrate this process of finding a good line on the temperature data. Look at the figure on the next page. We place a string over our scatterplot and place it tight so that it seems to be a reasonable fit to the positive trend that we see. After some experimentation, we arrive at the thick line that is drawn on the scatterplot.

To find the formula for our line, we choose two points on the line. These are indicated by two large stars. Using the gridlines in the figure, we see the points are (32, 52) and (50, 67). Using the slope formula, we find that our line has slope

$$b = \frac{67 - 52}{50 - 32} = \frac{15}{18} = .83.$$

We can find the intercept a of our line by using the formula

$$Y = a + bX.$$

In the formula, we substitute the value of the slope b and the coordinates for one of the two points and solve the resulting equation for the intercept a. Here we substitute the slope $b = .83$ and the point (32, 52) ($X = 32, Y = 52$)

$$52 = a + .83(32)$$

Solving for a in the above equation, we obtain $a = 25.4$. So our line using the string method is

$$Y = 25.4 + .83X$$

Activity 8-2: Air Fares (cont.)

Consider the data from Activity 6-7 concerning distances and air fares, a scatterplot of which appears on the next page:

A natural goal is to try to use the distance of a destination to predict the air fare for flying there, and the simplest model for this prediction is to assume that a straight line summarizes the relationship between distance and air fare.

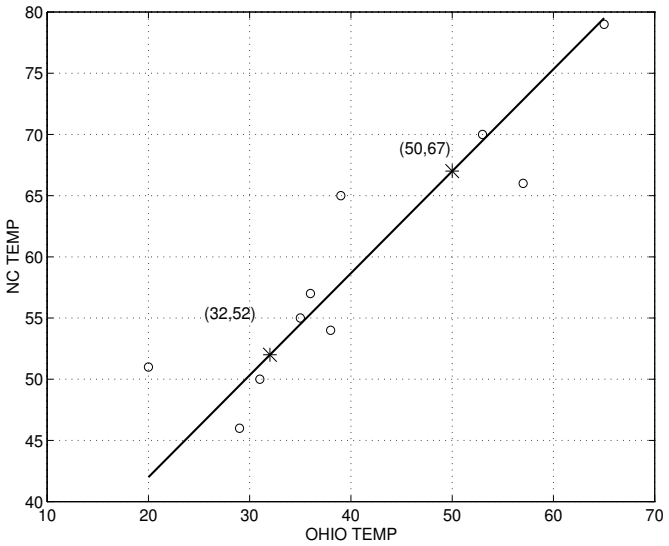


Figure 4: A best line through the points using the "string method".

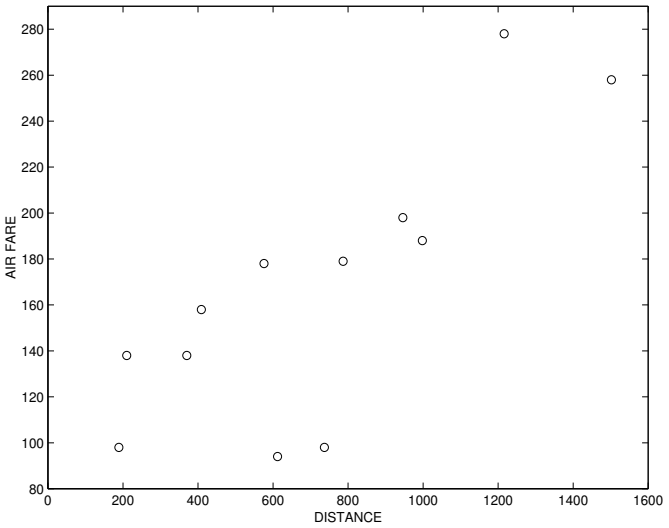


Figure 5: Scatterplot of distance and air fare.

(a) Place a thread over the scatterplot above so that the thread forms a straight line which roughly summarizes the relationship between distance and air fare. Then draw this line on the scatterplot.

(b) Roughly what air fare does your line predict for a destination which is 300 miles away?

(c) Roughly what air fare does your line predict for a destination which is 1500 miles away?

The equation of a line can be represented as $Y = a + bX$, where X denotes the variable being predicted (which is plotted on the vertical axis), Y denotes the variable being used for the prediction (which is plotted on the horizontal axis), a is the value of the y-intercept of the line, and b is the value of the slope of the line. In this case X represents distance and Y air fare.

(d) Use your answers to (b) and (c) to find the slope b of your line, remembering that slope

$$b = \frac{\text{rise}}{\text{run}} = \frac{\text{change in Y}}{\text{change in X}}.$$

(e) Use your answers to (d) and (b) to determine the intercept of your line, remembering that, if (x_1, y_1) is one point and b is the slope, then

$$a = y_1 - bx_1.$$

(f) Put your answers to (d) and (e) together to produce the equation of your line. It is good form to replace the generic symbols in the equation with the actual variable names, in this case *distance* and *air fare*, respectively.

The Least squares Line

This string method of finding a line is helpful in understanding what we mean by a line that is a good fit to the points in a scatterplot. However, it's a little cumbersome to use this method for every example. Also, although most people will find similar lines that fit a given scatterplot, there will be small differences in the slopes and intercepts of the equations of the lines that people actually find. It would be useful to have an automatic method of finding a good fitting line.

Fortunately, there is an automatic line that we can compute. This is called the **least squares line**. This is a line that is "best" using the least squares criteria. What does this mean? Consider the scatterplot of the temperature data given in the figure shown above. A possible best line is drawn on top of the scatterplot. Suppose that we consider the distance of each point from the line. These distances are represented by the short vertical lines that we see on the figure. One can measure the goodness of a particular line by the sum of the squares of these vertical distances. The least squares line is the line which makes the sum of squares of distances as small as possible (or short, least squares).

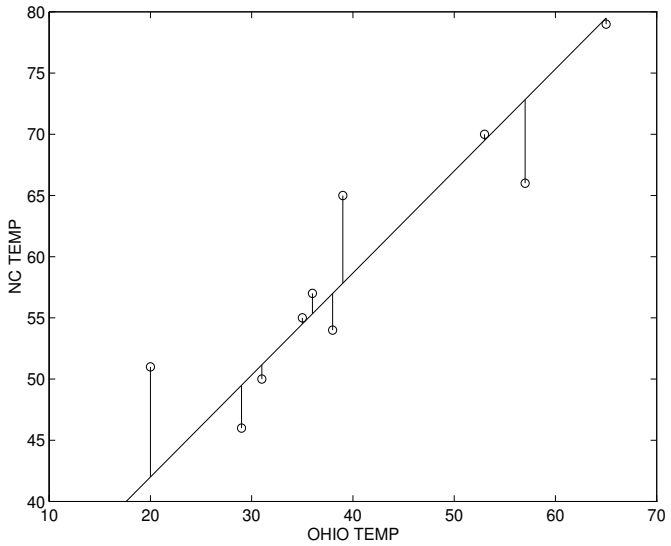


Figure 6: Vertical distances of ponts from one possible "best line".

The least squares line can be computed using quantities that we have already discussed in this chapter. Recall $\bar{x}$ and $\bar{y}$ are the means of the X and Y data, respectively. The respective standard deviations of the two datasets are s_X and s_Y. The correlation coefficient computed from the

scatterplot is r. Then the slope of the least squares line is given by

$$b = r\frac{s_Y}{s_X}$$

After we have computed the slope b, the y-intercept of the line is given by

$$a = \bar{y} - b\bar{x}$$

We will illustrate the computation of the least squares line for the temperature data. From earlier computations, we have

$$\bar{x} = 40.3, s_X = 13.2, \bar{y} = 59.3, s_Y = 9.8, r = .91$$

So the slope of the line is given by

$$b = .91 \times \frac{9.8}{13.2} = .68$$

The y-intercept is

$$a = 59.3 - .68 \times 40.3 = 31.9$$

So the least squares line relating the Ohio temperature (X) with the North Carolina temperature (Y) is

$$Y = 31.9 + .68X$$

Prediction

The least squares line is useful for predicting the value of the Y variable given a value of the X variable. To illustrate, suppose that my friend in Findlay, Ohio observes that the high temperature on a particular day is 50 degrees. Can she make an intelligent guess or prediction at the high temperature in Raleigh, North Carolina on that day?

Let us use the least squares line computed in the previous section. Recall that the Y variable was the North Carolina temperature and the X variable was the Ohio temperature. The least squares line was $Y = 31.9 + .68X$, or equivalently

$$\text{TEMP in NC} = 31.9 + .68\,\text{TEMP in OHIO}$$

We can use this formula to predict the North Carolina temperature given a value of the Ohio temperature. In particular, if the high Ohio temperature is 50 degrees, then one can predict that the high North Carolina temperature is

$$31.9 + .68 \times 50 = 65.9$$

degrees.

This predicted value can be found graphically. On the figure above, we find the Ohio temperature on the horizontal axis. Then we draw a vertical line up to the least squares line and a horizontal line to the left until we hit the y-axis. The value on the y-axis is the North Carolina predicted temperature.

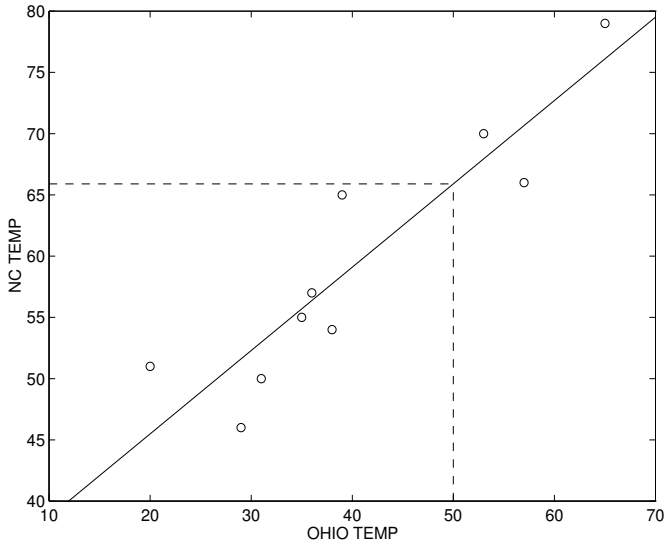


Figure 7: Temperature data with least squares line shown. The dotted lines illustrate using the line for prediction.

Activity 8-3: Air Fares (cont.)

The table below gives the mean and standard deviation of distance and air fare. Also the table displays the correlation between the two variables.

	mean	std. dev.	correlation
air fare (Y)	166.92	59.45	.795
distance (X)	712.67	402.69	

(a) Use these statistics and the formulas given above to calculate the least squares coefficients a and b; record them below:

(b) Write the equation of the least squares line for predicting air fare from distance (using the variable names DISTANCE and AIR FARE rather than the generic X and Y).

One of the primary uses of regression is for prediction, for one can use the regression line (another name for the least squares line) to predict the value of the Y -variable for a given value of the X-variable simply by plugging that value of X into the equation of the regression line. This is, of course, equivalent to finding the Y-value of the point on the regression line corresponding to the X-value of interest.

(c) What air fare does the least squares line predict for a destination which is 300 miles away?

(d) What air fare does the least squares line predict for a destination which is 1500 miles away?

(e) Draw the least squares line on the scatterplot on page 165 by plotting the two points that you found in (c) and (d) and connecting them with a straight line.

(f) Just from looking at the regression line that you have drawn on the scatterplot, guess what value the regression line would predict for the air fare to a destination 900 miles away.

(g) Use the equation of the regression line to predict the air fare to a destination 900 miles away, and compare this prediction to your guess in (f).

Cautions about Prediction

A couple of cautions should be made about prediction. First, although we predict the North Carolina temperature to be 65.9 degrees on that particular day, this is not a sure thing. Since the points don't fall on the line, the relationship between the temperatures in the two states is not precise. It would not be unexpected for the actual North Carolina temperature to be five degrees warmer or cooler than the value to be predicted. Second, we can only make predictions for the range of values of the X variable in the data. In this example, the Ohio temperatures in the data range between 20

and 65 degrees. It would be permissible only to make predictions for this range, since we have not observed any temperature data outside these values. As an extreme example, suppose that the Ohio temperature on a frigid day in January was -20. It would be ill-advised to use our least squares formula to predict the North Carolina temperature on this day.

Activity 8-4: Air Fares (cont.)

(a) What air fare would the regression line predict for a flight to San Francisco, which is 2842 miles from Harrisburg? Would you take this prediction as seriously as the one for 900 miles? Explain.

The actual air fare to San Francisco at that time was 198 dollars. That the regression line's prediction is not very reasonable illustrates the danger of extrapolation; i.e., of trying to predict for values of distance beyond those contained in the data. Since we have no reason to believe that the relationship between distance and air fare remains roughly linear beyond the range of values contained in our data set, such extrapolation is not advisable.

(b) Use the equation of the regression line to predict the air fare if the distance is 900 miles. Record the prediction in the table below, and repeat for distances of 901, 902, and 903 miles.

distance	900	901	902	903
predicted air fare				

(c) Do you notice a pattern in these predictions? By how many dollars is each prediction higher than the preceding one? Does this number look familiar (from your earlier calculations)? Explain.

This exercise demonstrates that one can interpret the slope coefficient of the least squares line as the predicted change in the Y-variable (air fare, in this case) for a one-unit change in the X-variable (distance).

(d) By how much does the regression line predict air fare to rise for each additional 100 miles that a destination is farther away?

The Fit and the Residual

A common theme in statistical modeling is to think of each data point as being comprised of two parts: the part that is explained by the model (often called the **fit**) and the "left-over" part (often called the **residual**) that is either the result of chance variation or of variables not measured. In the context of least squares regression, the fitted value for an observation is simply the Y-value that the regression line would predict for the X-value of that observation. The residual is the difference between the actual Y-value and the fitted value (residual = actual -fitted), so the residual measures the vertical distance from the observed Y-value to the regression line.

Let's illustrate the computation of residuals for the temperature example. We'll focus on the computation of residuals for two days — the Ohio and North Carolina temperatures for these two days are listed in the table below.

Ohio Temp	NC Temp	Fitted Value	Residual
31	50		
20	51		

Consider the first day where the Ohio temperature was 31 degrees and the North Carolina temperature was 50 degrees. Recall that the equation of the least squares line was computed to be

$$Y = 31.9 + .68X,$$

where X is the Ohio temperature and Y the NC temperature. The residual is the difference between the actual NC temperature and the NC temperature that is predicted from the least squares line. The actual NC temperature that we observed on this day is

$$50.$$

We find the fitted or predicted NC temperature using the equation of the line. If the Ohio temperature is 31° (that is, $X = 31$), then the fitted Y value is

$$31.9 + .68 \times 31 = 52.98.$$

So the residual is equal to

$$50 - 52.98 = -2.98.$$

The fitted value and the residual are placed in the table below. The residual computation is illustrated in the figure on the next page. The observed data point is the right circle that is filled in. The large "x" on the least squares line indicates the fitted Y value. The vertical line from the observed data point to the fitted value corresponds to the residual. In this case, the residual has a *negative* value, which indicates that the observed data point falls *below* the line.

Similarly, we can compute the residual for the second day. We'll do this in outline form.

1. The observed NC temperature on this day is $51°$.
2. Since the Ohio temperature on this day is $20°$, the fitted NC temperature is $31.9 + .68 \times 20 = 45.5°$.
3. The residual is the difference $51° - 45.5° = 5.5°$

This residual is *positive* in sign which indicates that the data point is *above* the least squares line. Note that the second day residual (5.5) is larger in size than the first day residual (-2.98). This means that the second day NC temperature is farther from the line than the first day NC temperature — we did a better job predicting the NC temperature on the first day than the second day.

Ohio Temp	NC Temp	Fitted Value	Residual
31	50	52.98	-2.98
20	51	45.5	5.5

Activity 8-5: Air Fares (cont.)

(a) If you look back at the original listing of distances and air fares, you find that Atlanta is 576 miles from Baltimore. What air fare would the regression line have predicted for Atlanta? (This is the fitted value for Atlanta.)

(b) The actual air fare to Atlanta at that time was \$178. Determine the residual value for Atlanta by subtracting the predicted fare from the actual one.

(c) Record your answers to (a) and (b) in the table below. Then calculate Boston's residual and Chicago's fitted value without using the equation of the regression line, showing your calculations.

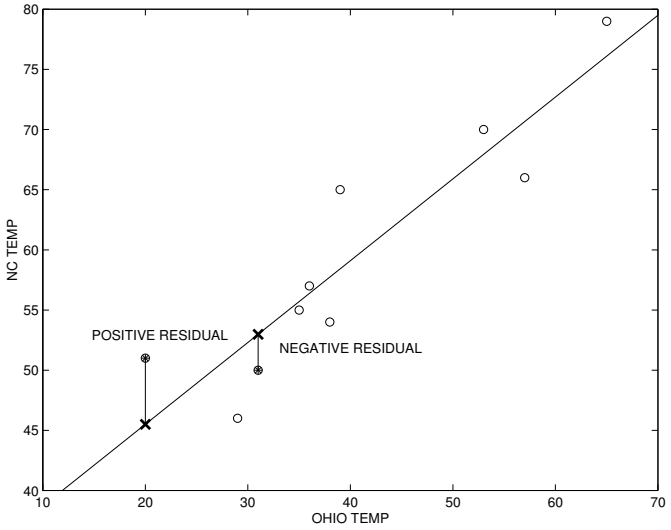


Figure 8: Illustration of the computation of two residuals.

destination	distance	air fare	fitted	residual
Atlanta	576	178		
Boston	370	138	126.70	
Chicago	612	94		-61.10
Dallas/Fort Worth	1216	278	226.00	52.00
Detroit	409	158	131.27	26.73
Denver	1502	258	259.57	-1.56
Miami	946	198	194.30	3.70
New Orleans	998	188	200.41	-12.41
New York	189	98	105.45	-7.45
Orlando	787	179	175.64	3.36
Pittsburgh	210	138	107.92	30.08
St. Louis	737	98	169.77	-71.77

(d) Which city has the largest (in absolute value) residual? What were its distance and air fare? By how much did the regression line err in predicting its air fare; was it an underestimate or an overestimate? Circle this observation on the scatterplot containing the regression line on the next page.

(e) For observations with positive residual values, was their actual air fare greater or less than the

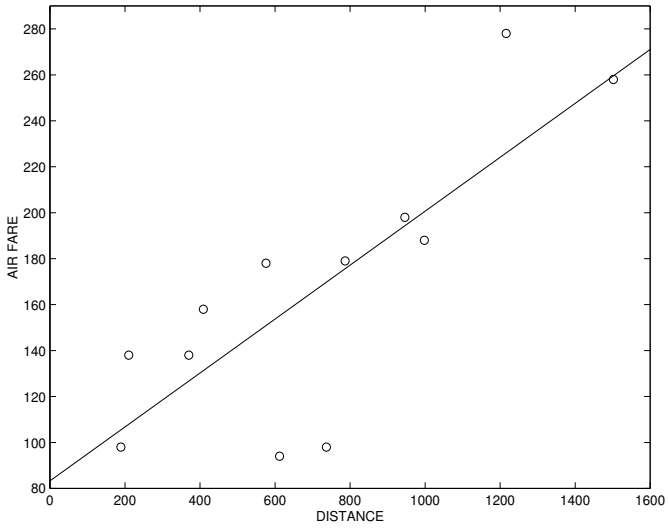


Figure 9: Scatterplot of air fare and distance with least squares line plotted on top.

predicted air fare?

(f) For observations with negative residual values, do their points on the scatterplot fall above or below the regression line?

(g) Use the computer to find the residuals and to calculate summary statistics concerning the residuals. Record the mean and standard deviation of the residuals below.

Recall the idea of thinking of the data as consisting of a part explained by the statistical model and a "left-over" (residual) part due to chance variation or unmeasured variables. One can obtain a numerical measure of how much of the variability in the data is explained by the model and how much is "left over" by comparing the residual standard deviation with the standard deviation of the dependent (Y) variable.

(h) Recall that you found the standard deviation of the air fares in Activity 8.3 and the standard deviation of the regression residuals in (g). Divide the standard deviation of the regression

residuals by the standard deviation of the air fares. Then square this value and record the result below.

(i) Recall that you also found the correlation between distance and air fare in Activity 8.3. Square this value and record the result below.

(j) What do you notice about the sum of your answers in (h) and (i)?

It turns out that the square of the correlation coefficient (written r^2) measures the proportion of variability in the Y-variable that is explained by the regression model with the X-variable. This percentage measures how closely the points fall to the least squares line and thus also provides an indication of how confident one can be of predictions made with the line.

(k) What proportion of the variability in air fares is explained by the regression line with distance? (You have already done this calculation above.)

Activity 8-6: Students' Measurements

Refer back to the data collected in Topic 2 on students' heights and foot length.

(a) Use the computer to produce a scatterplot of height vs. foot length. Based on the scatterplot does it look like a line would summarize well the relationship between height and foot length?

(b) Have the computer determine the equation of the least squares line for predicting height from foot length. Record the equation below and have the computer draw the line on the scatterplot. Does the line seem to summarize the relationship well?

(c) Interpret the value of the slope coefficient. In other words, explain what the value of the slope signifies in terms of predicting height from foot length.

(d) What proportion of the variability in heights is explained by the least squares line involving foot lengths?

(e) Use the equation of the least squares line to predict the height of a student whose foot length is 25 centimeters. Show your calculation below.

HOMEWORK ACTIVITIES

Activity 8-7: Students' Measurements (cont.)

Consider again the data on students' heights and foot sizes, but this time analyze men's and women's data separately.

(a) Use the computer to determine the regression equation for predicting a male student's height from his foot length. Record the value of the equation.

(b) Use the computer to determine the regression equation for predicting a female student's height from her foot length. Record the value of the equation.

(c) Compare the slopes of these two equations in (a) and (b), commenting on any differences between them.

(d) What height would the regression equation in (a) predict for a man with a foot length of 25 centimeters?

(e) What height would the regression equation in (b) predict for a woman whose foot length is 25 centimeters?

(f) What height would the regression equation in Activity 8-6 predict for a student of unknown gender with a foot length of 25 centimeters?

(g) Comment on how much the predictions in (d), (e), and (f) differ.

Activity 8-8: Cars' Fuel Efficiency (cont.)

Refer back to the data in Activity 6-1 concerning the relationship between cars' weight and fuel efficiency. The means and standard deviations of these variables and the correlation between them are reported below:

	mean	std. dev.	correlation
weight	3208	590	-0.959
mpg	27.56	6.70	

(a) Use this information to determine (by hand) the coefficients of the least squares line for predicting a car's miles per gallon rating from its weight.

(b) By how many miles per gallon does the least squares line predict a car's fuel efficiency to drop for each additional 100 pounds of weight? (Use the slope coefficient to answer this question.)

(c) What proportion of the variability in cars' miles per gallon ratings is explained by the least squares line with weight?

Activity 8-9: Governor Salaries (cont.)

Reconsider the data from Activity 6-14 concerning governor salaries and average pay in the states.

(a) Use the computer to determine the regression equation for predicting a state's governor salary from its average pay. Also have the computer calculate residuals. Record the equation of the regression line.

(b) What proportion of the variability in governor salaries is explained by this regression line with average pay?

(c) Which state has the largest positive residual? Explain what this signifies about the state.

(d) Which state has the largest (in absolute value) negative residual? Explain what this signifies about the state.

(e) Which state has the largest fitted value? Explain how you could determine this from the raw data without actually calculating any fitted values.

Activity 8-10: Basketball Rookie Salaries

The table below pertains to basketball players selected in the first round of the 1991 National Basketball Association draft. It lists the draft number (the order in which the player was selected) of each player and the annual salary of the contract that the player signed. The two missing entries are for players who signed with European teams.

pick no.	salary	pick no.	salary	pick no.	salary
1	$3,333,333	10	$1,010,652	19	$828,750
2	$2,900,000	11	$997,120	20	$740,000
3	$2,867,100	12	$1,370,000	21	$775,000
4	$2,750,000	13	$817,000	22	$180,000
5	$2,458,333	14	$675,000	23	$550,000
6	$1,736,250	15	*	24	$610,000
7	$1,590,000	16	$1,120,000	25	*
8	$1,500,000	17	$1,120,000	26	$180,000
9	$1,400,000	18	$875,000	27	$605,000

(a) Use the computer to look at a scatterplot of the data and to calculate the regression line for predicting salary from draft number. Record this regression equation below.

(b) What proportion of the variability in salary is explained by the regression model with draft number?

(c) Calculate (by hand) the fitted value and residual for the player who was draft number 12.

(d) What yearly salary would the regression line predict for the player drafted at number 15? How about for number 25?

(e) By how much does the regression line predict the salary to drop for each additional draft number? In other words, how much does a player stand to lose for each additional draft position which passes him by?

Activity 8-11: Fast Food Sandwiches (cont.)

Reconsider the data from Activity 6-4 about fast food sandwiches. The mean serving size is 7.557 ounces; the standard deviation of serving sizes is 2.008 ounces. The mean calories per sandwich is 446.9 with a standard deviation of 143.0. The correlation between serving size and calories is 0.849.

(a) Use this information to determine the least squares line for predicting calories from serving size. Record the equation of this line.

(b) Reproduce the labeled scatterplot from Activity 6-4 and sketch the least squares line on it. How well does the line appear to summarize the relationship between calories and serving size?

(c) What proportion of the variability in calories is explained by the least squares line with serving size?

(d) Do the roast beef sandwiches tend to fall above the line, below the line, or about half and half? What about the chicken sandwiches? How about turkey? Comment on what your findings reveal.

Activity 8-12: Electricity Bills

The following table lists the average temperature of a month and the amount of the electricity bill for that month:

month	temp	bill	month	temp	bill
Apr-91	51	$41.69	Jun-92	66	$40.89
May-91	61	$42.64	Jul-92	72	$40.89
Jun-91	74	$36.62	Aug-92	72	$41.39
Jul-91	77	$40.70	Sep-92	70	$38.31
Aug-91	78	$38.49	Oct-92	*	*
Sep-91	74	$37.88	Nov-92	45	$43.82
Oct-91	59	$35.94	Dec-92	39	$44.41
Nov-91	48	$39.34	Jan-93	35	$46.24
Dec-91	44	$49.66	Feb-93	*	*
Jan-92	34	$55.49	Mar-93	30	$50.80
Feb-92	32	$47.81	Apr-93	49	$47.64
Mar-92	41	$44.43	May-93	*	*
Apr-92	43	$48.87	Jun-93	68	$38.70
May-92	57	$39.48	Jul-93	78	$47.47

(a) Before you examine the relationship between average temperature and electric bill, examine the distribution of electric bill charges themselves. Create (by hand) a dotplot of the electric bill charges, and use the computer to calculate relevant summary statistics. Then write a few sentences describing the distribution of electric bill charges.

(b) Use the computer to produce a scatterplot of electric bill vs. average temperature. Does the scatterplot reveal a positive association between these variables, a negative association, or not much association at all? If there is an association, how strong is it?

(c) Use the computer to determine the equation of the least squares (regression) line for predicting the electric bill from the average temperature. Record the equation of the line.

(d) Use this equation to determine (by hand) the fitted value and residual for March of 1992.

(e) Use the computer to calculate the residuals and fitted values for each month. Which month(s) have unusually large residual values? Were their electric bills higher or lower than expected for their average temperature?

(f) Create (by hand) a dotplot of the distribution of residuals and comment on its key features.

Activity 8-13: Turnpike Tolls

If one enters the Pennsylvania Turnpike at the Ohio border and travels east to New Jersey, the mileages and tolls for the turnpike exits are as displayed in the scatterplot below. The regression line for predicting the toll from the mileage has been drawn on the scatterplot; its equation is: TOLL $= -0.006 + 0.0400$MILEAGE. The correlation between toll and mileage is 0.999.

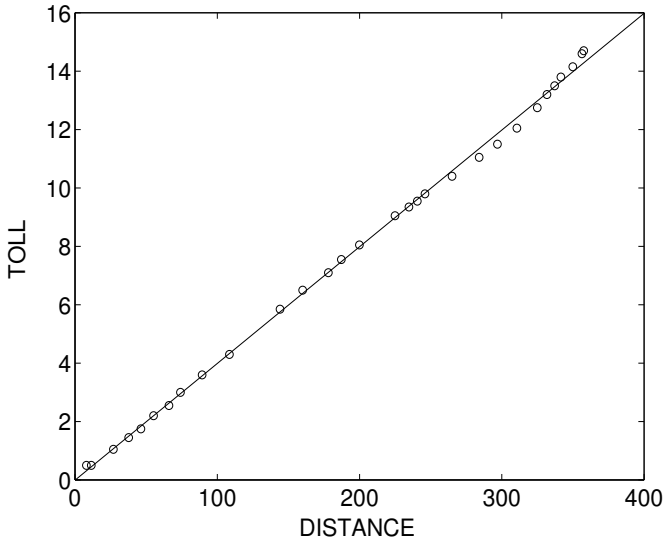


Figure 10: Scatterplot of mileages and tolls for all exits of the Pennsylvania Turnpike.

(a) What proportion of the variability in turnpike tolls is explained by the regression line with mileage?

(b) Use the regression equation to predict the toll for a person who needs to drive 150 miles on the turnpike.

(c) By how much does the regression equation predict the toll to rise for each additional mile that you drive on the turnpike?

(d) About how many miles do you have to drive in order for the toll to increase by one dollar?

Activity 8-14: Beatles' Hit Songs

In topic 5, the times of songs from three albums of the Beatles were compared. Here we analyze characteristics of the entire set of singles that were released by the Beatles during their career. There

were 58 Beatles singles that made the Billboard hit chart. The first song that reached number 1 on the chart was "I Want to Hold Your Hand" in 1964 and their last song to make number 1 was "Long and Winding Road" in 1970. For each Beatles single, two variables were recorded. The first variable which we call PEAK is the highest position on the Billboard hit chart, and the second variable WEEKS is the number of weeks that the song appeared on the Billboard Top 100 chart. One of the author's personal favorites, "Strawberry Fields", reached number 8 on the charts and stayed on the Top 100 for 9 weeks, so PEAK = 8 and WEEKS = 9.

The figure above displays a scatterplot of the PEAK and WEEKS variables for all 58 singles. To better understand the relationship between the two variables, we compute the least squares line which is given by

$$\text{WEEKS} = 11.54 - 0.124\text{PEAK}.$$

This line is placed on top of the scatterplot in the figure.

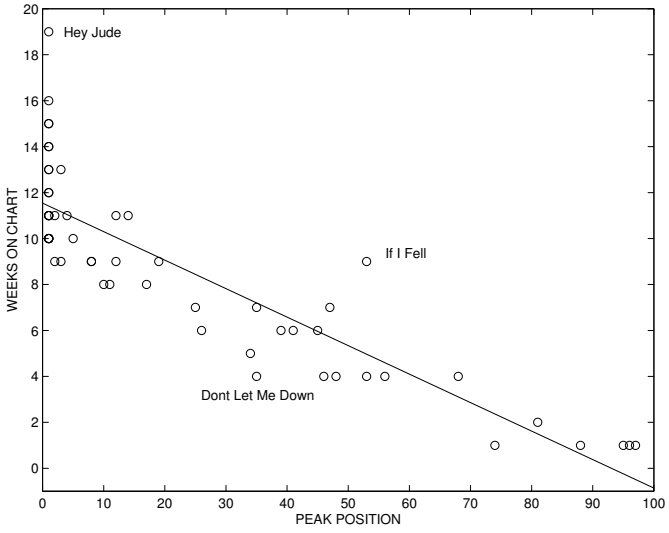


Figure 11: Scatterplot of weeks on chart and peak position for Beatles' singles.

(a) Describe the general relationship between WEEKS and PEAK that you see in the scatterplot.

(b) Suppose that a Beatles' song peaks at number 20 on the hit chart. Use the least squares line to predict how many weeks this song will stay on the Billboard Top 100.

(c) The point corresponding to the song "Hey Jude" is labelled on the scatterplot. This song peaked at number 1 and stayed 19 weeks on the hit chart. Compute the residual for this song.

(d) Two other songs, "If I Fell" and "Don't Let Me Down", are also labelled on the plot. By just looking at the plot, estimate the residuals for each of these songs. What is distinctive about these two songs that makes them have large residuals?

(e) Where in the plot are the negative residuals (the points that fall below the line) located? Where are the positive residuals located? This pattern in the residuals suggests that a straight line is not the best fit to this particular data set.

Activity 8-15: Climatic Conditions (cont.)

Reconsider the climatic data presented in Activity 7-10.

(a) Choose any pair of variables (preferably a pair which is strongly correlated) and use the computer to determine the least squares line for predicting one variable from the other. Record the equation of this line, being sure to identify which variables you are considering.

(b) Select one particular value of the independent variable and use the least squares line to predicting the value of the dependent variable.

(c) Which city has the largest (in absolute value) residual from your least squares line? What is the value of this residual? Interpret what this residual says about the city.

(d) What proportion of the variability in the dependent variable is explained by the least squares line?

WRAP-UP

This topic has led you to study a formal mathematical model for describing the relationship between two variables. In studying least squares regression, you have encountered a variety of related terms and concepts. These ideas include the use of regression in prediction, the danger of extrapolation, the interpretation of the slope coefficient, the concepts of fitted values and residuals, and the interpretation of r^2 as the proportion of variability explained by the regression line. Understanding all of these ideas is important to applying regression techniques thoughtfully and appropriately.

Topic 9: Relationships with Categorical Variables

Introduction

You have been studying how to explore relationships between two variables, but the variables that you have analyzed have all been measurement variables. This topic asks you to study some basic techniques for exploring relationships among categorical variables. (Remember that a categorical variable is one which records simply that category into which a person or thing falls on the characteristic in question.) These techniques involve the analysis of two-way tables of counts.

Let's return to the example described in Topic 1 in which I was shopping for a used car by looking at car ads in the local newspaper. In this setting, we may be interested in the relationship between year and manufacturer name. How many of the vintage cars in the ad are foreign? If I'm interested in purchasing a Chevrolet, are there many Chevrolets that are nearly new? Could I make a general statement like "most of the old cars are foreign and the new cars are typically American"?

We can learn about the relationship between two categorical variables by means of a **two-way table**. To illustrate the construction of a two-way table, let us consider a new example. In professional basketball, points are scored in different ways. A team scores points by making baskets during the flow of the game. Two points are scored by baskets made inside of the 3-point line and three points are scored by baskets made outside of the 3-point line. Teams also score points by making foul shots (or free throws) from the free throw line. Suppose that we are interested in learning about the relationship between a player's ability to make two-point shots and his ability to make free throws. If a player makes a high percentage of two-point shots, is it reasonable to believe that he will also be successful in shots from the free throw line? Likewise, if a player doesn't shoot well in two-point shots, will he also be weak from the free throw line?

To investigate the relationship between two-point shooting and foul shots, we need some data. Consider the shooting statistics of players in the National Basketball Association for the 1994-95 year. To learn about a player's free throw shooting ability, he should have enough opportunities. (It

is certainly hard to learn about someone's shooting ability if he has only taken 10 foul shots during the season.) So we limit our study to the 236 players who had at least 80 free throw attempts. All of these players had at least 80 two-point shot attempts. For each player, we classify his shooting ability as "poor", "average" or "good". We will say that a player is a "poor" two-point shooter if his shooting percentage is 44% or smaller, is a "average" two-point shooter if his percentage is between 45% and 49%, and "good" if his shooting percentage is 50% or higher. Likewise, we classify his free throw shooting ability into three groups. He is a "poor" free throw shooter if his percentage of successful shots is 69% or smaller, "average" if he makes between 70 - 79% of his free throws, and "good" if he makes at least 80% of his free throws.

For each basketball player, two categorical variables are recorded — his ability to shoot two-point shots (poor, average, or good) and his ability to shoot free throws (poor, average or good). For space limitations, I can't list the data for the entire collection of 236 players, but part of the data (including some famous players) is listed in the below table.

Name	Two-point Shooting Ability	Free Throw Shooting Ability
Danny Ainge	average	good
Charles Barkley	average	average
Shawn Bradley	average	poor
Clyde Drexler	average	good
Patrick Ewing	good	average
Grant Hill	average	average
Michael Jordan	poor	good
Scott Kerr	good	average

Basketball shooting data for eight NBA players during the 1994-95 season.

This data can be organized by means of a two-way table. We construct a table with three rows and three columns where a row corresponds to a category of two-point shooting ability and a column corresponds to the level of free throw shooting ability. We tally the data by placing a mark in the table for each observation corresponding to the values of the two categorical values. For example, note the Danny Ainge is an average two-point shooter and good in shooting free throws. We place a mark in the square in the table corresponding to the "average" row and the "good" column. If we tally the eight observations in the table above, we obtain the result shown in the table below.

	Free Throw Shooting Ability		
Two-Point Shooting Ability	poor	average	good
poor			|
average	|	||	||
good		||	

Tallying data in a two-way table.

We use the computer to tally the observations for all 236 basketball players. If these tallies are converted into counts, we obtain the following two-way table..

	Free Throw Shooting Ability		
Two-Point Shooting Ability	poor	average	good
poor	16	38	18
average	25	40	41
good	30	22	6

Two-way table of free throw shooting and two-point shooting.

The numbers in the table give the counts of players having each possible combination of two-way shooting ability and free throw shooting ability. For example, we see that 38 players are poor two-point shooters and average free throw shooters and 6 players are poor in both characteristics.

In the previous section, we talked about obtaining a count table for a single categorical variable to understand the proportions of individuals in the different categories. One can obtain the counts for each variable in a two-way table by the computation of **marginal totals**. For each row, we add the counts in all of the columns; we place the resulting sum in a column named "TOTAL". In a similar fashion, for each column (including the new total column), we add the counts of all the rows and place the result in a new "TOTAL" column. We obtain the following table — we'll call it a two-way table with marginal totals added.

	Free Throw Shooting Ability			
Two-Point Shooting Ability	poor	average	good	TOTAL
poor	16	38	18	72
average	25	40	41	106
good	30	22	6	58
TOTAL	71	100	65	236

Two-way table with marginal row and column totals added.

The marginal row and column totals are helpful in understanding the distribution of each categorical variable. What is the proportion of poor free throw shooters in the NBA? We see that, of the 236 players, 71 are poor free throw shooters (shoot under 70%) for a proportion of 71/236 = 30%. Is it common for a NBA player to have a two-point shooting percentage of 50% or higher? This refers to the "good" category; we note that 58 are good two-point shooters for a proportion of 58/236 = 25%. Since 25% is a relatively small percentage, we would say that shooting at least 50% is not very common.

Although the two-way table can be used to learn about the marginal totals of each variable, it is most useful in learning about the **relationship** or the association between the two variables. In this example, we are interested in the relationship between a player's shooting ability in two-point

situations and his shooting ability from the free throw line. Is it true that an excellent shooter from the field will also be excellent from the free throw line?

To understand the association in a two-way table, we compute **conditional proportions** in the table. Suppose that we wish to compare the free throw shooting ability of the poor, average, and good two-point shooters. For the group of poor two-point shooters, represented by the first row of the table, we compute the proportion of poor, average, and good free throw shooters. There are 72 poor shooters in the first row; in this group the proportion of poor free throw shooters is 16/72 = .22 or 22%. The proportion of average free throw shooters in this group is 38/72 = .53 or 53% and the proportion of poor shooters from the foul line is 18/72 = .25 or 25%. These numbers are displayed in the first row of the table below. We call these numbers conditional proportions since they are computed conditional on the fact that only poor two-point shooters are considered.

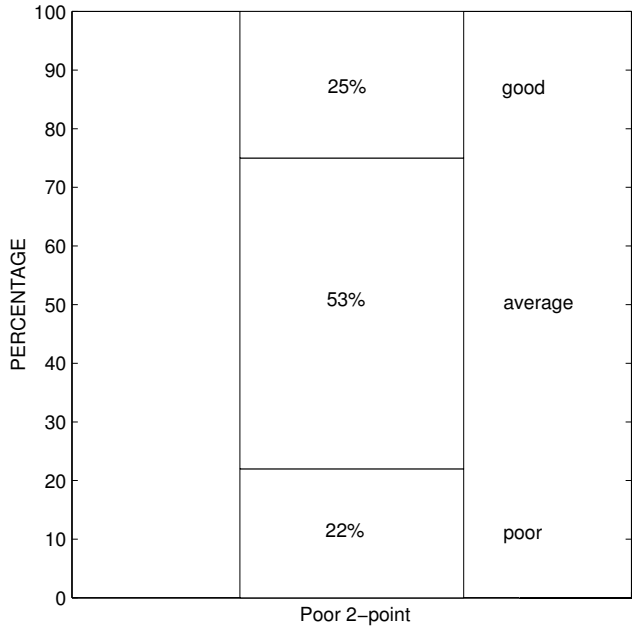
Similarly, we can compute the conditional proportions for the second and third rows of the table. For the "average" field shooters, we divide the counts 25, 40 and 41 by the total in the second row 106 to get the proportions of poor, average, and good free throw shooters in this average group. Finally, we compute the proportions of the different categories of free throw shooting among the "good" field shooters in the third row. When we are finished, we obtain the **row percent** form of the two-way table displayed in the following table. The additional "TOTAL" column is displayed to emphasize the fact that we are computing row percentages and so the percentages for each row should sum to 100. (Note that the numbers in the second row don't quite add up to 100. This is due to rounding errors — each percentage is rounded to the nearest whole number.)

	Free Throw Shooting Ability			
Two-Point Shooting Ability	poor	average	good	TOTAL
poor	22	53	25	100
average	23	37	39	100
good	52	38	10	100

Two-way table with row percentages as the entries.

We learn about the relationship between the two types of shooting ability by inspection of the row percentages. Of the poor field shooters, represented by the first row in the table, roughly half are average free shooters and the remaining are equally split between the poor and good categories. The average field shooters tend to be average or good free throw shooters (76% in these two categories). Looking at the third row, we see that over half (52%) of the good field shooters are poor free throw shooters; only 10% of this group are good in shooting free throws.

One can see the differences between the row percentages by the use of a plot called a **segmented bar chart**. We first illustrate the construction of a single plot of this type. Consider the percentages of poor, average, and good free throw shooters among the poor two-point shooters. These are given



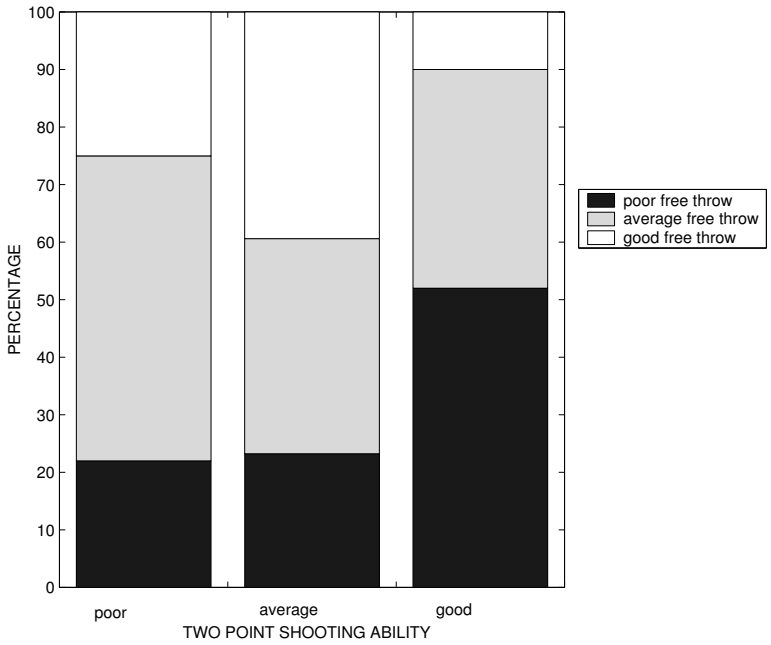
Segmented bar graph of free throw shooting ability for poor field shooters.

in the first row of the above table:

Free Throw Shooting Ability			
poor	average	good	TOTAL
22	53	25	100

We graph these row percentages (which sum to 100) by a single vertical bar which is divided into pieces such that the heights of the individual pieces correspond to the category percentages. The construction of a segmented bar graph is illustrated in the figure below. We start with a vertical bar with bottom and top equal to 0 and 100 percent, respectively. The first percentage is 22%, so we draw one horizontal line at 22%. The next percentage is 53%. We want the next piece of the bar to contain 53%, so we draw a second horizontal line at 22% + 53% = 75%. The remaining piece of the bar corresponds to the last percentage 25%. We complete the graph by shading the three pieces using different colors, and adding a legend to the graph which indicates which shaded region corresponds to which category.

We use one segmented bar graph for each row of the table. One bar graph is used to display the free throw shooting abilities of the poor field shooters, a second bar graph is used to display the free throw shooting abilities of the average field shooters, and a third graph is used for the poor field shooters. The set of segmented bar graphs is displayed in the figure below. This graph shows how



Segmented bar graphs of free throw shooting ability.

the distribution of poor, average, and good free throw shooting differs between the three types of two-point shooters.

The above discussion and graph focused on the row percentages in the table. One can also learn about the connection between the two variables by the computation of **column percentages**. We divide the players into poor, average and good free throw shooters, which correspond to the three columns of the table. For each column, we compute the percentage of poor, average and good two-point shooters. For example, there are 71 poor shooters from the foul line. Of these players, 16/71 = 22% are poor field shooters, 25/71 = 35% are average field shooters and 30/71 = 42% are good field shooters. We repeat this procedure for the second and third columns of the table. The result of this computation is the two-way table in a **column percent** form. The entries of the table, like the row percentage table, are conditional proportions. We are now conditioning on the columns of the table. The extra "TOTAL" row in the table is the sum of the percentages in each column. Since percentages are computed separately for each column, the sum of the percentages for each column (across rows) should be equal to 100.

	Free Throw Shooting Ability		
Two-Point Shooting Ability	poor	average	good
poor	23	38	28
average	35	40	63
good	42	22	9
TOTAL	100	100	100

Two-way table with column percentages as the entries.

From this column percentage table, we can compare the field shooting ability of the poor, average, and good free throw shooters. The poor free throw shooters are generally average or good in shooting from the field (the total percent in the average and good categories is 78%). The average shooters from the foul line are generally poor or average from the field (total percent of 78%), and the good shooters from the charity stripe are primarily average shooters from two-point land.

So what do we conclude? Are good field shooters (from two-point land) also good free throw shooters? Actually, the opposite appears to be true. From the table of row percentages, we see that the good field shooters are generally poor free throw shooters and the average field shooters are the best shooters from the free throw line. Why is this true? This association can be partly explained in terms of the different positions of the players. The players who play center or power forward in the NBA attempt most of their two-point shots close to the basket. It is easier to make close shots (most dunk shots are successful) and so most of these players are good two-point shooters. However, these tall players have a much harder time making shots when they are further away from the basket. In particular, they typically are poor shooters from the free throw line. The players that are traditionally thought to be good shooters are the shooting guards in the NBA. These players have a good shooting touch, but their two-point shooting percentages are relatively low since they take most of their shots far from the basket. These same players will be good shooters from the foul line since they have a good shooting touch.

PRELIMINARIES

1. Do you think that a student's political leaning has any bearing on whether he/she wants to see the penny retained or abolished?

2. Do you think that Americans tend to grow more liberal or more conservative in their political ideology as they grow older, or do you suspect that there is no relationship between age and political ideology?

3. Record for each student in the class whether he/she has read the novel Jurassic Park by Michael Crichton and whether he/she has seen the movie Jurassic Park directed by Stephen

Spielberg.

student	read book?	seen movie?	student	read book?	seen movie?	student	read book?	seen movie?
1			9			17		
2			10			18		
3			11			19		
4			12			20		
5			13			21		
6			14			22		
7			15			23		
8			16			24		

4. Is there a difference between the proportion of American men who are U.S. Senators and the proportion of U.S. Senators who are American men? If so, which proportion is greater?

5. Do you think it would be more common to see a toy advertisement in which a boy plays with a traditionally female toy or one in which a girl plays with a traditionally male toy?

6. Do you suspect that it is any more or less common for a physician to be a woman if she is in a certain age group? If so, in what age group would you expect to see the highest proportion of woman physicians?

IN-CLASS ACTIVITIES

Activity 9-1: Penny Thoughts (cont.)

Reconsider the data collected in Topic 1 on students' political inclinations and opinions about whether the penny should be retained or abolished.

(a) For each student, tally which of the six possible category pairs he/she belongs to:

category pair	tally
retain, liberal	
retain, moderate	
retain, conservative	
abolish, liberal	
abolish, moderate	
abolish, conservative	

(b) Represent the counts of students falling into each of these category pairs in a two-way table. For example, the number that you place in the upper left cell of the table should be the number of students who classified themselves as liberal and also believe that the penny should be retained.

	retain	abolish
liberal		
moderate		
conservative		

Activity 9-2: Age and Political Ideology

In a survey of adult Americans in 1986, people were asked to indicate their age and to categorize their political ideology. The results are summarized in the following table of counts:

	liberal	moderate	conservative	total
under 30	83	140	73	296
30-49	119	280	161	560
over 50	88	284	214	586
total	290	704	448	1442

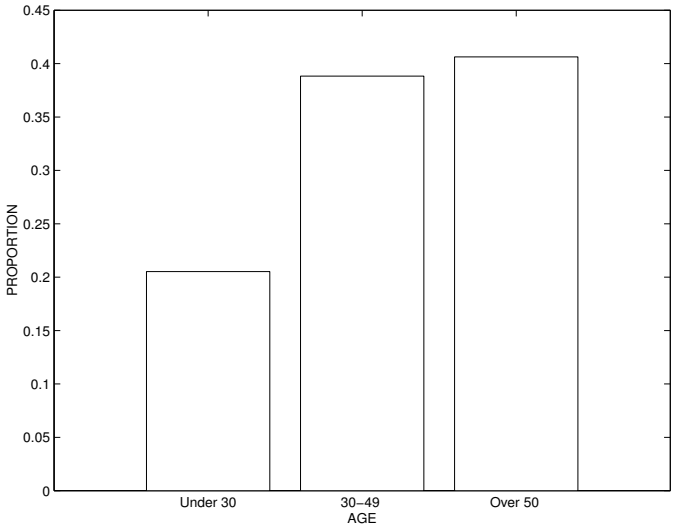
This table is called a two-way table since it classifies each person according to two variables. In particular, it is a 3 x 3 table; the first number represents the number of categories of the row variable (age), and the second number represents the number of categories of the column variable (political ideology). As an example of how to read the table, the upper-left entry indicates that of the 1442 respondents, 83 were under 30 years of age and regarded themselves as political liberals. Notice that the table also includes row and column totals.

(a) What proportion of the survey respondents were under age 30?

(b) What proportion of the survey respondents were between 30 and 50 years of age?

(c) What proportion of the survey respondents were over age 50?

You have calculated the marginal distribution of the age variable. When analyzing two-way tables, one typically starts by considering the marginal distribution of each of the variables by



Bar graph of age variable.

themselves before moving on to explore possible relationships between the two variables. You saw in Topic 1 that marginal distributions can be represented graphically in bar graphs. A bar graph illustrating the marginal distribution of the age variable appears above.

To study possible relationships between two categorical variables, one examines conditional distributions; i.e., distributions of one variable for given categories of the other variable.

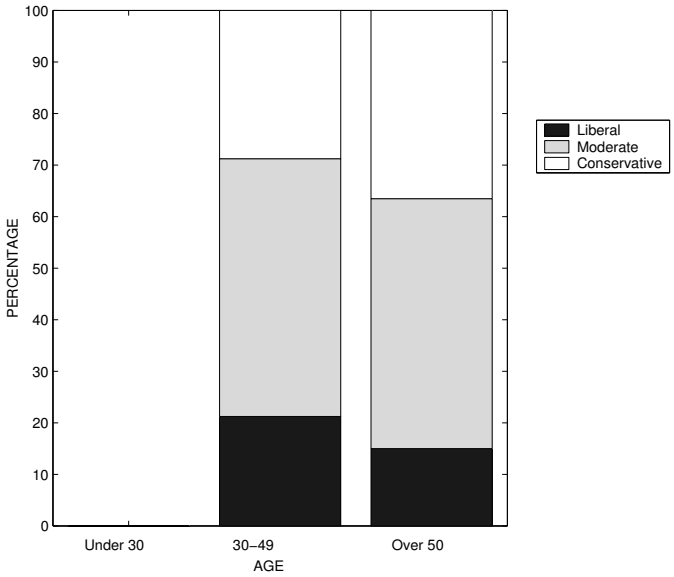
(d) Restrict your attention (for the moment) just to the respondents under 30 years of age. What proportion of the young respondents classify themselves as liberal?

(e) What proportion of the young respondents classify themselves as moderate?

(f) What proportion of the young respondents classify themselves as conservative?

One can proceed to calculate the conditional distributions of political ideology for middle-aged and older people. These turn out to be:

	middle-aged	older
proportion liberal	.2125	.1502
proportion moderate	.5000	.4846
proportion conservative	.2875	.3652



Segmented bar graph of political ideology data.

Conditional distributions can be represented visually with segmented bar graphs. The rectangles in a segmented bar graph all have a height of 100%, but they contain segments whose length corresponds to the conditional proportions.

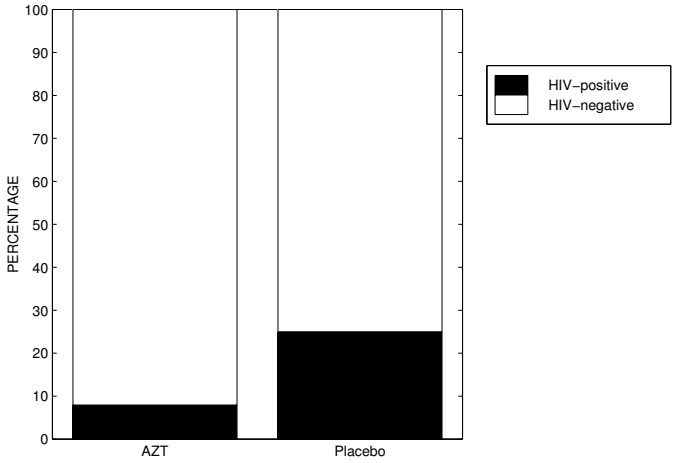
(g) Complete the segmented bar graph above by constructing the conditional distribution of political ideology among young people.

(h) Based on the calculations that you have performed and the display that you have created above, write a few sentences commenting on whether there seems to be any relationship between age and political ideology. In other words, does the distribution of political ideology seem to differ among the three age groups? If so, describe key features of the differences.

When dealing with conditional proportions, it is very important to keep straight which category is the one being conditioned on. For example, the proportion of American males who are U.S. Senators is very small, yet the proportion of U.S. Senators who are American males is very large.

Refer to the original table of counts to answer the following:

(i) What proportion of respondents under age 30 classified themselves as political moderates?



Segmented bar graph of HIV status.

(j) What proportion of the political moderates were under 30 years of age?

(k) What proportion of the 1442 respondents identified themselves as being both under 30 years of age and political moderates?

Activity 9-3: Pregnancy, AZT, and HIV

In an experiment reported in the March 7, 1994 issue of Newsweek, 164 pregnant, HIV-positive women were randomly assigned to receive the drug AZT during pregnancy and 160 such women were randomly assigned to a control group which received a placebo ("sugar" pill). The following segmented bar graph displays the conditional distributions of the child's HIV status (positive or negative) for mothers who received AZT and for those who received a placebo.

(a) Use the graph to estimate the proportion of AZT-receiving women who had HIV-positive babies and the proportion of placebo-receiving women who had HIV-positive babies.

The actual results of the experiment were that 13 of the mothers in the AZT group had babies who tested HIV-positive, compared to 40 HIV-positive babies in the placebo group.

(b) Use this information to calculate the proportions asked for in (a). Compare your calculations to your estimates based on the graph.

(c) The proportion of HIV-positive babies among placebo mothers is how many times greater than the proportion of HIV-positive babies among AZT mothers?

(d) Comment on whether the difference between the two groups appears to be important. What conclusion would you draw from the experiment?

Activity 9-4: Hypothetical Hospital Recovery Rates

The following two-way table classifies hypothetical hospital patients according to the hospital that treated them and whether they survived or died:

	survived	died	total
hospital A	800	200	1000
hospital B	900	100	1000

(a) Calculate the proportion of hospital A's patients who survived and the proportion of hospital B's patients who survived. Which hospital saved the higher percentage of its patients?

Suppose that when we further categorize each patient according to whether they were in good condition or poor condition prior to treatment we obtain the following two-way tables:
good condition:

	survived	died	total
hospital A	590	10	600
hospital B	870	30	900

poor condition:

	survived	died	total
hospital A	210	190	400
hospital B	30	70	100

(b) Convince yourself that when the "good" and "poor" condition patients are combined, the totals are indeed those given in the table above.

(c) Among those who were in good condition, compare the recovery rates for the two hospitals. Which hospital saved the greater percentage of its patients who had been in good condition?

(d) Among those who were in poor condition, compare the recovery rates for the two hospitals. Which hospital saved the greater percentage of its patients who had been in poor condition?

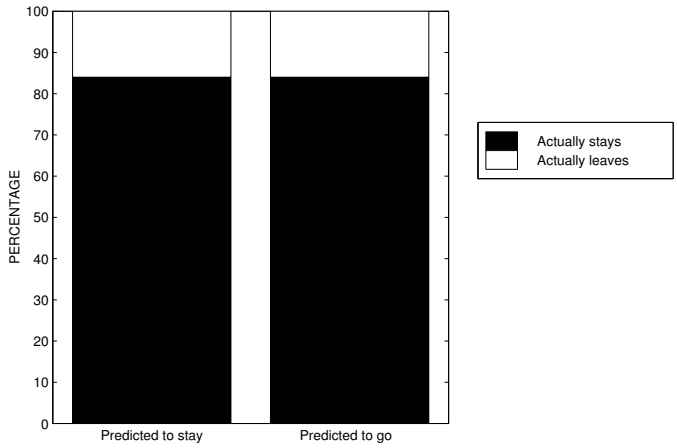
The phenomenon that you have just discovered is called Simpson's paradox (I think it's named for Lisa, but it could be for Bart or Homer), which refers to the fact that aggregate proportions can reverse the direction of the relationship seen in the individual pieces. In this case, hospital B has the higher recovery rate overall, yet hospital A has the higher recovery for each type of patient.

(e) Write a few sentences explaining (arguing from the data given) how it happens that hospital B has the higher recovery rate overall, yet hospital A has the higher recovery rate for each type of patient. (Hints: Do good or poor patients tend to survive more often? Does one type of hospital tend to treat most of one type of patient? Is there any connection here?)

(f) Which hospital would you rather go to if you were ill? Explain.

Activity 9-5: Hypothetical Employee Retention Predictions

Suppose that an organization is concerned about the number of its new employees who leave the company before they finish one year of work. In an effort to predict whether a new employee will leave or stay, they develop a standardized test and apply it to 100 new employees. After one year, they note what the test had predicted (stay or leave) and whether the employee actually stayed or left. They then compile the data into the following table:



Segmented bar graph of employee retention.

	actually stays	actually leaves	row total
predicted to stay	63	12	75
predicted to leave	21	4	25
column total	84	16	100

(a) Of those employees predicted to stay, what proportion actually left?

(b) Of those employees predicted to leave, what proportion actually left?

(c) Is an employee predicted to stay any less likely to leave than an employee predicted to leave?

The segmented bar graph displaying the conditional distribution of employee retention between those predicted to stay and those predicted to leave follows:

(d) Considering your answers to (a), (b), and (c) and this segmented bar graph, does the test provide any information at all about whether an employee will leave or stay?

Two categorical variables are said to be **independent** if the conditional distributions of one variable are identical for every category of the other variable. In this case the employee outcome is independent of the test prediction.

(e) Sketch what the segmented bar graph would look like if the test was perfect in its predictions.

(f) Sketch what the segmented bar graph would look like if the test was very useful but not quite perfect in its predictions.

HOMEWORK ACTIVITIES

Activity 9-6: Gender-Stereotypical Toy Advertising

To study whether toy advertisements tend to picture children with toys considered typical of their gender, researchers examined pictures of toys in a number of children's catalogs. For each picture, they recorded whether the child pictured was a boy or girl. (We will ignore ads in which boys and girls appeared together.) They also recorded whether the toy pictured was a traditional "male" toy (like a truck or a toy soldier) or a traditional "female" toy (like a doll or a kitchen set) or a "neutral" toy (like a puzzle or a toy phone). Their results are summarized in the following two-way table:

	traditional "boy" toy	traditional "girl" toy	neutral gender toy
boy child shown	59	2	36
girl child shown	15	24	47

(a) Calculate the marginal totals for the table.

(b) What proportion of the ads showing boy children depicted traditionally male toys? traditionally female toys? neutral toys?

(c) Calculate the conditional distribution of toy types for ads showing girl children.

(d) Construct a segmented bar graph to display these conditional distributions.

(e) Based on the segmented bar graph, comment on whether the researchers' data seem to suggest that toy advertisers do indeed tend to present pictures of children with toys stereotypical of their gender.

Activity 9-7: Gender-Stereotypical Toy Advertising (cont.)

Reconsider the data concerning toy advertising presented in Activity 9-6. Let us refer to ads which show boys with traditionally "female" toys and ads which show girls with traditionally "male" toys as "crossover" ads.

(a) What proportion of the ads under consideration are "crossover" ads?

(b) What proportion of the crossover ads depict girls with traditionally male toys?

(c) What proportion of the crossover ads depict boys with traditionally female toys?

(d) When toy advertisers do defy gender stereotypes, in which direction does their defiance tend?

Activity 9-8: Female Senators

The following table classifies each U.S. Senator of 1994 according to his/her gender and political party:

	Male	Female	row total
Republicans	42	2	44
Democrats	51	5	56
column total	93	7	100

(a) What proportion of the Senators are women?

(b) What proportion of the Senators are Democrats?

(c) Are most Democratic Senators women? Support your answer with an appropriate calculation.

(d) Are most women Senators Democrats? Support your answer with an appropriate calculation.

Activity 9-9: Jurassic Park Popularity

Consider the data collected above concerning whether or not students have read the novel and/or seen the movie Jurassic Park.

(a) Tabulate the data in a two-way table such as the following:

	seen movie	not seen movie
read book		
not read book		

(b) Calculate the conditional distributions of movie watching for those who have read the book and for those who have not read the book.

(c) Construct a segmented bar graph to display these conditional distributions.

(d) Write a brief paragraph summarizing whether the data seem to suggest an association between reading the book and seeing the movie.

Activity 9-10: Gender of Physicians (cont.)

The following data address the question of whether percentages of female physicians are changing with time. The table classifies physicians according to their gender and age group.

	under 35	35-44	45-54	55-64
male	93,287	153,921	110,790	80,288
female	40,431	44,336	18,026	7,224

(a) For each age group, calculate the proportion of its physicians who are women.

(b) Construct a segmented bar graph to represent these conditional distributions.

(c) Comment on whether your analysis reveals any connection between gender and age group. Suggest an explanation for your finding.

Activity 9-11: Children's Living Arrangements

The following table classifies the living arrangements of American children (under 18 years of age) in 1993 according to their race/Hispanic origin and which parent(s) they live with.

	both	just mom	just dad	neither
white	40,842,340	9,017,140	2,121,680	1,060,840
black	3,833,640	5,750,460	319,470	745,430
Hispanic	4,974,720	2,176,440	310,920	310,920

Analyze these data to address the issue of whether a relationship exists between race/Hispanic origin and parental living arrangements. Write a paragraph reporting your findings, supported by appropriate calculations and visual displays.

Activity 9-12: Civil War Generals

The following table categorizes each general who served in the Civil War according his background before the war and the Army (Union or Confederate) for which he served.

	military	law	business	politics	agriculture	other
Union	197	126	116	47	23	74
Confederate	127	129	55	24	42	48

Analyze these data to address the question of whether Union and Confederate generals tended to have different types of backgrounds. Perform calculations and construct displays to support your conclusions.

Activity 9-13: Berkeley Graduate Admissions

The University of California at Berkeley was charged with having discriminated against women in their graduate admissions process for the fall quarter of 1973. The table below identifies the number of acceptances and denials for both men and women applicants in each of the six largest graduate programs at the institution at that time:

	men accepted	men denied	women accepted	women denied
program A	511	314	89	19
program B	352	208	17	8
program C	120	205	202	391
program D	137	270	132	243
program E	53	138	95	298
program F	22	351	24	317
total				

(a) Start by ignoring the program distinction, collapsing the data into a two-way table of gender by admission status. To do this, find the total number of men accepted and denied and the total number of women accepted and denied. Construct a table such as the one below:

	admitted	denied	total
men			
women			
total			

(b) Consider for the moment just the men applicants. Of the men who applied to one of these programs, what proportion were admitted? Now consider the women applicants; what proportion of them were admitted? Do these proportions seem to support the claim that men were given preferential treatment in admissions decisions.

(c) To try to isolate the program or programs responsible for the mistreatment of women applicants, calculate the proportion of men and the proportion of women within each program who were admitted. Record your results in a table such as the one below.

	proportion of men admitted	proportion of women admitted
program A		
program B		
program C		
program D		
program E		
program F		

(d) Does it seem as if any program is responsible for the large discrepancy between men and women in the overall proportions admitted?

(e) Reason from the data given to explain how it happened that men had a much higher rate of admission overall even though women had higher rates in most programs and no program favored men very strongly.

Activity 9-14: Baldness and Heart Disease

To investigate a possible relationship between heart disease and baldness, researchers asked a sample of 663 male heart patients to classify their degree of baldness on a 5-point scale. They also asked a control group (not suffering from heart disease) of 772 males to do the same baldness assessment. The results are summarized in the table:

	none	little	some	much	extreme
heart disease	251	165	195	50	2
control	331	221	185	34	1

(a) What proportion of these men identified themselves as having little or no baldness?

(b) Of those who had heart disease, what proportion claimed to have some, much, or extreme baldness?

(c) Of those who declared themselves as having little or no baldness, what proportion were in the control group?

(d) Construct a segmented bar graph to compare the distributions of baldness ratings between subjects with heart disease and those from the control group.

(e) Summarize your findings about whether a relationship seems to exist between heart disease and baldness.

(f) Even if a strong relationship exists between heart disease and baldness, does that necessarily mean that heart disease is caused by baldness? Explain your answer.

Activity 9-15: Softball Batting Averages

Construct your own hypothetical data to illustrate Simpson's paradox in the following context: Show that it is possible for one softball player (Amy) to have a higher percentage of hits than another (Barb) in the first half of the season and in the second half of the season and yet to have a lower percentage of hits for the season as a whole. I'll get you started: suppose that Amy has 100 at-bats in the first half of the season and 400 in the second half, and suppose that Barb has 400 at-bats in the first half and 100 in the second half. You are to make up how many hits each player had in each half of the season, so that the above statement holds. (The proportion of hits is the number of hits divided by the number of at-bats.)

of season	first half of season	second half a whole	season as
Amy's hits			
Amy's at-bats	100	400	500
Amy's proportion of hits			
Barb's hits			
Barb's at-bats	400	100	500
Barb's proportion of hits			

Activity 9-16: Employee Dismissals

Suppose that you are asked to investigate the practices of a company that has recently been forced to dismiss many of its employees. The company figures indicate that, of the 1000 men and 1000 women who worked there a month ago, 300 of the men and 200 of the women were dismissed. The company employs two types of employees - professional and clerical, so you ask to see the breakdown for each type. Even though the company dismissed a higher percentage of men than women, you know that it is possible (Simpson's paradox) for the percentage of women dismissed within each employee type to exceed that of the men within each type. The company representative does not believe this, however, so you need to construct a hypothetical example to convince him of the possibility. Do so, by constructing and filling in tables such as the following.

overall (professional and clerical combined):

	dismissed	retained
men	300	700
women	200	800

Professional only:

	dismissed	retained
men		
women		

Clerical only:

	dismissed	retained
men		
women		

Activity 9-17: Politics and Ice Cream

Suppose that 500 college students are asked to identify their preferences in political affiliation (Democrat, Republican, or Independent) and in ice cream (chocolate, vanilla, or strawberry). Fill in the following table in such a way that the variables political affiliation and ice cream preference turn out to be completely independent. In other words, the conditional distribution of ice cream preference should be the same for each political affiliation, and the conditional distribution of political affiliation should be the same for each ice cream flavor.

	chocolate	vanilla	strawberry	row total
Democrat	108			240
Republican		72	27	
Independent		32		80
column total	225			500

Activity 9-18: Penny Thoughts (cont.)

Refer to the two-way table that you created in Activity 9-1 classifying students according to their opinion about the U.S. penny and their political leanings. Analyze this table to address the question of whether a relationship exists between these two variables. Write a paragraph summarizing your findings.

Activity 9-19: Variables of Personal Interest (cont.)

Think of a pair of categorical variables that you would be interested in exploring the relationship between. Describe the variables in as much detail as possible and indicate how you would present the data in a two-way table.

WRAP-UP

With this topic we have concluded our investigation of relationships between variables. This topic has differed from earlier ones in that it has dealt exclusively with categorical variables. The most important technique that this topic has covered has involved interpreting information presented in two-way tables. You have encountered the ideas of marginal distributions and conditional distributions, and you have learned to draw bar graphs and segmented bar graphs to display these distributions. Finally, you have discovered and explained the phenomenon known as Simpson's Paradox, which raises interesting issues with regard to analyzing two-way tables.

These first two units have addressed exploratory analyses of data. In the next unit you will begin to study background ideas related to the general issue of drawing inferences from data. Specifically, you will take a figurative step backward in the data analysis process by considering issues related to the question of how to collect meaningful data in the first place. You will also begin to study ideas of randomness that lay the foundation for procedures of statistical inference.

Topic 10: Random Sampling

Introduction

To this point in the course, you have been analyzing data using exploratory methods. With this topic you will take a conceptual step backward by considering questions of how to collect data in the first place. The practice of statistics begins not *after* the data have been collected but *prior* to their collection. You will find that utilizing proper data collection strategies is critical to being able to draw meaningful conclusions from the data once it has been collected and analyzed. You will also discover that randomness plays an important role in data collection.

Preliminaries

1. What percentage of adult Americans do you think believe that Elvis Presley is still alive?

2. Do you believe that Elvis Presley is still alive?

3. Who won the 1936 U.S. Presidential election? Who lost that election?

4. Suppose you want to learn about the proportion of students at your school that own their own cars. Would it be reasonable to take a sample of students from the parking lot of your school? Why or why not?

5. U. S. Today Weekend magazine is interested in learning about the public support for school that is on a year-round schedule without any summer holiday. Out of 2298 visitors to the U. S. Today Weekend website, 51% agreed that students should attend school year-round. Would you feel confident that over half of Americans are in support of school year-round? Why or why not?

IN-CLASS ACTIVITIES

Activity 10-1: Elvis Presley and Alf Landon

On the twelfth anniversary of the (alleged) death of Elvis Presley, a Dallas record company sponsored a national call-in survey. Listeners of over 1000 radio stations were asked to call a 1-900 number (at a charge of $2.50) to voice an opinion concerning whether or not Elvis was really dead. It turned out that 56% of the callers felt that Elvis was still alive.

(a) Do you think that 56% is an accurate reflection of beliefs of all American adults on this issue? If not, identify some of the flaws in the sampling method.

A famous bad poll

In 1936, *Literary Digest* magazine conducted the most extensive (to that date) public opinion poll in history. They mailed out questionnaires to over 10 million people whose names and addresses they had obtained from phone books and vehicle registration lists. More than 2.4 million people responded, with 57% indicating that they would vote for Republican Alf Landon in the upcoming Presidential election. (Incumbent Democrat Franklin Roosevelt won the election, carrying 63% of the popular vote.)

(b) Offer an explanation as to how the *Literary Digest*'s prediction could have been so much in error. In particular, comment on why its sampling method made it vulnerable to overestimating support for the Republican candidate.

Population and sample

These examples have (at least) two things in common. First, the goal in each case was to learn something about a very large group of people (all American adults, all American registered voters) by studying a portion of that group. That is the essential idea of sampling: to learn about the whole by studying a part.

Two extremely important terms related to the idea of sampling that we will use throughout the course are **population** and **sample**. In the technical sense with which we use these terms, population means the entire group of people or objects about which information is desired, while sample refers to a (typically small) part of the population that is actually examined to gain information about the population.

(c) Identify the population of interest and the sample actually used to study that population in the Elvis and *Literary Digest* examples.

Biased samples

Another thing that the two examples have in common is that both illustrate a very poor job of sampling; i.e., of selecting the sample from the population. In neither case could one accurately infer anything about the population of interest from the sample results. This is because the sampling methods used were **biased**. A sampling procedure is said to be biased if it tends systematically to overrepresent certain segments of the population and systematically to underrepresent others.

These examples also indicate some common problems that produce biased samples. Both are **convenience samples** to some extent since they both reached those people most readily accessible. Another problem is **voluntary response**, which refers to samples collected in such a way that members of the population decide for themselves whether or not to participate in the sample. The related problem of nonresponse can arise even if an unbiased sample of the population is contacted.

Choosing a simple random sample

In order to avoid biased samples, it seems fair and reasonable to give each and every member of the population the same chance of being selected for the sample. In other words, the sample should be selected so that every possible sample has an equal chance of being the sample ultimately selected. Such a sampling design is called **simple random sampling.**

While the principle of simple random sampling is probably clear, it is by no means simple to implement. One thought is to rely on physical mixing: write the names on pieces of paper, throw them into a hat, mix them thoroughly, and draw them out one at a time until the sample is full. Unfortunately, this method is fraught with the potential for hidden biases, such as different sizes of pieces of paper and insufficient mixing.

A better alternative for selecting a simple random sample (hereafter to be abbreviated SRS) is to use a computer-generated table of random digits. Such a table is constructed so that each position is equally likely to be occupied by any one of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 and so that the occupant of any one position has no impact on the occupant of any other position. A table of random digits is presented here:

Row	Random digits									
1	17139	27838	19139	82031	46143	93922	32001	05378	42457	94248
2	20875	29387	32682	86235	35805	66529	00886	25875	40156	92636
3	34568	95648	79767	16307	71133	15714	44142	44293	19195	30569
4	11169	41277	01417	34656	80207	33362	71878	31767	04056	52582
5	15529	30766	70264	86253	07179	24757	57502	51033	16551	66731
6	33241	87844	41420	10084	55529	68560	50069	50652	76104	42086
7	83594	48720	96632	39724	50318	91370	68016	06222	26806	86726
8	01727	52832	80950	27135	14110	92292	17049	60257	01638	04460
9	86595	21694	79570	74409	95087	75424	57042	27349	16229	06930
10	65723	85441	37191	75134	12845	67868	51500	97761	35448	56096
11	82322	37910	35485	19640	07689	31027	40657	14875	07695	92569
12	06062	40703	69318	95070	01541	52249	56515	59058	34509	35791
13	54400	22150	56558	75286	07303	40560	57856	22009	67712	19435
14	80649	90250	62962	66253	93288	01838	68388	55481	00336	19271
15	70749	78066	09117	62350	58972	80778	46458	83677	16125	89106
16	50395	30219	03068	54030	49295	48985	01624	72881	88310	18172
17	48993	89450	04987	02781	37935	76222	93595	20942	90911	57643
18	77447	34009	20728	88785	81212	08214	93926	66687	58252	18674
19	24862	18501	22362	37319	33201	88294	55814	67443	77285	36229
20	87445	26886	66782	89931	29751	08485	49910	83844	56013	26596

Activity 10-2: Sampling Students

To illustrate the process of taking a simple random sample, we'll consider the following problem of learning about the opinions of college students. Suppose that 100 students live in a particular college dormitory, called Trax. Suppose that two residents of Trax want the university to purchase an ice machine for the use of all the dorm residents. The college administration is willing to buy this ice machine only if they are convinced that a majority (over 50%) of the Trax residents are in favor of having an ice machine.

In this situation, the two college students wish to learn about the opinions of the Trax students regarding the installation of the ice machine. Here the **population** of interest is the group of 100 students who live in this particular dorm. We'll assume that each student is either in favor (which we represent by the symbol $\smile$) or not in favor (represented by the symbol $\frown$) of the addition of the ice machine. We represent the opinions of the population by the following diagram. There are

ten floors of the dorm, numbered 0, 1, ..., 9, and students live in single rooms on each floor that are also numbered 0, 1, ..., 9. (So the room number 24 refers to the fifth room on the second floor.) The opinions of all of the students are represented by ⌣'s and ⌢'s. You can check that there are exactly 60 students in favor of the machine and 40 students against. The proportion of the population that is favor of the machine is therefore 60/100 = .6.

	ROOM									
FLOOR	0	1	2	3	4	5	6	7	8	9
9	⌣	⌣	⌢	⌣	⌢	⌢	⌣	⌣	⌣	⌢
8	⌢	⌣	⌢	⌣	⌣	⌣	⌢	⌣	⌣	⌢
7	⌣	⌣	⌣	⌢	⌣	⌢	⌣	⌣	⌢	⌣
6	⌣	⌣	⌢	⌣	⌣	⌢	⌣	⌣	⌣	⌣
5	⌣	⌣	⌢	⌣	⌢	⌣	⌣	⌣	⌣	⌣
4	⌣	⌣	⌢	⌣	⌣	⌣	⌢	⌢	⌣	⌣
3	⌢	⌢	⌣	⌣	⌢	⌣	⌣	⌣	⌢	⌣
2	⌢	⌣	⌣	⌣	⌣	⌣	⌢	⌢	⌢	⌢
1	⌣	⌢	⌢	⌢	⌢	⌢	⌢	⌢	⌣	⌢
0	⌣	⌣	⌢	⌣	⌣	⌢	⌢	⌢	⌣	⌢

Now it seems that it is easy for the students to find out if a majority of the dorm is in favor of the ice machine. They can ask each of the 100 students their opinion about the machine and then know exactly how many of the students are in favor. However, this is an impossible task. It can be difficult to find students in their rooms and it would take an unreasonable amount of time and effort to survey all 100 students in the dorm. This is typical of the usual statistical inference problem. Although one is interested in the composition of a population, it is usually impossible to survey every single member of the population.

However, the students have the time to survey students from a simple random sample taken from the dorm population. Suppose they want to take a SRS of size 15 from the population of 100 students. We can use the table of random digits to select this sample as follows:

- **Label each population member** We first assign a two-digit label to each student in the dorm. Since each student lives in a separate room, we'll use the room numbers 00, 01, ..., 99 as the labels.

- **Select numbers from the random digit table** We start anywhere in the random digit table and systematically read the digits going across a single row or a single column. Suppose we start in the fifth row of the table — the following digits are read:

Row 5: 15529 30766 70264 86253 07179 24757 5750

- **Find the sample corresponding to the digits selected** Since the population members are labelled using two digits, we break this sequence of digits into groups of two:

  | 15 | 52 | 93 | 07 | 66 | 70 | 26 | 48 | 62 | 53 | 07 | 17 | 92 | 47 | 57 | 57 | 50 |

  These numbers correspond to our sample. The first student sampled will be 15, which is the student living in room 15, the next student sampled lives in room 52, the third student sampled lives in room 93, and so on. When we select our sample, we should be careful not to select the same student twice.

In the diagram below, we have placed a box around the students that are in our sample of 15.

	ROOM									
FLOOR	0	1	2	3	4	5	6	7	8	9
9	⌣	⌣	⌢	⌣	⌢	⌢	⌣	⌣	⌣	⌢
8	⌢	⌣	⌢	⌣	⌣	⌣	⌢	⌣	⌣	⌢
7	⌣	⌣	⌣	⌢	⌣	⌢	⌣	⌣	⌢	⌣
6	⌣	⌣	⌢	⌣	⌣	⌢	⌣	⌣	⌣	⌣
5	⌣	⌣	⌢	⌣	⌢	⌣	⌣	⌣	⌣	⌣
4	⌣	⌣	⌢	⌣	⌣	⌣	⌢	⌢	⌣	⌣
3	⌢	⌢	⌣	⌣	⌢	⌣	⌣	⌣	⌢	⌣
2	⌢	⌣	⌣	⌣	⌣	⌣	⌢	⌢	⌢	⌢
1	⌣	⌢	⌢	⌢	⌢	⌢	⌢	⌢	⌣	⌢
0	⌣	⌣	⌢	⌣	⌣	⌢	⌢	⌢	⌣	⌢

In the table below, we have listed the members of the sample and their opinions about the ice machine.

Student	15	52	93	07	66	70	26	48	62	53	17	92	47	57	50
Response	⌢	⌢	⌣	⌢	⌣	⌣	⌢	⌣	⌢	⌣	⌢	⌢	⌢	⌣	⌣

In our sample of 15 students, we count that 7 are in favor of the machine and 8 are against. The proportion of students in the sample in favor is $7/15 = .47$. By the way, note that, although the proportion of students in favor *in the whole dorm* is .6, the corresponding proportion of students *in the sample* is only .47.

(a) Using the procedure described above, take five new SRS's of size 15 from the population of dorm students. Take your five samples respectively from row numbers 4, 10, 12, 16, 20 of the random digit table. In the table below, write down the responses (⌣or ⌢) for each student in the sample. Also, find the number in favor, and the proportion in favor — also put these numbers in the table.

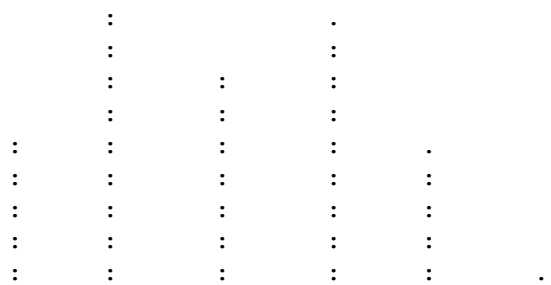
		Student responses															# in	Prop.
Sample	Row	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	favor	in favor
1	4																	
2	10																	
3	12																	
4	16																	
5	20																	

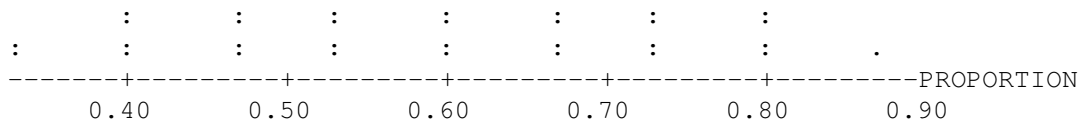
(b) Compare the results of your five samples. Did the proportion of students in favor change from sample to sample? Find the mean and the range of the proportions from samples 1, ..., 5.

Pattern of proportions from many SRS's

In part (b) you were asked to take five SRS's from the population of Trax residents. What if we took 100 SRS's? This would be too much work for you, but this is easily done by computer. The computer was asked to take a SRS of size 15 and find the proportion of students in agreement. This was repeated for a total of 100 times and the following sample proportions were obtained. A dotplot of these 100 proportions is also presented.

0.40	0.73	0.53	0.73	0.60	0.33	0.60	0.60	0.47	0.60
0.53	0.47	0.53	0.53	0.73	0.53	0.53	0.47	0.60	0.60
0.60	0.73	0.67	0.53	0.67	0.53	0.40	0.60	0.67	0.67
0.53	0.73	0.53	0.47	0.67	0.60	0.73	0.53	0.80	0.67
0.53	0.73	0.53	0.53	0.73	0.73	0.67	0.73	0.53	0.53
0.67	0.67	0.67	0.73	0.60	0.73	0.67	0.60	0.60	0.33
0.47	0.47	0.53	0.60	0.60	0.47	0.67	0.80	0.60	0.67
0.47	0.87	0.53	0.67	0.60	0.53	0.53	0.47	0.80	0.67
0.73	0.67	0.53	0.67	0.53	0.40	0.67	0.47	0.40	0.47
0.47	0.67	0.80	0.67	0.67	0.60	0.47	0.47	0.80	0.60





(c) Look at the dotplot of proportions from the 100 samples. What is the shape of this distribution? What is an average value of this distribution and what is its spread?

Property of a SRS

We have demonstrated an important property of a SRS in the above activity. Remember that the actual proportion of the student population in favor of the ice machine was .6. When you took repeated SRS's from the population, you observed considerable variation in the values of the sample proportions. But you found that the average sample proportion value (over 100 SRS's) was .6, which is equal to the population proportion which we are interested in. This observation is generally true. We are unsure if one sample proportion will be close to the population proportion. But if we take many simple random samples, the sample proportion values will, on average, be close to the population proportion.

Activity 10-3: Sampling Students (continued).

In Activity 10-2, we looked at the distribution of proportions of random samples. This activity investigates the consequences of choosing a *non-random sample*. Again we consider the problem in Activity 10-2 of sampling students to learn about the proportion of the population who are in favor of the installation of the ice machine. The two students doing the study have learned in their introductory statistics class that it is best to take a simple random sample. However, the process of taking a SRS seems to take a lot of work. They think that there must be an easier way of taking a "good" sample. These two students live on the ground floor of the dormitory and they devise the following easier method of sampling. They will randomly select 15 residents from only the floors 0, 1, 2 as their sample. This is convenient since they know the residents on the first three floors of Trax pretty well and can more easily locate these students to ask the survey question.

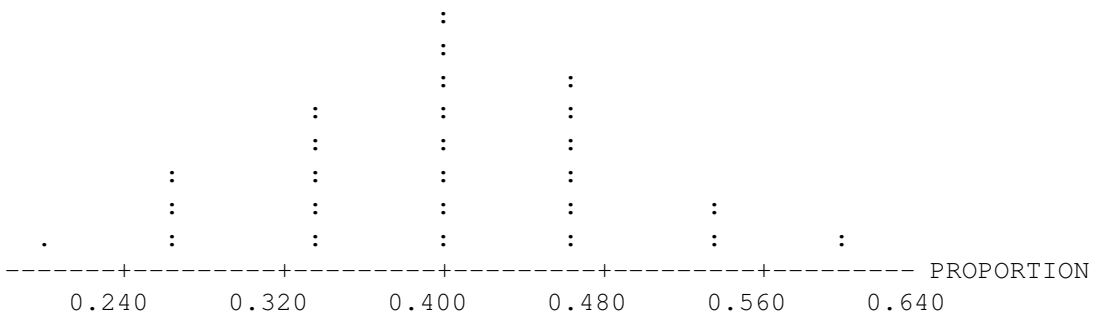
This is certainly not a simple random sample, since residents from the top seven floors of the dorm can't be selected. We'll see below what is wrong with this type of convenience sample.

(a) Use the random digit table to take a random sample from only floors 0, 1, 2 of the dormitory.

Put the responses from your sample in the table below. (Hint: In this case, you are selecting students from rooms 00, 01, ..., 29. When you use the random digit table, most of the numbers you select will not be used. It is more efficient to assign three two-digit labels to each student before you select the random digits. For example, labels 00, 30, 60 can correspond to student 00, labels 01, 31, 61 to student 01, and so on.)

Student responses															# in	Prop.
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	favor	in favor

To see what happens when one takes samples repeatedly from the first three floors, the computer was used to take 100 samples of this type. A dotplot of the sample proportions from these 100 samples is shown below.



(b) Looking at the dotplot of sampling proportions, comment on (1) the shape of the distribution, (2) an average proportion value, and (3) the variation of the proportions.

(c) Remember that the students want to learn about the proportion of all students in the dorm that are in favor of the ice machine. What is this proportion? (Look back at Activity 10-2.)

(d) Is the average sample proportion value close to the population proportion value? If not, how is it different?

(e) This activity demonstrates that the opinions of the residents of the first three floors are different from the opinions of the entire dorm. Suppose that the ice machine will be installed on the bottom floor. Can you provide a plausible explanation why residents of the bottom floors may have different opinions about the ice machine than residents of the top floors?

Activity 10-4: Sampling U.S. Senators

Consider the members of the U.S. Senate of 1994 as the population of interest. These Senators are listed below, along with their sex, party, state, and years of service (as of 1994) in the Senate. Notice that each has been assigned a two-digit identification number.

ID#	name	sex	party	state	years
01	Akaka	m	Dem	Hawaii	4
02	Baucus	m	Dem	Montana	16
03	Bennett	m	Rep	Utah	1
04	Biden	m	Dem	Delaware	21
05	Bingaman	m	Dem	New Mexico	11
06	Bond	m	Rep	Missouri	7
07	Boren	m	Dem	Oklahoma	15
08	Boxer	f	Dem	California	1
09	Bradley	m	Dem	New Jersey	15
10	Breaux	m	Dem	Louisiana	7
11	Brown	m	Rep	Colorado	3
12	Bryan	m	Dem	Nevada	5
13	Bumpers	m	Dem	Arkansas	19
14	Burns	m	Rep	Montana	5
15	Byrd	m	Dem	West Virginia	35
16	Campbell	m	Dem	Colorado	1
17	Chafee	m	Rep	Rhode Island	18
18	Coats	m	Rep	Indiana	5
19	Cochran	m	Rep	Mississippi	16
20	Cohen	m	Rep	Maine	15
21	Conrad	m	Dem	North Dakota	7
22	Coverdell	m	Rep	Georgia	1
23	Craig	m	Rep	Idaho	3
24	D'Amato	m	Rep	New York	13
25	Danforth	m	Rep	Missouri	18
26	Daschlee	m	Dem	South Dakota	7
27	DeConcini	m	Dem	Arizona	17
28	Dodd	m	Dem	Connecticut	13
29	Dole	m	Rep	Kansas	25
30	Domenici	m	Rep	New Mexico	21
31	Dorgan	m	Dem	North Dakota	1
32	Durenberger	m	Rep	Minnesota	16
33	Exon	m	Dem	Nebraska	15
34	Faircloth	m	Rep	North Carolina	1
35	Feingold	m	Dem	Wisconsin	1

ID#	name	sex	party	state	years
36	Feinstein	f	Dem	California	1
37	Ford	m	Dem	Kentucky	20
38	Glenn	m	Dem	Ohio	20
39	Gorton	m	Rep	Washington	13
40	Graham	m	Dem	Florida	7
41	Gramm	m	Rep	Texas	9
42	Grassley	m	Rep	Iowa	13
43	Gregg	m	Rep	New Hampshire	1
44	Harkin	m	Dem	Iowa	9
45	Hatch	m	Rep	Utah	17
46	Hatfield	m	Rep	Oregon	27
47	Heflin	m	Dem	Alabama	15
48	Helms	m	Rep	North Carolina	21
49	Hollings	m	Dem	South Carolina	28
50	Hutchison	f	Rep	Texas	1
51	Inouye	m	Dem	Hawaii	31
52	Jeffords	m	Rep	Vermont	5
53	Johnston	m	Dem	Louisiana	22
54	Kassebaum	f	Rep	Kansas	16
55	Kempthorne	m	Rep	Idaho	1
56	Kennedy	m	Dem	Massachusetts	32
57	Kerry, J	m	Dem	Massachusetts	9
58	Kerry, R	m	Dem	Nebraska	5
59	Kohl	m	Dem	Wisconsin	5
60	Lautenberg	m	Dem	New Jersey	12
61	Leahy	m	Dem	Vermont	19
62	Levin	m	Dem	Michigan	15
63	Lieberman	m	Dem	Connecticut	5
64	Lott	m	Rep	Mississippi	5
65	Lugar	m	Rep	Indiana	17
66	Mack	m	Rep	Florida	5
67	Matthews	m	Dem	Tennessee	1
68	McCain	m	Rep	Arizona	7
69	McConnell	m	Rep	Kentucky	9
70	Metzenbaum	m	Dem	Ohio	18

ID#	name	sex	party	state	years
71	Mikulski	f	Dem	Maryland	7
72	Mitchell	m	Dem	Maine	14
73	Moseley-Braun	f	Dem	Illinois	1
74	Moynihan	m	Dem	New York	17
75	Murkowski	m	Rep	Alaska	13
76	Murray	f	Dem	Washington	1
77	Nickles	m	Rep	Oklahoma	13
78	Nunn	m	Dem	Georgia	22
79	Packwood	m	Rep	Oregon	25
80	Pell	m	Dem	Rhode Island	33
81	Pressler	m	Rep	South Dakota	15
82	Pryor	m	Dem	Arkansas	15
83	Reid	m	Dem	Nevada	7
84	Riegle	m	Dem	Michigan	18
85	Robb	m	Dem	Virginia	5
86	Rockefeller	m	Dem	West Virginia	9
87	Roth	m	Rep	Delaware	23
88	Sarbanes	m	Dem	Maryland	17
89	Sasser	m	Dem	Tennessee	17
90	Shelby	m	Dem	Alabama	7
91	Simon	m	Dem	Illinois	9
92	Simpson	m	Rep	Wyoming	15
93	Smith	m	Rep	New Hampshire	3
94	Specter	m	Rep	Pennsylvania	13
95	Stevens	m	Rep	Alaska	26
96	Thurmond	m	Rep	South Carolina	38
97	Wallop	m	Rep	Wyoming	17
98	Warner	m	Rep	Virginia	15
99	Wellstone	m	Dem	Minnesota	16
00	Wofford	m	Dem	Pennsylvania	3

Some characteristics of this population of senators are:

Gender	Count
males	93
females	7

Party	Count
Democrats	56
Republicans	44

Years of service

mean	std. dev.	min	QL	median	QU	max
12.54	8.72	1	5	13	17	38

(a) For each of the four variables that are recorded about the Senators, identify whether it is a categorical or measurement variable. If it is categorical, specify whether it is also binary.

- gender:
- party:
- state:
- years of service:

(b) Use the table of random digits to select a simple random sample of 10 U.S. Senators. Do this by entering the table at any point and reading off the first ten two-digit numbers that you happen across. (If you happen to get repeats, keep going until you have ten different two-digit numbers.) Record the names and other information of the Senators corresponding to those ID numbers:

	ID	senator	gender	party	state	years
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

(c) Enter in the table below the numbers of men, women, Democrats, and Republicans in your sample.

	Count
males	
females	

Party	Count
Democrats	
Republicans	

(d) What is your home state? Is there a senator from your home state in the sample?

(e) Create (by hand) a dotplot of the years of service in your sample. Then use the computer to calculate summary statistics for the distribution of years of service and record your findings below.

	mean	std. dev.	min	QL	median	QU	max
Years of service							

(f) Does the proportional breakdown of men/women in your sample equal that in the entire population of Senators? How about the breakdown of Democrats/Republicans? Does the mean years of service in your sample equal that of the population?

(g) If your answer to any part of question (f) is "no," does that mean that your sampling method is biased like the *Literary Digest*'s was? Explain.

(h) Describe specifically how you would have to alter this procedure to use the table of random digits to select an SRS of size 10 from the population of the 435 members of the U.S. House of Representatives.

Models, parameters and statistics

The ideas of population and sample are crucial in statistics. We will use other words to describe characteristics of populations and sample. A **model** is any particular description of a population. One simple description of a population is a number called a **parameter**. A **statistic**, on the other hand, is a number that is associated with a sample. (To help you keep this straight, notice that population and parameter start with the same letter, as do sample and statistic.)

We will be very careful to use different symbols to denote parameters and statistics. For example, we use the following symbols to denote proportions, means, and standard deviations:

	(population) parameter	(sample) statistic
proportion:	p	$\hat{p}$
mean:	M	$\bar{x}$
standard deviation:	h	s

(Note: $\hat{p}$ is read "p-hat" and $\bar{x}$ is read "x-bar".)

In the student survey example of Activities 10-1 and 10-2, the proportion of all the students in the dorm in favor of the ice machine is the parameter p. The proportion of students in favor from one particular sample of 15 is the statistic $\hat{p}$. This example was unusual since we knew that the parameter value $p = .6$. In the typical statistical inference problem, the parameter (characteristic of the population) will be unknown, the statistic (characteristic of the sample) will be known, and the goal is to learn about the parameter from the statistic.

(i) Identify each of the following as a parameter or a statistic, indicate the symbol used to denote it, and specify its value in this context.

- the proportion of men in the entire 1994 Senate
- the proportion of Democrats among your 10 Senators
- the mean years of service among your 10 Senators
- the standard deviation of the years of service in the 1994 Senate

Activity 10-5: Sampling U.S. Senators (cont.)

(a) In Activity 10-4, each member of the class took a SRS of ten Senators. Here we combine the results of your SRS with the results of nine of your classmates. In the table below, record for each SRS the sample proportion of Democrats in the sample and the sample mean of the years of service of those Senators in your sample.

sample	1	2	3	4	5	6	7	8	9	10
prop. Dem.										
mean years.										

(b) Did you get the same sample proportion of Democrats in each of your ten samples? Did you get the same sample mean years of service in each of your ten samples?

This simple question illustrates a very important statistical property known as sampling variability: the value of sample quantities vary from sample to sample.

(c) Create (by hand) a dotplot of your sample proportions of Democrats. (The value marked with the arrow is .56, the population proportion of Democrats.)

(d) From the dotplot, make a good guess at the mean of the sample proportions.

These activities have illustrated that the value of a sample statistic (the sample proportion of Democrats, in this case) varies from sample to sample if one repeatedly takes simple random samples from the population of interest.

Two caveats are in order, however. First, one still gets the occasional "unlucky" sample whose results are not close to the population. Second, the sample may not be close to the population when a nonrandom sample is taken. Remember that the *Literary Digest* had a huge sample of 2.4 million people, yet their results were not close to the truth about the population.

HOMEWORK ACTIVITIES

Activity 10-6: Sampling Gears

Suppose that a company receives an allotment of gears that are shipped in boxes of ten. To insure the quality of each shipment, four gears are randomly selected (without replacement) from each box and inspected. Suppose that one box contains 3 defective gears and 7 good ones which is pictured below (the symbol $\odot$ represents a good gear and $\otimes$ represents a defective gear).

Gear number	0	1	2	3	4	5	6	7	8	9
State	$\odot$	$\otimes$	$\otimes$	$\odot$	$\odot$	$\odot$	$\odot$	$\odot$	$\odot$	$\otimes$

(a) Suppose that the box of 10 gears represents the population of interest. What symbol represents the proportion of defective gears in this population?

(b) Using the random digit table, select a SRS of size four from the box. (The gear numbers given in the table can represent the labels. A SRS can be chosen by looking at single digits from any row of the table.) Write down the numbers and the states of the gears that you chose in the table below.

Gear number				
State				

(c) Compute the proportion of defective gears in your sample. What is the symbol for this proportion?

(d) Suppose that a second SRS is taken from the box. Do you expect that the proportion of defectives in this new sample will be the same as the proportion you computed in part (c)? Why or why not?

Activity 10-7: Emotional Support

In the mid-1980's Shere Hite undertook a study of women's attitudes toward relationships, love, and sex by distributing 100,000 questionnaires through women's groups. Of the 4500 women who returned the questionnaires were returned, 96% said that they give more emotional support than they receive from their husbands or boyfriends.

(a) Comment on whether Hite's sampling method is likely to be biased in a particular direction. Specifically, do you think the 96% figure overestimates or underestimates the truth about the population of all American women?

An ABC News/Washington Post poll surveyed a random sample of 767 women, finding that 44% claimed to give more emotional support than they receive.

(b) Which poll surveyed the larger number of women?

(c) Which poll's results do you think are more representative of the truth about the population of all American women? Explain.

Activity 10-8: Alternative Medicine

In a spring 1994 issue, Self magazine reported that 84% of its readers who responded to a mail-in poll indicated that they had used a form of alternative medicine (e.g., acupuncture, homeopathy, herbal remedies). Comment on whether this sample result is representative of the truth concerning the population of all adult Americans. Do you suspect that the sampling method has biased the result? If so, is the sample result likely to overestimate or underestimate the proportion of all adult Americans who have used alternative medicine? Explain your answers.

Activity 10-9: Courtroom Cameras

An article appearing in the October 4, 1994 issue of The Harrisburg Evening-News reported that Judge Lance Ito (who is trying the O.J. Simpson murder case) had received 812 letters from around the country on the subject of whether to ban cameras from the courtroom. Of these 812 letters, 800 expressed the opinion that cameras should be banned.

(a) What proportion of this sample supports a ban on cameras in the courtroom? Is this number a parameter or a statistic?

(b) Do you think that this sample represents well the population of all American adults? Comment on the sampling method.

Activity 10-10: Parameters vs. Statistics

(a) Suppose that you are interested in the population of all students at this college and that you are using the students enrolled in this course as a (non-random) sample. Identify each of the following as a parameter or a statistic, and indicate the symbol used to denote it:

- the proportion of the college's students who participate in inter-collegiate athletics
- the proportion of students in this course who participate in inter-collegiate athletics
- the mean college grade point average for the students in this course
- the standard deviation of the college grade point averages for all students at the college

(b) Suppose that you are interested in the population of all college students in the United States and that you are using students at this college as a (non-random) sample. Identify each of the following as a parameter or a statistic, and indicate the symbol used to denote it:

- the proportion of this college's students who have a car on campus
- the proportion of all U.S. college students who have a car on campus
- the mean financial aid amount being received this academic year for all U.S. college students
- the standard deviation of the financial aid amounts being received this academic year for all students at this college

Activity 10-11: Non-Sampling Sources of Bias

(a) Suppose that simple random samples of adult Americans are asked to complete a survey describing their attitudes toward the death penalty. Suppose that one group is asked, "Do you believe that the U.S. judicial system should have the right to call for executions?" while another group is asked, "Do you believe that the death penalty should be an option in cases of horrific murder?". Would you anticipate that the proportions of "yes" responses might differ between these two groups? Explain.

(b) Suppose that simple random samples of students on this campus are questioned about a proposed policy to ban smoking in all campus buildings. If one group is interviewed by a person wearing a t-shirt and jeans and smoking a cigarette while another group is interviewed by a non-smoker wearing a business suit, would you expect that the proportions declaring agreement with the policy might differ between these two groups? Explain.

(c) Suppose that an interviewer knocks on doors in a suburban community and asks the person who answers whether he/she is married. If the person is married, the interviewer proceeds to ask, "Have you ever engaged in extra-marital sex?" Would you expect the proportion of "yes" responses to be close to the actual proportion of married people in the community who have engaged in extra-marital sex? Explain.

(d) Suppose that simple random samples of adult Americans are asked whether or not they approve of the President's handling of foreign policy. If one group is questioned prior to a nationally televised speech by the President on his/her foreign policy and another is questioned immediately after the speech, would you be surprised if the proportions of people expressing approval differed between these two groups? Explain.

(e) List four sources of bias that can affect sample survey results even if the sampling procedure used is indeed a random one. Base your list on the preceding four questions.

Activity 10-12: Survey of Personal Interest

Find a newspaper, magazine, or televised account of the results of a recent survey of interest to you. Write a description of the survey in which you identify the variable(s) involved, the population, the sample, and the sampling method. Also comment on whether the sampling method seems to have been random and, if not, whether it seems to be biased in a certain direction.

WRAP-UP

The material covered in this topic differs from earlier material in that you are beginning to consider issues related to how to collect data in the first place. This topic has introduced four terms that are central to formal statistical inference- population and sample, parameter and statistic.

One of the key ideas to take away from this topic is that a poor method of collecting data can lead to misleading (if not completely meaningless) conclusions. Another fundamental idea is that of random sampling as a means of selecting a sample that will (most likely) be representative of the population that one is interested in.

At this stage you have also begun to investigate (informally) properties of randomness. Probability is the language for expressing the likelihoods of random events. Probability can be used to describe the uncertainty in results of random samples from a known population. Probability can also be used in problems of statistical inference, where we want to express our knowledge about a population parameter such as p from information collected in a simple random sample. In the next topic, we begin our study of probability by describing two distinct interpretations of probability statements.

Topic 11 - What is a Probability?

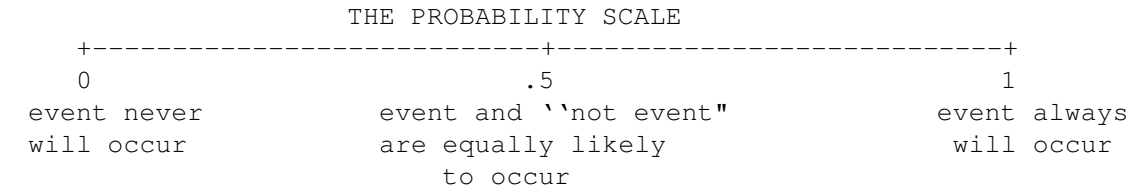
Introduction

What is a probability? What does it mean to say that a probability of a fair coin is one half, or that the chance I pass this class is 80 percent, or that the probability that the Steelers win the Super Bowl this season is .1?

First, think of some event where the outcome is uncertain. Examples of such outcomes would be the roll of a die, the amount of rain that we get tomorrow, the state of the economy in one month, or who will be the President of the United States in the year 2051. In each case, we don't know for sure what will happen. For example, we don't know exactly how much rain we will get tomorrow.

A **probability** is a numerical measure of the likelihood of the event. It is a number that we attach to an event, say the event that we'll get over an inch of rain tomorrow, which reflects the likelihood that we will get this much rain.

A probability is a number from 0 to 1. If we assign a probability of 0 to an event, this indicates that this event *never* will occur. A probability of 1 attached to a particular event indicates that this event *always* will occur. What if we assign a probability of .5? This means that it is just as likely for the *event to occur* as for the *event to not occur.*



There are two basic interpretations, or ways of thinking, about probabilities. These are the **relative frequency** viewpoint and the **subjective** viewpoint. These interpretations will help us assign probabilities to uncertain outcomes.

PRELIMINARIES

1. What does it mean if tomorrow's weather forecast gives

(a) a 100% chance of rain?

(b) a 70% chance of rain?

(c) a 50% chance of rain?

2. Suppose that 200 babies are born this month in your local hospital. How many of them do you expect to be girls?

3. What is a typical shooting percentage from the three-point line for a good woman college basketball player?

4. What is the probability that you will get a B or higher in this class?

5. Suppose that your two best friends took this same class last semester with the same instructor, and they both received A's. Does this information change the probability that you gave in the previous question? What is your new probability of getting a B or higher in this class?

6. What is the probability that John Tyler (former United States president) was born after the year 1800?

The Relative Frequency Interpretation of Probability

We are interested in learning about the probability of some event in some process. For example, our process could be rolling two dice, and we are interested in the probability in the event that the sum of the numbers on the dice is equal to 6.

Suppose that we can perform this process repeatedly under similar conditions. In our example, suppose that we can roll the two dice many times, where we are careful to roll the dice in the same manner each time.

I did this dice experiment 50 times. Each time I recorded the sum of the two dice and got the following outcomes:

4	10	6	7	5	10	4	6	5	6	11	11	3	3	6
7	10	10	4	4	7	8	8	7	7	4	10	11	3	8
6	10	9	4	8	4	3	8	7	3	7	5	4	11	9
5	5	5	8	5										

To approximate the probability that the sum is equal to 6, I count the number of 6's in my experiments (5) and divide by the total number of experiments (50). That is, the probability of observing a 6 is roughly the relative frequency of 6's.

$$\text{PROBABILITY (SUM IS 6) is approximately} \frac{\#\text{ of } 6'\text{s}}{\#\text{ tosses}}$$

$$= \frac{5}{50} = .1$$

In general, the probability of an event can be approximated by the *relative frequency* , or proportion of times that the event occurs.

$$\text{PROBABILITY (EVENT) is approximately} \frac{\#\text{ of times event occurs}}{\#\text{ experiments}}$$

Two comments should be made about this definition of probability:

- The observed relative frequency is just an approximation to the true probability of an event. However, if we were able to perform our process more and more times, the relative frequency will eventually approach the actual probability. We could demonstrate this for the dice example. If we tossed the two dice 100 times, 200 times, 300 times, and so on, we would observe that the proportion of 6's would eventually settle down to the true probability of .139.

- This interpretation of probability rests on the important assumption that our process or experiment can be repeated many times under similar circumstances. In the case where this assumption is inappropriate, the subjective interpretation of probability is useful.

IN-CLASS ACTIVITIES

Activity 11-1: Is it a Boy or a Girl?

Suppose that a baby is born in my hometown. What's the probability that the baby will be a boy? We know that, roughly, there are equal numbers of boys and girls born in the United States. So you might guess that the probability of a boy is one-half. The point of this activity is to demonstrate this fact by looking at hospital records of births that are published in a local newspaper.

In thirteen days in May I looked up the hospital records in my local paper. The paper announces the birth and sex of those babies born the previous day in the community hospitals. For each day, I write down the number of boys and number of girls that are born. This data is recorded in the table below.

DAY	# OF BOYS	# OF GIRLS	DAY	# OF BOYS	# OF GIRLS
1	2	5	8	4	1
2	1	2	9	2	6
3	5	2	10	4	1
4	2	1	11	0	1
5	3	5	12	4	3
6	2	0	13	2	1
7	0	1			

One way to interpret a probability is as a *long-term proportion*. Suppose we're interested in the probability of the outcome "the baby is a boy". If we observe some births and observe the sex of each child, then we can approximate this probability by the proportion of boys:

$$\text{PROB(boy)} \approx \frac{\text{number of boys}}{\text{number of births}}$$

As we observe more births, then this proportion will get closer to the true probability of a boy being born.

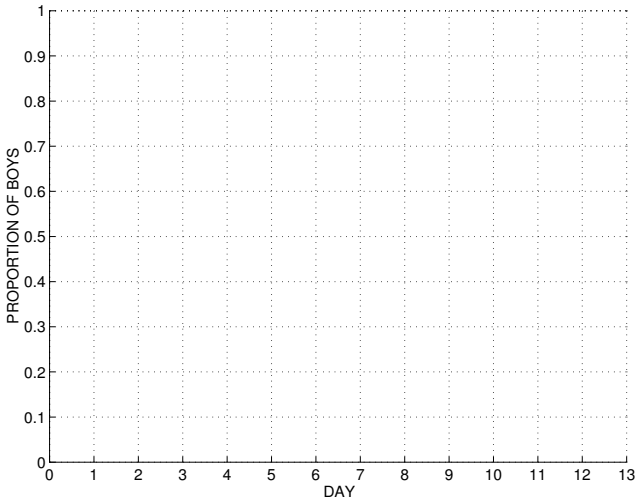
(a) Let's demonstrate this long-run behavior by computing proportions of boys for more and more hospital data. We'll perform our calculations in the table below. There are three columns to fill in. In the "CUM. BOYS" column, put the cumulative or total number of boys born in that day and all previous days. In "CUM. BIRTHS", put the cumulative births born that day or earlier. Based on these two entries, compute "PROP. BOYS", the proportion of boys born on that day and previous days.

I have filled in the first two rows for you. Using your calculator, fill in the rest of the rows.

DAY	# OF BOYS	# OF GIRLS	CUM. BOYS	CUM. BIRTHS	PROP. BOYS
1	2	5	2	7	.29
2	1	2	3	10	.30
3	5	2			
4	2	1			
5	3	5			
6	2	0			
7	0	1			
8	4	1			
9	2	6			
10	4	1			
11	0	1			
12	4	3			
13	2	1			

(b) Graph your data on the graph above. You graph the proportion of boys against the corresponding day. (I have graphed the first two days for you.)

(c) What pattern do you see in the graph? As you collect data from more and more days, what happens to the proportion of boys? Does this confirm that the probability of a boy is indeed one half?



The Subjective Interpretation of Probability

The relative frequency notion of probability is useful when the process of interest, say tossing a coin, can be repeated many times under similar conditions. But we wish to deal with uncertainty of events from processes that will occur a single time. For example, you are likely interested in the probability that you will get an A in this class. You will take this class only one time; even if you retake the class next semester, you won't be taking it under the same conditions as this semester. You'll have a different instructor, a different set of courses, and possibly different work conditions. Similarly, suppose you are interested in the probability that your team wins the championship in football next year. There will be only a single football season in question, so it doesn't make sense to talk about the proportion of times your team would win the championship under similar conditions.

In the case where the process will happen only one time, how do we view probabilities? Return to our example in which you are interested in the event "get an A in this class". You assign a number to this event (a probability) which reflects your personal belief in the likelihood of this event happening. If you are doing well in this class and you think that an A is a certainty, then you would assign a probability of 1 to this event. If you are experiencing difficulties in the class, you might think "getting an A" is close to an impossibility and so you would assign a probability close to 0. What if you don't know what grade you will get? In this case, you would assign a number to this event between 0 and 1. The use of a calibration experiment is helpful for getting a good measurement at your probability.

Comments about the subjective interpretation of probability:

- A subjective probability reflects a person's opinion about the likelihood of an event. If our event is "Joe will get an A in this class", then my opinion about the likelihood of this event

is probably different from Joe's opinion about this event. Probabilities are personal and they will differ between people.

- Can I assign any numbers to events? The numbers you assign must be proper probabilities. That is, they must satisfy some basic rules that all probabilities obey. Also, they should reflect your opinion about the likelihood of the events.
- Assigning subjective probabilities to events seems hard. Yes, it is hard to assign numbers to events, especially when you are uncertain whether the event will occur or not. We will learn more about assigning probabilities by comparing the likelihoods of different events.

Activity 11-2: Probability Phrases

A probability is a numerical measure of the likelihood of a statement. It can be difficult to assign probabilities to statements. For example, it may be difficult to assign your probability of a nuclear war in the next hundred years. But we use words, such as *unlikely*, *occasionally*, *even-chance*, and *rarely*, to indicate the chances that a particular event will occur.

For example, consider the following sentences:

- I'm *sure* it will rain tomorrow.
- He'll *never* finish his college degree.
- I *have no idea* whether I'll be able to pay my bills.
- The Braves *have a good chance* of winning the pennant.
- I *may* be able to finish that report within two weeks.
- It would *take a miracle* for him to recover from cancer.

Each word in italics expresses a *degree of belief* about a particular event or circumstance. We would like to assign numbers (called probabilities) to these words. Before we try to do this, we'll rank different words from least likely to most likely. An *unlikely* word is a word such as *never* which indicates a small chance of happening. A *likely* word is a word such as *probable* which indicates a large chance of happening.

(a) For each pair of words below, circle the word that you believe indicates a greater chance of happening.

It may be helpful to use these words in a particular context. For example, if you are describing the amount of precipation in your hometown, you might use the sentence "It _____ rains

around here," where the probability phrase goes in the blank. To answer the first question below, ask yourself: Which statement indicates a greater likelihood of happening – "It *sometimes* rains around here" or "It *often* rains around here"? If you think the word *often* means a larger likelihood, you circle the word *often* below.

(1) sometimes / often

(2) always / very-frequent

(3) seldom / even-chance

(4) unlikely / possible

(5) sometimes / even-chance

(6) very-frequent / often

(b) Now make a list of the eight words *sometimes, often, always, very-frequent, seldom, even-chance, unlikely, possible*, where the most likely word is on top and the least likely word is at the bottom.

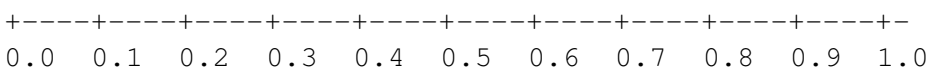
(c) Now assign numbers (each between 0 to 1) to the eight words – put these numbers to the right of the words. The numbers you assign should be consistent with the ordering of the words you made in part (b).

Activity 11-3: Assigning Numbers to Words.

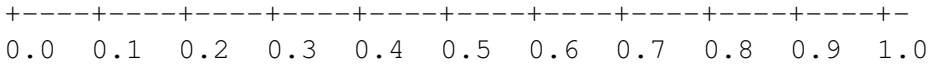
In the previous activity, we focused on ordering "degree of belief" words from the least likely to the most likely. Here we focus on the task of actually assigning numbers, called probabilities, to these words. For each of the following statements, think about the probability or likelihood that it is true. Place a mark on the probability scale which indicates your degree of belief. Remember ...

- a probability of 0 means that you are sure the statement is false
- a probability of 1 means that you think the statement must be true.
- a probability of .5 means that the statement and the "not statement" are equally likely.

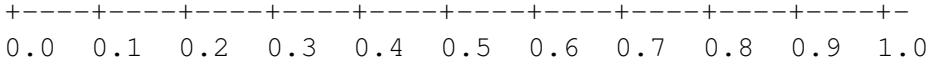
(a) I will get a “head” when I toss a fair coin.



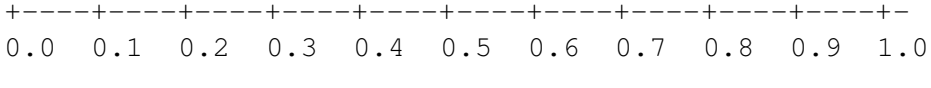
(b) There is life on the planet Mars.



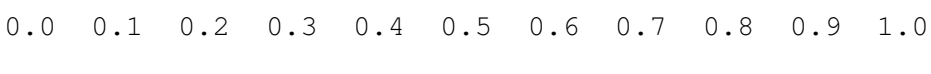
(c) A woman will be elected for president in the next 20 years.



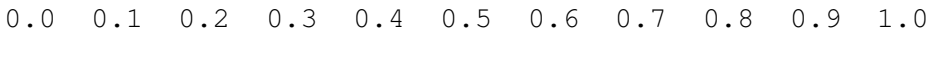
(d) I will get an A as a final grade in this class.



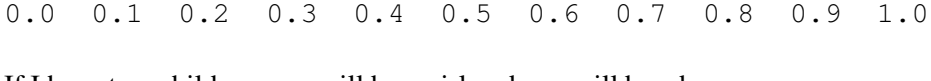
(e) I will marry someone taller than myself.



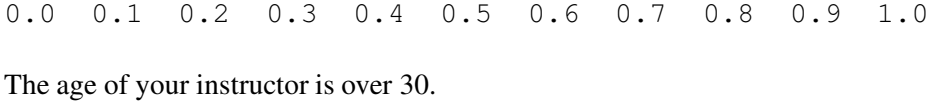
(f) I will find my first job (after graduation) in Ohio.



(g) My women’s basketball team will win the conference championship this year.



(h) If I have two children, one will be a girl and one will be a boy.



(i) The age of your instructor is over 30.

(j) The population of London is over 10 million.

```
+----+----+----+----+----+----+----+----+----+----+-
0.0  0.1  0.2  0.3  0.4  0.5  0.6  0.7  0.8  0.9  1.0
```

(k) It will rain tomorrow.

```
+----+----+----+----+----+----+----+----+----+----+-
0.0  0.1  0.2  0.3  0.4  0.5  0.6  0.7  0.8  0.9  1.0
```

Activity 11-4: Will the Chosen Ball be Black?

Suppose that you are going to draw a ball out from a bag consisting of white balls and black balls. Let's assume that you are going to shake the bag before you do your selection, so you can assume that each ball in the bag has the same chance of being chosen. You are interested in

Probability(a black ball is chosen).

This activity illustrates that we all can come up with the same probabilities on outcomes if we have the same information about the random experiment.

(a) Find the probability of choosing a black ball if

(i) the bag contains 10 balls, 5 of which are black and 5 are white

(ii) the bag contains 10 balls, 3 of which are black and 7 are white

(iii) the bag contains entirely white balls

(iv) you have no idea how many white and black balls are in the bag

(b) Suppose that the bag contains 5 black and 5 white balls and you choose a black ball. If you don't return this ball to the bag, find the probability that a second ball chosen from the bag is black.

Activity 11-5: When Was John Tyler Born? (from DeGroot)

The point of this activity is to demonstrate that probabilities that we assign are conditional on our current knowledge. As we learn and gain more information, our probabilities about statements can change. Later, we will introduce a formula, Bayes' rule, which will tell us how to compute new probabilities when we get information.

(a) Consider the year of birth of John Tyler, former President of the United States. The table below lists four statements about Tyler's year of birth.

YEAR OF BIRTH	PROBABILITY
no later than 1750	
between 1751 and 1775	
between 1776 and 1800	
after 1800	

- Which of the four statements do you believe is most likely?
- Which of the statements do you believe is least likely?
- Give probabilities to the four events that are consistent with the answers you made above. Put your answers in a probability table like shown above.

(b) You are now given the information that John Tyler was the tenth president of the United States. Use this information to reevaluate the probabilities you made above. Before you assign probabilities, answer the first two questions stated above.

(c) You are now given the information that George Washington, the first President of the United States, was born in 1732. Again reevaluate your probabilities and answer all three questions.

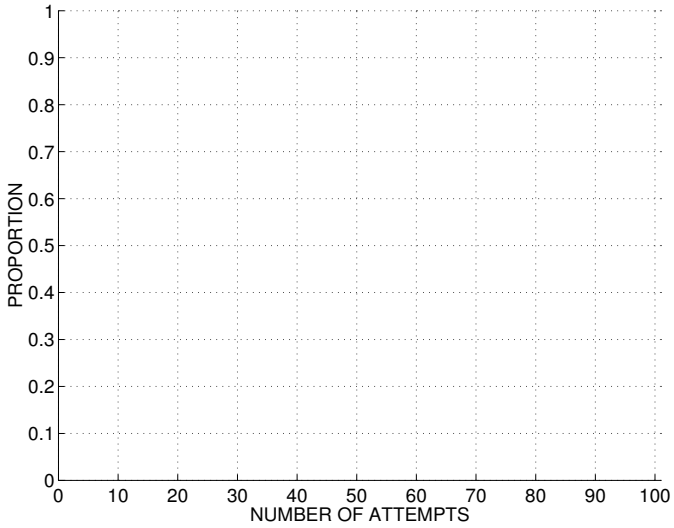
(d) You are now given the information that John Tyler was inagurated as President in 1841. Answer the same three questions.

HOMEWORK ACTIVITIES

Activity 11-6: What's Katie Voigt's Shooting Percentage?

In basketball, the basic measure of a player's shooting ability is the field goal percentage. This "average" is the proportion of shots that a player makes during a basketball season. In college and professional basketball, two field goal percentages are recorded: (1) the percentage of all attempted field goals made, and (2) the percentage of all field goals made that were attempted behind the three-point line. Generally, a shooting percentage of 50% and a three-point shooting percentage of 40% are both considered excellent marks for a player who plays the guard position.

One way of thinking about a shooting percentage is that it is the approximate probability of making a shot in a single attempt. Katie Voigt was a four-year starter on the University of Wisconsin's basketball team. Suppose that we are interested in estimating Katie's three-point shooting percentage, or shooting probability. We keep track of Katie's success in shooting for 100 three-point



attempts in one practice session. In the table below, we record the total number of attempts and the total number of made shots after each group of 5 shots.

(a) In the table, compute Katie's proportion of made shots after the numbered games 5, 10, 15, ... Put your proportions in the "PROPORTION" column.

ATTEMPTS	MADE	PROPORTION.	ATTEMPTS	MADE	PROPORTION
5	1		55	37	
10	4		60	40	
15	8		65	45	
20	12		70	49	
25	16		75	53	
30	19		80	58	
35	23		85	61	
40	27		90	63	
45	31		95	66	
50	34		100	68	

(b) Graph your proportion values against the number of attempts on the grid on the next page.

(c) Describe the pattern of the points in the first 50 attempts.

(d) Describe the pattern of the points for attempts 51-100..

(e) What do you think Katie's shooting percentage would be if she attempted 100,000 points? Why?

Activity 11-7: Tossing a Cup

Take an ordinary drink cup. Suppose you drop it on the floor from a height of 3 feet. We are interested in the probability that it lands with the larger end down.

(a) Before you try dropping the cup, guess at this probability.

(b) Drop the cup twenty times on the floor and keep track in the table below what you observe (L means that the larger end landed down and S means that the smaller end landed down).

TRIAL	L or S	TRIAL	L or S
1		11	
2		12	
3		13	
4		14	
5		15	
6		16	
7		17	
8		18	
9		19	
10		20	

(c) After the 5th, 10th, 15th, and 20th trials, compute the proportion of times the larger end of the cup landed down. Put your answers in the table below.

# OF TRIALS	PROPORTION
5	
10	
15	
20	

(d) If you were able to toss the cup 100 times, how many times would you expect the cup to land with the larger end down?

Activity 11-8: Dropping Two Tacks

Suppose you drop two thumbtacks on your desk. We are interested in the number of tacks that land up ($\perp$).

(a) If you drop two tacks and we're only interested in the number that land up, list all of the different possibilities of this experiment.

(b) Which of the different possibilities do you think is most likely? (Don't drop them yet!)

(c) Drop two tacks twenty times. Record the number of tacks landing up for each trial in the table below.

TRIAL	# OF TACKS $\perp$	TRIAL	# OF TACKS $\perp$
1		11	
2		12	
3		13	
4		14	
5		15	
6		16	
7		17	
8		18	
9		19	
10		20	

(d) Summarize your results in the table below. Put the number of trials in which 0, 1, or 2 tacks landed up and find the proportion of each case.

# OF TACKS $\perp$	COUNT	PROPORTION
0		
1		
2		

(e) If you toss two tacks, use the table to find the (approximate) probability that at least one tack lands up.

Activity 11-9: Risk of Losing a Job

Discover magazine had a special issue (May, 1996) on the subject of risk. In one article, they gave the probabilities of various calamities that one could face in his or her lifetime. Specifically, it was stated that the risk of an adult losing his or her job in the next year was 1 in 33, which corresponds to a probabability of .0294. Using this information, make intelligent guesses at

(a) the probability that a doctor will lose his/her job in the next year

(b) the probability that a lawyer will lose his/her job in the next year

(c) the probability that a bus driver will lose his/her job in the next year

(d) the probability that a farm worker will lose his/her job in the next year

This activity illustrates that it can be especially difficult to specify probabilities which are close to zero.

Activity 11-10: Tomorrow's High Temperature?

We will see in Topic 12 that all probabilities must satisfy certain rules. In this activity, you are asked to specify particular probabilities about tomorrow's high temperature and then in Activity 12-12, you will check if your probabilities are consistent with the rules.

Consider the high temperature (in degrees Fahrenheit) in your town tomorrow. Make intelligent guesses at the following probabilities:

(a) the probability that the high temperature will be over 60 degrees _____

(b) the probability that the high temperature will be between 40 and 60 degrees _____

(c) the probability that the high temperature will be under 40 degrees _____

(d) the probability that the high temperature will be under 60 degrees _____

(e) the probability that the high temperature will be over 70 degrees _____

Activity 11-11: Weather Forecasting

(a) Using the relative frequency notion of probability, explain what it means when a weather forecaster says that there is 75% chance of rain tomorrow.

(b) If a weather forecaster says that there is 50% chance of rain tomorrow, is it correct to say that the forecaster has no idea whether it will rain or not? Explain.

Activity 11-12: Which is the Correct Interpretation?

There are two ways of viewing a probability:

- as a long-run relative frequency
- as one's subjective opinion about the likelihood of an event

For each of the following scenarios, a probability will be described. Say whether this probability is best measured as a *relative frequency* or as a *subjective opinion*. Explain briefly the reason for your answer.

(a) The probability that a Democrat will win the presidential election in the year 2000 is .7.

(b) If I use a particular strategy in playing blackjack, the probability that I win a game is .6.

(c) The probability a randomly selected student from your school has blue eyes is .4.

(d) The probability that Northwest Ohio will experience more snow this winter than last winter is .7.

(e) The probability of tossing three heads in three tosses of a fair coin is 1/8.

(f) The probability that an adult (over 16) will lose his or her job in the next year is 1/33.

(g) The probability that I will get an A in this class is .7.

Activity 11-13: How Large is Pennsylvania? (from DeGroot)

Consider the area, in square miles, of Pennsylvania.

(a) The table below lists four statements about the state's size.

PERIOD	PROBABILITY
less than 5,000 square miles	
between 5,000 and 50,000 square miles	
between 50,000 and 100,000 square miles	
over 100,000 square miles	

- Which of the four statements do you believe is most likely?
- Which of the statements do you believe is least likely?
- Give probabilities to the four events that are consistent with the answers you made above. Put your answers in a probability table like shown above.

(b) You are now given the following information: The area of Alaska, the largest of the fifty states, is 586,400 sq. miles, and the area of Rhode Island, the smallest state, is 1,214 sq. miles. Use this information to reevaluate the probabilities you made above. Before you assign probabilities, answer the first two questions.

(c) You are now given the information that when area is considered, Pennsylvania is the thirty-third largest of the 50 states. Again reevaluate your probabilities and answer all three questions.

(d) You are now given the information that the area of New York, the 30th largest state, is 49,576 sq. miles. Answer the same three questions.

WRAP-UP

In this topic, we're focused on two basic interpretations of probabilities – the relative frequency viewpoint and the subjective viewpoint. In a given situation, one or the other interpretation may be appropriate for use. The subjective viewpoint is the best way of thinking of a probability when the particular random process (such as the performance of your favorite sports team this year) is essentially a one-time occurrence. The relative frequency notion of probability is more useful for random experiments which can be repeated many times under similar conditions. Simple chance experiments such as tossing a coin or rolling a die are of this type and we can use the relative frequency interpretation in measuring probabilities of outcomes for these experiments.

No matter how we think about probabilities, all probabilities that we assign must satisfy certain rules. In the next topic, we'll discuss issues related to assigning probabilities to outcomes of a random experiment.

Topic 12: Assigning Probabilities

Introduction

In this topic, we'll explore how one actually assigns probabilities to different events. Before we talk about probabilities, it is important to completely specify all of the possible results of a random process or experiment. The collection of possible results is called the **sample space** of the experiment. The next step is to assign numbers called **probabilities** to the different outcomes. We can't assign just any numbers to these outcomes — we'll see that probabilities must satisfy **some basic rules**.

The remainder of the topic focuses on different methods of assigning probabilities to results of a random experiment. In some cases, the different outcomes specified in the sample space are **equally likely**; in this case, it is relatively easy to assign probabilities. For many situations, outcomes will not be equally likely and this method will not work. In the situation where the random process is repeatable under similar conditions, one can **simulate** the process many times, and assign probabilities by computing proportions of outcomes. In other situations, the random process is not repeatable. In this case, we will discuss the use of a **calibration experiment** to help in the assignment of subjective probabilities.

PRELIMINARIES

1. If you toss a coin three times, how many different outcomes are possible?
2. If you roll a die, what's the chance that you will roll a 2?
3. If you draw a card from a standard 52 card deck, what is the chance that you will draw an ace?
4. What is the chance that you *will not* draw an ace?
5. If you toss a fair coin 10 times, would you be surprised to see five straight heads?
6. Would you be surprised to see eight tails (out of 10)?

Listing All Possible Outcomes (The Sample Space)

Suppose that we will observe some process or experiment in which the outcome is not known in advance. For example, suppose we plan to roll two dice and we're interested in the sum of the two numbers appearing on the top faces. Before we can talk about probabilities of various sums, say 3 or 7, we have to understand what outcomes are possible in this experiment.

If we roll two dice, each die could show 1, 2, 3, 4, 5, 6. So the sum of the two faces could be any whole number from 2 to 12. We call this set of possible outcomes in the random experiment the *sample space* . Here the sample space can be written as

$$\text{Sample space} = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

Let's consider the set of all possible outcomes for other basic random experiments. Suppose we plan to toss a coin 3 times and the outcome of interest is the number of heads. The sample space in this case is the different numbers of heads you could get if you toss a coin three times. Here you could get 0 heads, 1 heads, 2 heads or 3 heads, so we write the sample space as

$$\text{Sample space} = \{0, 1, 2, 3\}$$

Don't forget to include the outcome 0 – if we toss a coin three times and get all tails, then the number of heads is equal to 0.

The concept of a sample space is also relevant for experiments where the outcomes are non-numerical. Suppose I draw a card from a standard deck of playing cards. If I'm interested in the suit of the card, there are four possible outcomes and the sample space is

$$\text{Sample space} = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}.$$

If I am interested in the suit and the face of the card, then there are many possible outcomes. One can represent the sample space by the following table:

	FACE OF CARD												
SUIT	2	3	4	5	6	7	8	9	10	J	Q	K	Ace
♠	x	x	x	x	x	x	x	x	x	x	x	x	x
♡	x	x	x	x	x	x	x	x	x	x	x	x	x
♢	x	x	x	x	x	x	x	x	x	x	x	x	x
♣	x	x	x	x	x	x	x	x	x	x	x	x	x

Each "x" in the table corresponds to a particular outcome of the experiment. For example, the first "x" in the third row of the table corresponds to a draw of the 2 of Diamonds. We see from this table that there are 52 possible outcomes.

Once we understand what the collection of all possible outcomes looks like, we can think about assigning probabilities to the different outcomes. But be careful – incorrect probability assignments can be made because of mistakes in specifying the entire sample space.

IN-CLASS ACTIVITIES

Activity 12-1: Specifying Sample Spaces

Different "experiments" are described below — in each case, the outcome of the experiment or process is unknown. Give a reasonable sample space for each experiment. In other words, write down all of the different outcomes that are possible. To help you out, one experiment outcome is listed in italics.

(a) You toss a coin four times and count the number of heads. [*One sample outcome is 3.*]

(b) A box contains four slips of paper numbered 1, 2, 3, 4. You randomly pick one slip from the box. [*One sample outcome is 3.*]

(c) Suppose you pick two slips of paper from the box described above. We'll assume this is done without replacement, which means that you can't pick the same slip of paper twice. You record the sum of the numbers on the two slips. [*One sample outcome is 7.*]

(d) You ask a friend to give you the first names of all the Beatles. You write down how many names he correctly identifies. [*One sample outcome is 1 name.*]

(e) You toss a die repeatedly until you get two 6's. You are interested in how many tosses it takes. [*One sample outcome is 5 tosses.*]

(f) You walk from your home to your first school class. How much time could it take? [*One sample outcome is 15 minutes.*]

(g) Suppose an undergraduate student is sampled and you find out his/her age. [*One sample outcome is 18 years old.*]

(h) When will a man again land on the moon? [*One sample outcome is the year 2050.*]

Probability Rules

Suppose that a random process results in a number of different outcomes. A sample space is a list of all such outcomes. As an example, suppose I'm interested in the amount of time (in minutes) it takes to drive to work this morning. Based on my past experience, I know that there are four different possibilities. So my sample space looks like the following:

OUTCOME
it takes under 30 minutes
it takes between 30 and 35 minutes
it takes between 35 and 40 minutes
it takes over 40 minutes

I wish to assign probabilities to these four outcomes. Before we actually attach numbers to these outcomes, we should first ask: Are there any rules that probabilities must satisfy?

Yes, probabilities must follow three general rules:

- **Rule 1:** Any probability assigned must be a nonnegative number.
- **Rule 2:** The probability of the sample space (the collection of all possible outcomes) is equal to 1.
- **Rule 3:** If you have two outcomes that can't happen at the same time, then the probability that either outcome occurs is the sum of the probabilities of the individual outcomes.

How do we use these rules to assign probabilities in the above "drive to work" example? The first rule tells us that **probabilities can't be negative**, so it makes no sense to assign -1, say, to the outcome "it takes over 30 minutes". The second and third rules tell us that the probabilities that we assign to a collection of *nonoverlapping* outcomes must add to 1. Nonoverlapping outcomes means that they can't occur at the same time. For example, the outcomes "takes over 20 minutes" and "takes under 30 minutes" are *overlapping* since they both can happen (if, say, the trip takes 24

minutes). The outcomes "takes under 20 minutes" and "takes over 25 minutes" are nonoverlapping, since at most one of these events can happen at the same time.

With these rules in mind, here are three hypothetical assignments of probabilities, corresponding to four people, Max, Joe, Sue, and Mary.

Four Sets of Probabilities

OUTCOME	Max	Joe	Sue	Mary
it takes under 30 minutes	.3	.2	.4	0
it takes between 30 and 35 minutes	$-.1$	.3	.4	.2
it takes between 35 and 40 minutes	.4	.4	.1	.8
it takes over 40 minutes	.4	.3	.1	0

Who has made legitimate probability assignments in the above table? There are problems with the probabilities that Max and Joe have assigned. Max can't give the outcome "it takes between 30 and 35 minutes" a negative probability, no matter how unlikely this outcome. Joe has made a mistake, since the sum of his probabilities for the four nonoverlapping outcomes is 1.2, which is not equal to 1.

Sue and Mary have given sets of legitimate probabilities, since they are all nonnegative and they sum to 1. But there are differences between these two sets of probabilities, which reflect different opinions of these two people about the length of time to work. Sue is relatively optimistic about the time to work, since .8 of her probability is on the outcomes "under 30 minutes" and "between 30 and 35 minutes". In contrast, Mary believes that a trip under 30 minutes will never occur (it has a probability of 0) and it is very probable that it will take between 35 and 40 minutes.

Activity 12-2: Assigning Probabilities to Rolls of a Die

Suppose that you are going to toss a die. The probabilities that you assign to the six possible rolls 1, 2, 3, 4, 5, 6 depend on what you know about the die. In each of the following parts, some information will be given about the die. Based on this information, assign probabilities to the numbers 1, ..., 6 and put your answers in the table given.

(a) The six sides of the die are 1, 2, 3, 4, 5, 6 and the die is balanced — so each of the six numbers is equally likely to occur.

ROLL	1	2	3	4	5	6
PROBABILITY						

(b) The die is balanced with three sides showing 2 and the remaining three sides showing 3.

ROLL	1	2	3	4	5	6
PROBABILITY						

(c) The die has six sides with numbers 1, 2, 3, 4, 5, 6. But the die is unbalanced in such a way that only 5 shows when the die is tossed.

ROLL	1	2	3	4	5	6
PROBABILITY						

(d) The die is balanced with two sides showing 4 and the remaining four sides showing 6.

ROLL	1	2	3	4	5	6
PROBABILITY						

Computing Probabilities With Equally Likely Outcomes

Before we can compute any probabilities for outcomes in a random process, we have to define the sample space, or collection of all possible outcomes. If we have listed all outcomes and it is reasonable to assume that the outcomes are *equally likely*, then it is easy to assign probabilities.

Let's consider a simplified lottery game. Suppose that Ohio has a game where you try to guess a random two-digit number that is selected. This "winning" random number is selected by the following process. There are two boxes, labeled box A and box B, that each contain 10 ping pong-balls labeled using the digits 0 through 9. A random number is selected by letting the first digit be the number of the ball selected from box A, and the second digit is the number of the ball selected from box B.

What is the sample space? There are 100 possible winning two-digit numbers that are listed below.

00	01	02	03	04	05	06	07	08	09
10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89
90	91	92	93	94	95	96	97	98	99

By the way the two-digit number is selected, no particular number listed above has any more or less chance of being selected than another number. So it is reasonable to assign the same probability

to each number in the above list. What probability should be assigned? There are 100 possible winning numbers. If we wish to assign the same probability to each number and keep the total probability of all the numbers equal to 1, then each number should be given the probability 1/100 = .01.

In general, if there are N possible outcomes in an experiment and the outcomes are *equally likely*, then you should assign a probability of *1/N* to each outcome.

Once we have assigned probabilities to outcomes in the sample space, one can find a probability of a more complicated event by (1) finding the outcomes that are contained in the event and (2) summing the probabilities of the outcomes. To illustrate, suppose we wish to find the probability of the event $\{$sum of the two digits is equal to $8\}$. Looking at the sample space, we note that the sum of the digits is eight for the numbers

$$08, 17, 26, 35, 44, 53, 62, 71, 80.$$

Since the probability of each two-digit number is 1/100, and there are 9 outcomes listed, the desired probability is

$$\text{Prob}(\text{sum of digits is equal to } 8) = \frac{9}{100} = .09.$$

CAUTION: This recipe for assigning probabilities works only when the outcomes are equally likely. It is easy to misuse this. For example, suppose you toss a coin three times and you're interested in the number of heads. The possible numbers of heads (the sample space) are

0 head, 1 head, 2 heads, 3 heads

There are four outcomes in this case. But it is incorrect to assume that the probabilities of each outcome is 1/4 = .25. These four outcomes are *not* equally likely. In fact, the probability of 1 head can be shown to be three times the probability of 3 heads.

Activity 12-3: Drawing a Card

Suppose you draw one card at random from an ordinary card deck. There are 52 possible cards that you can select — the table below shows all selections, with the row indicating the suit and the column the face.

	FACE OF CARD												
SUIT	2	3	4	5	6	7	8	9	10	J	Q	K	Ace
♣	x	x	x	x	x	x	x	x	x	x	x	x	x
♢	x	x	x	x	x	x	x	x	x	x	x	x	x
♡	x	x	x	x	x	x	x	x	x	x	x	x	x
♠	x	x	x	x	x	x	x	x	x	x	x	x	x

(a) Assuming that all possible card selections are equally likely, what probability should be assigned to each selection?

(b) Find the probability that you choose the 5 of ♢.

(c) Find the probability that you choose a ♠.

(d) Find the probability that you choose a face card (J, Q, K or Ace).

(e) Find the probability that you choose a black card (♠ or ♣).

Computing Probabilities by Simulation

In the case where all of the outcomes of an experiment are equally likely, then it is easy to assign probabilities. But, when outcomes are *not* equally likely, it can be hard to allocate probabilities to events. However, there is a general method which will give us probabilities when the random process can be repeated many times under similar conditions. Examples of such processes that can be repeated include rolling two dice to learn about the probability of getting a sum of 6 and tossing 20 coins to learn about the probability of obtaining exactly 10 heads. In each case, we compute probabilities by the relative frequency notion of probability. The probability of a particular outcome (say getting a sum of 6 in a roll of two dice) is approximated by the proportion of times the outcome occurs in our experiment.

Let's illustrate this method of assigning probabilities by considering the experiment of tossing a fair coin 20 times. We're interested in the total number of heads observed. First, we think about the sample space — how many heads could we get if we toss the coin 20 times? There are 21 possible outcomes in this experiment – the sample space consists of $\{0, 1, 2, ..., 20\}$, where each number in the set is a plausible number of heads in 20 tosses of the coin.

We use the computer to simulate 20 coin tosses. In the first simulation, we observe the following sequence of heads (H) and tails (T).

Coin Tosses																				# of Heads
H	T	T	T	H	H	H	H	T	H	T	H	H	T	H	H	H	T	H	T	12

We note that there are 12 heads in this sequence. Is this a typical value for the number of heads of 20 tosses of a fair coin? To help answer the question, we run this experiment 9 more times; the tosses for the 10 experiments are shown below.

Coin Tosses																				# of Heads
H	T	T	T	H	H	H	H	T	H	T	H	H	T	H	H	H	T	H	T	12
T	H	T	T	T	H	H	H	T	H	H	H	H	T	H	H	H	H	T	H	13
H	H	H	H	T	T	H	T	T	H	H	T	T	T	T	H	T	H	T	T	9
T	T	H	T	H	T	T	H	T	T	T	T	H	H	T	H	H	T	H	H	9
T	H	H	T	H	T	H	T	T	T	T	H	H	H	T	H	H	H	T	T	10
H	T	T	H	H	T	H	H	T	T	T	T	T	H	T	T	T	H	T	H	8
H	H	T	H	H	T	H	H	H	T	H	T	H	H	T	H	H	H	H	H	15
T	T	T	T	H	H	H	T	H	H	H	T	H	H	T	H	T	H	T	T	10
T	T	H	T	H	H	H	H	T	T	T	H	H	T	T	H	T	T	H	T	9
H	H	T	T	T	T	T	H	H	T	H	T	H	T	H	H	H	H	H	H	11

To help understand any pattern in these observed numbers of heads, we use a stemplot display. The stems on the left are the elements of the sample space $\{0, 1, 2, ..., 20\}$. Then we indicate by leafs of 0's the outcomes in our 10 experiments.

```
 0                  10 EXPERIMENTS
 1
 2
 3
 4
 5
 6
 7
 8 0
 9 000
10 00
11 0
12 0
13 0
14
15 0
16
17
18
19
20
```

We're starting to see some clumping of the values about 10, but we haven't performed enough experiments to see a strong pattern. Let's continue to toss coins until we've done 50 experiments.

```
 0                  50 EXPERIMENTS
 1
 2
 3
 4
 5 0
 6 0
 7
 8 0000
 9 000000000000
```

```
10 0000000000000
11 0000000
12 0000
13 000
14 00
15 00
16 00
17
18
19
20
```

We're starting to see a clearer pattern in the stemplot. Let's continue and take a total of 200 experiments.

```
 0      0                 200 EXPERIMENTS
 0      1
 0      2
 0      3
 0      4
 3      5 000
 8      6 00000000
 9      7 000000000
20      8 00000000000000000000
43      9 0000000000000000000000000000000000000000000000
41     10 0000000000000000000000000000000000000000
33     11 000000000000000000000000000000000
17     12 00000000000000000
12     13 000000000000
 8     14 00000000
 4     15 0000
 2     16 00
 0     17
 0     18
 0     19
 0     20
```

It appears that we have done enough simulations – we are starting to see a pretty good pattern in the stemplot. From this graph, we can compute approximate probabilities for different outcomes. To make the below calculations easy, I have listed the frequencies of each outcome to the left of the stemplot.

- What is an average number of heads in 200 tosses? We see from the graph that the distribution of probabilities is roughly symmetric about 9 or 10 – either number could be considered an average value.

- What is the probability of tossing exactly 10 heads in 200 tosses? From the graph, we see that we observed 41 "10 heads" in 200 tosses, so the probability

Probability(10 heads) is approx 41/200 = .205

It is interesting to note that, although 10 heads is an *average* value, it is not a *likely* value, since the probability is only about 20%.

- What are unlikely values? We see from the display that 4 or fewer or 17 or more heads were never observed in our simulation. So the probability of these outcomes must be small.

Activity 12-4: Tossing Coins

(a) Suppose that you are going to toss a fair coin 10 times. Imagine that you are doing this (don't actually toss a coin). In the spaces below, write down the outcomes of the 10 tosses, where H stands for a head and T for a tail.

TOSS	1	2	3	4	5	6	7	8	9	10
RESULT										

(b) For your imaginary set of 10 tosses, find

- the number of heads
- the number of switches from H to T or from T to H (for example, the number of switches in HTTHHTTTHH is 4.)
- the length of the longest run of heads (this would be 2 in the above sequence)

(c) Now let's try real coin-tossing. Your instructor will give you a penny and tell you how to toss the coin. Toss your coin 10 times and record the results (H or T) in the first row of the table. Continue this 10-toss experiment for a total of 20 experiments. You should have filled in all of the boxes with the exception of the last two columns.

EXP	TOSSES										# OF HEADS	SWITCHES	LONGEST RUN
1													
2													
3													
4													
5													
6													
7													
8													
9													
10													
11													
12													
13													
14													
15													
16													
17													
18													
19													
20													

(d) For each set of 10 tosses, record the number of heads and put the number in the # OF HEADS column. Graph the 20 numbers on the dotplot below.

```
 +----+----+----+----+----+----+----+----+----+----+- NUMBER OF HEADS
0.0       2.0       4.0       6.0       8.0      10.0
```

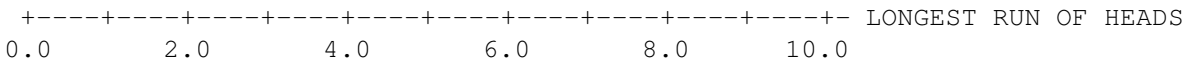
Describe the pattern of values that you see on the dotplot.

(e) For each set of 10 tosses, record the number of switches and put it in the SWITCHES column. Graph your 20 values of number of switches on the dotplot below.

```
 +----+----+----+----+----+----+----+----+----+----+- NUMBER OF SWITCHES
0.0       2.0       4.0       6.0       8.0      10.0
```

Describe the pattern of values that you see on the dotplot.

(f) For each set of 10 tosses, record the length of the longest run of heads and put it in the LONGEST RUN. Graph your 20 run lengths on the dotplot.



Describe the pattern of values that you see on the dotplot.

(g) Go back to your set of imaginary tosses. On the three dotplots above, plot (with a large X) the number of heads, the number of switches, and the length of the longest run for your imaginary data. Were your values of these three variables consistent with the values that you obtained by real coin-tossing? Explain.

Probabilities of "OR" and "NOT" Events

Sometimes we are interested in computing probabilities of more complicated events. Here, we introduce two properties of probabilities. The first property is useful for finding the probability of one event or another event. The second property tells us how to compute the probability that an event does not occur.

The addition property (for computing probabilities of "or" events)

We will illustrate this property with an example. Suppose Ohio has a two-digit lottery game and the winning number will be chosen at random from all possible two-digit numbers

$$\{00, 01, 02, 03, ..., 97, 98, 99\}.$$

There are 100 possible winning numbers and since each has the same chance of being chosen, we assign a probability of 1/100 = .01 to each number.

Suppose we want to find the probability that the winning number has the same two digits or the winning number is between 89 and 96 inclusive. If these two events ("same two digits" and "between 89 and 96") are nonoverlapping, then we can find the probability of "same two digits" or "between 89 and 96" by adding:

$$\begin{aligned}&\text{Prob("same two digits" or "between 89 and 96")}\\ &= \text{Prob("same two digits")+Prob("between 89 and 96")}\end{aligned}$$

Are these two events nonoverlapping? Nonoverlapping means that it is impossible for the two events to occur at the same time. Here "same two digits" means the winning number is from the set $\{00, 11, 22, 33, 44, 55, 66, 77, 88, 99\}$. "Between 89 and 96" means that the number is in the set $\{89, 90, 91, 92, 93, 94, 95, 96\}$. Note that these two sets have nothing in common; in other words, it is impossible for the winning number to have the same two digits and be between 89 and 96. So we can add the probabilities to find the probability of the "or" event. The probability of "same two digits" is 10/100 and the probability of "between 89 and 96" is 8/100. Therefore the probability of interest is

$$\text{Prob("same two digits" or "between 89 and 96")} = 10/100 + 8/100 = 18/100 = .18$$

What if we wanted to find the probability of "same two digits" or "an even second digit"? Here we can't use this addition property, since these two events are overlapping. It is possible for the winning to have the same two digits and have an even second digit – the number 44 (and other numbers) is in both events. So this property cannot be used in this case.

This property is also applicable in the case where you want to find the probability of a collection of different outcomes. Suppose you toss a coin five times and you wish to find the probability that the number of heads is 2 or fewer. You can think of the event "2 or fewer heads" as an "or" event:

$$\{\text{ 2 or fewer heads }\} = \{0 \text{ heads}\} \text{ or } \{1 \text{ head}\} \text{ or } \{2 \text{ heads}\}$$

By definition, the three outcomes $\{0 \text{ heads}\}$, $\{1 \text{ heads}\}$ and $\{2 \text{ heads}\}$ are nonoverlapping, since you can only observe at most one of these outcomes when you toss the coin three times. So the addition property can be

$$\text{Prob(2 or fewer heads)} = \text{Prob(0 heads)} + \text{Prob(1 head)} + \text{Prob(2 heads)}$$

The complement property (for computing probabilities of "not" events)

Let's return to our lottery example. What if you're interested in the probability that the winning number does not have the same two digits? The property for "not" events is called the complement property:

$$\text{Probability(“not” an event)} = 1 - \text{Probability(event)}$$

In this case, we can write

$$\text{Probability(not same digits)} = 1 - \text{Probability(same digits)}$$

We have already found the probability that the winning number has the same two digits, so the probability of interest is

$$\text{Probability(not same digits)} = 1 - 10/100 = 90/100$$

The complement property is especially useful in the case where it hard to compute the probability of an event, but it is relatively easy to compute the probability of “not” the event. For example, suppose we wish to compute the probability of tossing at least one head in 10 tosses of a coin. In this case, it would make sense to first perform the easier computation, the probability of “not at least one head” or “no heads”. Then we apply the complement property to find the probability of the event of interest.

$$\text{Probability(at least one head)} = 1 - \text{Probability(no heads)}$$

Activity 12-5: Drawing a Card (cont.)

Let’s return to the experiment of drawing a single card from a standard card deck. Remember there were 52 possible cards that could be selected and each card has probability 1/52 of being chosen.

(a) Consider the events $\{$draw a face card$\}$ and $\{$draw a 5 or smaller$\}$. Are these events overlapping?

(b) Are the events $\{$draw a face card$\}$ and $\{$draw a spade$\}$ overlapping?

(c) Use the addition property to find the probability of $\{$draw a face card$\}$ or $\{$draw a 5 or smaller$\}$.

(d) Can you use the addition rule to find the probability of $\{$draw a face card$\}$ and $\{$draw a spade$\}$? Explain.

(e) Find the probability of $\{$draw a red card$\}$ or $\{$draw a spade$\}$.

(f) Find the probability of not drawing a 5 or smaller.

(g) Find the probability that the Ace of hearts is not drawn.

Measuring Probabilities Using a Calibration Experiment

Probabilities that are viewed from a subjective viewpoint are generally hard to measure. It is easy to measure probabilities of events that are extremely rare or events that are extremely likely to occur. For example, your probability that the moon is made of green cheese (a rare event) is probably close to 0 and your probability that the sun will rise tomorrow (a sure event) is likely 1. But consider your probability for the event "There will be a white Christmas this year". You can remember years in the past where there was snow on the ground on Christmas. Also you can recall past years with no snow on the ground. So the probability of this event is greater than 0 and less than 1. But how do you obtain the exact probability?

To measure someone's height we need a measuring instrument such as a ruler. Similarly, we need a measuring device for probabilities. This measuring device that we use is called a **calibration experiment**. This is an experiment which is simple enough so that probabilities of outcomes are easy to specify. In addition, these stated probabilities are objective; you and I would assign the same probabilities to outcomes of this experiment.

The calibration experiment that we use is called a **chips-in-bowl** experiment. Suppose we have a bowl with a certain number of red chips and white chips. We draw one chip from the bowl at random and we're interested in

Probability(red chip is drawn)

This probability depends on the number of chips in the bowl. If, for example, the bowl contains 1 red chip and 9 white chips, then the probability of choosing a red is 1 out of 10 or 1/10 = .1. If the bowl contains 3 red and 7 chips, then the probability of red is 3/10 = .3. If the bowl contains only red chips (say 10 red and 0 white), then the probability of red is 1. At the other extreme, the probability of red in a bowl with 0 red and 5 white is 0/5 = 0.

Let's return to our event "There will be a white Christmas this year". To help assess its probability, we compare two bets – one with our event and the second with the event "draw a red chip" from the calibration experiment. This is best illustrated by example. Consider the following two bets:

- BET 1: You get $100 if there is a white Christmas and nothing if there is not a white Christmas.

- BET 2: You get $100 if you draw red in a bowl of 5 red and 5 white and nothing otherwise.

Which bet do you prefer? If you prefer BET 1, then you think that your event of a white Christmas is more likely than the event of drawing red in a bowl with 5 red, 5 white. Since the probability of a red is 5/10 = .5, this means that your probability of a white Christmas exceeds .5. If you prefer BET 2, then by similar logic, your probability of a white Christmas is smaller than .5.

Say you prefer BET 1 and you know that your probability is larger than .5, or between .5 and 1. To get a better estimate at your probability, you make another comparison of bets, where the second bet has a different number of red and white chips.

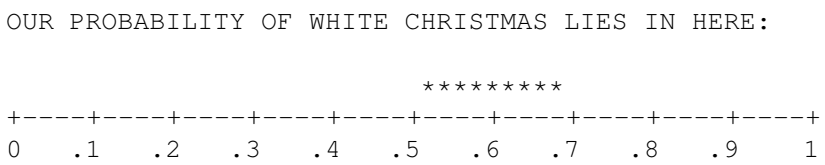
Next you compare the two bets

- BET 1: You get $100 if there is a white Christmas and nothing if there is not a white Christmas.

- BET 2: You get $100 if you draw red in a bowl of 7 red and 3 white and nothing otherwise.

Suppose that you prefer BET 2. Since the probability of red in a bowl of 7 red and 3 white is 7/10 = .7, this means that your probability of a white Christmas must be smaller than .7.

We've now made two judgments between bets. The first judgment told us that our probability of white Christmas was greater than .5 and the second judgment told us that our probability was smaller than .7. What is our probability? We don't know the exact value yet, but we know that it must fall between .5 and .7. We can represent our probability by an interval of values on a number line.



What if we wanted to get a more accurate estimate at our probability? We need to make more comparisons between bets. For example, we could compare two bets, where the first bet used our event and the second used the event "draw red" from a bowl of chips with 6 red and 4 white. After a number of these comparisons, we can get a pretty accurate estimate at our probability.

Activity 12-6: Using Chip-in-Bowl Experiments

(a) Consider the following "chip-in-bowl" experiments. What is the probability of *drawing a red* if the bowl contains

(1) 5 red and 5 white?

(2) 2 red and 8 white?

(3) 7 red and 3 white?

(4) 0 red and 10 white?

(b) Consider the statement

A: "A woman will be elected for president in the next twenty years."

We want to determine your personal probability that A is true, call this PROB(A).

Consider the following two bets:

BET 1: i. If a woman is elected president, then you win \$20.

ii. If a woman is not elected president, you win nothing.

BET 2: i. If you draw a red in a chip-in-bowl experiment with 5 red and 5 white chips, you win \$20.

ii. If you draw a white, you win nothing.

(1) Which bet (1 or 2) do you prefer?

(2) Based on your answer to (1), do you think PROB(A) $>$ .5 or PROB(A) $<$ .5?

(c) Let's continue to make this comparison for more chip-in-bowl experiments. Each row of the table below gives two choices. The left choice is BET 1: you win \$20 if a woman is elected president and win nothing if this event does not happen. The choice on the right is BET 2: you win \$20 if you draw a red from a bowl with a certain number of reds and whites; otherwise you win nothing. For each pair of bets, circle the choice which you prefer

\$20 if a woman is elected president nothing if a woman is not elected president	\$20 if draw red in bowl with 0 red and 10 white chips nothing if draw white
\$20 if a woman is elected president nothing if a woman is not elected president	\$20 if draw red in bowl with 2 red and 8 white chips nothing if draw white
\$20 if a woman is elected president nothing if a woman is not elected president	\$20 if draw red in bowl with 4 red and 6 white chips nothing if draw white
\$20 if a woman is elected president nothing if a woman is not elected president	\$20 if draw red in bowl with 6 red and 4 white chips nothing if draw white
\$20 if a woman is elected president nothing if a woman is not elected president	\$20 if draw red in bowl with 8 red and 2 white chips nothing if draw white
\$20 if a woman is elected president nothing if a woman is not elected president	\$20 if draw red in bowl with 10 red and 0 white chips nothing if draw white

(d) Based on your choices you made above, you should have a better idea about your probability of a woman elected president in the next 20 years elected. Mark on the below number line your probability. (This value should be consistent with the choices that you made.)

```
+----+----+----+----+----+----+----+----+----+----+-
0.0  0.1  0.2  0.3  0.4  0.5  0.6  0.7  0.8  0.9  1.0
```

(e) Repeat this entire exercise using a different statement.

Interpreting Odds

Often probabilities are stated in the media in terms of odds. For example, the following is a paragraph from Roxy Roxborough's *Odds and Ends* column in *USA Today*.

> EYE OF THE TIGER: Sensational PGA tour rookie Tiger Woods has been installed an early 9-5 favorite to win the most cash at the Skins Game at the Rancho La Quinta Golf Course in La Quinta, Calif., Nov. 30-Dec. 1. The charismatic Woods will have to contend with Fred Couples, the close 2-1 second choice, as well as the still dangerous Tom Watson, 3-1, and power stroker John Daly, 9-2.

We read that Tiger Woods is the favorite to win this golf tournament at a 9-5 odds. What does this mean?

An odds of an event is the ratio of the probability that the event *will not occur* to the probability that the event *will occur* . In our example, the event is "Tiger Woods will win". We read that the odds of this event are 9 to 5 or 9/5. This means that

$$ODDS = \frac{\text{Probability(Tiger Woods will not win)}}{\text{Probability(Tiger Woods will win)}} = \frac{9}{5}.$$

This means that it is more likely for Woods to lose the tournament than win the tournament.

How can we convert odds to probabilities? There is a simple recipe. If the odds of an event are stated as A to B (or $A - B$ or A/B), then the probability of the event is

$$\text{Probability(event)} = \frac{B}{B + A}.$$

So, for example, if the odds of Woods winning are 9-5 or 9/5, then the probability of Woods winning is

$$\text{Probability(event)} = \frac{5}{5 + 9} = .3571.$$

Let's interpret other odds. If I were to play a golf tournament, then I would be given high odds, say 1000-1, of winning. The probability that I would win the tournament would be given by

$$\text{Probability(winning)} = \frac{1}{1+1000} = .001.$$

I have a very small chance of winning the tournament. So high odds can be translated to small probabilities. As another example, if the odds that a horse will win a race is 1-1 (called even odds), the probability the horse will win is given by

$$\text{Probability(winning)} = \frac{1}{1+1} = .5.$$

So odds of 1-1 correspond to a 50% probability.

Activity 12-7: Who's Going to Win the Women's World Cup in Soccer?

In 1999, the Women's World Cup soccer championship was held. For sports events such as this one, the casinos will state odds on the different possible outcomes. Essentially, these odds are subjective probabilities which reflect the opinions of the people who place bets on these events.

In May 1999, an internet gambling site gave the following odds for these teams winning the 1999 Women's World Cup.

TEAM	ODDS	PROBABILITY
U.S.A	4-5	
China	5-2	
Norway	7-2	
Germany	8-1	
Brazil	15-1	
Denmark	20-1	
Russia	25-1	
Sweden	40-1	

(a) For each team in the table above, use the given odds to compute the probability that it will win the 1999 World Cup.

(b) Find the probability that the USA, China, or Denmark will win the championship.

(c) Find the probability that the USA will *not* win the championship.

(d) Actually, the numbers in the table are *betting odds*, not *true odds.* We won't explain how betting odds are obtained, but the odds and corresponding probabilities that you computed in the table don't satisfy basic probability rules.

If we were to add up the probabilities, what would you expect the sum to be? Now add up the numbers and check to see if you are right.

(e) On the World Wide Web or a newspaper, find the odds for some sporting event. Write the table of odds below and write a short paragraph explaining what the odds are telling you.

HOMEWORK ACTIVITIES

Activity 12-8: Specifying Sample Spaces (cont.)

In each of the following, an experiment is described and an *incorrect* sample space is stated. Explain why the sample space is incorrect in each case and write down a corrected sample space.

(a) A coin is tossed repeatedly until a head is observed. You record the number of tosses.

$$\text{Sample space} = \{1, 2, 3, 4\}$$

(b) You ask each of four students if she or he lives on-campus. You record the number of students who say that they live on-campus.

$$\text{Sample space} = \{1, 2, 3, 4\}$$

(c) Three books, labeled A, B, and C, are mixed up and placed on a shelf. You noticed the order of the books on the shelf.

$$\text{Sample space} = \{\text{ABC, ACB, BAC, BCA}\}$$

(d) You measure the time (in minutes) it takes a student to complete a multiple-choice exam.

$$\text{Sample space} = \{10, 15, 20, 25, 30, 35, 40\}$$

(e) Each day next week, you plan to buy one lottery ticket, and there is a small chance of winning \$100 on each ticket. You record the amount of money (in dollars) you win for all tickets you buy that week.

$$\text{Sample space} = \{100, 200, 300\}$$

Activity 12-9: Birthmonths.

Suppose four people are in a room. Is it possible that two people were born in the same month, not necessarily the same year?

(a) Guess at the probability that two persons have the same birthmonth.

(b) We can simulate this experiment using a deck of cards.

- Separate the deck into four groups by suit (♠, ♡, ♢, ♣). Remove the aces, so you will have four groups of 12 cards.
- Shuffle each group and then place one card from each group on your desk. Each card represents one birth month.
- Note if you have any matching cards (like two kings or two fives).
- Return the four cards to the groups.

(c) Repeat this experiment a total of 20 times. Record the number of times (out of 20) you found any matches, and found no matches in the table below. Also record in the table the proportion of experiments in which you found matches and the proportion in which you found no matches.

OUTCOME	COUNT	PROPORTION
Match		
No Match		

(d) How did the observed proportion of matching compare with your guess in part 1?

Activity 12-10: Odds in a Horse Race

Suppose that in the next Kentucky Derby, there are three favorites. The odds of "Lucky" winning is 2 to 1, "Best Shot" is going off at odds of 4 to 1, and "Streaky" has odds of 6 to 1. The table below gives four possible winners of the race and the stated odds.

WINNER	ODDS	PROBABILITY
"Lucky"	2 to 1	
"Best Shot"	4 to 1	
"Streaky"	6 to 1	
Another horse		

(a) For each of the three horses in the table, find the probability that he/she wins. Put your answers in the table.

(b) Based on your probabilities in part (a), find the probability that another horse wins the race.

(c) Find the probability that either "Lucky" or "Streaky" wins the race.

(d) Find the probability that "Best Shot" does not win the race.

Activity 12-11: Are the Probabilities Legitimate?

In each part below, a hypothetical set of probabilities is given. Check if it is an allowable set of probabilities. If a probability or group of probabilities is missing, put in numbers so that it is an allowable probability distribution.

(a) I pass through four traffic lights on my commute to school. The following table gives the probability distribution for the number of red lights that I will hit during my ride.

NUMBER OF RED LIGHTS	0	1	2	3	4
PROBABILITY	.2	.5	.2	.1	0

(b) How many years will it take me to complete my undergraduate degree? Based on my current knowledge, I think it will take 3, 4, 5, or 6 years with the following probabilities:

NUMBER OF YEARS	3	4	5	6
PROBABILITY	.1	.3	.4	.3

(c) Our church is planning a car wash to raise money for youth programs. Based on our success in raising money in past car washes, I construct the following probability distribution for the amount of money we will collect:

MONEY COLLECTED	Under \$20	\$20 – \$50	\$50– \$100	Over \$100
PROBABILITY	.2	.4		

Activity 12-12: Tomorrow's High Temperature? (cont.)

In Activity 11-10, you were asked to make intelligent guesses regarding the high temperature in your town on the next day. Record your probability estimates in the table below.

EVENT	PROBABILITY
over 60°	
between 40° and 60°	
under 40°	
under 60°	
over 70°	

Answer the questions below to check if your probabilities follow the basic rules.

(a) The probabilities for "over $60°$" and "under $60°$" should sum to _____.

(b) The probability for "between $40°$ and $60°$" should be (choose less than, equal to, or greater than) _____ the probability for "under $60°$".

(c) Based on your table, what should be the probability of "under $70°$"?

(d) Based on your table, what should be the probability that the high temperature tomorrow will *not* be between 40-$60°$?

(e) The probability for "under $60°$" should be (choose less than, equal to, or greater than) _____. the sum of the probabilities for "under $40°$" and "between $40°$ and $60°$".

Activity 12-13: Using Chip-in-Bowl Experiments (cont.)

In each part below, I am interested in assessing the probability of an event. I describe some bets and say which bets I prefer. Based on this information, circle the possible probabilities of the event. (More than one answer is possible.)

(a) I am interested in the probability that there will be more than 20 inches of snow in Findlay this winter. Consider the bets:

- BET 1: get $20 if there is more than 20 inches of snow, get nothing if there is less snow
- BET 2: get $20 if draw red in bowl with 5 red and 5 white chips, get 0 if draw white

I prefer BET 1. My probability of more than 20 inches of snow can be:

0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1

(b) What's my probability that I will win the Ohio lottery in my lifetime? Consider the bets:

- BET 1: get $20 if I win the lottery, get nothing if I don't win
- BET 2: get $20 if draw red in bowl with 1 red and 9 white chips, get 0 if draw white

I prefer BET 2. My probability of winning the lottery can be:

0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1

(c) I'm interested in my probability that man will land on the moon in the next 20 years. Consider the three bets:

- BET 1: get $20 if man lands on the moon, get nothing if we don't land on the moon
- BET 2: get $20 if draw red in bowl with 3 red and 7 white chips, get 0 if draw white
- BET 3: get $20 if draw red in bowl with 5 red and 5 white chips, get 0 if draw white

I prefer BET 1 to BET 2. Also, I prefer BET 3 to BET 1. My probability that man will land on the moon can be:

0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1

Activity 12-14: Roulette

One of the most popular casino games is Roulette. In this game, there is a wheel with 38 numbered metal pockets (1 to 36 plus 0 and 00). The wheel is spun moving a metal ball and the ball comes to rest in one of the 38 pockets. The wheel is balanced so that the ball is equally likely to fall in any one of the 38 possible numbers. You play this game by betting on various outcomes — you win if the particular outcome is spun.

What is the probability of winning ...

(a) if you bet on the number 32?

(b) if you bet on even numbers (2, 4, etc.)?

(c) if you bet on the first twelve numbers (1 to 12)?

(d) if you bet on four numbers, such as 20, 21, 23, 24?

Activity 12-15: Drawing a Ball Out of a Box

Suppose a box contains 4 red, 3 white, and 2 black balls. We can represent the box by the set {R, R, R, R, W, W, W, B, B}. Suppose that we choose one ball from the box at random.

(a) Find the probability that the ball you choose is red.

(b) Find the probability that the ball is *not* white.

(c) Find the probability that the ball is either red or white.

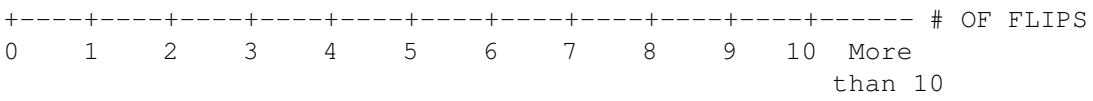
(d) You should note that the outcomes "choose red", "choose white", and "choose black" are not equally likely. Could you add balls to the box to make these three outcomes equally likely? Explain.

Activity 12-16: How Many Births Until a Girl?

Suppose that a couple decides to continue to have children until a girl is born. How many children will they have? We can simulate this experiment by means of coin flips. Suppose a single flip corresponds to the sex of a baby — heads is a boy and tails is a girl. You keep tossing until you get a tail. When you finally get a tail, record the total number of flips required. For example, if you roll a H, H, T, you got a tail (girl) on the third flip — the number of flips is 3. Perform this experiment 20 times and record your results for the 20 trials in the table below.

TRIAL	# OF FLIPS	TRIAL	# OF FLIPS
1		11	
2		12	
3		13	
4		14	
5		15	
6		16	
7		17	
8		18	
9		19	
10		20	

(a) Plot your 20 "# of flips" on the dotplot below.



(b) From looking at the dotplot, how many children does it take to get a girl, on the average?

(c) What is the (approximate) probability that at least 3 children are born?

(d) What is the probability that a girl is born on the first try?

Activity 12-17: A Simplified Lottery Game

Suppose that you play a two digit lottery game. A two digit number will be chosen at random from the 100 numbers 00, 01, 02, ..., 98, 99. You win $50 if the number you pick is the same as the winning number. You win $5 if the first digit of your number matches the first digit of the winning number. Otherwise, you win nothing.

Suppose you decide to choose the number 34.

(a) What is the probability that you win $50? That is, what is the probability the winning number is 34?

(b) What is the probability that you win $5? (Hint: count the number of two digit numbers that have the same first digit as 34.)

(c) Using the answers from (a) and (b), find the probability that you don't win.

Activity 12-18: Sitting in a Theater

Archie, Alice, Bob and Carla are going to a Broadway musical. Suppose that they are randomly assigned four adjacent seats in the theater. We're interested in the probability that Archie and Alice are sitting next to each other.

(a) One way of approximating this probability is by simulation:

- Take four cards from a playing deck — two aces and two cards with faces that are not aces. The aces will represent Archie and Alice and the other two cards Bob and Carla. Shuffle the four cards and then lay them down in a line. If the two aces are next to each other, Archie and Alice have adjacent seats.
- Repeat this process (shuffle the cards, lay them down, observe if the aces are adjacent) 20 times. Count the number of times the aces are adjacent. Approximate the probability of interest.

(b) Another way to compute this probability is by listing all of the possible outcomes of assigning sets to people. Let A and A represent Archie and Alice and B and B represent Bob and Carla. (We don't need to distinguish the people within a pair since we are only concerned if Archie and Alice are sitting next to each other.) Then there are six possible assignments of people to seats:

	Seat		
1	2	3	4
A	A	B	B
A	B	A	B
A	B	B	A
B	A	A	B
B	A	B	A
B	B	A	A

Each of the above six arrangements of people to seats is equally likely.

(i) Find the probability that Archie and Alice are in the first two seats.

(ii) Find the probability that Archie and Alice are in adjacent seats. Compare your answer with the approximate answer in part (a).

WRAP-UP

After completing this topic, you may think that probabilities are hard to compute. You're right — probabilities can be difficult to find for a number of reasons. It can be difficult to specify the sample space, and probabilities are hard to find when outcomes of experiments are complicated and when the outcomes in a sample space are not equally likely. The goal of this topic was primarily to introduce you to different methods for assigning probabilities. In this book we are more interested in interpreting tables of probabilities that have already been computed. The next topic will focus on the correct interpretation and summarization of a single probability table.

Topic 13: Probability Distributions

Introduction

In the last topic, we discussed methods for assigning probabilities to outcomes of a random experiment. In this topic, we'll discuss the special situation where the result of the random process is a number. A table which lists the possible numbers in the experiment and the corresponding probabilities is called a probability distribution. We will discuss how to graph and summarize this distribution. Many of the ideas will be familiar to you, since we discussed similar topics in the analysis of a data distribution in Topic 2.

PRELIMINARIES

1. What is your favorite Julia Roberts movie among *Pretty Woman*, *Hook*, *The Pelican Brief* and *Stepmom*?
2. Suppose 1000 people were asked to rate the movie *Hook* on a scale from 1 (the worst) to 10 (the best). What do you think would be an average rating of this movie?
3. What fraction of people do you think would give *Hook* a rating of 10 (the best)?
4. Suppose the identity of four babies gets confused in a hospital and the babies are randomly matched with the four mothers as they are sent home. How many correct matches of mothers with babies do you expect to get?
5. Suppose you are interested in collecting a series of six posters, where one is placed in a cereal box. You keep buying boxes of the same type of cereal until you get a complete set of six posters. How many boxes of cereal do you expect to purchase to get a complete set?

What is a Probability Distribution?

Suppose as in the previous two topics that we have some random process or experiment where the outcome is unknown. Examples of such processes are the result of a coin toss, the pick of a lottery

number, the age that a randomly selected college undergraduate student will first be married, or the height of a randomly chosen American male between the ages of 25 and 34. We assume that the outcome of the random process is a number. Three of the four examples are of this type: a lottery number pick is a number like 234, the age that a college student gets married is certainly a number, and the height of an American male (in inches) will typically be a number between 64 and 76.

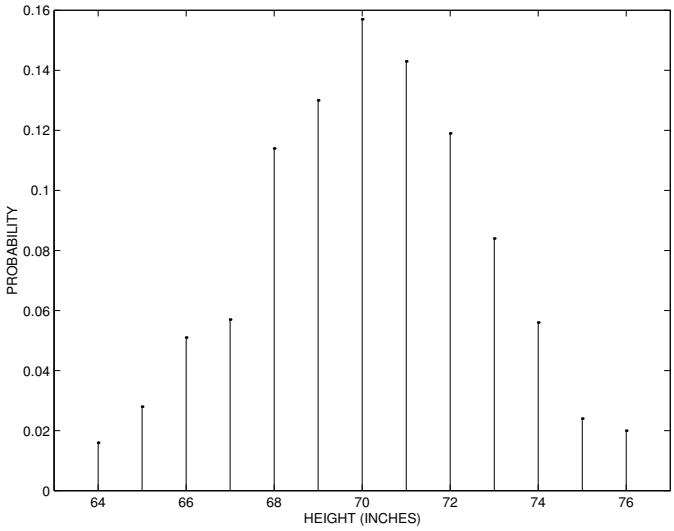
Let's consider the example of measuring the height of a randomly selected American male between the ages of 25 and 34. We describe probabilities of number outcomes (such as the man's height) by use of a **probability distribution**. This is a table which lists all of the possible number outcomes (the sample space) and the associated probabilities. In our example, if the height is measured to the nearest inch, then the sample space would consist of the whole numbers 64, 65, 66, ..., 76. A probability distribution for the heights, shown below, is a listing of these numbers together with a list of probabilities. These probabilities tell us the likelihood of finding different height values when a male is randomly selected. The chance that this man has a height of 64 inches or shorter is .016, the probability the man has a height of 65 inches is .028, and so on.

Height	Probability
64 and under	.016
65	.028
66	.051
67	.057
68	.114
69	.130
70	.157
71	.143
72	.119
73	.084
74	.056
75	.024
76 and over	.020

Probability distributions, just like probabilities, must satisfy some basic rules:

- We can't assign negative probabilities to any number outcomes. It is okay to give an outcome a zero probability — this just means that the particular outcome can not occur.
- The total probability of the sample space must be equal to one. So that means that the sum of the probabilities in a probability distribution must add up to one. It's a good idea to check this whenever you are given a probability distribution.

We can display the above probabilities by means of a line graph shown above. The different



height numbers are placed on the horizontal axis and a vertical line is drawn above each height corresponding to the probability.

From the line graph, we can learn quite a bit about the probability distribution of heights of American males. The features that we look for in the graph are analogous to the features of a data distribution that we discussed in Topic 2.

- **Shape of distribution** We see in this case that the probabilities are mound shaped — most of the probability is concentrated in the middle of the graph and there is little probabilitiy for small and large heights.

- **Typical values** We see from the line graph that the largest probabilities are assigned to the heights 69, 70, and 71. This means that the most likely height of a randomly selected male is about 70 inches, or 5 feet, 10 inches.

- **Spread of values** The probabilities are pretty spread out. All heights between 64 and 76 inches have significant probabilities. Heights close to 70 have probabilities in the 12-16 percent range and small and large heights have probabilities around 2 percent.

CLASSROOM ACTIVITIES

Activity 13-1: Ratings of Julia Roberts Movies

On the Internet Movie Database (http://us.imdb.com//), people are given the opportunity to rate movies that they have watched. The possible ratings are 1 (lowest) to 10 (highest). The distribution

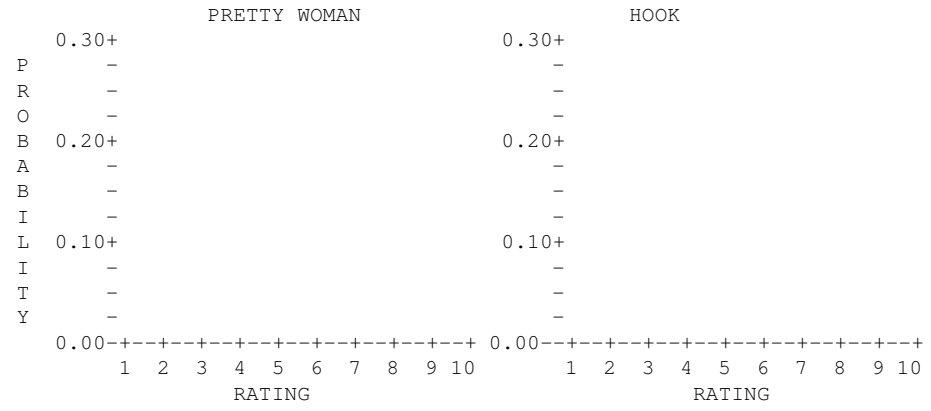
of ratings for four Julia Roberts movies *Pretty Woman*, *Hook*, *The Pelican Brief*, and *Stepmom* are given in the table below. Under the movie title, the column of the table lists the proportions of people that gave each of the 10 possible ratings.

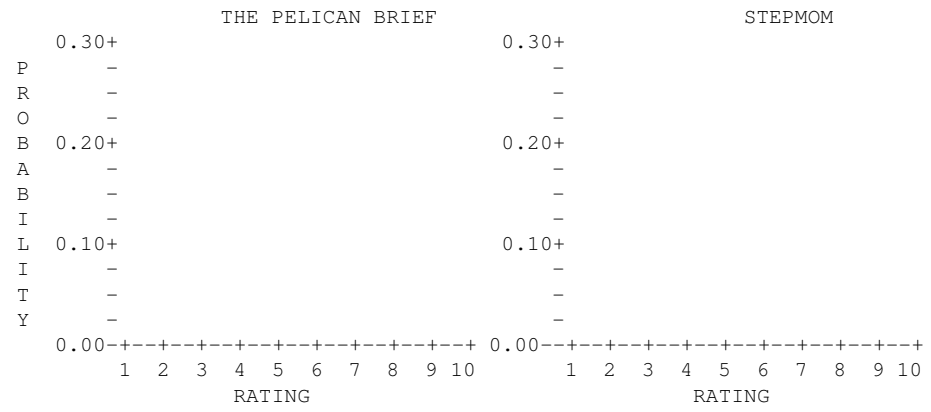
Rating	Pretty Woman	Hook	The Pelican Brief	Stepmom
1	.04	.05	.03	.06
2	.03	.05	.03	.03
3	.03	.06	.04	.04
4	.06	.08	.07	.04
5	.09	.13	.13	.07
6	.16	.18	.21	.09
7	.21	.20	.23	.18
8	.19	.14	.17	.21
9	.09	.06	.07	.14
10	.09	.05	.04	.15

We can view these proportions as approximate probabilities. If we consider a person who looks up this movie database and is interested in a particular Roberts movie, the numbers represent the approximate probabilities that the person will give each possible rating.

(a) Check if the probabilities for each movie actually correspond to real probabilities.

(b) Graph the probabilities for all four movies using line plots.





(c) Describe using a few sentences the picture of each of the probability distributions. (Talk about shape, typical values, and spread of the values.)

Summarizing a Probability Distribution

It is helpful to graph a probability distribution first to learn about some basic features of the distribution such as its shape and the location of likely and not-so-likely values. Next, as in the case of a data distribution, we want to compute several numbers that help to summarize the distribution and communicate the pattern that we see in the line graph.

Let's return to our probability distribution for the heights of American males. How tall are American men? Perhaps you think that a height of 6 feet (72 inches) or taller is notable. What is the probability that a man selected at random will be 72 inches or taller?

Look again at the probability table. If we are interested in the probability of 72 inches or larger, then we focus only on the heights that are 72 or bigger:

Height	Probability
72	.119
73	.084
74	.056
75	.024
76 and over	.020

We find this probability by adding the probabilities of 72, 73, 74, 75, and 76 and over:

$$\text{Prob(height is 72 or greater)} = .119 + .084 + .056 + .024 + .020 = .303.$$

The chance that a randomly selected American is 72 inches or taller is about 30 percent.

Similarly, we can compute other probabilities of interest:

- What's the probability that a height will be *at most* 68 inches? The phrase *at most* means that value (68) or smaller. We find this probability by adding up the probabilities corresponding to the heights 68 or smaller (64 and under, 65, 66, 67, 68):

$$\text{Prob(height is at most 68)} = .016 + .028 + .051 + .057 + .114 = .266.$$

- What is the probability that a random man's height will lie between 70 and 72 inches? To find this, we add up the probabilities corresponding to 70, 71 and 72:

$$\text{Prob(height is between 70 and 72)} = .157 + .143 + .119 = .419.$$

An Average Value of a Probability Distribution

Above we computed special probabilities that were informative about the distribution of men's heights. As in the case of a data distribution, we'd like to compute a single number which tells us the location of the center of the probability distribution.

We find the **average value** of a probability distribution as follows:

- For each outcome, we multiply the number by its corresponding probability – we'll call the results "products".

- We sum the products to get the average value.

The computation of the average value of the height probability distribution is illustrated in the table below. In each row of the table, we multiply the height by its probability — the results of this multiplication are placed in the "Product" column. We add up all the values in this column – the result, 70.058, is the average value.

Height	Probability	Product
64 and under	.016	$64 \times .016 = 1.024$
65	.028	1.820
66	.051	3.366
67	.057	3.819
68	.114	7.752
69	.130	8.970
70	.157	10.990
71	.143	10.153
72	.119	8.568
73	.084	6.132
74	.056	4.144
75	.024	1.800
76 and over	.020	1.520
SUM		70.058

What's the interpretation of this average value? We'll briefly describe two interpretations.

- The average is a *useful summary* of the probability distribution much like a mean or median is a good summary of a data distribution. Looking back at the line graph of the probability distribution of heights, note that 70.058 is approximately at the center of the distribution. When the probability distribution is approximately symmetric, as in this case, the average value will be located at the center of the distribution.

- The average can be viewed as a *long-run mean* when we simulate many values from the probability distribution. In our example, suppose that we are able to randomly select many men and measure their heights. The sample mean of all of these heights will be approximately equal to the average of the probability distribution.

  Let's illustrate this. Using the computer, I sampled the following 50 heights from the above probability distribution. These numbers represent the measured heights of 50 men who were randomly sampled from the population of American males of ages 35-44.

  74 74 71 74 70 75 73 68 75 74
  71 68 70 71 71 71 69 68 67 68
  70 73 75 72 68 68 73 71 64 67
  72 72 70 67 66 73 69 67 69 71
  71 72 67 70 71 67 73 70 71 64

  The mean of these 50 heights is

$$\frac{74 + 74 + 71 + ... + 71 + 64}{50} = 70.3.$$

Note that this sample mean, 70.3, is approximately equal to the average of the probability distribution 70.058.

Activity 13-2: Ratings of Julia Roberts Movies (cont.)

(a) For the movie *Pretty Woman*, what proportion of people gave it a high rating (8 or higher)? What proportion gave it a low rating (3 or lower)? Find the proportion of high and low ratings for the three other movies. Put your answers in the table that follows.

Movie	Proportion of Low Ratings	Proportion of High Ratings
Pretty Woman		
Hook		
The Pelican Brief		
Stepmom		

(b) For each movie, compute the average rating of the probability distribution. Put the averages in the table below.

Movie	Average Rating
Pretty Woman	
Hook	
The Pelican Brief	
Stepmom	

(c) From the numbers you computed above, which movie appears to be the most popular? Which appears to be the least popular?

(d) Could you conclude from this survey which of the four Roberts movies is the most popular among Americans who watch movies? Are the people who give ratings on the internet survey representative of all movie watchers? Discuss how this survey could be giving misleading information.

(e) Describe a better way of taking a survey to learn about the popularity of Roberts movies among all Americans who watch movies.

(f) Find the movie database on the World Wide Web and find two movies — one that you think is very popular and one that you think generally is not liked. Find the survey ratings for each movie. Convert the counts you find into approximate probabilities and compare the ratings.

Activity 13-3: The Minnesota Cash Lotto Game

Gopher 5 is the Minnesota cash lotto game. You play this game by picking five distinct numbers between 1 and 39. You win if you match 3, 4, or 5 of the five winning numbers that are drawn. The description of the game on the World Wide Web gives the following prizes (for a $1 play) and odds for the three winning possibilities:

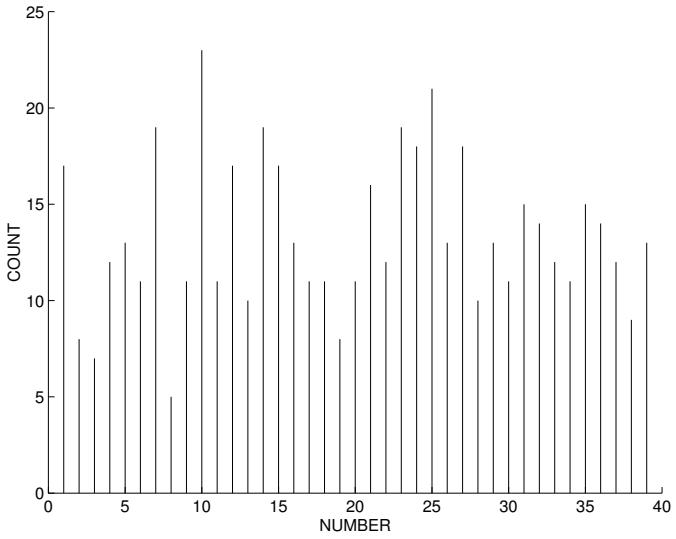
MATCH	PRIZE	ODDS
5 of 5	$100,000	575,757 to 1
4 of 5	$250	3,387 to 1
3 of 5	$10	103 to 1

One natural question to ask when you play this game is : are there "special" numbers that have a good probability of being chosen? The web site publishes all of the winning numbers that have been drawn for the last two years. For each of 104 weeks, five numbers were drawn for a total of $104 \times 5 = 520$ winning numbers.

(a) There are 39 possible numbers in the game. If a total of 520 winning numbers have been drawn, how many of each possibility (1, 2, ..., 39) do you expect to be drawn?

(b) The graph below shows the count of each winning number for the past two years. Are there particular numbers which have been drawn often, or more than you would expect? List these numbers.

(c) Are there particular numbers that have been drawn a relatively small number of times? If so, list these numbers.



(d) If you played this game tomorrow, which numbers would you select? Give any rationale for your selection.

One way to understand the benefit of playing this lottery game is to compute an average winning for a $1 play.

MATCH	PRIZE	ODDS	PROBABILITY	PRODUCT
5 of 5	$100,000	575,757 to 1		
4 of 5	$250	3,387 to 1		
3 of 5	$10	103 to 1		
None	0			
SUM				

(e) In the above table, use the stated odds to find the probability of each prize. Put your probabilities in the table.

(f) What is the probability of no match (and winning nothing)? Put this number in the PROBABILITY column.

(g) To find an average winning

- For each possibility, multiply the prize by the corresponding probability to get a product:

$$\text{PRODUCT} = \text{PRIZE} \times \text{PROBABILITY}$$

Put these products in the PRODUCT column.

- Sum the values in the product column — this is your average winning.

Suppose that you played this lottery game many times. The average you computed above is approximately the mean amount you win for all of these games.

(h) Is this a fair game? (Remember you played $1 to play.) Explain.

Activity 13-4: Mothers and Babies

On May 12, 1996, four boys were born in River City Hospital. However, due to a labeling problem, the identities of all four babies were confused. The nurses decided to send the babies home with the mothers using a random assignment, hoping that at least a couple of babies were sent home with their right mothers.

This raises an interesting question. If all the babies are randomly matched with mothers, how many matches to you expect to get? We'll learn about the likely number of matches using a simulation experiment.

(a) You will be given four red cards and four black cards; the reds and blacks have the same faces, such as 3, 4, 5, 6. We'll let the reds represent the mothers and the blacks the children. To perform one simulation of this random baby assignment

- Put the four red cards face up in a row on your desk.
- Shuffle the four black cards and place them in a row directly below the row of red cards.
- If the red card and black card beneath it are the same face, then you have a match of mother and baby. Count the total number of matches. (This will be a number from 0 to 4.)

(b) Repeat this simulation 20 times. Place your results in the table below:

Simulation	# of Matches	Simulation	# of Matches
1		11	
2		12	
3		13	
4		14	
5		15	
6		16	
7		17	
8		18	
9		19	
10		20	

(c) Summarize your simulation results in the below table. Put the number of times you observed 0 matches, 1 match, ..., 4 matches in the COUNT column.

# of Matches	COUNT
0	
1	
2	
3	
4	

(d) What is the (approximate) probability that all babies were given to the right mothers — that is, that the number of matches is 4?

(e) What is the most likely number of matches?

(f) Is it possible that the number of matches is 3? Why or why not?

(g) There is another way of finding probabilities for this problem. We can list all of the possible assignments of babies to mothers. Let's call the mothers A, B, C, D and the babies a, b, c, d. The 24 possible assignments are listed in the table below. Next to each assignment, I have written the number of matches.

MOTHERS				
A	B	C	D	
BABIES				# of Matches
a	b	c	d	4
a	b	d	c	2
a	c	b	d	2
a	c	d	b	1
a	d	b	c	1
a	d	c	b	2
b	a	c	d	2
b	a	d	c	0
b	c	a	d	1
b	c	d	a	0
b	d	a	c	0
b	d	c	a	1
c	a	b	d	1
c	a	d	b	0
c	b	a	d	2
c	b	d	a	1
c	d	a	b	0
c	d	b	a	0
d	a	b	c	0
d	a	c	b	1
d	b	a	c	1
d	b	c	a	2
d	c	a	b	0
d	c	b	a	0

If the babies are randomly matched with mothers, then each of the 24 possible assignments has probability 1/24. Using this table, find the probability that there is exactly 0 matches, 1 match, 2 matches, 3 matches, and 4 matches. Compare your answers with the probabilities you computed using simulation.

Activity 13-5: The Collector's Problem.

Suppose your favorite cereal is Wheaties. During the current spring promotion each box of cereal contains one poster of a famous woman athlete — either Steffi Graf (tennis), Venus Williams (tennis), Chamique Holdsclaw (basketball), Michelle Kwan (skating), Karrie Webb (golf), or Jackie

Joyner Kersee (track). You would like to collect a complete set of six posters. There are two ways you can get a complete set. You can keep on buying $2 boxes of Wheaties until you get posters for all six players. Alternately, you can buy a complete set of posters from a dealer specializing in sports collectibles for $20. What should you do? We'll use dice to simulate the process of collecting posters by buying boxes. By summarizing results of the simulations, we'll discover the better way to get a complete set.

(a) What is the smallest number of Wheaties boxes you have to buy to get a complete set?

(b) How many boxes do you think you have to buy so that you are pretty confident of getting a complete set?

(c) We'll discover how many boxes we need to buy using a simple simulation. Before doing this simulation, we have to make a couple of assumptions:

- Each of the six posters has the same chance of being in each cereal box.
- The posters in different cereal boxes are independent. This means (for example) if the first box contains a Steffi Graf poster, then that won't change the probability that the next box contains a Steffi Graf poster or any other poster.

(d) Here's how you'll do the simulation. Let each number on the die represent the poster for a particular player (1 – Graf, 2 – Williams, 3 – Holdsclaw , 4 – Kwan, 5 – Webb, 6 – Joyner Kersee). When you roll a die, tally the number that you observe on the scoresheet on the following page. Keep tallying in the boxes 1–6 until you have a tally in each of the six boxes (you've collected a complete set!) The number of Wheaties boxes you bought is the total number of tallies — record this number in the # BOXES column.

Repeat this simulation until you have 20 values of # BOXES.

EXPERIMENT

1	DIE # TALLY	1	2	3	4	5	6	# BOXES
2	DIE # TALLY	1	2	3	4	5	6	# BOXES
3	DIE # TALLY	1	2	3	4	5	6	# BOXES
4	DIE # TALLY	1	2	3	4	5	6	# BOXES
5	DIE # TALLY	1	2	3	4	5	6	# BOXES
6	DIE # TALLY	1	2	3	4	5	6	# BOXES
7	DIE # TALLY	1	2	3	4	5	6	# BOXES
8	DIE # TALLY	1	2	3	4	5	6	# BOXES
9	DIE # TALLY	1	2	3	4	5	6	# BOXES
10	DIE # TALLY	1	2	3	4	5	6	# BOXES
11	DIE # TALLY	1	2	3	4	5	6	# BOXES
12	DIE # TALLY	1	2	3	4	5	6	# BOXES
13	DIE # TALLY	1	2	3	4	5	6	# BOXES
14	DIE # TALLY	1	2	3	4	5	6	# BOXES
15	DIE # TALLY	1	2	3	4	5	6	# BOXES
16	DIE # TALLY	1	2	3	4	5	6	# BOXES
17	DIE # TALLY	1	2	3	4	5	6	# BOXES
18	DIE # TALLY	1	2	3	4	5	6	# BOXES
19	DIE # TALLY	1	2	3	4	5	6	# BOXES
20	DIE # TALLY	1	2	3	4	5	6	# BOXES

(e) Graph the values of # BOXES on the dotplot below.

```
-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 6   8  10  12  14  16  18  20  22  24  26  28  30  32 # BOXES
```

(f) What is the probability that you get a complete set by buying 10 or fewer boxes?

(g) What is the probability that it takes more than 20 boxes to get a complete set?

(h) Find the average number of boxes you need to buy to get a complete set. (Show how you are computing the average.)

(i) Back to our original question. Is it better to buy boxes or to just buy a complete set for $20. Explain.

Activity 13-6: Playoffs

Suppose that Boris and Joe are playing a championship chess match. Boris is generally a better player than Joe — if they played a large number of games, then Boris would win approximately 58% of the time.

We can simulate a chess game between Boris and Joe using two dice. Let one person roll a die representing Boris and a second person roll a die representing Joe. Both people roll their dice simultaneously. If the number on the first person's die is *at least as large* as the number on the second person's die, then Boris wins; otherwise Joe wins. For example, if

- person 1 rolls a 3 and person 2 rolls a 4 $\Rightarrow$ Joe wins
- person 1 rolls a 5 and person 2 rolls a 1 $\Rightarrow$ Boris wins
- person 1 rolls a 4 and person 2 rolls a 4 $\Rightarrow$ Boris wins

(a) Play this chess game 20 times. Put the tally of the winner of each game in the TALLY row. When you are finished, record in the table the total number of wins for Joe and Boris and the proportion of wins for each player. You should check that the proportion of wins for Boris is approximately the probability of winning, .58.

Single games

	Joe wins	Boris wins
Tally		
Count		
Proportion		

There is one problem with the above match consisting of a single chess game. Boris is the significantly better player, but there is a sizeable probability (42%) that he will lose this single game. Perhaps it would be better if Boris and Joe played a longer match. Suppose that they play a "best of three" match — the first player to win two chess games wins the match.

(b) Simulate this chess match 20 times. That is, first simulate one match using a pair of dice. Continue tossing dice until one player has won two games — that player is the winner of the match. Then repeat this process for a total of 20 matches. After each match, record the winner in the TALLY row of the table below. When you are done, put the total number of wins and the proportion of wins for each player in the table.

Best-of-three matches

	Joe wins	Boris wins
Tally		
Count		
Proportion		

(c) Since we're enjoying doing this so much, let's try a longer match. The two chess players play a "best-of-five" match where the winner is the first to win three games. Simulate this type of match 20 times and record the winners below. Again complete the COUNT and PROPORTION rows.

Best-of-five matches

	Joe wins	Boris wins
Tally		
Count		
Proportion		

(d) What have we learned?

In parts (a), (b), (c), we computed the approximate probability that Boris wins for single game, best-of-three, and best-of-five matches. Put these numbers in the table below.

Type of match	Probability Boris wins match
Single game	
Best-of-three	
Best-of-five	

Describe what happens to the probability of Boris winning when you have longer matches. What would happen if they play a best-of-seven match? Based on the information in the table, make an intelligent guess at the probability of Boris winning.

(e) In baseball, the World Series is a best-of-seven match between two baseball games that have roughly equal abilities. The winner of the World Series is called the "best team in baseball". Based on what you've learned above, is this a reasonable statement to make?

HOMEWORK ACTIVITIES

Activity 13-7: How Many Keys?

It's late at night and you are trying to open your front door. Your key chain has five keys, two of which will open this door. Since it is dark, you are trying keys in a random order, not trying the same key twice. How many keys will it take before you find the right one to get you in your house?

In this exercise, one can list all of the possible outcomes of this "experiment". Let K, K represent the correct keys and W, W, W the incorrect ones. There are 10 possible orderings of the five keys K,K,W,W,W which are listed in the table below.

1st try	2nd try	3rd try	4th try	5th try	# of Tries
K	K	W	W	W	1
K	W	K	W	W	
K	W	W	K	W	
K	W	W	W	K	
W	K	K	W	W	
W	K	W	K	W	
W	K	W	W	K	
W	W	K	K	W	
W	W	K	W	K	
W	W	W	K	K	

(a) If all possible key orderings are equally likely, what probability should be assigned to each ordering?

(b) For each key order, find the number of keys needed to get the door open. Write these numbers in the # of Tries column of the table. (The first value has been done for you.)

(c) Find the probability that the door opens on the first key. Also, find the probability that the door opens on the second key, on the third key, and on the fourth key. Place these probabilities in the below table.

# of Keys to open door	Probability
1	
2	
3	
4	

(d) What is the most likely number of keys needed to open the door?

(e) What is the probability that at most 3 keys are needed?

Activity 13-8: Tossing Four Coins

Suppose you toss a fair coin four times. We're interested in counting the number of heads. I used the computer to simulate this experiment 40 times. Each row of the table gives the results of the four tosses. For example, the first experiment resulted in the tosses H, T, T, T, where H stands for a head and T for a tail. Here the number of heads I got was 1. (I've put this in the table for you.)

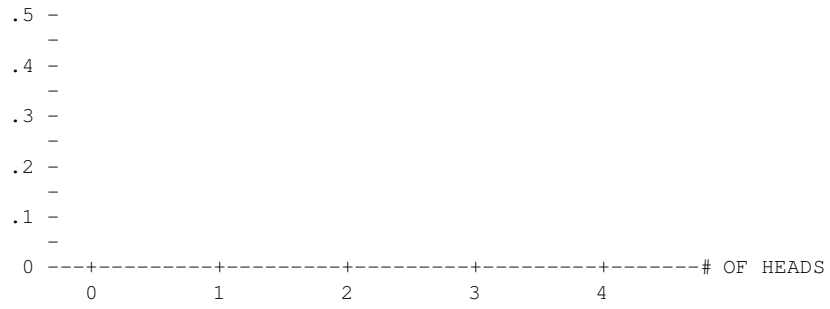
Tosses				# of Heads	Tosses				# of Heads
H	T	T	T	1	T	H	H	T	
H	T	H	T		T	T	H	H	
T	H	T	H		H	H	H	T	
T	T	H	T		T	H	T	H	
H	H	T	H		T	H	H	T	
H	H	H	H		T	T	T	T	
H	H	T	H		H	T	T	T	
T	T	H	T		T	T	H	H	
H	T	T	H		T	T	H	H	
T	H	H	T		T	T	H	T	
H	H	T	T		T	H	H	T	
T	T	H	H		H	T	T	H	
T	H	T	T		T	H	T	H	
H	H	T	T		H	T	T	H	
H	T	T	T		H	H	T	H	
H	H	H	H		T	H	H	T	
H	T	H	H		T	H	T	H	
T	T	T	T		H	H	T	H	
T	H	T	T		T	T	T	T	
T	T	H	H		H	T	H	H	

(a) For the remaining 39 experiments (19 on the left side of the table and 20 on the right), find the number of heads observed.

(b) In the below table, I have put all of the possibilities (outcomes) for the number of heads (0, 1, ..., 4). Find the counts and proportions of these five possibilities from the above simulations.

Outcome	Count	Proportion
0 Heads		
1 Head		
2 Heads		
3 Heads		
4 Heads		

(c) Graph the proportions (approximate probabilities) using a line graph below.



(d) From this table, what is the most likely outcome? What is its approximate probability?

(e) What is the probability of tossing at least 2 heads?

(f) What is the probability of tossing fewer than 3 heads?

(g) Suppose someone tells you that he's sure to get 2 heads if he tosses a coin four times. Explain why he is wrong.

Activity 13-9: Going to the Car Wash

Suppose a jar contains 3 dimes and 2 quarters. You grab two coins out of the jar to get money to wash the car. If we represent the contents of the jar by the set $\{10, 10, 10, 25, 25\}$, there are ten possible pairs of coins that you can choose (shown below), and each pair has the same chance of being chosen.

(10, 10) (10, 10) (10, 25) (10, 25) (10, 25)
(10, 10) (10, 25) (10, 25) (10, 25) (25, 25)

Consider the total amount of money (in cents) that you choose from the jar.

(a) Find the probability distribution for the amount of money you choose. Place the probabilities in the table below.

Money (cents)	Probability
20	
35	
50	

(b) If a car wash costs 25 cents, find the probability that you select enough money to pay for the wash.

(c) Find the "average" amount of money that you get from the jar.

(d) Find the probability that you select less than 50 cents from the jar.

Activity 13-10: Tossing a Die Until You Observe a 5 or 6

Suppose that you keep tossing a fair die until you get a 5 or 6. We had the computer do this experiment a large number of times; each time I keep track of how many tosses it takes to get a 5 or 6. (For example, in my first experiment, my tosses were 3, 1, 4, 6 — here it took 4 tosses to get a 6.) The results of 100 experiments are shown below:

```
4  8  1  1  1  3  2  1  9  8  2  1 13
3  7  3  1  2  1  1  2  6  2  3  1  1
1  4  8  1  3  2  1  1  3  4  1  1  1
2  7  2  1  4  3  1  2  4  4  3  5  2
1  1  7  1  2  2  3  1  2  4  6  1  7
2  3  4  1  2  2  1  8  1  1  1  6 11
1  1  2  1  1  1  1  4  3  3  2  2  1
3  1  4  3  2  2  4  5  5
```

(a) Find the probability distribution for the number of tosses.

(b) Find the probability that it takes more than two tosses to get a 5 or 6.

(c) Find the probability that it takes at most 2 tosses.

(d) Find the average number of tosses.

Activity 13-11: Baseball Takes Forever to Play?

How long is a baseball game? The *Baseball Weekly* lists complete accounts of major league baseball games. The description of an individual game includes the time of the game from the first pitch to the last out. I jotted down the times for 86 games that were played during a week in June, 1996. The following table gives the number of games that were under 2 1/2 hours in length, between 2 1/2 – 3 hours, and so on.

TIME OF GAME	COUNT	PROPORTION
under 2 1/2 hours	11	
2 1/2 – 3 hours	37	
3 – 3 1/2 hours	27	
3 1/2 – 4 hours	8	
over 4 hours	3	
TOTAL	86	

(a) Find the proportion of games in each time interval — place your answers in the PROPORTION column. These numbers can be viewed as probabilities. If you choose a game at random, then these proportions represent the (approximate) probabilities that the game is under 2 1/2 hours, between 2 1/2 – 3 hours, etc.

(b) If you go to a major league game, use the table to find the probability that the game lasts no longer than 3 hours.

(c) Find the probability that the game lasts over 3 1/2 hours.

(d) From the table, what is an "average" length of a major league game. Explain how you obtained this average.

Activity 13-12: Roulette (cont.)

Refer to the game of roulette described in Activity 12-14. Suppose that you place a $20 bet on the first twelve numbers 1–12. There are two possible outcomes of this game. Either the ball falls in a number from 1 to 12 and you win $40 (a 2 to 1 bet). Otherwise, you lose your $20. In other words, your *winnings* in this game in dollars is either 40 or -20. The table below states your two possible winnings and the associated probabilities of each outcome.

Winnings (in dollars)	Probability
40	12/38
-20	26/38

(a) Find the expected winnings by placing this particular bet. (Multiply each winning by its associated probability and add the products.)

(b) Interpret the expected winnings computed in part (a) if you were to place a large number of $20 bets on numbers 1–12.

(c) Specifically, suppose that you place 100 $20 bets on 1–12 on the roulette table. What do you expect is the *average* winnings of the 100 bets? What do you expect is the *total* winnings of all the bets?

Activity 13-13: One Big Bet or Many Small Bets?

Suppose that you have $20 to bet on the roulette wheel. You are interested on betting on the numbers 1–12. There are two possible betting strategies. We'll call the first strategy "one big bet". Here you bet the whole amount ($20) on a single roll of the wheel. In this case, you will either win $40 (if a 1–12 comes up) or lose $20 (if the spin of the wheel is a number besides 1–12).

An alternative strategy is to place "many small bets". Instead of betting $20 on one spin, you place 20 consecutive $1 bets on 20 spins of the wheel. For each $1 bet, you will either win $2 (if the spin is 1–12) or lose $1 (if 0, 00, 13–36 comes up). In this scenario, you could win $40 if you won all 20 bets or lose $20 if you lose all of the bets.

Which strategy is better? You are interested in your total winnings from the $20 that you bet during the day. By simulating many bets on roulette games, we can see what our winnings look like in the long run if we use the "one large bet" and "many small bets" strategies. By looking at the distribution of daily winnings for the two strategies, we can decide which method is bets.

First we will simulate many days in which the "one large bet" strategy is used. Suppose on 100 consecutive days, you bet $20 on a single bet of 1–12. I simulated this game on the computer and kept track of the daily winnings. The table below gives a count table of your winnings for the 100 days.

Winnings of one large bet for 100 days

WINNINGS (in dollars)	COUNT	PROBABILITY
40	35	
-20	65	
TOTAL	100	

Next we simulate the "many small bets" strategy. Each day, 20 $1 bets are placed and, when the day is through, the total winnings are recorded. I repeat this process for 100 days — a count table of the daily winnings is below.

Winnings of many small bets for 100 days

WINNINGS (in dollars)	COUNT	PROBABILITY
13	2	
10	2	
7	8	
4	5	
1	17	
-2	18	
-5	22	
-8	14	
-11	7	
-14	4	
-20	1	
TOTAL	100	

(a) For each table, compute the proportions of each possible value of WINNINGS and place your values in the table. These proportions represent the (approximate) probabilities of these different winning values.

(b) Compare the distributions of winnings for the two strategies.

(c) For the "one large bet" strategy, what was the most likely daily winnings? What is the probability that you are a winner on a single day?

(d) For the "many small bets" strategy, what is the most likely winnings after a single day? What is the probability that you are a winner using this strategy?

(e) From your work in parts (a)–(d), which strategy (one large bet or many small bets) do you prefer? Why?

WRAP-UP

In this topic, we have discussed methods for graphing and summarizing probability distributions. We will later see that these methods will be useful in statistical inference. Unknown population quantities such as proportions will have associated probability distributions and we will perform inferences about the population by summarizing the relevant probability distribution. In this topic, we have considered the case where only one random outcome is of interest. In the next topic we will consider probabilities when there are two numerical outcomes of interest.

Topic 14: Two-Way Probability Tables

Introduction

In Topic 13, we focused on computing probabilities in the case in which only one outcome was observed. Suppose instead that we observe two outcomes in an experiment. For example, if we roll two dice, we might observe the larger of the two rolls and the sum of the two dice.

We describe probabilities for two outcomes using a **two-way probability table.** The rows of the table correspond to possible values of the first outcome and the columns correspond to values of the second outcome. The numbers in the body of the table are the probabilities. In the dice example, a particular probability in the table will tell us the probability that the larger roll is 5 and the sum of the dice is 7. In this topic, we will get some experience constructing and interpreting this type of probability table.

PRELIMINARIES

1. If you roll one die, what are the possible outcomes?

2. What is the probability of rolling a 1? Give an explanation for your answer.

3. If you roll two dice, say a red die and a green die, how many possible outcomes are there? List all of the outcomes.

4. Suppose you roll two dice and observe the sum of the two numbers. (For example, if you roll a 2 and a 3, the sum is 5.) List all of the possible values of the sum of the two dice.

5. In the 1992 Presidential Election, do you think the proportion of people of voting age who actually voted was under 50% or over 50%?

6. Make a guess at the proportion of people that you think voted in the 1992 election.

7. What proportion of young voters (age 18-20) do you think voted in the 1992 election?

8. What proportion of middle-age voters (age 35-44) do you think voted in the 1992 election?

A Two-Way Probability Table

Suppose we have a random process where an outcome is observed and two things are measured. For example, suppose we toss a fair coin three times and we observe the sequence

H T H

where H is a head and T a tail. Suppose we record

- the number of heads
- the number of runs in the sequence (a run is a sequence of heads, like HHH, or a sequence of tails, like TT)

In the three tosses above, we observe 2 heads and 3 runs in the sequence.

We are interested in talking about probabilities involving both measurements "number of heads" and "number of runs". These are described as **joint probabilities**, since they reflect the likelihoods of the outcomes of two variables.

To construct this type of probability distribution, we first describe the collection of possible outcomes for the two variables. The number of heads in three tosses could be 0, 1, 2, or 3, and the number of runs in a sequence could be 1, 2, or 3. We represent these outcomes by the following **two-way table:**

	# OF HEADS			
# OF RUNS	0	1	2	3
1				
2				
3				

Next we have to place probabilities in the above table. If we toss a coin three times, there are 8 possible outcomes. Since the coin is fair, each of the outcomes has the same probability. In the table below, we list the eight outcomes, the number of heads and the number of runs in the outcome and the probability of the outcome.

OUTCOME	# OF HEADS	# OF RUNS	PROBABILITY
H H H	3	1	1/8
H H T	2	2	1/8
H T H	2	3	1/8
H T T	1	2	1/8
T H H	2	2	1/8
T H T	1	3	1/8
T T H	1	2	1/8
T T T	0	1	1/8

Now we are ready to fill in the probability table. Start with the box in the first row and first column. What is the probability that the number of runs is equal to 1 *and* the number of heads is equal to 0? Looking at the outcome table, we see that this happens once (for outcome TTT). So the probability of 1 run and 0 heads is equal to the probability of TTT, which is 1/8. What's the probability of 1 run and 1 head? We see from the outcome table that this never happens, so the probability in this box is 0. Next look at the box in the second row and second column. To find the probability of 2 runs and 1 head, we see that this happens twice in the outcome table (HTT and TTH). So the probability in this box is 2/8. If we continue this for all boxes, we get the following probability table.

	# OF HEADS			
# OF RUNS	0	1	2	3
1	1/8	0	0	1/8
2	0	2/8	2/8	0
3	0	1/8	1/8	0

For the following questions, it might be helpful to convert this probability table to a count table. Suppose that we tossed three coins 800 times. Then we would expect to get 1 run and 0 head 1/8th of the experiments, or 100 times, we expect to get 2 runs and 1 head 2/8th of the time, or 200 experiments, and so on. By converting probabilities to counts, we get the following count table. This represents what we think would happen if we did this coin experiment many times. Note that I have added an extra row and extra column to the table. The TOTAL column gives the number of counts in each row and the TOTAL row gives the number of counts in each column.

	# OF HEADS				
# OF RUNS	0	1	2	3	TOTAL
1	100	0	0	100	200
2	0	200	200	0	400
3	0	100	100	0	200
TOTAL	100	300	300	100	800

Marginal and Conditional Probabilities

Using the two-way count table, we'll ask some questions which explore the connection between the number of heads and the number of runs in this experiment.

- **Probabilities About Number of Runs**

  If you toss three coins, how many runs are likely to occur? Look at the TOTAL column of the table. This tells us that, out of 800 experiments, we expect to get 200 "one run", 400 "two runs", and 200 "three runs". The most likely possibility is "two runs" which has a probability of 400/800 = .5. We call this a **marginal probability** which is a probability about one of the two variables.

- **Conditional Probabilities of Number of Heads Given Number of Runs**

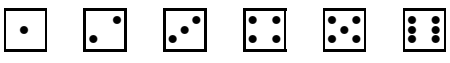
  If we are told that we get two runs in our coin tossing, what does that tell you about the number of heads? Look only at the "two runs" row of the table. Two runs happened 400 times in our hypothetical experiments. Of these 400 counts, only 1 and 2 heads occurred with respective frequencies 200 and 200. So if we are given that two runs occur, then we know that the only two possibilities are 1 and 2 heads with respective probabilities 200/400 and 200/400. What we have just done is compute a **conditional probability**. We've computed the probability of a certain number of heads *conditional* on the information that there were two runs in the sequence.

- **Conditional Probabilities of Number of Runs Given Number of Heads**

  What if we are told that the experiment resulted in exactly 2 heads? Have you learned anything about the number of runs in the sequence? Focus now on the "2 heads" column of the table. The number of 1 run, 2 runs, and 3 runs in this column are respectively 0, 200, and 100. So if we know that there were 2 heads in the sequence, then the probabilities of 1 run, 2 runs, and 3 runs are 0/300, 200/300, and 100/300. So it is most likely that 2 runs will occur, and 1 run in the sequence is impossible. Here we are computing probabilities of number of runs conditional on the knowledge that we have gotten two heads.

IN-CLASS ACTIVITIES

Activity 14-1: Rolling Dice



You will be given two dice to roll for this activity. Roll two dice 20 times — each time you roll two dice, record two numbers

- the sum of the two numbers (the "sum of dice")
- the larger of the two numbers (the "maximum roll")

Tally each of the 20 results in the two-way table below. For example, if you roll a 4 and a 5, the sum is 9 and the maximum is 5 — you put a tally mark (|) in the box in the row which corresponds to a sum of 9 and the column which corresponds to a maximum of 5.

	MAXIMUM ROLL						
SUM OF DICE	1	2	3	4	5	6	TOTAL
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
TOTAL							

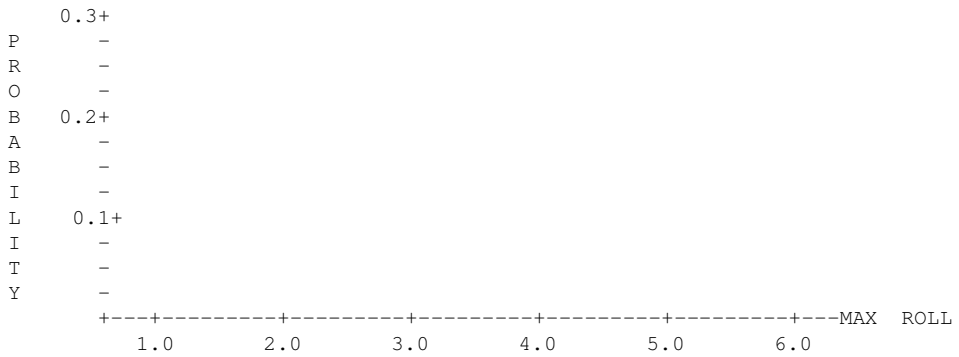
We can learn more about the relationship between the sum of the dice and the maximum roll by doing the experiment more times. On the computer, I simulated 5000 dice rolls — the counts of the different results are presented below. We will use this table to find different probabilities of interest.

	MAXIMUM ROLL						
SUM OF DICE	1	2	3	4	5	6	TOTAL
2	134	0	0	0	0	0	134
3	0	273	0	0	0	0	273
4	0	152	291	0	0	0	443
5	0	0	277	284	0	0	561
6	0	0	145	291	284	0	720
7	0	0	0	267	282	292	841
8	0	0	0	138	270	273	681
9	0	0	0	0	297	251	548
10	0	0	0	0	135	257	392
11	0	0	0	0	0	271	271
12	0	0	0	0	0	136	136
TOTAL	134	425	713	980	1268	1480	5000

(a) Using the column totals given in the bottom of the table, find the (approximate) probabilities that the maximum roll of the two dice is 1, 2, 3, 4, 5, 6. Place the probabilities in the table below.

MAXIMUM ROLL	PROBABILITY
1	
2	
3	
4	
5	
6	

(b) Graph the probabilities by a line graph below. What does the graph tell you about the relative chances of getting a 1, 2, etc. as the maximum roll?



(c) Find the probability that the maximum roll is 1 or 2.

(d) Find the probability that the maximum roll is at most 4.

For the remaining questions, look again at the two-way table of 5000 simulations of rolls.

(e) With regards to the sum of the two dice, which outcome is most likely? What is the probability of this outcome? (Look at the row totals on the right.)

(f) What values for the sum of two dice are relatively unlikely to occur?

(g) How many times did we get a maximum roll of 6 *and* the sum equal to 7? What is the probability of this possibility?

(h) What is the probability that the maximum roll is 4 and the sum of the two dice is 4?

(i) If the sum of the two dice happens to be 7, what's the probability that the maximum roll was 4? (Here you only are interested in the rolls in which the sum is 7 — only look at the row of the table corresponding to "sum equal to 7". In this row, find the proportion of times the maximum roll is 4.)

(j) If the sum of the two dice is 8, which values of the maximum roll are most likely? (Only look at the "sum equal to 8" row of the table.)

(k) In the above two parts, we're interested in the chances of different maximum rolls given values of the sum. Let's turn this around. Suppose that you're told that the maximum roll is 4. (Look at the "max equal to 4" column of the table.) What values for the sum are possible? List all of the possible values for the sum and find their associated probabilities.

Sum of Dice				
Probability				

(l) What's the probability of getting a sum of dice equal to 7 if you know the maximum roll on the dice was 5?

Activity 14-2: Voting Behavior in the Presidential Election

In this activity, we'll look at how people in different age groups voted in the 1992 Presidential Election in which Bill Clinton defeated George Bush.

(a) I've listed eight age categories below. Circle the age group that you think was *most likely* to vote in the 1992 election. Also circle the age group that you think was *least likely* to vote.

Most likely to vote

18-20	21-24	25-34	35-44	45-54	55-64	65-74	75 and over

Least likely to vote

18-20	21-24	25-34	35-44	45-54	55-64	65-74	75 and over

The *count table* below gives the number of people who voted and who did not vote in the 1992 election for different age groups.

Counts

AGE	VOTE	DIDN'T VOTE	TOTAL
18-20	3749	5978	9727
21-24	6693	7951	14644
25-34	22120	19483	41603
35-44	25269	14447	39716
45-54	19292	8766	28058
55-64	15107	5982	21089
65-74	13607	4838	18445
75 and over	8030	4371	12401
TOTAL	113867	71816	185683

Looking at the table ...

(b) If a voting-age person was selected at random, what's the probability that he was between 21-24 and didn't vote in the election?

(c) What is the probability the person is between 18 and 20?

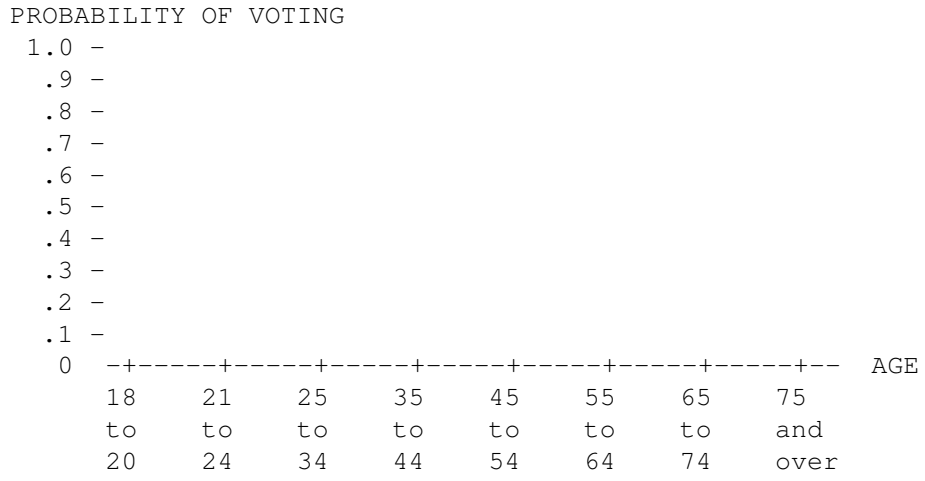
(d) What is the probability the person is between 21 and 44?

(e) Find the probability that a person voted in the election. Compare this answer with your guess in the Preliminaries section of this topic.

(f) Suppose you know the person is between 18-20 — what is the probability that he voted?

(g) For each age group, compute the probability of voting. Put your answers in the table below and graph your values on the below grid.

AGE	PROB(VOTE)
18-20	
21-24	
25-34	
35-44	
45-54	
55-64	
65-74	
75 and over	



(h) Describe in words what you have learned. What is the general pattern of voting among Americans in this election and how does voting depend on age?

Activity 14-3: Playing Yahtzee.

In this activity we play a simplified version of the dice game Yahtzee. One rolls a set of five dice three times to get particular patterns on the dice. We will consider nine possible patterns. The first eight will be described in increasing order of their value. A "one-pair" (abbreviated "1P") is a sequence of rolls such as 1, 2, 2, 3, 4, where exactly one number (here 2) occurs twice. A more valuable pattern, a "two-pair" (2P), is when two numbers each occur twice, such as 2, 3, 4, 3, 4. A "3-of-a-kind' (3K) is when exactly one number occurs three times, such as 5, 4, 4, 4, 6. A "full house" (FH) is when one number occurs twice and a second number occurs three times, such as 6, 6, 3, 3, 6. Also, sequences of consecutive die numbers are of interest. A "small straight" (SS) is

when exactly four numbers occur in sequence such as 2, 3, 4, 5, 5. A "large straight" (LS) is when five numbers occur in sequence, such as 1, 2, 3, 4, 5. A "4-of-a-kind" (4K) is the pattern where one die number occurs four times such as 5, 5, 5, 3, 5 and a "yahtzee" (Y) is when the same number appears on all five dice. If none of these patterns occur, we record a "nothing" — this pattern is the least valuable.

A player wishes to obtain the most valuable pattern of dice using three rolls. After each roll, she has the option of saving particular dice and only rolling the remaining.

(a) Practice tossing five dice a number of times (say, 20). Keep track of the patterns that you observe by tallying (drawing "|") in the table below.

PATTERN	0	1P	2P	3K	FH	SS	LS	4K	Y
TALLY									

(b) What was the most common pattern you saw?

(c) Were there any patterns that you did not see? If so, which ones?

(d) On the computer I rolled 5 dice 1000 times and recorded the pattern each time. The results of the 1000 tosses are given in the table below.

PATTERN	0	1P	2P	3K	FH	SS	LS	4K	Y
COUNT	28	388	235	149	27	117	34	22	0

From this table, write down the nine outcomes in order from the most likely to the least likely. Next to each outcome, write down its probability (using the counts in the table).

(e) Let's now consider playing the simplified Yahtzee game that consists of three rolls. Suppose that I play using the following strategy. For the first and second rolls, I will save die numbers that are repeated and reroll the remaining dice. The exception to this rule is that the numbers

of a small or large straight will be kept. On the computer, I play this Yahtzee game with this strategy 1000 times. Each time the game is played two outcomes are recorded — the pattern of the first roll and the pattern at the end of the game (after three rolls). The results of these 1000 games are recorded in the following two-way table. To help understand this table, note the number 100 which occurs at the "1P" row and "2P" column. This means that, for 100 games, the first roll of the Yahtzee game was one pair and the final roll was 2-pair.

	FINAL ROLL									
FIRST ROLL	0	1P	2P	3K	FH	SS	LS	4K	Y	TOTAL
0	0	1	13	3	2	5	1	3	0	28
1P	0	17	100	80	99	17	5	60	10	388
2P	0	0	97	0	138	0	0	0	0	235
3K	0	0	0	43	31	0	0	66	9	149
FH	0	0	0	0	27	0	0	0	0	27
SS	0	0	0	0	0	73	44	0	0	117
LS	0	0	0	0	0	0	34	0	0	34
4K	0	0	0	0	0	0	0	17	5	22
Y	0	0	0	0	0	0	0	0	0	0
TOTAL	0	18	210	126	297	95	84	146	24	1000

(f) Look at the last row of the table. This row gives the counts for all patterns for the final roll. As before, write down the patterns in order from most likely to least likely. Next to each pattern write the associated probability.

(g) How many times did I roll first a small straight (SS) and finished with a large straight?

(h) How many times did I first roll a 2-pair (2P) and finished with a full house?

(i) If I get a 3-of-a-kind on the first roll, what patterns are possible on the final roll? (These are the ones which have counts that are not 0.) Given an initial 3-of-a-kind, find the probability of each possible outcome at the end.

(j) If I get a small straight first, what is the chance of getting a large straight after the final roll?

Independent Events

Suppose that you are playing bridge. Out of a bridge deck of 52 cards, you are dealt 13 cards. You are interested in the number of points that you hold in your hand. In bridge, aces count 4 points, kings 3 points, queens 2 points, and jacks 1 point. All remaining cards are not given any points. Suppose that during a particular year, you are dealt 1000 hands.

For each of the 1000 hands, you record

- the number of aces
- the total number of points

In the following table, we classify the 1000 hands by the number of points and the number of aces.

	Number of Aces			
Number of Points	0	1	2	3
0-7	191	98	0	0
8-12	78	281	81	2
13-20	8	93	127	36
21-40	0	0	2	3

To motivate the notion of independence, let's compute two probabilities.

- What is the probability that you get between 8-12 points?

  Looking at the two-way table, we see that, out of the 1000 hands, we obtained between 8-12 points a total of $78 + 281 + 81 + 2 = 442$ times. So

  $$\text{Prob}(8 - 12 \text{ points}) = \frac{442}{1000} = .442.$$

- If your hand contains 2 aces, what is the chance that you have between 8-12 points?

  If we know the hand contains 2 aces, then we only consider the counts in the "2 aces" column of the table. We obtained 2 aces a total of $81 + 127 + 2 = 210$ times, and 81 of those times we also obtained 8-12 points. So

  $$\text{Prob}(8 - 12 \text{ points knowing we have 2 aces}) = \frac{81}{210} = .386.$$

Consider the two events

- "you get between 8-12 points"
- "you get 2 aces"

We say that these two events are **independent** if the knowledge of one event does not change your beliefs about the second event. In this case, your degree of belief about "8-12 points" is measured by the probability of "8-12 points". To see if the event "8-12 points" is independent of "2 aces", you see if the probability of "8-12 points" changes if you know that you have "2 aces". If this probability *does not* change, the two events are independent. Otherwise, we say that the two events are **dependent**.

In this case, we computed that

- the probability of "8-12 points" is .442
- the probability of "8-12 points" knowing that we have "2 aces" is .386

Since these two probabilities are different (.442 is not equal to .386), the two events "8-12 points" and "2 aces" are dependent.

Activity 14-4: Independent Events.

Let's consider a different experiment. Suppose you toss a fair coin two times. You keep track of the result of the first toss (head or tails) and the result of the second toss. This procedure is repeated 1000 times; the table below records the outcomes of the 1000 experiments.

	Second Toss	
First Toss	H	T
H	261	234
T	266	239

Consider the two events

- "the first toss is heads"
- "the second toss is tails"

(a) Using the table, find the probability that the second toss is tails.

(b) Find the probability that the second toss is tails if you know the first toss is heads.

(c) Compare the answers to parts (a) and (b). Are they approximately equal? Use this answer to decide if these two events are independent.

To further illustrate the notion of independence, consider the event

"you get an A on the next test in this class" (or "Ace" for short)

(d) Give your probability for "Ace".

For parts (e)-(h) below, I list another event, called "B". For each event B,

- give your new probability for "Ace" if you know that event B is true
- decide if "Ace" and this event B are independent.

We will do one of these for you. Suppose you are good at statistics and you think you will "Ace" a test with probability .9. Consider the event B: "You take the test with a bad cold".

- If this event is true (You have a bad cold), then it would adversely affect your performance on the test – your probability of "Ace" would drop to, say, .8.
- Since "bad cold" changes your probability of "Ace", then the events "Ace" and "bad cold" are **dependent**.

(e) B: "The next test is easy"

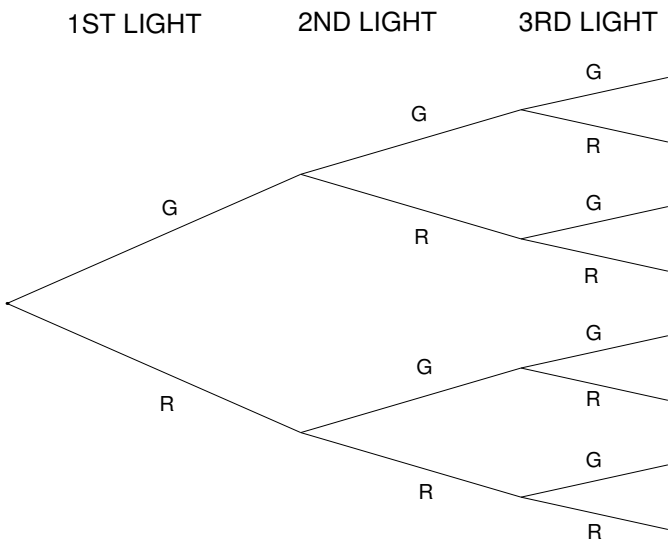
(f) B: "You got one hour of sleep the night before the test."

(g) B: "The weather is especially cold on test day."

(h) B: "You found a ride home this weekend"

Multiplying to Find Probabilities of "And" Events

Suppose that on your drive to school, you go through three traffic lights. Based on past experience, you believe that the probability that you will be stopped by a red light at the first traffic light is .3, by a red light at the second light is .5, and by a red light at the third light is .2.



Suppose you record whether you encounter a red light (R) or a green light (G) at each of the first, second, and third lights. All of the possible outcomes can be displayed by means of the tree diagram below.

The first set of branches on the left represents the possible states of the first light, the second set of branches represents the states of the second light, and so on.

Label each branch of the tree on the diagram above with its corresponding probability. For example, put the number .3 next to the letter R on the leftmost branch. This indicates that the probability of getting a red on the first light is .3. When you are done, you should have labeled all 16 branches.

Suppose you are interested in the probability that you hit a green on the first light and get stopped by a red on the second and third lights. In other words, you wish to find

$$\text{Prob(G and R and R)}$$

Suppose that you believe that the events "Green on first light" and "Red on second light" are independent. This means that the chance that you get a red light on the second light won't be affected by what happens on the first light. (Actually, in some cases, this may be a unreasonable assumption. Can you think of a scenario where the outcomes of traffic lights are not independent?) Likewise, you believe that the events "Green on first light" and "Red on third light" are independent, and also the results of the second and third lights are independent.

If you can make the above independence assumption, then you can find the above "and" type of probability by multiplying the individual probabilities. So the above probability is given by

$$\text{Prob(G and R and R)} = \text{Prob(G) Prob(R) Prob(R)} = (.7)\ (.5)\ (.2)$$

By a similar argument, the probability of going through all lights is the probability of getting three green lights:

$$\text{Prob(G and G and G)} = \text{Prob(G) Prob(G) Prob(G)} = (.7)\ (.5)\ (.8)$$

Activity 14-5: Traffic Lights

(a) Find the probability that you get a red light on all three lights.

(b) Find the probability that you get a red on the first and third and a green on the second light.

(c) Find the probability of each of the eight possible results GGG, GGR, GRG, etc. Put your answers to the right of the corresponding tree branch.

Suppose you are interested in the probability of getting a red at exactly one light. You find this using three steps:

- You write down all of the possible ways of getting stopped by exactly one red light. Here there are three possible ways: RGG, GRG, GGR.
- You find the probability of each outcome. That is you find the probability of RGG, the probability of GRG, and the probability of GGR.
- You add the probabilities to get the answer.

(d) Using these method, find the probability of getting stopped by one red light.

(e) Find the probability of getting stopped by exactly two red lights.

HOMEWORK ACTIVITIES

Activity 14-6: Live Births by Race and Age of Mother

Of 1000 live births in the United States in 1991 born to white and black mothers, the table below gives the expected number of births classified by race and the age of the mother. For example, note

that the expected number of white births of mothers between 20–24 is 212 — this means that 212 out of 1000 or approximately 21% of the births were in this category. If we divide these numbers in the table by 1000, the fractions can be interpreted as probabilities. If a birth in 1991 born to either a white or black mother is selected at random, the counts divided by 1000 give the probabilities of the birth falling in the different categories.

	Age of Mother					
Race	< 20	20–24	25–29	30–34	35–39	40 and up
White	91	212	255	188	69	11
Black	40	56	42	25	9	2

(a) Find the probability that a birth is from a white mother.

(b) Find the probability that the mother's age is between 30-34.

(c) There can be more risk in a baby born to a mother who is 30 or over. Find the probability of a mother 30 and over.

(d) Find the probability that a white mother is under 20.

(e) Find the probability that a black mother is under 20.

(f) From the probabilities you computed in parts (d) and (e), is there any relationship between race and the age of the mother? Explain.

Activity 14-7: Rolling Dice (cont.)

Suppose that you roll a white die and a green die. One outcome of this experiment is $\{2$ on white, 3 on green$\}$. There are 36 possible outcomes of this experiment and they are all equally likely. If we roll the dice 36 times, the table below displays the expected number of counts of each outcome.

	Roll of Green					
Roll of White	1	2	3	4	5	6
1	1	1	1	1	1	1
2	1	1	1	1	1	1
3	1	1	1	1	1	1
4	1	1	1	1	1	1
5	1	1	1	1	1	1
6	1	1	1	1	1	1

Find the probability that

(a) the green die is 5.

(b) the white die is larger than 4.

(c) the numbers on the white and green dice are the same.

(d) the sum of the dice is 8.

(e) the number on the white die is greater than the number on the green die

Activity 14-8: Participation in College Sports by Gender

In the school year 1995-1996, there were a total of 104,207 participants in the NCAA college sports of basketball, cross country, golf, soccer and tennis. The table below classifies these participants by gender and sport.

	DISCIPLINE				
GENDER	Basketball	Cross Country	Golf	Soccer	Tennis
Male	15160	10113	7163	16885	7961
Female	13343	9949	2083	13394	8156

(a) Suppose you choose one of these athletes at random. What's the chance that she is a woman golfer?

(b) In these five sports, what proportion of athletes are women?

(c) In which sport (among the five listed) is there the greatest participation? Why?

(d) For each of the five sports, compute the proportion of woman athletes. Looking at these proportions, are there any disciplines which appear to have a relatively high or low proportion of females?

Activity 14-9: Time of Baseball Game and Runs Scored

Activity 13-11 investigated the lengths of baseball games. What variables during a game influence its length? You might think of a number of variables that make a baseball game short or long, such as the number of pitches thrown, the number of hits, or the number of runs in the game. Here we look at the relationship between the length of a game and the total number of runs scored. For each of 86 games recorded in the *Baseball Weekly*, I jot down the elapsed time (from first pitch to last out) and the runs scored. This data is summarized in the following two-way table. For a particular time interval and range of runs scored, the table gives the proportion of the 86 games in that category. For example, we see from the table that the proportion of games which lasted less than 170 minutes and in which between 10–15 runs were scored was .256 or approximately 26%.

	Runs Scored		
Time of Game	1–9	10–15	> 15
< 170 minutes	.256	.128	.012
170 – 200 minutes	.128	.174	.140
> 200 minutes	.012	.047	.105

(a) If we call a "quick" game one that lasts less than 170 minutes, what proportion of games were quick?

(b) What proportion of games were 200 minutes or less?

(c) If the game was low scoring (between 1–9 runs scored), what is the (approximate) probability it was quick?

(d) If the game was high scoring (over 15 runs scored), what is the probability it was quick?

(e) From your answers to parts (c) and (d), comment on the relationship between runs scored and the length of a game.

Activity 14-10: Classifying Poor by Race and Region

Who are the American poor? *The World Almanac* classifies the people below the poverty line in 1992 by the region of the country (Northeast, Midwest, South, and West) and their race (White, Black or Hispanic). The numbers presented in the almanac have been converted into proportions, which can be interpreted as approximate probabilities. If a poor person is selected at random, then this table gives the probabilities that this person falls in one of the 12 categories of the table.

	Race		
Region	White	Black	Hispanic
Northeast	.103	.040	.028
Midwest	.129	.056	.011
South	.202	.143	.048
West	.153	.015	.073

(a) What is the probability a poor person lives in the Midwest?

(b) What is the probability a poor person is black?

(c) Given that a poor person is white, find the probabilities that he/she lives in each of the four regions of the country. Where is the most likely residence of a poor white?

(d) Given that a poor person is black, find the probabilities of living in each of the four regions. Where is a poor black likely to live? Where is he/she unlikely to live?

(e) Answer the questions in part (d) assuming that the poor person is Hispanic in origin.

(f) Write a few sentences on what you have learned about the pattern of poor people in the United States.

Activity 14-11: Independent Events (cont.)

In the almanac, various weather data is reported for cities of the United States. For a number of cities, suppose we observe (1) the average temperature in January, (2) the average temperature in July, and (3) the average number of days of precipitation during the year.

The following table categorizes 99 cities by the January temperature and the number of days of precipitation.

January	Number of days of precipitation			
Temperature	0-50	51-100	101-130	131 and over
12-19	0	4	7	4
20-29	0	11	10	9
30-39	0	4	16	5
40 and over	6	8	15	0

(a) If you choose a city at random, what is the probability that it has a January temperature exceeding 40? Round your answer to two decimal places.

(b) Suppose you are told that this city has between 51-100 days of precipitation. What is the probability that its January temperature exceeds 40?

(c) Are the events "January temperature exceeds 40" and "between 51-100 days of precipitation" independent or dependent? Why?

The table below categories 100 cities by the average January temperature and the average July temperature.

January	July Temperature			
Temperature	48-69	70-74	75-79	80 and over
12-19	8	6	1	0
20-29	4	17	6	3
30-39	4	3	14	4
40 and over	2	2	3	23

(d) Find the probability that a city has an average January temperature of 30 or over.

(e) Suppose you are told that the average July temperature of the city was 75 or over. Without doing any computation, do you think the probability of a January temperature of 30 or over would be different from the value you found above? Explain.

(f) From the table, compute the probability of "July temperature 75 or over" given that January temperature is 30 or over.

(g) Are the events "July temperature 75 or over" and "January temperature 30 or over" independent or dependent?

Activity 14-12: Independent Events (cont.)

In each of the following,

- Make an intelligent guess at the probability of the event A.
- Suppose you know that the event B is true. Make a new guess at the probability of the event A.
- Decide if the events A and B are independent.

(a)	Event A: You will fall down when you walk to work today.	Event B: The sidewalks are icy from freezing rain last night.

Guess at probability of A: _____ New guess at probability of A: _____
A, B independent? _____

(b)	Event A: Man will land on the moon in the next 10 years.	Event B: Scientists will soon discover a cure for the AIDS virus.

Guess at probability of A: _____ New guess at probability of A: _____
A, B independent? _____

(c)	Event A: Your son will play basketball on the high school team.	Event B: Your son will grow to be 7 feet tall by the time he's in high school.

Guess at probability of A: _____ New guess at probability of A: _____
A, B independent? _____

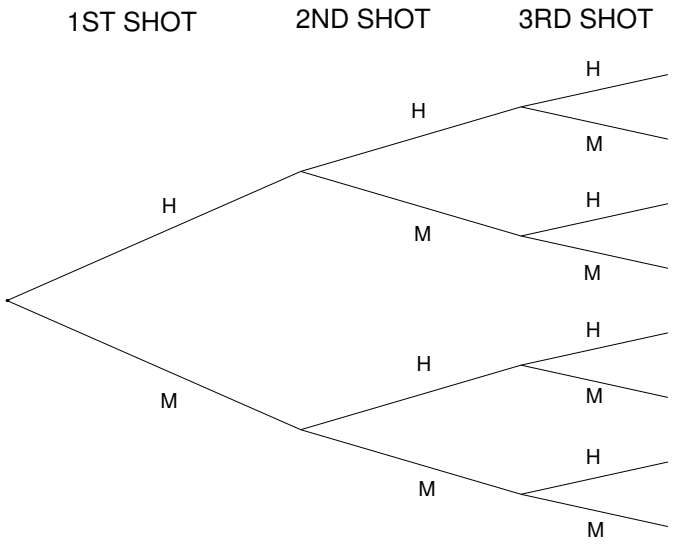
(d)	Event A: You will live longer than 80 years	Event B: Your favorite color is red.

Guess at probability of A: _____ New guess at probability of A: _____
A, B independent? _____

Activity 14-13: Multiplying Probabilities

Suppose a basketball player gets three attempts to make free throws during a game. You know from past data that he makes 70% of his free throw shots. Assume that the results of his three shots are independent. That means, for example, that the probability that he makes his second shot will be unaffected by what he does on his first shot. The probability that he misses the third shot will be the same whether he made or missed his first two shots, and so on.

(a) The tree diagram below can be used to record all the possible outcomes of his three foul shots. The first branch on the left represents the results of the first shot (Hit or Miss), the second set of branches the outcome of the second shot, and the rightmost branches the result of the third shot. Next to each outcome (H or M), write the corresponding probability.



(b) Circle the outcome in the tree diagram where the player makes all three of his shots. Find the probability of this outcome by multiplying the probabilities across the tree.

(c) Find the probability that the player misses his first two shots and makes his last shot.

(d) Suppose you wish to find the probability that he makes exactly one shot. One way he could do this is HMM – that is, make the first shot and miss the second and third shots. Write down other ways he could make exactly one shot.

(e) Find the probability of making one shot by (1) finding the probability of each outcome listed in part (d) and (2) adding these probabilities to find the probability of interest.

(f) Find the probability that he makes exactly two shots.

WRAP-UP

In this topic, we have practiced on interpreting probability tables when there are two numerical outcomes of interest. There is much information in a two-way probability table. One can learn about one of the numerical outcomes by summing the probability table over rows or columns. Also we can compute conditional probabilities by restricting attention to only one row or only one column. These conditional probabilities that we compute help us understand how one outcome is dependent on the knowledge of the second outcome. We will use two-way probability tables in the introduction of statistical inference in the next topic.

Topic 15: Learning About Models Using Bayes' Rule

Introduction

We are ready to begin our study of statistical inference. We begin with something that is unknown. In the activities to follow, we don't know if we have a particular disease, if a die being used in the casino is fair, the exact number of fish in our backyard pond, or the number of defectives in a shipment of electronic components. A **model** is a possible description of this unknown phenomena. In each application, there will be a number of possible models. We assign **prior probabilities** to these models — these probabilities reflect our opinions about which models are more or less likely to be true. To learn about the correct model, **data** will be observed. In the activities, the data will consist of the results of a blood test, the results of a roll of the die, the results of a capture-recapture experiment, and the number of defectives observed in a sample. This data will change our opinions about the likelihoods of the different models. **Bayes' rule** is the formula which tells us how to compute the **new model probabilities.** We begin with one of the best illustrations of Bayes' rule: what do you learn about your disease status on the basis of a blood test?

PRELIMINARIES

1. What is the chance that a randomly selected person in your community has tuberculosis?

2. Suppose that this randomly selected person takes a blood test for the disease, and the result is positive, which indicates that the person may have the disease. What is your new probability that this person has tuberculosis?

3. Suppose that a person is holding two dice. The first die is the usual fair type which has the numbers 1, 2, 3, 4, 5, 6 on the six sides, and the second die is a special die with two 1's, two 2's, and two 3's on the sides.

(a) Suppose that the person chooses one of the two type at random. What is the probability she chooses the special die?

(b) Suppose that this chosen die is tossed once and a 4 is observed. What is the probability now that she chose the special die?

4. Suppose that you wake up one morning in the summer not knowing the weather forecast. What is the probability that it will rain on that particular day?

5. After you wake up, you look out your bedroom window and notice black clouds and lightning in the sky. What is the probability now that it will rain on that day?

IN-CLASS ACTIVITIES

Activity 15-1: Do You Have a Rare Disease?

This activity illustrates a particular screening program for a medical disease. People that are considered at risk for the disease are subjected to various diagnostic tests to help in early identification of the disease. Unfortunately, the results of these tests are often not well understood, both by the patient and the doctor. Here we will use Bayes' rule to compute the probability of having the disease given that the diagnostic test gives a positive result. We will see that the resulting probability of having the disease can differ from one's intuitive belief about the likelihood of having the disease.

Suppose in this particular case that you are given a blood test for tuberculosis. Since this disease is rare, only a very small proportion of the population has the disease. Let's assume that this disease proportion is .001 — this means that 1 out of every 1000 people have tuberculosis. The blood test that you take has two possible results — positive, which is some indication that you may have tuberculosis, or negative. It is possible that the test will give the wrong result. If you really have tuberculosis, it will sometimes give a negative reading. Let's assume that this error rate is 10% or .10. So even when you have the disease, 10% of the time the test will say that you are healthy. Likewise, the test will give a false positive result with probability .10. That means that the test will occasionally (10% of the time) say that you have the disease when you really are healthy.

Suppose that you have a blood test and the result is positive. Should you be concerned that you have tuberculosis?

In this example, you are uncertain if you have tuberculosis. There are two possible alternatives: either you have the disease, or you don't have the disease. We will refer to these alternatives as *models*. In the following, we will describe the two models using the terms "have disease" and "don't have disease".

We use probabilities to reflect one's opinions about models. In this example, you can initially assign probabilities to the models "have disease" and "don't have disease" which indicate your belief in the plausibility of having this disease. We call these *prior probabilities* since they reflect your opinion *before* or *prior* to the blood test.

If you take a blood test and the result is positive, then your feelings about the relative possibility of the two models would change. Specifically, a positive blood test result would make you more concerned that you indeed have tuberculosis. Your new opinions about "have disease" and "don't have disease" will be reflected in your *posterior probabilities* – these are the probabilities assigned to the models *after* or *posterior* to the blood test.

How does one obtain prior and posterior probabilities? You assign prior probabilities to the two models "have disease" and "don't have disease" which reflect your opinions about the relative likelihoods of these two possibilities. We will suggest below one probability assignment that reflects the knowledge that you are a "representative" person in the population. Then, using Bayes' rule, these prior probabilities will be combined with the error rates of the blood test to compute posterior probabilities. Bayes' rule is a general recipe for updating one's probabilities about models when more data is obtained.

The prior probabilities

Let's start with the prior probabilities. Before you have a blood test, you want to assign probabilities to the models "have disease" and "don't have disease" which reflect the plausibility of these two models. You think that your chance of having tuberculosis is similar to the chance of a randomly selected person from the population. Since you know that the tuberculosis proportion in the population is .001, it would be reasonable in this case to assign the event "have disease" a probability of .001. By a property of probabilities, this implies that the event "don't have disease" has a probability of $1 - .001 = .999$.

What if you don't think your chance of having tuberculosis is the same as that of a randomly chosen person in the population? Then it will be more difficult to make the probability assignments. Suppose that you think you are more susceptible to tuberculosis because of your age. Then it may be more reasonable to think of your health as similar to the health of people that are approximately your age. You would need to find the tuberculosis rate of people in your age group. If that disease rate was .004 (instead of .001), then you could assign the model "have disease" the prior probability of .004 and the model "don't have disease" the probability .996.

In the following, we'll return to the first scenario. We will assume that you think that your likelihood of getting the disease is about the same as the population disease rate. The table below summarizes what we know so far. The column "MODEL" lists the two possible models and the

Graph of prior probabilities of two models.

column "PRIOR" gives the corresponding prior probabilities. The prior probabilities are displayed using a segmented bar chart above. Note from the graph that the bar appears entirely white since the chance of having tuberculosis is so small.

The Prior Probabilities

		PRIOR
MODEL	have disease	.001
	don't have disease	.999

The data

The new information that we obtain to learn about the different models is called *data*. In this example, the data is the result of the blood test. The different data results are called *observations*. Here the two possible observations are a positive result which we denote by a plus sign $(+)$ or a negative result which we denote by a minus sign $(-)$.

In the statement of the problem, we are told how the data (the blood test) is related to the models (have disease or don't have the disease). For each model, we are given the probabilities of each test result. If we "have the disease", the probability of a mistake in the test (a negative blood test result) is .10. By a property of probabilities, this means that the probability of a "$+$" result is .90. It may be easier to think of these probabilities in terms of counts in a hypothetical population. Suppose 1000 people actually have the rare disease. Then we would expect that 10% of 1000 or 100 people would get a "$-$" test result; the remaining 900 people would get the right "$+$" result.

Similarly, if we don't have the disease, the test will give an incorrect positive result with probability .10. That means that in this case the test will be negative with probability .90. Again think of a group of 1000 people who don't have the disease. We expect that 10% or 100 of these people will get the incorrect "$+$" result; the remaining 900 people would get a negative result.

The numbers which relate data with models are called *likelihoods* – they are the probabilities of each possible data value conditional on each possible model. In this problem there are two sets of likelihoods. There is one set of likelihoods (probabilities of the two test results) in the case where you have the disease. There is a second set of likelihoods in the case where you don't have the disease.

All of the likelihood values are placed in the table below. The column "MODEL" lists the two possible models and the two possible data outcomes "+" and "−" are placed under the heading "DATA". The first row of the table gives the likelihoods of the two data outcomes for the "have disease" model and the second row gives the likelihoods for the "don't have disease" model.

The Likelihoods

		DATA	
		+	−
MODEL	have disease	.90	.10
	don't have disease	.10	.90

Bayes' rule

Remember that you got a positive blood test result ("+"). Bayes' rule is the recipe for revising your probabilities of the two models after getting this new information. We can perform Bayes' rule by thinking of a large group of people and classifying these people by the model and the test result.

Consider a representative group of 10,000 people from the population. (We are using the large number 10,000 since the calculations are easier to do by hand. Actually, any number for the group size will work.) We wish to classify these people by their disease status and the result of the blood test. We will make this classification by placing counts in the following two-way table with the variables "MODEL" and "DATA". We will call this table a "Bayes' box" – it will be our usual device for computing posterior probabilities using Bayes' rule.

(a) The Bayes box for this problem is shown below. To start, put the number of people (10,000) in the lower right cell of the table (see the arrow).

A Bayes' box for the blood testing example.

		DATA		
		+	−	TOTAL
MODEL	have disease			
	don't have disease			
	TOTAL			

Next we put counts in the Bayes' box which correspond to the two possible models. Here we put counts in the "TOTAL" column that correspond to our knowledge about the chances of the two

disease states in our group of 10,000. Remember the prior probabilities of "have disease" and "don't have disease" were .001 and .999, respectively. Out of 10,000 people,

(b) How many people do you expect to have the disease?

(c) How many people do you expect to be healthy?

(d) Put these numbers in the TOTAL column of the Bayes' box.

		DATA		
		+	−	TOTAL
MODEL	have disease			
	don't have disease			
	TOTAL			10,000

Next, we allocate these marginal counts to the 'DATA' cells of the table using the likelihoods. Recall that, if you have the disease, then the probability is .1 that you will get a negative test result and .9 that you will get a positive result. If 10 people have the disease,

(e) How may do you expect to have a "−" result?

(f) How many do you expect to have a "+" result?

(g) Put these numbers in the first row of the Bayes' box.

		DATA		
		+	−	TOTAL
MODEL	have disease			10
	don't have disease			9990
	TOTAL			10,000

Next, look at the "don't have disease" row. The chance that a healthy person gets a positive test result is .1 and a negative test result is .9. Out of a group of 9990 healthy people,

(h) How many do you expect to get a "+" blood test result?

(i) How many do you expect to get a "−" blood test result?

(j) Put these numbers in the second row of the Bayes' box.

		DATA		
		+	−	TOTAL
MODEL	have disease	9	1	10
	don't have disease			9990
	TOTAL			10,000

(k) Complete the table by finding the totals of each of the data columns "−" and "+" and placing them in the "TOTAL" row.

		DATA		
		+	−	TOTAL
MODEL	have disease	9	1	10
	don't have disease	999	8991	9990
	TOTAL			10,000

Now we are ready to answer our original question. If you get a positive test result, how likely is it that you have the disease? This can be answered by looking at the Bayes' box and restricting attention only to the group of people that had a positive test result. That is, we look only at the "+" column of the table. Copy the counts from the Bayes' box into the "+" column of the table below.

		DATA	
		+	PROPORTION
MODEL	have disease		
	don't have disease		
	TOTAL		

(l) How many people had a positive test result?

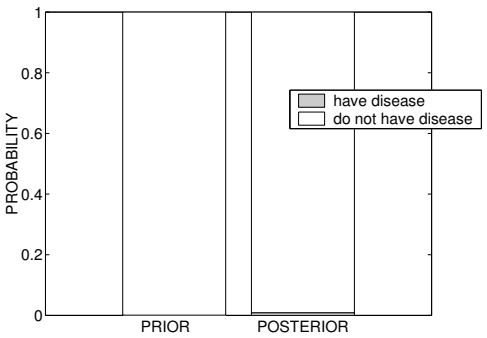
(m) Of all of the people who had a positive test result, find the proportion of these people who had the disease and the proportion of people who don't have the disease. Put these proportions in the above table. These proportions represent the probabilities after having seen the test result.

Let's summarize what you have learned. Initially, you were pretty sure that you did not have tuberculosis.

(n) The prior probability that you had the disease was _____.

(o) After getting a positive blood test result, the probability that you have the disease has increased to _____

Segmented bar graphs comparing the prior and posterior probabilities of the two models are shown in the figure above. You should note that posterior probability of tuberculosis is much larger than the prior probability (before the blood test), but the actual size of the probability is still small. You would need additional evidence, such as a second positive blood test result, to be concerned that you actually have the disease.



Graph of prior and posterior probabilities of two models.

Activity 15-2: Is the Die Fixed?

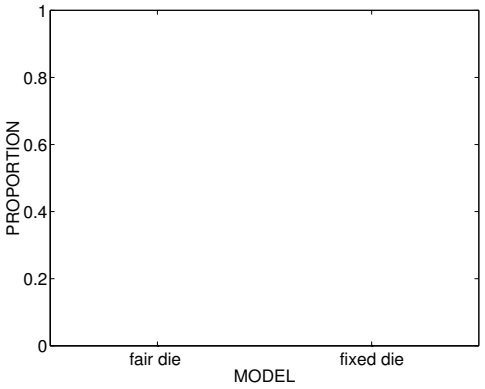
Suppose you decide to take a vacation in Las Vegas. One day you go into one of the popular casinos and watch the game of craps being played. Craps is a game in which a pair of dice is thrown and players win or lose money on the basis of the rolls of dice.

As you watch this game, some questions come to mind. In particular, you wonder how likely it is to win money. The chances that you win or lose money may depend on the exact rules of the game. Also, the chances may depend on the fairness of the dice. You know that each die has 1 through 6 dots on the six sides. But you are not sure of the "true" probabilities that correspond to the six possible rolls.

You think there are two plausible sets of probabilities for one particular die that you are watching. It may be a "fair die" where each of the six possibilities (1, 2, 3, 4, 5, 6) has the same probability of being tossed. But, based on watching the rolls of this die, you think it may be "fixed" and only roll a 1, 2, or 3, where each of the three outcomes has the same probability.

We'll decide if **the die is fair or the die is fixed** by using Bayes' rule.

(a) Let's think of 3000 dice that are similar to the one that you are observing. You are pretty confident that the casino is honest and the die is fair. But there is a small chance that the die is fixed and only will show a 1, 2, or 3 on a single roll. To reflect this opinion, we'll divide the 3000 dice into two groups — the fair dice and the fixed dice. Suppose that you think the chance that the die is fair is .9 and the chance that the die is fixed is .1. In that case, how many of our 3000 dice would be expected to be fair and how many would be fixed? Put your answers in the table below — we use the word "MODEL" to describe the state of each die (fair or fixed) and "COUNT" refers to the number of dice of each type. Also find the proportion of dice that are fair and the proportion of dice that are fixed. Put these numbers in the "PROPORTION" column.



Probabilities of fair and fixed die before "roll of 1".

		COUNT	PROPORTION
MODEL	fair die		
	fixed die		
	TOTAL		

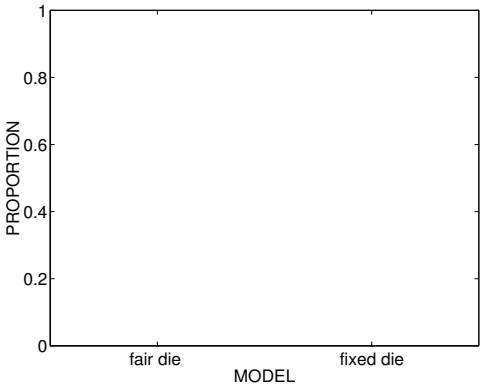
(b) Graph these proportion values on the graph above.

(c) Next, let's consider what rolls are to be expected from the group of **fair** dice. The number on each die can be 1, 2, ..., 6. If the dice are fair, then you expect that each number is equally likely to occur. Using this assumption, write in the table below the number of ones, twos, threes, fours, fives and sixes that you expect from this group of fair dice. (As a check of your work, add up all of the ones, twos, etc. and put the sum in the total column — this should be equal to the number of fair dice from part (a).)

NUMBER ON DIE						
1	2	3	4	5	6	TOTAL

(d) Now consider the group of **fixed** dice from part (a). You think that each die in this group can only roll a 1, 2 or 3 — rolls 4, 5 and 6 are not possible. You don't suspect any other problem with this type of die, so it reasonable to assume that each of the three possibilities is equally likely. In the below table, write down the number of ones, twos, threes you expect. You don't expect any fours, fives, of sixes, so write down 0 in each of these boxes. (As before, add up all of the counts and put the result in the box under the "TOTAL" column.)

NUMBER ON DIE						
1	2	3	4	5	6	TOTAL



Probabilities of fair and fixed dice after "roll of 1".

(e) Let's combine the two tables from parts (c) and (d). The table below has two rows of empty boxes — one for the fair dice and one for the fixed dice. Copy the values from the table in part (c) and put those in the "fair dice" column. In a similar fashion, copy the values from part (d) into the "fixed dice" column.

		NUMBER ON DIE						
		1	2	3	4	5	6	TOTAL
MODEL	fair die							
	fixed die							

Now we are ready to observe some data. We watch the particular die in question. It is rolled and **we observe a 1.** We'll use this single die roll to update our probabilities that the die is fair or fixed.

Looking at the above table which includes the fair dice and the fixed dice ...

(f) What are the total number of dice where a 1 was rolled?

(g) Of these dice where a 1 was rolled, how many correspond to fair dice?

(h) Of the dice where a 1 was rolled, what is the proportion of fair dice?

(i) Of the dice where a 1 was rolled, what is the proportion of fixed dice?

(j) Graph the proportions of fair and fixed dice on the graph above:

(k) Compare the two bar graphs that you made. The first graph reflected your probabilities about the two possibilities (fair die or fixed die) before you observed any rolls. The second graph

reflects your probabilities after seeing a die roll of 1. How have your beliefs about the die changed? Would it be accurate to say that you are more confident that the die is fair? Explain.

Activity 15-3: How Many Fish?

This activity is a simple illustration of a capture-recapture experiment. The objective of this experiment is to learn about the total number of fish in a large body of water. One first captures and tags a particular group of fish. The tagged fish are returned to the water and, after some time, a new group of fish is captured. The number of tagged and nontagged fish in the captured sample is informative about the number of fish in the lake.

In our simple setting, suppose you are interested in learning about the number of fish in the pond in your back yard. It's a small pond, so you don't expect many fish to live in it. In fact, you believe that the number of fish is either 1, 2, 3, or 4.

To learn about the number of fish, you will perform a capture-recapture experiment. You first catch one of the fish, tag it, and return it to the pond. After a period of time, you catch another fish and observe that it is tagged. (This fish is also tossed back into the pond.) What have you learned about the number of fish in the pond?

In this example, the unknown quantity, the model, is the number of fish in the pond. There are four possible models "one fish", "two fish", "three fish" and "four fish". (This is sounding like a Dr. Seuss book!) You think that the four models are all equally plausible and so you assign a probability of 1/4 to each possibility.

The models and the prior probabilities are listed in the table below.

The Prior Probabilities

		PRIOR
MODEL	one fish	1/4
	two fish	1/4
	three fish	1/4
	four fish	1/4

In our experiment, we catch a fish from the pond and it either will be tagged, or not tagged. These are the two possible data outcomes. If we know the number of fish in the lake (the model),

we can compute the probabilities of these two outcomes. Suppose that the true model is "one fish" – that is, there is exactly one fish in the pond. Then we are certain to catch a tagged fish. So the probability of a "tagged" will be 1 and the probability of "not tagged" will be 0. We write these numbers in the likelihoods table below.

Let us consider the likelihoods for the other three models. If there are really two fish in the pond, then one is tagged and the other will be not tagged. The probability of catching a tagged fish will be 1/2 and the probability of grabbing a untagged fish will be 1/2. In a similar fashion, if there are three fish, the chance of "tagged" will be 1/3 and, if there are four fish, the chance of "tagged" will be 1/4. We can now fill up all of the likelihoods in the table.

The Likelihoods

		DATA	
		tagged	not tagged
MODEL	one fish	1	0
	two fish	1/2	1/2
	three fish	1/3	2/3
	four fish	1/4	3/4

Now we have all of the information collected to form our Bayes' box. This box will classify many hypothetical ponds by the number of fish (the model) and the data result.

Specifically, suppose that there are 1000 ponds just like the one in your back yard. We put this number in the lower right hand corner of the box. (You will fill in the rest of the box in the questions below.)

A Bayes' box for the fish example.

		DATA		
		tagged	not tagged	TOTAL
MODEL	one fish			
	two fish			
	three fish			
	four fish			
	TOTAL			1000

(a) First, allocate these 1000 ponds to the four models (rows) by your prior probabilities. You think that the four models "one fish", "two fish", "three fish" and "four fish" are equally likely. So how many ponds do you expect will contain one fish, two fish, three fish, and four fish? Put these counts in the TOTAL column of the Bayes' box on the right.

(b) Now fill up the rows of the table from the information in the likelihoods table. For example, look at the first row — model "one fish". If the pond contains only one fish, then we would always observe a tagged fish. (We can also see this from the likelihoods table.) So put all of the the counts in the first row in the "tagged" column and put a 0 in the "untagged" column.

Continue and fill in all of the counts in the Bayes' box for the two fish, three fish, and four fish models.

(c) Remember that you did catch a fish that was tagged. We use the Bayes' box to update our uncertainty about the number of fish in light of this evidence that the fish we caught was tagged. Look at the above Bayes' box – since your data was "tagged", we look only at the "tagged" column of the table. Put the counts from the "tagged" column in the table below.

		DATA tagged	PROPORTION
MODEL	one fish		
	two fish		
	three fish		
	four fish		
	TOTAL		

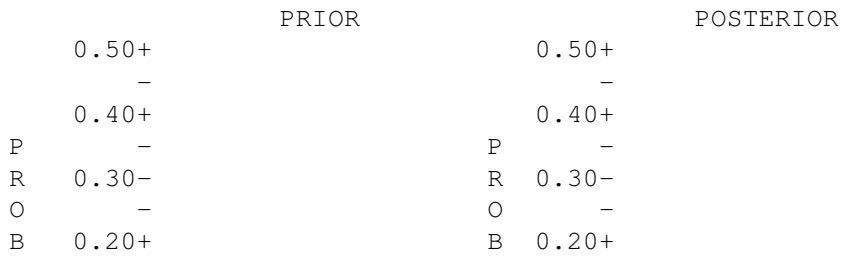
(d) What is the total number of ponds (out of the 1000) where you caught a tagged fish? (Sum the counts in the above table.)

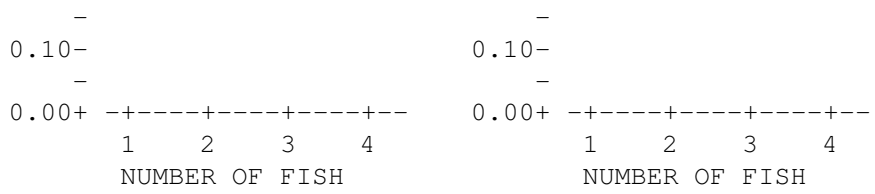
(e) For these tagged fish, find the proportion of fish that came from ponds with one fish, ponds with two fish, etc. Put this information in the PROPORTION column of the table. These proportions are the probabilities of these four models after we observed a tagged fish.

(f) In the table below, put the model probabilities before and after observing the tagged fish.

		PRIOR PROBABILITIES	POSTERIOR PROBABILITIES
MODEL	one fish		
	two fish		
	three fish		
	four fish		

(g) Plot the two sets of probabilities on the graphs below.





(h) Before you caught any fish, how many fish did you think were in the pond? After catching a tagged fish, how have your probabilities changed? Why did the probabilities change?

Activity 15-4: How Many Defectives in the Box?

In all of our activities in this chapter, we have updated our probabilities about models using a Bayes' box. In the next two examples, we illustrate a different method of computing the posterior probabilities.

Suppose that a company manufactures a special type of electronic component to be installed in automobiles. The quality of the components is very important to the company. Indeed, of all of the components that will be produced this year, the company would like only a small proportion of them to be defective.

The components are shipped in boxes of four. Periodically, to ensure that the components are of high quality, a worker opens a box, chooses one component at random, and performs a thorough inspection. This inspection is expensive and time-consuming, so it is not cost-effective to inspect more than one component from the box.

On a given day, suppose that the worker opens a box, chooses a component and finds it to have no defects. Can the worker make an intelligent guess at the total number of defectives in the box? Specifically, is she pretty confident that the box does not contain any defectives?

The model

As in the earlier activities, we need to first identify the model which is unknown. Here the model is the total number of defectives in the box. Since there are four components in the box, there are five possible numbers of defectives: 0, 1, 2, 3, or 4. We write the possible models in a column.

MODEL
0 defective
1 defective
2 defectives
3 defectives
4 defectives

The prior

Next, probabilities are assigned to the different models. Remember that the company wishes to produce components of high quality. Generally, the customers have been satisfied with the components and there have been relatively few complaints about defects in recent history. So the company expects that most of the components in a box will be acceptable — in fact, it is unlikely that there is more than one defective in a box. Based on this past experience, the company assigns the following probabilities to the models. We label this probabilities by the word "PRIOR", since they are probabilities that are assigned *prior* or before any data is observed.

MODEL	PRIOR
0 defectives	.7
1 defective	.1
2 defectives	.1
3 defectives	.05
4 defectives	.05

A few comments should be made about these probabilities. First, the model "0 defectives" is given a probability of .7; this reflects the company's belief that this is the most likely possibility. Note that the probabilities of the remaining models "1 defective"–"4 defectives" are approximately the same. This reflects the company's opinion about the behavior of the machine that is making the components for the box. It is possible that the machine slips into a "broken" mode and is more likely to produce poor components. So, if the machine produces one defective, it is pretty likely that the components that will be produced next will also be defective.

The likelihoods

The inspector wishes to adjust her probabilities about the five models on the basis of the data outcome, which is the result of the inspection. Recall that she selected one component at random from the box, gave it a complete inspection, and found it not defective. The next step is to compute the likelihoods. These are the probabilities of the data result "not defective" for each possible model. We'll compute these starting with the model "0 defectives". Suppose that there are no defective components in the box. What's the probability of choosing one that is not defective? Since all of

the components have no defects, this probability is 1. We place this number in the first cell of the "LIKELIHOOD" column of the table. (We will fill out the rest of the table in parts (a) and (b).)

data result = "not defective"

MODEL	PRIOR	LIKELIHOOD
0 defectives	.7	1
1 defective	.1	
2 defectives	.1	
3 defectives	.05	
4 defectives	.05	

(a) Now compute the likelihood for the next model "1 defective". If there is exactly one defective in the box (and therefore three nondefectives), what is the probability that the inspector will draw one out that is nondefective? (You can think of the box as the set $\{\bigcirc, \bigcirc, \bigcirc, \bigotimes\}$, where $\bigcirc$ is a nondefective and $\bigotimes$ is a defective.) Put this probability in the "LIKELIHOOD" column in the "1 defective" row.

(b) Compute the likelihoods for the remaining three models. Find the probability of picking a nondefective if there are 2 defectives in the box. ($\{\bigcirc, \bigcirc, \bigotimes, \bigotimes\}$). Put this number in the LIKELIHOOD column for 2 defectives. Likewise find the probability of choosing a nondefective if there are 3 defectives ($\{\bigcirc, \bigotimes, \bigotimes, \bigotimes\}$), and if there are 4 defectives in the box ($\{\bigotimes, \bigotimes, \bigotimes, \bigotimes\}$). Put these two probabilities in the LIKELIHOOD column.

(c) We wish to compute the inspector's posterior probabilities of the models. By Bayes' rule, the posterior probability of a given model (abbreviated by "POST") is proportional to the product of the prior probability ("PRIOR") and the likelihood. Briefly,

$$\text{POST is proportional to PRIOR} \times \text{LIKELIHOOD}.$$

We do the work in the table below which we'll call a Bayes' table. First, copy the numbers from the LIKELIHOOD column into the table below. Then you can compute the updated probabilities in three steps.

- **[MULTIPLY]** For each model, we *multiply* each prior probability by the corresponding likelihood. In the table below, we place these products in a new column called "PRODUCT".
- **[SUM]** Find the *sum* of these products. This sum is placed at the bottom of the "PRODUCT" column.
- **[DIVIDE]** Last, we *divide* each of the products by the sum that we just computed. The answers we get are the posterior probabilities — these are placed in the "POST" column.

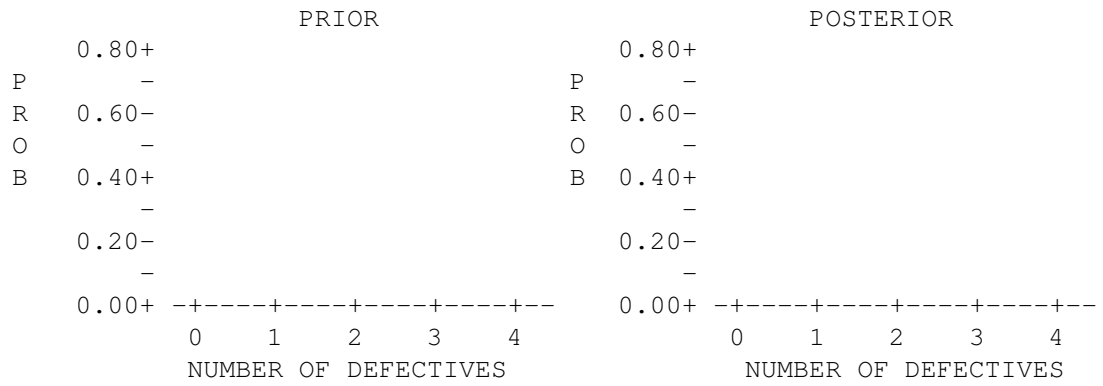
data result = “not defective”

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POST
0 defectives	.7			
1 defective	.1			
2 defectives	.1			
3 defectives	.05			
4 defectives	.05			
SUM				

(d) What has the inspector learned by analyzing one component which was not defective? Copy the prior and posterior probabilities for the five models into the table below. Graph the two sets of probabilities on the graph below.

data result = “not defective”

MODEL	PRIOR	POST
0 defectives		
1 defective		
2 defectives		
3 defectives		
4 defectives		



(e) Compare the shapes of the prior and posterior distributions of probabilities.

(f) After inspecting the one component, what model (0, 1, 2, 3, or 4 defectives) is *not* possible?

(g) Find the prior probability that the box contains 2 or fewer defectives.

(h) Find the posterior probability that the box contains 2 or fewer defectives.

(i) Compare your answers to (g) and (h). Describe in words how your probabilities have changed.

Activity 15-5: Our Team Scored First — Will They Win the Game?

You are going to a sports event in which your favorite team is playing. You know something about the quality of your team and the opposing team. Based on this information, you may be able to make a prediction of the game's outcome. Specifically, you can guess which team will win. The game begins and, after a portion of the game has been played, your team is leading. How does this new information adjust your opinion about the likelihood of your team winning the game?

Suppose you are watching the Super Bowl —- this is the game matching the two best teams in professional football. Call the two teams "D" and "P". Before the game started, team D was favored to win. You start watching the game, and team D scores the first touchdown in the first quarter. Presumably, you would now be more confident that team D would win the game. The television announcer gives you some extra information which supposedly helps in understanding how likely it now is that team D will win. He says that, in the 29 previous Super Bowls, the team that scored first won 21 of the games. How can you use this new information to update your opinion about team D winning the game?

Let's solve this problem by use of Bayes' rule. The model is the winner of this football game. Since a tie is not possible in the Super Bowl, there are two possible outcomes: "D wins" and "P wins". Before the game begins, you can assign prior probabilities to the models which reflects your belief about the likelihoods of each team winning. In this case, we'll assume that you are not a strong fan of either team D or P and your beliefs about the outcome of the Super Bowl are similar to the beliefs expressed by the bookies who place odds on the game outcome. So you think that D will win with probability .6. Therefore, P will win with probability .4. The models and corresponding prior probabilities are placed in the Bayes' rule table:

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POST
D wins	.6			
P wins	.4			
SUM				

The data in this problem is what we observed in the first quarter of the Super Bowl – "D scored first".

(a) To continue our Bayes' rule calculations, we need to compute our likelihoods. These are the probabilities of the data result "D scored first" for each model. Suppose that D did indeed win the game. What is the probability that D scored first? Recall that we were told that the team that scores first won 21 out of the previous 29 Super Bowls. Remember that each Super Bowl has a winner and a loser. There were 29 winners in the past — of these 29 winners, 21 happened to score first. So the probability of "D scored first" if D is a winner is approximately _____. Put this likelihood value in the table.

(b) We also need to compute the probability of the data "D scored first" if D actually is the loser. There were 29 losers in the previous Super Bowls and, using our information, only $29 - 21 = 8$ of these teams happened to score first. So the probability of our data if D is a loser is _____. Put this likelihood value in the table.

(c) Compute the Bayes' rule table. Remember the three steps — multiply, sum, divide:

- we *multiply* the values of the prior probabilities and the likelihoods to get products
- we add up the products to get a *sum*
- we *divide* each product by the sum to get our posterior probabilities

(d) Summarize what you have learned. Initially, you thought team D would win the game — we assigned this possibility a probability of _____. We observed that D scored first. Presumably this would increase our confidence that D would win — by looking at the Bayes' table, we see that the probability of "D wins" has increased to _____.

HOMEWORK ACTIVITIES

Activity 15-6: How Many Fish? (cont.)

You are interested in learning about the number of fish in your neighbor's backyard. As in Activity 15-3, you believe that there are 1, 2, 3, or 4 fish in this pond. But, since you think the pond is dirty and contains few fish, you assign prior probabilities of .4, .3, .2, and .1 to the above four models. (The probability that there is 1 fish is .4, the probability there are 2 fish is .3, and so on.) You use the same capture-recapture experiment as in Activity 15-3 to learn about the number of fish. You capture one fish and it is not tagged. Use a Bayes' box to find the posterior probabilities of the four models.

Activity 15-7: Which Bag?

Suppose that you have two bags in your closet One bag contains four white balls (we'll call this the "white bag") and the second bag contains two white and two black balls (the "mixed bag"). The closet is dark and you just grab one bag out.

(a) What is the probability that you have chosen the white bag? The mixed bag? Place your probabilities in the table below. We'll call these your prior probabilities since they are prior to any data or information that you have collected.

The Prior Probabilities

		PROBABILITY
MODEL	white bag	
	mixed bag	

(b) We will learn which bag we have chosen by drawing balls from the bag with replacement. Suppose we draw one ball and it turns out to be white. Use a Bayes' box to find the new probabilities of the two models "white bag" and "mixed bag". Place your new model probabilities in the probability table labeled "posterior probabilities".

(Hint: Think of 300 replications of the experiment of choosing a bag. How many times would you expect to choose a white bag? A mixed bag? Begin your calculations by putting these numbers in the TOTAL column of the Bayes' box.)

		DATA (Color of Ball Drawn)		
		white	black	TOTAL
MODEL	white bag			
	mixed bag			
	TOTAL			300

Posterior Probabilities

		PROBABILITY
MODEL	white bag	
	mixed bag	

(c) Return the white ball that you drew to the bag and choose a new ball. Suppose it also is white. Use the Bayes' box to find updated probabilities for your two models — put these in the table below labeled "posterior probabilities".

(Hint: The prior probabilities about "white bag" and "mixed bag" reflect your current beliefs about these two models. Since you've already observed one ball color, your prior probabilities here will be the posterior probabilities from part (b).)

Prior Probabilities

		PROBABILITY
MODEL	white bag	
	mixed bag	

DATA (Color of Ball Drawn)

		white	black	TOTAL
MODEL	white bag			
	mixed bag			
	TOTAL			300

Posterior Probabilities

		PROBABILITY
MODEL	white bag	
	mixed bag	

(d) Do this one more time. Return the ball to the bag and select a third one – again it turns out to be white. Compute new probabilities of "white bag" and "mixed bag". (Remember your prior probabilities here will be the posterior probabilities from part (c).)

Prior Probabilities

		PROBABILITY
MODEL	white bag	
	mixed bag	

DATA (Color of Ball Drawn)

		white	black	TOTAL
MODEL	white bag			
	mixed bag			
	TOTAL			300

Posterior Probabilities

		PROBABILITY
MODEL	white bag	
	mixed bag	

(e) Describe in a few sentences how your probabilities have changed as you get more information by drawing balls. Are you pretty sure which bag you have been drawing from? If so, why?

Activity 15-8: What Proportion of M&M's are Brown?

Suppose we are interested in the numbers of M&M's of different colors in a "snack-size" bag. Suppose we have a bag of 10 candies and 3 of them (30 percent) turn out to be brown. What have we learned about the proportion of all M&M candies that are brown? This is a problem of **statistical inference** where we want to learn about a population (all M&M candies) based on a sample (the M&M candies in the snack-size bag).

- **(FOUR MODELS)**

  Suppose that there is a machine in Hackettstown, New Jersey that produces M&M's. This machine will produce, in the long-run, a proportion p of brown M&M. We don't know the exact value of p. But we think that, on average, the machine produces 10% browns, or it produces 20% browns, or 30% browns, or 40% browns. In other words, p is either .1, .2, .3, or .4.

- **(INITIAL PROBABILITIES)**

  We think that the proportion of browns p could be .1, .2, .3, .4, but we don't have any additional knowledge to indicate that any particular values of p are more or less likely. So we assign the probabilities .25, .25, .25, .25 to these values.

- **(DATA)**

  We're going to open up a bag of candies and look at the colors of the first 10. Our *data* consists of the number of browns we find. (This will be a number from 0 to 10.)

- **(SIMULATING MODELS AND DATA)**

  We're interested in the relationship between the long-run proportion of browns the machine produces (the model or parameter) and the number of browns that I find when I look at 10 candies (the data or statistic). To learn about this relationship, we perform the following simulation.

  There are four possible M&M machines:

  - Machine 1 – 10% browns are produced
  - Machine 2 – 20% browns are produced
  - Machine 3 – 30% browns are produced
  - Machine 4 – 40% browns are produced

We perform a simulation by first choosing a machine at random (the model), and then letting the machine we choose produce 10 M&M's. We count the number of browns produced (the data). Here's the result of one simulation that I ran on the computer — machine 2 was selected and this particular sample produced 3 brown candies out of 10. M&M's.

- **(RESULT OF MANY SIMULATIONS)** We repeat this process 1000 times, getting 1000 values of (model, data), where model is the particular machine and data is the number of browns observed from the 10 candies. We record the results of these 1000 simulations using the following table:

	0 brn	1 brn	2 brn	3 brn	4 brn	5 brn	6 brn	7 brn	8 brn	9 brn	10 brn
Machine 1	90	96	48	12	4	0	0	0	0	0	0
Machine 2	26	63	78	48	19	6	1	0	0	0	0
Machine 3	7	37	59	69	56	24	9	2	0	0	0
Machine 4	2	9	32	61	64	42	25	8	2	1	0

To understand this table, note that the count in the cell in the first row and the first column is 90. In these 1000 simulations, machine 1 was picked and 0 browns were produced 90 times.

- **(QUESTIONS)**

  (a) If you sample from Machine 1 which produces 10% browns (in the long-run), what is the most likely number of browns produced in your sample of 10? (Restrict your attention to the row of the table corresponding to Machine 1.)

  (b) How often does Machine 1 produce 2 or more browns? What is the chance of Machine 1 producing 2 or more?

  (c) What is the probability that Machine 4 produces 2 or more browns out of 10?

  (d) The above three questions focused on the sampling behavior of particular machines. But remember our original question: We sampled 10 candies and found 3 brown. What does this tell us about the machine or the long-run proportion of browns? In the following, restrict your attention to the column of the table corresponding to 3 brown.

  (e) Of all of the simulations in which you got 3 brown, how many came from Machine 1? From Machine 2? From Machine 3? From Machine 4?

  (f) Of all of the "3 brown" simulations, which machine was most likely? What is its probability?

  (g) If you get 3 browns, what is the chance that the machine was 1 or 2?

(h) Suppose instead that you found 5 brown M&M's. What is the most likely model and what is its probability? What is the least likely model and its corresponding probability?

Activity 15-9: Likelihoods

In statistical inference, we are interested in learning about **models** from observed **data**. Our initial beliefs about the models are described by our **prior probabilities**. We use Bayes' rule as our basic tool for changing our model probabilities after seeing the data – our new opinions about the models are reflected in the **posterior probabilities**. By Bayes' rule, we find the posterior probability of a model by multiplying its prior probability by its likelihood. In abbreviated notation

$$\text{POST is proportional to PRIOR} \times \text{LIKELIHOOD.}$$

What is a likelihood? This quantity relates the data that we observe with the models. Specifically, the likelihood for a model is the probability of the observed data assuming that the model is true.

$$\text{LIKELIHOOD for MODEL} = \text{Probability(DATA if MODEL is true)}$$

We compute this probability for each model, obtaining a collection of likelihoods. This activity will give you some practice in computing likelihoods for two simple examples.

Example 1: Is the coin fair?

You are holding a coin in your hand which you purchased from a magic shop. You haven't inspected it; you think that it is either a fair coin with a heads and a tails, a two-headed coin with heads on both sides, or a two-tailed coin. If we think of a model as the type of coin, there are three possible models:

$$\text{MODELS} = \{\text{FAIR, TWO-HEADED, TWO-TAILED}\}$$

$$\text{DATA} = \text{“COIN FLIP IS HEADS” (or “HEADS” for short)}$$

The likelihoods for the data result "HEADS" are the probabilities of "HEADS" for each possible model. We'll put the likelihoods in the following table. We first put all the models in the MODEL column.

Data result is "HEADS"

MODEL	LIKELIHOOD
FAIR	
TWO-HEADED	
TWO-TAILED	

For each model, the likelihood is the probability of our data ("HEADS") *assuming that model is true*. Below, fill in the blanks and put your answers in the above table.

(a) Suppose the model is FAIR (coin with two different sides) – the probability that I get "HEADS" is _____.

(b) Suppose the model is TWO-HEADED – the probability that I get "HEADS" is _____.

(c) Suppose the model is TWO-TAILED – the probability that I get "HEADS" is _____.

We have just found the likelihoods if our data result is "HEADS". We also could have observed a "TAILS" in our coin flip. Corresponding to this observation, there is a new set of likelihoods. Find the probability of "TAILS" if the coin is fair; find the probability of "TAILS" if the coin is two-headed, and so on. Put your answers in the likelihood table.

Data result is "TAILS"

MODEL	LIKELIHOOD
FAIR	
TWO-HEADED	
TWO-TAILED	

Example 2: How many men?

Suppose you know there are exactly three people in a room, but you don't know the gender of any of the people. Suppose that a model consists of the number of men and the number of women in the room. One possible model is (one man, two women).

(d) Write down all of the possible models (you should list four).

(e) To learn more about the mix of men and women in the room, you are going to choose one person at random from the room – let's suppose that this person turns out to be a man. So

DATA = "person chosen is male".

Let's consider the model (one man, two women), which indicates that there are really one man and two women in the room. The likelihood for this model is the probability of the DATA or choosing a male person *if* the room consists of one man, two women. Since there are 3 people in the room and each person is equally likely to be chosen,

LIKELIHOOD = Prob("person chosen is male" if (one man, two women)) = 1/3

(f) In the table below, there are two columns – the MODEL and LIKELIHOOD. For each MODEL, compute the probability of choosing a man – put your answer in the same row in the LIKELIHOOD column. (I have filled in one row for you based on the above computation.)

Data result = "person chosen is male".

MODEL	LIKELIHOOD
(one man, two women)	2/3

(g) There is another possible data result – I could have chosen a woman. For this data, find the likelihoods. You put all of the possible models in the MODEL column. In the LIKELIHOOD column find the probability of choosing a woman if the corresponding model is true.

Data result = "person chosen is female".

MODEL	LIKELIHOOD
(one man, two women)	1/3

(h) We use likelihoods to update our probabilities about models. Suppose that, before taking any data, we believe that all of the models are equally likely. We choose a person at random and she is female. Use the table below to find the posterior probabilities of the models. (You will be using the likelihoods that you just computed in part (g).)

Data result = "person chosen is female".

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POSTERIOR
(one man, two women)				

Activity 15-10: Is the Student Guessing?

Suppose that an instructor is teaching a class with two types of students. One group of students works hard and generally is well-prepared for tests. The second student group is lazy and is ill-prepared for tests. The instructor gives true/false tests and the instructor estimates that each "good" student has a probability .8 of answering a particular question on the test correctly. In contrast, the

instructor believes that a "weak" student just guesses at the true/false questions and therefore will get a particular item correct with probability .5. Suppose the instructor is unsure if one student is good or weak. He gives the student a five question test and she gets four out of five correct. What does the instructor think now about the student's ability?

We start with 1000 students that we think are representative of the student in question. Since we don't know if the student is good or weak, we'll assume that 500 of the hypothetical students are good and the remaining 500 are weak.

How would 500 good students perform on the five question true/false test? Remember these students have a 80% chance of getting a single question correct. Using the computer, the test results of these good students are simulated with the following results:

# Questions Correct	0	1	2	3	4	5
COUNT	0	3	34	116	198	149

For these good students ...

(a) What is the most likely number of questions correct?

(b) What is the least probable number correct?

(c) What is the probability of a good student getting at least 4 correct?

Now let's look at the test performance for the 500 ill-prepared students. These students are guessing at each question — they have a 50% chance of getting a single question correct. Again using the computer, we have these students take the test with the results below:

# Questions Correct	0	1	2	3	4	5
COUNT	23	72	153	156	76	20

For these weak students ...

(d) What is the most likely number of questions correct?

(e) What is the probability of a good student getting at least 4 correct?

Let's put our results into a Bayes' box. The *model* here is the ability of the student (good or weak) and the *data* is the number of questions correct on the test.

The Bayes' Box

	DATA (# Correct)					
MODEL	0	1	2	3	4	5
good student						
weak student						
TOTAL						

(f) Put the test results (number correct) of the good students in the "good student" row of the Bayes' box.

(g) Put the results of the weak students in the "weak student" row.

Now we've got all the information we need to implement Bayes' rule. Remember the instructor observed that our student got 4 out of 5 correct on the True/False test.

(h) How many students in the Bayes' box got exactly 4 questions correct?

(i) Of these students who got 4 correct, how many were good students?

(j) Of these students who got 4 correct, what was the proportion of good students?

(k) Of these students who got 4 correct, what was the proportion of weak students?

(l) Summarize what you have learned. What was the probability the student was good before she took the test? What is the probability of "good" after she took the test and got 4 right?

(m) Suppose the instructor is also unsure about the ability of a second student. As in the case of the first student, this second student is equally likely to be good or poor. This student takes the test and gets exactly 2 correct. Find the new probability that the student is a good student.

Activity 15-11: Testing for a Disease (cont.)

Suppose you take a blood test for a second rare disease which is prevalent in 1 person out of 100. The blood test has two possible results — positive (+) or negative (−). The test can make mistakes: 1 in 10 of those free of the disease have positive test results, and 1 in 5 of those having the disease have negative results. Suppose that you take the test and get a negative result. What is the new probability that you have the disease?

Find this probability by constructing a Bayes' box which is given below. The model is your condition — you have the disease or you don't have the disease. The data is the test result which could be positive or negative.

The Bayes' box

	DATA		
MODEL	+	−	TOTAL
have disease			
don't have disease			
TOTAL			1000

To construct this Bayes' box:

(a) Start with a large number of people who are similar to you. Here 1000 will work — this is put in the TOTAL row of the table.

(b) Using the incidence of the disease, fill in how many people you expect to have the disease or not have the disease. Put these numbers in the TOTAL column.

(c) For the "have disease" people, put the number of people you expect to have a $+$ result and the number you expect to have a $-$ result. Put these numbers in the "have disease" row of the table.

(d) For the "don't have disease" people, likewise enter the number you expect to have a positive and negative test results.

(e) Since you got a negative test result, find the number of people who get a negative result (sum over the "$-$" column).

(f) Find the proportions of these "$-$" people who have the disease and don't have the disease.

Activity 15-12: Is the Coin Fair?

Suppose you are watching a person flip a coin in a magic shop. You know that this shop sells two-headed coins, so there is a good chance that the person is flipping a two-headed coin. In fact, you believe that the coin is equally likely to be fair (with a head and a tail on the two sides) or two-headed. Suppose you watch the person flip the coin twice and get two heads. What is the new probability that the coin is fair?

Here there are two models: "coin is fair" and "coin is two-headed". If you flip the coin twice and we keep track of the order, there are four possible data outcomes: HH, HT, TH and TT. We're interested in the probabilities of the models "coin is fair" and "coin is two-headed" given that we observed the datum HH.

Find this probability using a Bayes' box which is displayed below.

The Bayes' box

	DATA				
MODEL	HH	HT	TH	TT	TOTAL
coin is fair					
coin is two-headed					
TOTAL					

- Start with 1000 coins similar to the one that is begin tossed. Since fair and two-sided are equally likely, 500 of the coins will be fair and the remaining are two sided.

- If the coin is fair, then the four possible outcomes HH, HT, TH, TT are equally likely. So allocate the 500 fair coins evenly among the four possible data values.
- If the coin is two-sided, there is only one possible outcome. Place all of the 500 two-sided coins in this box.
- To find the probabilities, look only at the coins which result in HH, and find the proportions of coins in this column which are two-sided or fair.

Activity 15-13: How Many Greens?

Suppose a box contains three balls. Some of the balls are green and the remaining are red. It is possible that all the balls are green or that all are red. Initially, you have no clue how many green balls are in the box. So there are four possible models: the box has 0 greens, 1 green, 2 greens or 3 greens.

(a) Suppose you think that all the models are equally likely. Assign probabilities to the four models and put your values in the PRIOR column below.

MODEL	PRIOR
0 Greens:	
1 Green:	
2 Greens:	
3 Greens:	

(b) Suppose that you choose a ball from the box and it is green. We'll call this data "choose a green". To find the new probabilities of the four models, we will use Bayes' rule in the table format. The first step is to find the likelihoods. These are the probabilities of our data "choose a green" for each of the possible models. We'll compute these one at a time.

(1) If there are 0 greens in the box, it looks like R,R,R, where R represents a red ball and G a green. Find the probability of "choose a green".

(2) If there is 1 green in the box (G,R,R), find the probability of "choose a green".

(3) If there are 2 greens in the box (G,G,R), find the probability of "choose a green".

(4) If there are 3 greens in the box (G,G,G), find the probability of "choose a green".

(5) Put these four likelihood values in the LIKELIHOOD column of the table. Also fill the PRIOR column from the values in the table above.

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POSTERIOR
0 Greens:				
1 Green:				
2 Greens:				
3 Greens:				
TOTAL				

(b) Find the new (posterior) probabilities for the models using the "multiply, sum, divide" recipe. The final probabilities should be in the POSTERIOR column.

(c) What is the new probability that there are at least 2 green balls in the box?

(d) What is the most likely number of greens in the box?

Activity 15-14: How Many Greens (cont.)

Return to Activity 15-13 in which you were interested in the number of green balls in a box. Redo the Bayes' rule calculations using a Bayes' box. To make the calculations easy, suppose that you start with 120 boxes similar to the one that you are drawing from. Allocate these 120 boxes to the cells of the Bayes' box below. Note that there are two possible data outcomes here: "choose a green" and "choose a red".

The Bayes' box

	DATA		
MODEL	choose a green	choose a red	TOTAL
0 Greens			
1 Greens			
2 Greens			
3 Greens			
TOTAL			120

(a) Assuming that you draw a green ball from the box, find the posterior probabilities of the four models and compare your answers with those obtained in Activity 15-13.

(b) Suppose, you drew a red ball from the box instead of a green. Use the Bayes' box to find the posterior probabilities of the models.

Activity 15-15: Does the Person Live in the Suburbs?

Often you can learn something about a person by knowing his political affiliation. For example, suppose you meet a person and find out that he is a registered Democrat. You learn from that information. Perhaps you can make an intelligent guess at the person's opinions about different

issues, such as the role of the government in public schools. Maybe this information helps you guess at the person's age, background, or where he lives. Here we illustrate learning about the location of a person's home by knowing his political affiliation.

Suppose that you are shopping in a large mall in a metropolitan area. The people who shop at this mall either live downtown or in the suburbs. Recently a market research firm surveyed mall shoppers — from this survey, they believe that 70% of the shoppers live in the suburbs and 30% live downtown.

You pick a shopper at random and are interested where he lives. There are two models — either he lives in the suburbs or he lives downtown.

(a) Based on the information above, assign prior probabilities to the two models.

MODEL	PRIOR
lives in suburbs	
lives downtown	

(b) You know that there is a relationship between one's political affiliation and where the person lives. You know that 40% of the adults who live in the suburbs are registered Democrats and 80% of the downtown residents are Democrats.

Suppose that you ask the person his political affiliation and he tells you that he's a Democrat. What are the new probabilities that he lives in the suburbs or downtown? You'll find these using Bayes' rule.

(c) Next we find the likelihoods. Find the probability that he is a Democrat if he lives in the suburbs. Also find the probability he's a Democrat if he lives downtown. Put these two probabilities in the LIKELIHOOD column below.

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POSTERIOR
lives in suburbs				
lives downtown				

(d) Complete the above Bayes' table and find the posterior probabilities that the person lives in the two locations. (Multiply prior and likelihood values to get products, sum the products, and divide each product by the sum.)

(e) Describe briefly how your probabilities about the shopper's residence has changed by knowing his political affiliation.

Activity 15-16: Is the Defendant the Father?

A man is accused to be the father of a child in a paternity suit. The defendant is found to have a genetic marker that appears in 1% of the adult population. This genetic marker is also found in the child. This particular marker could only be transmitted to the child through the father and the child is 100% certain of acquiring the marker if his father has it.

(a) There are two possible models — either the defendant is the father or he isn't the father. Suppose that a juror is undecided between guilt or innocence and assigns "is the father" a prior probability of .5. Given the above genetic evidence, find the posterior probability of guilt.

(b) Suppose another juror is pretty sure initially that the defendant is not guilty and thinks that the he is the father with the small probability .1. Find this juror's posterior probability of guilt.

Activity 15-17: If the Cubs are Leading After the Fifth Inning, Will They Win?

Suppose you are watching a baseball game. A game consists of nine innings and your favorite team is leading after five innings. At this point in the game, your team has scored more runs than the opposing team. Will they win the nine inning game?

The probability that your team wins the game depends on their ability and the ability of the opposing team. Based on your knowledge of the team and their past success in winning previous games, you can make an intelligent prediction before the game starts whether the team will win or lose. After you watch them play five innings, your opinion about the team winning can change. In particular, if the team is leading after five innings, you presumably would be more confident of a victory. Here we will compute the probability of a team winning using Bayes' rule.

We will focus on the 1989 Chicago Cubs. The Cubs had a very successful season in 1989, winning 57% of their games. Looking at the records of the 1989 season, we can study the relationship between the team's performance after five innings and the game result (win or loss). In the 91 games in which the Cubs won, they were leading after five innings for 75 of the games. Equivalently, they were leading after five innings for 82% of the games they won. In 67 games in which the Cubs lost, they were leading after five innings for 13 games. In other words, they were leading after five innings for 19% of the games they lost.

(a) In the game you are watching, the Cubs are playing a mediocre team. Before the game starts, you think the probability that the Cubs will win this particular game is .6. Now you watch the game and the Cubs are leading after five innings. What is your new probability that the Cubs will win the game?

Use a Bayes' table to perform this calculation. There are two models "the Cubs win" and "the Cubs lose". The data result is "the Cubs are leading after five innings". Show your calculations in a table with columns MODEL, PRIOR, LIKELIHOOD, PRODUCT and POST.

(b) Suppose that the Cubs are playing a good team and you think before the game that they will win with probability .4. If they are winning after five innings, what is the updated probability of winning this game?

Activity 15-18: Likelihoods (cont.)

Suppose a child is thinking about purchasing a "grab bag" at a toy store. There are three bags to choose from – the first bag contains one car and two dolls, the second bag contains two cars and one doll, and the third bag contains three cars.

(a) Suppose that the child chooses one of the bags – a *model* is the contents of the bag. List the three possible models.

(b) To learn which bag she is holding, the child will choose a toy at random from the bag. Suppose the toy chosen is a car. So the data is "choose a car". For each model, find the likelihood. (This is the chance of choosing a car for each possible composition of the grab bag.) Put your answers in a table with two columns labeled "MODEL" and "DATA".

(c) Suppose instead that the toy that the child picked out of the bag was a doll. Find the likelihoods for all of the models.

WRAP-UP

In this topic, we have been introduced to Bayes' rule, which is our main recipe for changing our probabilities for our collection of models when data is observed. Two methods have been used to compute posterior probabilities, a Bayes' box and a Bayes' table when we use the "multiply, sum, divide" method. In the next topic, we will use Bayes' rule to perform inferences about the proportion of a population.

Topic 16 - Learning About a Proportion

Introduction

In the previous topic, we considered the general inference problem of learning about a model based on data. Our process of learning was based on the use of probability to express our uncertainty about models, and on the use of Bayes' rule to update these probabilities when data is observed.

In this topic, we focus on learning about a population proportion. Suppose that the administration at your school is concerned about the small number of undergraduate university students who stay on campus over weekends. In this setting, the population of interest is the current undergraduate student body at your school. We can divide this population into those students who stay on campus on weekends and those who go away. Let p denote the proportion of all students who stay on campus. The administration doesn't know the actual value of p, since it is impossible to survey every undergraduate student. However, they can learn something about this proportion if they take a sample survey of students.

Suppose the administration takes a random sample of 100 students and asks each the question "Do you regularly stay on campus over weekends?" 34 students answer "yes". What has the administration learned about the true value of p?

In this topic, we'll use Bayes' rule to learn about a population proportion. After data is observed, then our knowledge about the proportion p is contained in its probability distribution. We perform inferences such as an interval estimate for p by computing appropriate summaries of this probability distribuiton.

PRELIMINARIES

1. Suppose that you purchase a music CD from your local store. What is the chance that the CD that you buy will be defective?

2. In professional baseball, what would be the batting average for an "average hitter"?

3. In baseball, what would be the batting average for a "weak hitter"?

4. In baseball, what would be the batting average for a "great hitter"?

5. Suppose that you spin a penny on its side. What do you think is the chance that it will fall heads?

6. Suppose that you spin the penny ten times and observe 8 heads. Does this mean that the chance of getting head is over .5?

IN-CLASS ACTIVITIES

Activity 16-1: Is the machine working?

Suppose a factory is manufacturing widgets. These parts are produced by machines that are prone to malfunction. Every hour that a given machine is in operation, four widgets are produced. The widgets that come out of the machine are inspected. If a particular widget is free of any defects, then it is considered *acceptable*; otherwise it is labeled *defective*.

A machine that is in good working order will generally produce acceptable widgets. Suppose that this "good" machine will produce 75% acceptable widgets and 25% defective widgets. However, this machine can malfunction after many hours of production. In this "broken" mode, the machine produces 50% acceptable widgets and 50% defective widgets.

A foreman is supervising the production of widgets made by a particular machine. She is inspecting the widgets and counting the number of defectives that the machine produces each hour. If the machine is producing too many defectives, then she may decide that it is broken. In this case, she will stop the machine and make the necessary repairs so the machine is returned to good working order.

During one hour, suppose that the foreman inspects the widgets produced and observes the following sequence:

WIDGET 1	WIDGET 2	WIDGET 3	WIDGET 4
acceptable	defective	defective	defective

Should the foreman stop the machine?

The model

The foreman is uncertain about the state of the machine and she would like to learn about its state by observing the quality of the four widgets that are produced. The state of the machine is the *model* and there are two possible models — *good machine* or *broken machine*. We can describe each model by the proportion of acceptable parts it produces — we call this proportion p.

- If the machine is in good working order, it produces 75% acceptable parts — $p = .75$.
- If the machine is broken, it produces 50% acceptable parts — $p = .5$.

Instead of good machine or broken machine, we will refer to the models by the values $p = .5$ and $p = .75$.

The prior

The foreman has some opinion initially about the chance that her machine is in good working order. From past records, she knows that the machine is working well 90% of the time. So the probability that the machine is good ($p = .75$) is .9 and so the probability the machine is broken ($p = .5$) is $1 - .9 = .1$. We write the two models and associated probabilities in table form:

MODEL (state of machine)	PRIOR
$p = .75$ (good working order)	.9
$p = .5$ (broken)	.1

The likelihoods

Let's suppose that the foreman observes the first widget manufactured from the machine and notes that it is acceptable. How does this change her opinion about the condition of the machine?

She will adjust her opinion about the machine being in good working order or broken by use of Bayes' rule. First, we need to compute the likelihoods. These are the probabilities of the data result "acceptable widget" for each model.

(a) Suppose the machine is in good working order ($p = .75$). What is the probability of observing an acceptable widget? _____ Put this value in the first row of the LIKELIHOOD column.

(b) If the machine is broken ($p = .5$), what is the probability of producing an acceptable part? _____ Put this value in the second row of the LIKELIHOOD column.

1st data result = "acceptable widget"

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POSTERIOR
$p = .75$	.9			
$p = .5$	.1			
SUM				

Bayes' rule

Now we can find our new probabilities of the two models using Bayes' rule. Remember our recipe (MULTIPLY, SUM, DIVIDE).

(c) Complete the PRODUCT and POSTERIOR columns of the table.

(d) Comment about the difference between the prior and posterior probabilities.

Now suppose the foreman observes the second widget off of the machine — it is defective. She wants to update her probabilities.

2nd data result = "defective widget"

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POSTERIOR
$p = .75$				
$p = .5$				
SUM				

(e) The prior probabilities now refer to the probabilities *before* observing this second part. These are the same as the posterior probabilities *after* observing the first part. Copy the probabilities from the POSTERIOR column in the first table and put them in the PRIOR column of the new table above.

(f) Compute the new likelihoods corresponding to observing a defective. These are the (1) probability of getting a defective if $p = .75$ and (2) the probability of getting a defective if $p = .5$. Put these numbers in the LIKELIHOOD column.

(g) Find the posterior probabilities using Bayes' rule. Complete the table.

(h) The foreman now observes a third part from the machine — it is defective. Find the new probabilities using Bayes' rule. Put your work in the table below. (Remember the PRIOR probabilities are the ones prior to observing the third part — after the first two parts — and the POSTERIOR probabilities are the values after all three parts are observed.)

3rd data result = "defective widget"

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POSTERIOR
$p = .75$				
$p = .5$				
SUM				

(i) The fourth part that comes from the machine is also defective. Find the new posterior probabilities of the two models in the below table.

4th data result = "defective widget"

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POSTERIOR
$p = .75$				
$p = .5$				
SUM				

In the above work, we updated our model probabilities one observation at a time. If we are interested only in the probabilities after observing the conditions of all four widgets, there is a quicker way of doing the Bayes' rule computation.

We start with a Bayes' table where the PRIOR column contains the probabilities before any data is observed.

data result = "acc, def, def, def"

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POSTERIOR
$p = .75$	.9			
$p = .5$	.1			
SUM				

Our data is the results of the four inspections, which we abbreviate as

$$\text{data} = \{\text{acc, def, def, def}\}$$

What is new is how the likelihoods are computed. For each model, the likelihood is the probability of the data $\{\text{acc, def, def, def}\}$. We find this by multiplying the probabilities of the individual outcomes.

$$\text{LIKELIHOOD} = \text{PROB}(\text{acc}) \times \text{PROB}(\text{def}) \times \text{PROB}(\text{def}) \times \text{PROB}(\text{def}).$$

(j) Compute the likelihood for each model (show your work below). Put your answers in the LIKELIHOOD column.

(1) Likelihood for $p = .75$:

$$\text{LIKELIHOOD} = \_\_\_ \times \_\_\_ \times \_\_\_ \times \_\_\_ =$$

(2) Likelihood for $p = .5$:

$$\text{LIKELIHOOD} = \_\_\_ \times \_\_\_ \times \_\_\_ \times \_\_\_ =$$

(k) Find the posterior probabilities.

(l) Compare your answer with the final posterior probabilities when we updated one observation at a time. (They should be approximately equal.)

(m) Suppose that the foreman will stop the machine if the probability that the machine is broken is over one half? Should she stop the machine? Why?

Activity 16-2: How good is this hitter?

Let's suppose that you are a scout for a major league baseball team. You are interested in the hitting ability of a particular player which we will call Mickey. You have some opinion about Mickey's skill as a hitter based on some general knowledge about baseball hitting and some reports that you've been given. You plan on watching him bat during a single baseball game. On the basis of Mickey's hitting success in this game, you plan to revise your opinion about his hitting ability and perhaps decide to offer him a contract to play for your team.

How does one measure a baseball hitter's batting ability? The usual measure is the batting average. This is a player's proportion of base hits if he is given a large number of chances to hit. What are values of typical batting averages? Most of the batting averages of players that play regularly in the major leagues fall between .2 and .35 — an average batting average is around .27. It is very rare for a hitter to hit under .2 or over .4. If a hitter's batting average is under .2, it is unlikely that he will remain on a major league team. It has been over 50 years since a player has had a batting average over .4 for an entire season.

We'll use this example to illustrate constructing a prior for Mickey's batting average p and making inferences about this proportion.

The prior

You are interested in constructing a prior distribution for Mickey's batting average. You do this in two steps. First, think of a plausible list of batting averages p. This list includes any batting averages for Mickey that you think are possible. If you think he might turn out to be a poor hitter, then you would include small values of p. In contrast, if you think he could be another Ted Williams (the last hitter to bat for .400 in a single season), you would include the value $p = .4$.

After you have a list of possible batting averages, you assign probabilities to the values in the list which reflect your beliefs in their relative likelihoods. You assign large probabilities to the values of the batting average p that you think are relatively likely. The batting averages which you think are unlikely are given relatively small probabilities.

For ease of presentation, we assume a simple prior distribution for Mickey's batting average. Since most batting averages fall between .2 and .4, we will let p be .2, .3 or .4 — as the following table indicates, these batting averages are descriptive in terms of Mickey's batting ability.

MODEL	DESCRIPTION
$p = .2$	poor hitter
$p = .3$	average hitter
$p = .4$	great hitter

What probabilities should be assigned to the three proportion values? As a scout, what do you know about batting averages? In the major leagues, most of the regular players have batting averages between .2 and .3. The average hitting proportion is .27, so .3 is likely more common than .2. A batting average of .4 would be considered unusually large. If Mickey's hitting ability is representative of all major league players, then the above information would be reflected in the following set of prior probabilities:

MODEL	PRIOR
$p = .2$	.4
$p = .3$	.59
$p = .4$	.01

If Mickey is an unusually hot prospect, then you might want to adjust these prior probabilities by giving the proportion $p = .3$ a larger value. For the remainder of this activity, we'll assume the above prior distribution.

The likelihoods and posterior probabilities

Now you watch Mickey play the baseball game. He gets 4 opportunities to bat and gets 2 hits. You would like to update your prior opinion about Mickey's batting ability on the basis of this data.

It is helpful to think of this data as the sequence of observations

H, H, O, O,

where H stands for a hit and O an out. (It doesn't matter what order we write the H's and O's — any order where Mickey gets 2 hits and 2 outs would work.)

The likelihood is the probability of this data result $\{\text{H, H, O, O}\}$ for a given model, or value of the proportion p. Suppose that $p = .2$ — Mickey has a .200 batting average. The probability that he gets a hit is .2 and therefore the probability he gets an out is $1 - .2 = .8$. We find the probability of the data result by multiplying the individual probabilities of hits and outs:

$$\begin{aligned} \text{PROB}(\{\text{H}, \text{H}, \text{O}, \text{O}\}) &= \text{PROB}(\text{H}) \times \text{PROB}(\text{H}) \times \text{PROB}(\text{O}) \times \text{PROB}(\text{O}) \\ &= .2 \times .2 \times .8 \times .8 \\ &= .0256 \end{aligned}$$

This is the likelihood value for the model $p = .2$; we enter this into the following Bayes' table:

data = "2 hits in 4 at-bats"

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POSTERIOR
$p = .2$	.4	. 0256		
$p = .3$	.59			
$p = .4$	.01			
SUM				

(a) Using a similar formula to the one above, find the likelihood of {H, H, O, O} if p = .3, and the likelihood if $p = .4$. Do your work in the space below. Put the likelihood values in the table above.

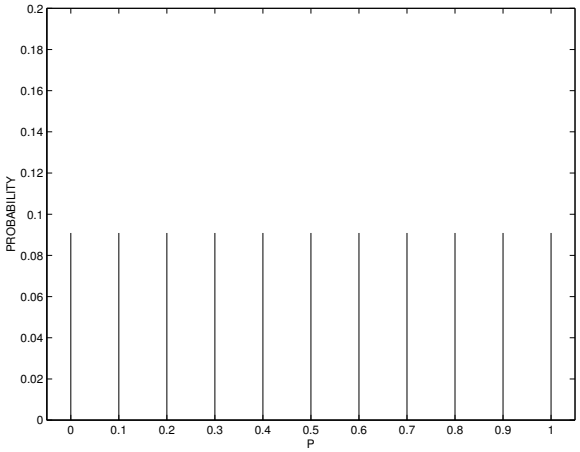
(b) Using the "product, sum, divide" recipe, find the posterior probabilities of the three batting averages.

(c) Compare the sets of prior and posterior probabilities for the three batting averages. Is it true that you now have a higher opinion about Mickey's batting ability? Explain.

Activity 16-3: Spinning a penny

Suppose that you spin a new penny on its side. (This is a different experiment than tossing a penny in the air.) If you could do this a large number of times, what do you guess is the proportion of heads? ____

We'll learn about the proportion of heads p by use of Bayes' rule. We will first construct a prior distribution for this unknown proportion. This prior distribution reflects the belief that any proportion value from 0 to 1 is equally likely. Then we will spin the penny five times and count the number of heads. We will use this data to update our probability distribution for the proportion of heads. This final probability distribution will be used to construct an interval of values that we are fairly confident contains the unknown value of p.



Prior distribution for the proportion of heads of a spinning penny.

The prior

You might have some idea about the proportion of heads in this penny spinning experiment. You might think that this experiment is similar to tossing a penny in the air and so the proportion is near .5. Alternatively, you might know something about how a penny is designed and believe that this spinning a penny is very different from tossing it — this might suggest that the proportion of heads is larger or smaller than one half. Perhaps you really are ignorant about the true proportion of heads. In this situation, there is a convenient method of constructing a prior distribution. Since the values of a proportion fall between 0 and 1, make a list of equally spaced values from 0 to 1. In this case, we'll let the proportion p be equal to the values 0, .1, .2, ..., .9, 1. Then we let each proportion value have the same probability. This prior information reflects your belief that the proportion could possibly be any value from 0 and 1 and you have no reason to think that any one value is more or less likely than another value.

This "vague" prior distribution is graphed below. Each of the eleven proportion values has a probability of 1/11 = .091. We will call this a "flat" prior due to the flat appearance of the bars in the graph.

The data

Now we need some data. Spin a penny on your desk five times and count the number of heads. Each time the penny should spin for a least a few seconds so that each trial of the experiment is done under similar conditions.

The number of heads I observed was ____

The posterior probabilities

To find the updated probabilities of the proportion values, we apply Bayes' rule. Remember we have eleven possible proportion values — p = 0, .1, .2, ..., 1. For each value, we compute the product

$$\text{LIKELIHOOD} \times \text{PRIOR},$$

where LIKELIHOOD is the probability of your data, say H, T, H, T, H, for the given proportion value, and PRIOR is the prior probability .091. For example, for the value $p = .4$ and data result H, T, H, T, H, this product is given by

$$(.4 \times .6 \times .4 \times .6 \times .4) \times (.091).$$

You need to compute this product for each of the eleven proportion values p = 0, .1, ..., 1. Then you would sum the products and divide each product by the sum to get posterior probabilities. This is extremely tedious to do by hand, but fortunately easy to do by computer. I have done all of the computations so we can focus on the interpretation of the posterior probabilities.

On the next page, all of the possible posterior distributions for our set of prior probabilities are displayed for five spins of the penny. You find your posterior distribution by matching it to your observed number of heads. For example, if you observed 2 heads, look at the distribution under the label "2 HEADS OUT OF 5 SPINS". Circle your posterior probabilities for your observed data.

(a) Graph your posterior probabilities on the empty graph on the following page. Describe in words the general pattern of the graph.

(b) What is the most likely proportion value?

(c) What's the probability that the proportion p is equal to .7?

(d) Find the probability that p is greater than .5.

0 HEADS IN 5 SPINS		1 HEADS IN 5 SPINS		2 HEADS IN 5 SPINS	
p	POSTERIOR	p	POSTERIOR	p	POSTERIOR
0.0	0.453	0.0	0.000	0.0	0.000
0.1	0.267	0.1	0.202	0.1	0.044
0.2	0.148	0.2	0.252	0.2	0.123
0.3	0.076	0.3	0.222	0.3	0.185
0.4	0.035	0.4	0.159	0.4	0.207
0.5	0.014	0.5	0.096	0.5	0.188
0.6	0.005	0.6	0.047	0.6	0.138
0.7	0.001	0.7	0.017	0.7	0.079
0.8	0.000	0.8	0.004	0.8	0.031
0.9	0.000	0.9	0.000	0.9	0.005
1.0	0.000	1.0	0.000	1.0	0.000

3 HEADS IN 5 SPINS		4 HEADS IN 5 SPINS		5 HEADS IN 5 SPINS	
p	POSTERIOR	p	POSTERIOR	p	POSTERIOR
0.0	0.000	0.0	0.000	0.0	0.000
0.1	0.005	0.1	0.000	0.1	0.000
0.2	0.031	0.2	0.004	0.2	0.000
0.3	0.079	0.3	0.017	0.3	0.001
0.4	0.138	0.4	0.047	0.4	0.005
0.5	0.188	0.5	0.096	0.5	0.014
0.6	0.207	0.6	0.159	0.6	0.035
0.7	0.185	0.7	0.222	0.7	0.076
0.8	0.123	0.8	0.252	0.8	0.148
0.9	0.044	0.9	0.202	0.9	0.267
1.0	0.000	1.0	0.000	1.0	0.457

(e) There is a blank table under your graph. There are three columns — P, POSTERIOR PROBABILITY, and CUMULATIVE PROBABILITY. Using your set of posterior probabilities, write down the values of p and associated probabilities from the most likely to least likely values. First, find the most likely value of p and put this value and its probability on the first line. Next, find the second most likely value of p — put it and its probability on the second line. Continue in this fashion until you have filled up the table.

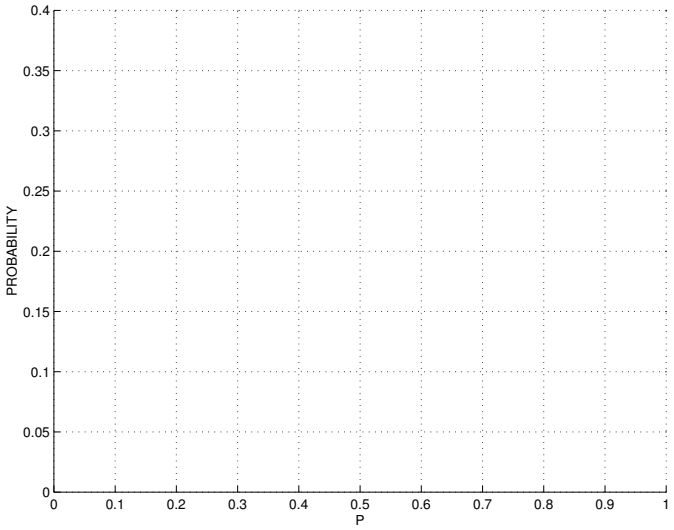
Last, fill in the cumulative probabilities. For each row, find the probability of its row and all the rows above it. If you have done this correctly, the last entry in this column should be approximately 1.

(f) Using this table, it is easy to construct an interval estimate for our proportion. We'll construct two intervals — a 50% one and a 90% one.

To find a 50% estimate, read down the cumulative probability column until you find a value equal or exceeding .5. Circle this value. Also, circle the values of the proportion p that are in this row and above. This set of proportion values is called a *50% probability set.* The probability that the proportion is in this set is at least 50%.

A 90% estimate is found in a similar fashion. Read down the cumulative probability column until you find a value .9 or above. Put a box around this value. The proportion values in this row and above make up a 90% probability set. Put a box around these values.

For each of the 50% and 90% sets, find the smallest and largest values of p. These smallest and largest values make up a probability interval for p. Put these values on the lines at the bottom of the page.



POSTERIOR PROBABILITIES FOR PENNY SPINNING EXAMPLE

	p	Probability	Cumulative Probability
most likely			
least likely			

50% set: _____

90% set: _____

Activity 16-4: Marriage ages

In Activity 2-8, we are interested in learning about the proportion of marriages in which the husband is older than the wife. Specifically, we will focus on Cumberland County, Pennsylvania, and learn about p, the proportion of *all* marriages in this particular county in which the husband is older.

(a) Make an intelligent guess at the value of the proportion p.

(b) Give an interval, like [.2, .9], in which you are pretty confident contains the unknown value of p. (Actually, I am positive that p falls in the interval [0, 1], but you can probably think of a shorter interval.)

Looking back at Activity 2.8, we are given the ages of a sample of 24 couples taken from marriage licenses filed in Cumberland County, Pennsylvania. Looking at the list, you will see that 2 couples had the same age, the husband was older for 16 couples, and the wife was older for the remaining 6 couples.

(c) Of the couples in the sample which have *different* ages, what is the proportion of couples in which the husband is older?

(d) Is it correct to say that your answer in (c) is the same as the proportion of *all* marriages in this county in which the husband is older? Explain.

To learn about the value of the proportion p, we'll use Bayes' rule for a large number of possible models. Suppose that the proportion can be any one of the values

$$p = 0, .01, .02, .03, ..., .97, .98, .99, 1.$$

We'll suppose that all 101 proportion values are reasonable, and will assign a flat uniform prior distribution where each value of p is assigned the probability 1/101.

Bayes' rule is easy to describe, but very tedious to apply in this setting. For each value of the proportion, we compute the likelihood which is the probability of the data result $\{$husband older for 16 couples, wife older for 6 couples$\}$. For example, if p is equal to .2, then we would compute

$$\text{LIKELIHOOD} = .2^{16}(1 - .2)^6.$$

We would compute this likelihood for each of the 101 proportion values. Then to compute our posterior probabilities, we do the usual "product, sum, divide" algorithm.

Using the Minitab program 'p_disc', I have computed the posterior probabilities of the 101 values of the proportion. The probabilities together with a graph are presented on the next page.

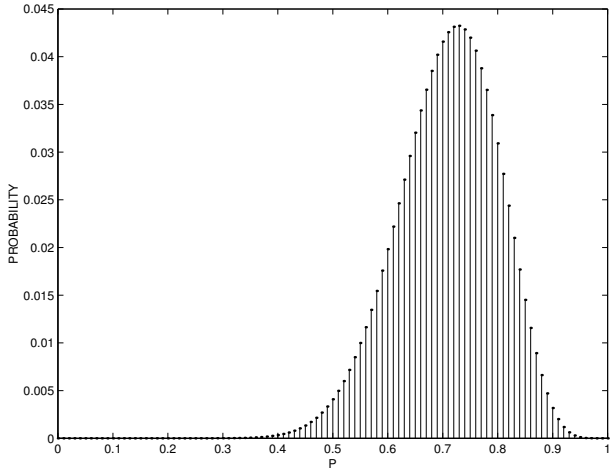
(e) Looking at the graph and table, what are the most likely value(s) of the proportion p? List these values below.

(f) What is the probability that the proportion p is larger than .5? (Look at the table. Find the probability that p is .5 or smaller — if we subtract this number from 1, we get the probability that p is larger than .5.)

(g) Suppose we're interested in a probability interval for this proportion. Look at the table of posterior probabilities. In the space below, put down the most likely values of p and their associated probabilities. (The most likely values come first, then the next most likely values, and so on.) I have put in the first four values for you. Continue until you have put down approximately 15 values of the proportion p and their probabilities.

PROPORTION P	PROBABILITY
.71, .72, .73, .74	.043, .043, .043, .043

(h) You have just constructed a probability interval. The proportion p lies between _____ and _____ with probability _____.



p	POSTERIOR	p	POSTERIOR	p	POSTERIOR	p	POSTERIOR
.00	.000	.26	.000	.51	.005	.76	.041
.01	.000	.27	.000	.52	.006	.77	.039
.02	.000	.28	.000	.53	.007	.78	.037
.03	.000	.29	.000	.54	.008	.79	.034
.04	.000	.30	.000	.55	.010	.80	.031
.05	.000	.31	.000	.56	.012	.81	.028
.06	.000	.32	.000	.57	.013	.82	.024
.07	.000	.33	.000	.58	.015	.83	.021
.08	.000	.34	.000	.59	.018	.84	.018
.09	.000	.35	.000	.60	.020	.85	.015
.10	.000	.36	.000	.61	.022	.86	.012
.11	.000	.37	.000	.62	.025	.87	.009
.12	.000	.38	.000	.63	.027	.88	.007
.13	.000	.39	.000	.64	.030	.89	.005
.14	.000	.40	.000	.65	.032	.90	.003
.15	.000	.41	.000	.66	.034	.91	.002
.16	.000	.42	.001	.67	.037	.92	.001
.17	.000	.43	.001	.68	.039	.93	.001
.18	.000	.44	.001	.69	.040	.94	.000
.19	.000	.45	.001	.70	.042	.95	.000
.20	.000	.46	.002	.71	.043	.96	.000
.21	.000	.47	.002	.72	.043	.97	.000
.22	.000	.48	.003	.73	.043	.98	.000
.23	.000	.49	.003	.74	.043	.99	.000
.24	.000	.50	.004	.75	.042	1.00	.000
.25	.000						

HOMEWORK ACTIVITIES

Activity 16-5: Does Frank have ESP?

Frank claims to have extra sensory perception (ESP). You are skeptical and so you ask Frank to participate in a card reading experiment. You have a deck of cards with equal numbers of ♠'s, ♡'s, ♣'s and ♢'s. You shuffle the deck and hold a card up with the face away from Frank. Frank will then tell you the suit of the card. For the next trial of the experiment, you put the card back into the deck, reshuffle, and draw a new card. The experiment is completed when ten cards have been used.

Suppose Frank was able to participate in an ESP experiment with a large number of cards. Let p be the proportion of cards in which he correctly identifies the suit. There are two models of interest: $p = .25$, which says that Frank does not have ESP and is essentially guessing at the suit of the card, and $p = .5$ which says that Frank does have some ability to detect the suit.

(a) Suppose that you are very skeptical that Frank has any ESP ability. Assign probabilities to the two models which reflect this skepticism (different answers are possible).

MODEL	PRIOR
$p = .25$	
$p = .5$	

(b) Suppose that Frank gets the first card correct — we'll represent this by the letter C. To update our probabilities, we compute likelihoods. These are the probabilities of "card correct" (C) for each model.

(1) If $p = .25$ (Frank is guessing), what is the probability of C?

(2) If $p = .5$ (Frank has some ESP), what is the probability of C?

Enter these two numbers in the LIKELIHOOD column of the Bayes' table below.

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POSTERIOR
$p = .25$				
$p = .5$				

(c) Using the "multiply, sum, divide" (MSD) recipe, complete the Bayes' table to find the posterior probabilities of the two models.

(d) Next, suppose that Frank gets the second card wrong — we'll call this W. Find the new model probabilities and show your work in the Bayes' table below. (First, copy the probabilities in

the POSTERIOR column of the first table into the PRIOR column of the second table. Next, compute the likelihoods: the probabilities of W for each model. Last, use the MSD recipe to find the posterior probabilities.)

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POSTERIOR
$p = .25$				
$p = .5$				

(e) In the table below, place all of your model probabilities. The first column contains the prior probabilities, the second column (POST1) contains the posterior probabilities after one observation (C) and the third column (POST2) contains the posterior probabilities after two observations (C, W). Describe how your probabilities of "Frank is guessing" and "Frank has ESP" have changed as you got data.

MODEL	PRIOR	POST1	POST2
$p = .25$			
$p = .5$			

Activity 16-6: How good is the shooter?

Suppose you are watching a basketball game. In basketball, there are different ways of scoring points. Most of the points are scored during the action of the game. If players are fouled, then points can also be scored by making shots from the free-throw line. Suppose a particular player called Bob gets fouled during the game and goes to the free-throw line.

You are not very familiar with Bob's shooting ability. Specifically, you are unsure if he is a poor free-throw shooter, an average shooter, or a great shooter. During the game, you watch Bob take four free-throw shots — he makes the first three and misses the last one. What is your new opinion about Bob's shooting ability?

First we explain what is meant by a poor shooter, an average shooter, or a great shooter. The quality of a shooter is measured by the proportion of shots he makes in many opportunities during the season. Call this proportion p. We think of a poor free-throw shooter as one who will only make, in the long run, 50% of his shots; that is, $p = .5$. An average shooter will make 67% of his shots ($p = .67$), and a great free-throw shooter will make 83% of his shots ($p = .83$).

(a) Here there are three models or values of the shooting proportion p. Suppose you think that all three models are equally plausible. Fill in probabilities in the PRIOR column which reflect this belief.

MODEL	PRIOR
$p = .5$	
$p = .67$	
$p = .83$	

(b) New information about the player's shooting ability is available after watching this basketball game. Remember Bob took four shots with the results G, G, G, M, where G represents a good shot and M represents a miss. To update our model probabilities, we first compute likelihoods. We compute the probability of the data {G, G, G, M} for each model.

If Bob's shooting ability is 50% ($p = .5$), then the probability of this data is

$$\begin{aligned}&\text{PROB(G)} \times \text{PROB(G)} \times \text{PROB(G)} \times \text{PROB(M)}\\ &= .5 \times .5 \times .5 \times (1 - .5) = .0625\end{aligned}$$

We enter this number in the LIKELIHOOD column in the table below for the $p = .5$ model.

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POSTERIOR
$p = .5$		.0625		
$p = .67$				
$p = .83$				

(c) Repeat this likelihood calculation for the $p = .67$ and the $p = .83$ models. Put your answers in the LIKELIHOOD column of the table.

(d) Compute the posterior probabilities of the three models. Put all of your work in the table.

(e) Compare the prior and posterior probabilities of the three models. What do you now think about the shooting ability of the player?

Activity 16-7: Is the coin fair?

Suppose that a friend is holding two pennies — one is the usual kind with a heads and tails and the second is a special two-headed coin with heads on both sides. He chooses one penny and starts tossing the coin. You wish to learn which coin is being tossed on the basis of the coin flips.

Let p denote the proportion that the coin lands heads in many tosses. There are two models for this proportion. If the coin has a heads and a tails, then $p = 1/2$. Otherwise, if the coin is two-headed, a head will always land and $p = 1$.

1. If you don't know at the beginning which coin is being tossed, put prior probabilities for the two models in the table below.

MODEL	PRIOR
$p = .5$	
$p = 1$	

2. Suppose that the coin is tossed once and a head is observed. Find the posterior probabilities for the models. (You can put your work in the table below.)

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POSTERIOR
$p = .5$				
$p = 1$				

3. The coin is now tossed a second time and a head is again observed. Find the (new) posterior probabilities.

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POSTERIOR
$p = .5$				
$p = 1$				

4. The coin is tossed a third time and a tail is observed. Without doing any calculation, find the new posterior probabilities.

Activity 16-8: Does *Sports Illustrated* have a high proportion of ads?

Suppose you like to read the magazine *Sports Illustrated*, but you've been frustrated with all of the ads that have appeared in recent issues. You wonder about the proportion of pages in the magazine that contain ads. Call this proportion p — this represents the proportion of all pages of current issues of *Sports Illustrated* that contain at least one ad.

(a) Based on your knowledge about the ad content of current magazines, guess at the value of p.

(b) Give an interval of values of p (like (.1, .9)) which you are pretty confident contains the true value.

For your models, let the proportion p take the 11 equally spaced values 0, .1, .2, ..., 1. Assume a flat prior in which each value of p has the same prior probability.

Next you need some data. The current issue of *Sports Illustrated* contains 78 pages. You take a random sample of 20 of these pages. In the table below, the page numbers are listed; next to each page is the observation if this page contained an ad or not.

PAGE	AD?	PAGE	AD?
2	No	30	No
3	No	36	Yes
5	Yes	37	Yes
9	Yes	42	No
10	No	49	No
12	Yes	50	No
13	No	63	No
20	No	65	No
21	No	66	No
24	No	74	Yes

(c) Based on this sample, guess at the value of p.

(d) To find the posterior probabilities, the program 'p_disc' is used. Output from this program is displayed below.

p	PROBABILITY
.0	.000
.1	.019
.2	.229
.3	.403
.4	.261
.5	.078
.6	.010
.7	.000
.8	.000
.9	.000
1.0	.000

Find the 3 most likely values of the proportion p.

(e) What is the probability that p is one of these three values in part (d)?

(f) Suppose a friend of yours claims that at least half of *Sports Illustrated*'s pages contains ads. What is the probability that p is at least .5? Is your friend right?

Activity 16-9: Why do people vacation in Rotterdam?

A study was carried out in Rotterdam, the Netherlands, to find out the reasons why tourists come to the city. A survey was given to 450 visitors to the Museum of Fine Arts. Of these visitors, 243 indicated that they came to Rotterdam for the purpose of visiting the art museum. The people

conducting the study were interested in learning about the proportion p of all Rotterdam tourists who came to the city to visit the art museum.

(a) Suppose we assume that the sample taken is representative of all Rotterdam tourists. For proportion models, we use the values $p = 0, .01, .02, ..., .99, 1$, and assume a flat prior on these 101 possible values. Here our data is {243 came to visit the museum, 207 didn't come to visit the museum} and a portion of the posterior probabilities (computed using the program 'p_disc') are presented below.

p	PROBABILITY	p	PROBABILITY
.44	.000	.54	.170
.45	.000	.55	.155
.46	.001	.56	.118
.47	.002	.57	.075
.48	.007	.58	.039
.49	.018	.59	.017
.50	.040	.60	.006
.51	.076	.61	.002
.52	.119	.62	.000
.53	.155	.63	.000

(b) What is the most likely value of the proportion p? What is its probability?

(c) What is the probability that less than 50% of the visitors come to Rotterdam for the art museum?

(d) Find the probability that p is in the interval [.52, .56].

(e) Based on this survey, it was concluded that museum visits may very well function as a pull factor on urban visitors. However, is it reasonable to assume (as we did) that the sample taken is representative of all visitors to Rotterdam? Explain how this sample could be biased.

Activity 16-10: Are people going to see "The Phantom Menace"?

In May 1999, there was quite a bit excitement generated by the debut of the movie "The Phantom Menace". To learn about the opinion of American adults about the release of this movie, the Gallup organization polled a sample of 1,014 adults selected at random from the country. One question asked about the adult's plans to see the movie. In the sample, 385 indicated that they planned to see the movie in the theater.

(a) Let p denote the proportion of all American adults that will see the movie in a theater. Based on the sample data, give your best estimate at the value of p.

Suppose the proportion p can take on any of the 101 values 0, .01, .02, ..., .99, 1, and a flat prior is assigned to these values. The posterior probabilities for p were computed using a computer and a portion of them are presented below.

p	PROBABILITY	p	PROBABILITY
.30	.000	.40	.109
.31	.000	.41	.038
.32	.000	.42	.009
.33	.001	.43	.001
.34	.008	.44	.000
.35	.038	.45	.000
.36	.112	.46	.000
.37	.214	.47	.000
.38	.262	.48	.000
.39	.209	.49	.000

(b) What is the most likely value of p? Does this value agree with your "best guess" that you made in part (a)?

(c) Find the probability that under 40% of adult Americans will watch the movie in theaters.

(d) Find a 95% probability interval for p.

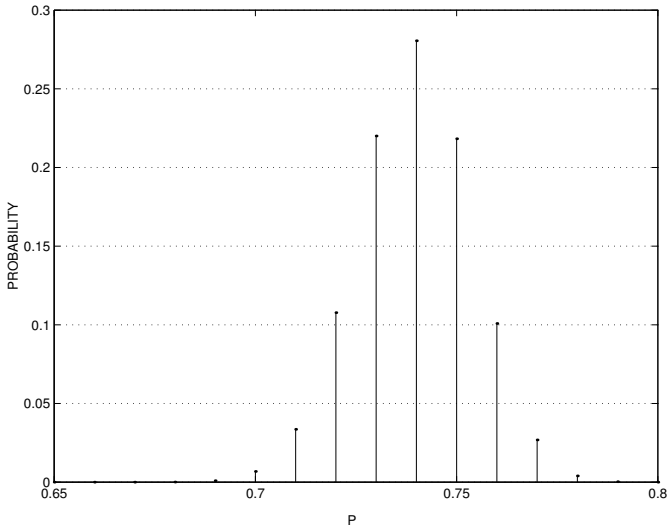
(e) In the Gallup poll article, it was stated that the margin of sampling error is plus or minus 3 percentage points. In other words, the value of the population proportion p is stated to be within .03 of the proportion computed from the sample. Verify this by using your answer from part (a) and your probability interval that you found in part (d).

Activity 16-11: Clean Air on Cruise Ships

The article "Clean Air on Cruise Ships" in the *Cornell Hotel and Restaurant Administration Quarterly* (February 1999) discusses the desire of American travelers to request smoke-free hotel rooms. In a survey reported on the website `cruisenet.com`, 950 people were surveyed who were interested in going on a cruise — 74% of this sample favored smoke-free cruise ships.

(a) How many people in the sample favor smoke-free cruise-ships?

To learn about the proportion p of *all* cruise travelers who favor smoke-free cruise ships, a flat prior was placed on the proportion values 0, .01, .02, ..., .99, 1. Using this prior and the above sample



Posterior distribution for proportion of travelers who favor a smoke-free cruise ship.

data, posterior probabilities were computed. The figure below displays the posterior probability distribution. Use the figure to answer the questions below.

(b) What is the probability that exactly 75% of all travelers favor smoke-free ships?

(c) Find the probability that p is .69 or smaller.

(d) Find the probability that p is .77 or larger.

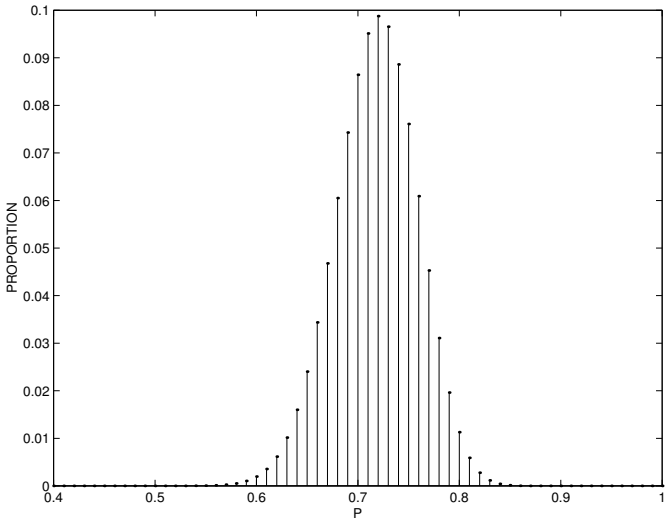
(e) Use the answers to (c) and (d) to find the probability that p falls between .70 and .76.

(f) Suppose that this survey data came from people responding to a survey that was posted at the web site cruisenet.com. Do you think this sample is representative of *all* American people that are interested in taking a cruise vacation? Explain.

Activity 16-12: Are Teachers Prepared in Technology?

The article "New Teachers and Technology: Are They Prepared" in *Technology and Learning* (April 1999), describes a mail survey that was sent to 1500 colleges of education to determine the role of technology in teacher preparation. This survey was sent to address the concern that schools of education are failing to integrate technology into their training. One hundred and twenty-two schools responded to the survey, and 88 of the schools require technology courses as part of the general teacher certification program.

Let p denote the proportion of all colleges of education that require technology courses as part of the general certification program. A flat prior was placed on the values $p = 0, .01, ..., .99, 1$. Based



Posterior distribution for proportion of colleges that require technology courses.

on the above data (88 schools requiring technology courses and 34 not), posterior probabilities for the proportion were computed. A graph of the probabilities is shown below.

(a) By looking at the graph, what value of p is most likely?

(b) What is the probability of the most likely value you found in (a)?

(c) By inspection of the graph, find an interval of values for the proportion p that contains approximately all of the posterior probability.

(d) Is it reasonable to say (on the basis of this sample) that over half of the colleges of education require some technology courses? Why?

(e) The article mentions that the schools that responded were probably *not* representative of all of the schools that were sent a survey. Why do you think that the survey was biased in this situation?

WRAP-UP

In this topic, we applied Bayes' rule for learning about a population population. One starts by making a list of plausible values for the proportion p, and then assigning probabilities to these values which reflected his or her beliefs about the likelihoods of the various proportion values. After the number of successes and failures in our dataset are observed, new (posterior) probabilities are computed using Bayes' rule. We perform inference about the proportion by summarizing this

probability distribution. One type of summarization is a probability interval which is a collection of values of p which contain a large amount of probability. In the next topic, we will introduce learning about a proportion using continuous models when p can conceivably take on any value from 0 and 1.

Topic 17: Learning About a Proportion Using Continuous Models

Introduction

In this topic, we continue our discussion about learning about a population proportion. In the previous topic, we started with listing a set of values for a proportion p, assigning prior probabilities to these values, and then obtaining posterior probabilities for the proportion values using Bayes' rule. Here we will make the more realistic assumption that the unknown proportion p can take on every possible value between 0 and 1 — we call these **continuous models** for p. We will use a beta continuous curve to represent our prior beliefs about the proportion and a beta curve will also represent our beliefs about p after data have been observed. We perform inference about the proportion by summarizing the beta curve. Specifically, we will find probabilities about the proportion p by computing areas under the beta curve.

PRELIMINARIES

1. Consider the body of undergraduate students at your school. Suppose that you are interested in the proportion of all these students that own their own credit card. We'll call this unknown proportion p.

   (a) What is the smallest possible value of p? What is the largest possible value for p?

   (b) List 13 different plausible values of the proportion p.

   (c) Do you think it is possible that p is equal to .4132? Why or why not?

   (d) On the number line below, mark all *possible* values of the proportion p.

```
 -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
-.2 -.1  0  .1  .2  .3  .4  .5  .6  .7  .8  .9  1.0 1.1 1.2  p
```

2. What is your favorite magazine? Make an intelligent guess at the proportion of pages in your favorite magazine that contain at least one ad.

3. Do you wear glasses and/or contacts? Put the responses (yes or no) for all of the students in the class in the table below.

Student	Glasses or contacts?	Student	Glasses or contacts?	Student	Glasses or contacts?
1		9		17	
2		10		18	
3		11		19	
4		12		20	
5		13		21	
6		14		22	
7		15		23	
8		16		24	

Probabilities for continuous values of the proportion

In Topic 16, we made the implicit assumption that the unknown proportion p could take on a finite set of values, such as 0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1. We constructed a prior distribution by assigning probabilities to these proportion values. After observing sample data, we used Bayes' rule to find posterior probabilities and this posterior probability distribution was used to perform inferences.

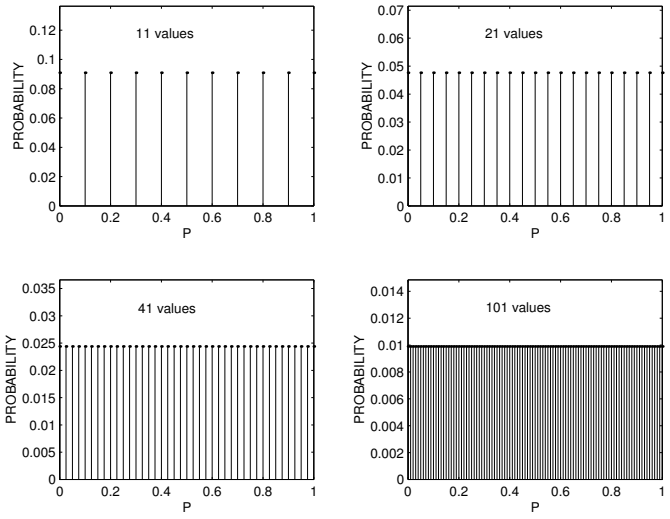
The above assignment of a prior distribution may appear to be a bit odd, since we are letting the proportion only take a specific set of values. We are saying, for example, that p could be equal to .1 or .2 — why can't this proportion be equal to a value between .1 and .2 such as .142?

To be less restrictive in our choice of prior distribution, we could allow the proportion to take on a greater number of values between 0 and 1. Let's return to our example where we are interested in the proportion p of students that possess their own credit cards. Suppose that we have no information about the value of this proportion. Here are four possible prior distributions that reflect this lack of knowledge.

- The proportion p can be any of the 11 values 0, .1, .2, ..., 1 and we assign the same probability (1/11) to each value.

- The proportion can be any of the 21 values 0, .05, .1, .15, ..., .95, 1 and we assign the probability 1/21 to each value.

- The proportion can be any of the 41 values 0, .025, .05, .075, ..., .975, 1 and a probability of 1/41 is given to each value.



Uniform probability distributions using different number of values.

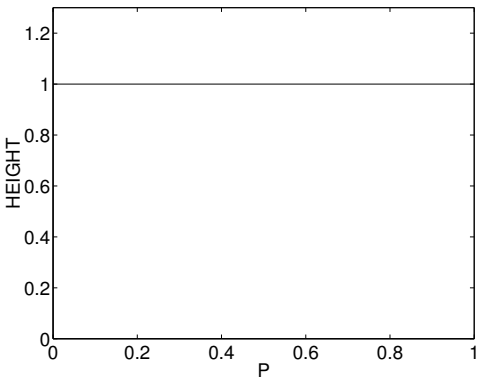
- The proportion can be any of the 101 values 0, .01, .02, ..., .99, 1 and we assign the probability of 1/101 to each value.

The four prior distributions described above are graphed using line plots in the figure below. Note that all of the lines in each graph are of the same height which reflects the fact that each possible proportion value is assigned the same probability.

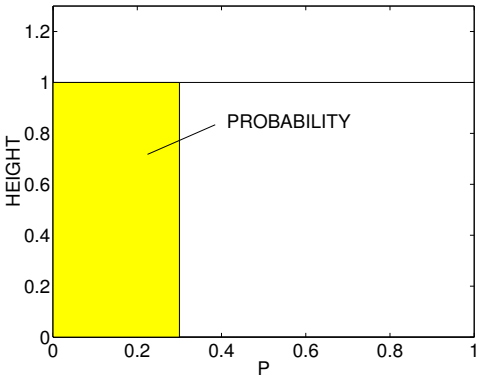
The bottom right graph represents the prior distribution when there are many possible proportion values such as .23, .54 and .75. But this distribution doesn't assign probability to *all* proportion values — for example, the values .523 or .82345 are not possible by this distribution.

What if the unknown proportion p is allowed to take on *all possible* values between 0 and 1? Then we won't be able to list all of the proportion values, since there are an *infinite* number of possibilities. In this situation, we'll need a new way of representing probabilities. Our old method of constructing a prior probability distribution won't work. To see this, suppose that we assigned a tiny number, say .000000001, to each proportion value. But if there is a huge (infinite) number of possible proportions, the sum of the probabilities over all the proportion values will exceed one. So we can't assign a small probability to each proportion value, and still have a proper probability distribution.

The bottom right graph in the figure suggests how we can represent probabilities when the proportion p is *continuous* — that is, when the proportion can take on all values between 0 and 1. When there are 101 possible proportion values, the top of the line graph of the probability distribution looks like a horizontal line. In the case when p is continuous, we represent its probabilities by means of a similar type of horizontal line graph shown below. We call this graph a *flat* prior curve



Uniform continuous probability curve.



Area corresponds to a probability.

since the shape of the plot is flat. We also call it a *uniform* prior curve, since the flat shape tells us that the probabilities for the proportion p are uniformly spread out between 0 and 1.

Suppose that we have little knowledge about the proportion of students p that own their own credit card, and we use this flat curve to represent our opinion about the location of p. Remember that the flat curve is saying that we think that any value of the proportion between 0 and 1 is plausible. How can we use this curve to find probabilities about p? We compute the probability that p falls in a particular interval by finding the *area* under the curve over the interval.

Let's illustrate this for a simple example. What is the probability that p is between 0 and .3? In other words, what is the chance that the proportion of students who own their own credit cards is smaller than .3? We locate the interval between 0 and .3 and then shade the region under the flat curve as shown below. The area of this region corresponds to the probability.

Recall how we find an area for a rectangular region. The area of a rectangle is the length of the base multiplied by its height:

$$\text{AREA} = \text{BASE} \times \text{HEIGHT}$$

For the shaded area in the figure, the base goes from 0 to .3, so its length is .3. The height of the rectangle is 1, so the area is given by

$$\text{AREA} = .3 \times 1 = .3.$$

So the probability that p is between 0 and .3 is .3.

Let's try a second example. What is the probability that the proportion falls between .5 and .9? Return to our figure of the flat probability curve. We draw the rectangle over the values of p between .5 and .9 and under the curve. The base of this rectangle has length $.9 - .5 = .4$ and the height is 1, so

$$\text{Prob}(p \text{ is between } .5 \text{ and } .9) = .4 \times 1 = .4.$$

IN-CLASS ACTIVITIES

Activity 17-1: Probabilities for Uniform Curves

Suppose you are interested in learning about the proportion of cars driven in America that are white. Assume that you've been hibernating in some cave for the last 100 years, so you know very little about cars, and even less about the possible colors of the cars that are driven today. So it reasonable to say in this case that your prior beliefs about the proportion of cars that are white can be represented by a uniform curve.

(a) Draw a picture of this uniform curve.

(b) Find the probability that the proportion is larger than .7. (Answer this question and the ones to follow by first drawing the area that you want to find, and then finding the probability by computing the area.)

(c) Find the probability that the proportion of white cars is between .2 and .4.

(d) Find the probability that the proportion of white cars is either larger than .8 or smaller than .3.

(e) Find the probability that the proportion is exactly equal to .4.

(f) Find the probability that the proportion is between 0 and 1.

Important remark

If you completed the last activity, you found that the area under the entire uniform probability curve was equal to one. This is generally true for all probability curves — the total area under the curve is one. This statement is analogous to the fact that, for a discrete probability distribution, the sum of the probabilities is equal to one.

Beta curves

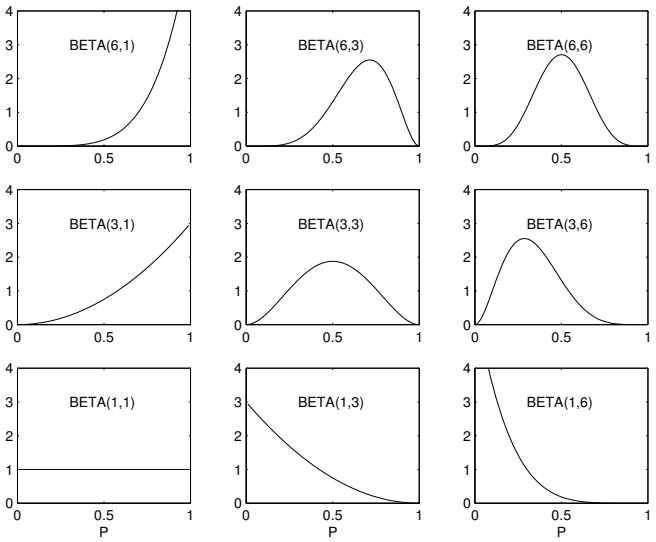
We can use a uniform or flat curve to represent our prior knowledge about a proportion p when we have little idea where this proportion is located. For example, suppose we are interested in the proportion p of adults in Tanzania (a country in east Africa) who drink coffee regularly. Many of us have little knowledge about the culture of this country and so we probably could not make an intelligent statement about Tanzania's coffee consumption. In this case, it may be reasonable to assign a uniform curve to the proportion p of coffee drinkers in this African country.

However, in many other situations, we do know something about the proportion of interest, and it wouldn't make any sense to use a uniform curve as our prior. In these cases, we will use a different type of curve to model our prior beliefs about a proportion. The particular curve that we will use is called a **beta curve**. Mathematically, this curve for the proportion p is expressed by the formula

$$BETA(p) = p^{a-1}(1-p)^{b-1}.$$

The numbers a and b in the formula determine the basic shape of the beta curve. We choose these two numbers to reflect our knowledge about the location of the proportion.

What does a beta curve look like? The figure on the next page shows nine different beta curves corresponding to different choices of a and b. Look first at the curve in the lower left corner of the figure. This is the uniform curve we saw earlier — this is a beta curve with values $a = 1$ and $b = 1$.



Many beta curves.

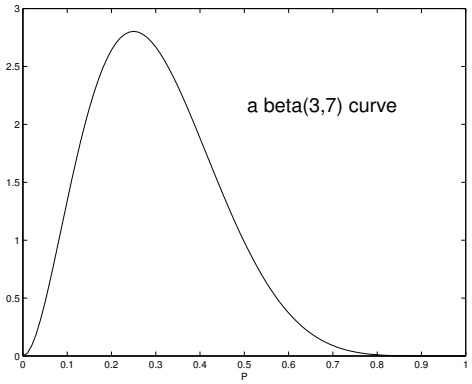
If we move from the uniform curve to the curves to the right, we are keeping the value of a at 1 and increasing the value of b from 1 to 6. Note that the curves on the right are higher for small values of p and smaller for large values of p. This means that we think that p is a small value near zero. Similarly, if we move from the uniform curve to the curves above it, then we are keeping the number b at 1 and increasing a from 1 to 6. These curves are high for large values for p — this indicates that we think that the proportion is a large value close to 1.

The message from this figure is that there are many shapes of beta curves. In practice, one chooses values of the beta numbers a and b so that the curve matches one's prior opinion about the location of the proportion p.

Finding probabilities using a beta curve

Let's return to our credit card example, where p represents the proportion of all undergraduate students at our college who own their own credit card. To be honest, I don't have a good idea how many students do own credit cards, so I searched the World Wide Web for some information about students' financial habits. I found results of a June 1997 survey prepared by Phoenix Home Life Mutual Insurance Company which asked students about their views about money management. In this survey, 1200 students, from seventh graders though college seniors, were interviewed by telephone and it was found that 24% of the students have a credit card for their personal use.

Does this mean that 24% of the students at our college own credit cards? Actually, probably not. The survey sampled high school and college students and it is not clear if the survey results



My beta curve for the proportion of students who own credit cards.

reflect the characteristics of only college students. Also, even if only college students were sampled, it is not obvious that the students at our college have financial habits that are similar to the general population of American college students.

After some thought, I decide that my prior beliefs about the proportion p can be represented using a beta curve with $a = 3$ and $b = 7$. This curve is shown in the figure below. Note that the highest point of the curve occurs around the value $p = .3$. This means that I think that the proportion of all students owning credit cards at our college is around .3. Note also that the curve is pretty spread out — the curve starts at $p = 0$, increases until about $p = .3$, and then decreases and hits zero at $p = .7$. This means that although I think the proportion is close to .3, I am unsure about the accuracy of this guess, and I think that p could be as small as 0 or as large as .7.

To better understand this beta curve, we can compute probabilities that the proportion falls in different intervals. As in the case of the uniform curve, the probability that p falls in the interval (.2, .4) will be the area under the beta curve between the values of .2 and .4. The chance that p is larger than .5 corresponds to the area under the curve between the values of .5 and 1. One important fact is that, since the proportion p is always between 0 and 1, the probability that the proportion lies in the interval (0, 1) is equal to 1. This means that the total area under any beta curve will be equal to one.

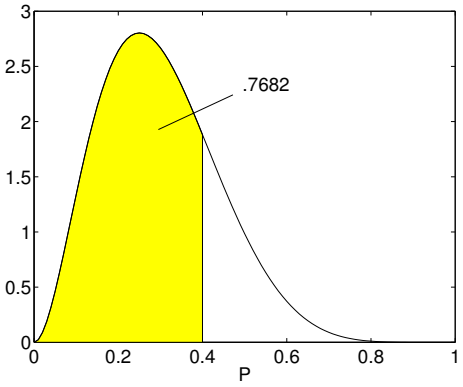
To find probabilities of intervals, or equivalently areas under the beta curve, we use a table of probabilities computed using the statistics package Minitab. A table of *cumulative probabilities* for the beta curve with $a = 3$ and $b = 7$ is presented below.

```
Beta with first shape parameter = 3.00000 and second = 7.00000

          x      P( X <= x)
     0.0000          0.0000
     0.1000          0.0530
```

```
0.2000        0.2618
0.3000        0.5372
0.4000        0.7682
0.5000        0.9102
0.6000        0.9750
0.7000        0.9957
0.8000        0.9997
0.9000        1.0000
1.0000        1.0000
```

The first column of this table (labeled 'x') lists different values of the proportion p and the second column gives the *cumulative probability*, which is the probability that the proportion is *less than or equal to* the corresponding value.

Let's illustrate using this table. Suppose we're interested in the probability that p is smaller than .4. Graphically, this is represented by the area under the beta curve between 0 and .4 which is shown in the figure below.

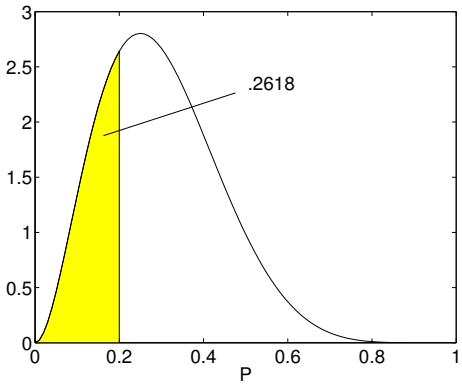


Probability that p is smaller than .4.

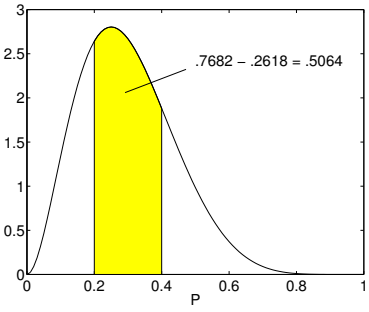
To find this area, we look for the value .4 in the first column of the table. The probability shown in the second column is 0.7682, which corresponds to the probability that the proportion of students owning credit cards is under 40%.

Suppose that we wish to find the probability that the proportion p is under .2. This is represented by the area under the curve between 0 and .2 which is shown in the figure below. To find this probability, we look for the value .2 in the first column of the table. The number in the second column is 0.2618 which corresponds to the probability that we're interested in.

Using the table, we can find the areas, or equivalently probabilities, for the proportion p between 0 and specified values. What if we're interested in the probability that p falls between two values, say .2 and .4? The corresponding area is shown below. This probability is not directly given in the table, but we observe that



Probability that p is smaller than .2.



Probability that p is between .2 and .4.

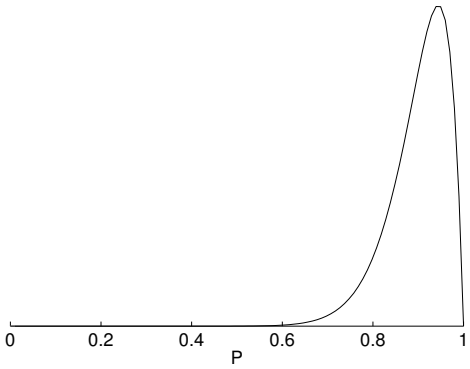
$$\text{Prob}(p \text{ is between .2 and .4}) = \text{Prob}(p \text{ smaller than .4}) - \text{Prob}(p \text{ smaller than .2}) .$$

In other words, we can find the probability that p is in an interval by subtracting two cumulative probabilities. We find the cumulative probabilities for the values .2 and .4 in the table. The interval probability is computed to be

$$\text{Prob}(p \text{ is between .2 and .4}) = 0.7682 - 0.2618 = .5064.$$

Activity 17-2: Employment Rate for Women

Suppose I am interested in the proportion p of adult women in the United States who are currently employed. I assume that this proportion is pretty large, but I don't know exactly the value of p. Also, it is impossible to compute the exact value of p since we can't check the employment status of the millions of adult women in the United States. To get some idea how large p is, I consult my handy encyclopedia. It tells me that the unemployment rate among all persons 16 and over in 1988 is 5.5%. So the proportion of current Americans employed is probably in the neighborhood of 95% or .95. But the encyclopedia rate is a 1988 statistic and it includes all people, men and women. I



My beta(18, 2) curve for the proportion of adult women in the U.S. that are employed.

think the employment proportion today is probably close to the 1988 number. Also, I think that the proportion of women employed is probably smaller than .95, since some women are staying at home raising families. I decide that my beliefs about the proportion p are described by a beta(18, 2) curve that is pictured below.

(a) Looking at this curve, what is the most likely proportion value?

(b) Based on this curve, what is a proportion value that is unlikely?

The table below gives a set of cumulative probabilities of the beta(18, 2) curve.

Beta with first shape parameter = 18.0000 and second = 2.00000

x	P(X <= x)	x	P(X <= x)
0.7000	0.0104	0.8600	0.2331
0.7100	0.0131	0.8700	0.2723
0.7200	0.0163	0.8800	0.3165
0.7300	0.0203	0.8900	0.3658
0.7400	0.0251	0.9000	0.4203
0.7500	0.0310	0.9100	0.4798
0.7600	0.0381	0.9200	0.5440
0.7700	0.0465	0.9300	0.6121
0.7800	0.0566	0.9400	0.6829
0.7900	0.0687	0.9500	0.7547
0.8000	0.0829	0.9600	0.8249
0.8100	0.0996	0.9700	0.8900
0.8200	0.1191	0.9800	0.9454
0.8300	0.1419	0.9900	0.9847

```
0.8400         0.1682    1.0000         1.0000
0.8500         0.1985
```

(c) Using this table, find the probability that p is smaller than .8

(d) Find the probability that p is smaller than .95.

(e) Find the probability that p is between 80 and 90 percent.

(f) Find the probability that p is between .77 and .85.

(g) Find the probability that p is larger than .9. [HINT: We haven't seen this type of problem yet. First, find the probability that p is smaller than .9. Then, since the total area under the entire beta curve is 1, the probability that p is larger than .9 will be one minus the probability that p is smaller than .9.]

(h) Find the probability that p is larger than .75.

Finding percentiles using a beta curve

Up to this point, we have focused on finding probabilities. Given several proportion values, say .1 and .4, we have found the probability that p falls between them. We are also interested in finding regions for p which contain a prescribed probability under the beta curve. These are called **probability intervals**, which were earlier discussed in Topic 16.

We first introduce the notion of a **percentile**. The qth percentile of the proportion p is the value of the proportion such that q percent of the probability falls to the left. Let's return to our ever popular beta(3, 7) curve which represented my opinion about the proportion p of students that use credit cards at my college.

The 20th percentile of p, which we will denote by p_{20}, is the value such that 20 percent of the probability lies to the left of p_{20}. The figure below shows the 20th percentile for a beta(3, 7) curve. We compute this number by using the Minitab command `invcdf`. On the first line, we put the

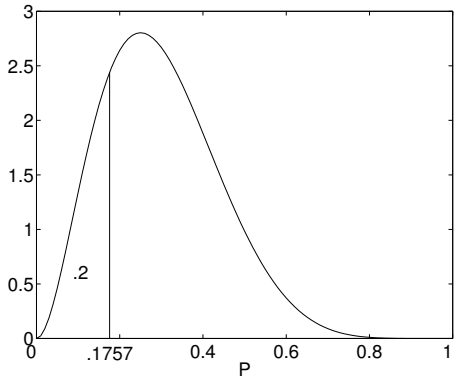


Illustration of the 20th percentile.

probability to the left (.2) and the second line we put the two numbers a and b of the beta curve (3, 7). We see below from the output that the 20th percentile of the proportion is .1757. This means that the probability that p is smaller than .1757 is equal to .2.

```
MTB > invcdf .2;
SUBC> beta 3 7.

Beta with first shape parameter = 3.00000 and second = 7.00000

 P( X <= x)          x
    0.2000        0.1757
```

Suppose that we wish to find a 90% probability interval for the proportion p. This is an interval of values for p which contains 90% of the probability content. A convenient way to find this interval is to find the "middle" 90% interval. This is the interval such that 5% of the probability lies to the left of the interval and 5% of the probability lies to the right. We find this interval by using Minitab to compute the 5th percentile p_5 and the 95% percentile p_{95}. From the output below, we see that $p_5 = .0977$ and $p_{95} = .5496$. As we see from the figure below, the probability that the proportion falls in the interval $(p_5, p_{95}) = (.0977, .5496)$ is .9, and so this is our 90% probability interval.

```
MTB > invcdf .05;
SUBC> beta 3 7.
Beta with first shape parameter = 3.00000 and second = 7.00000

 P( X <= x)          x
    0.0500        0.0977

MTB > invcdf .95;
SUBC> beta 3 7.
Beta with first shape parameter = 3.00000 and second = 7.00000
```

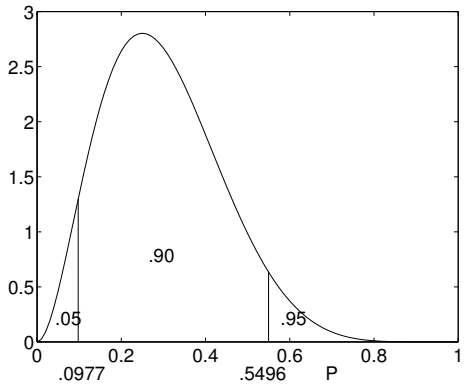
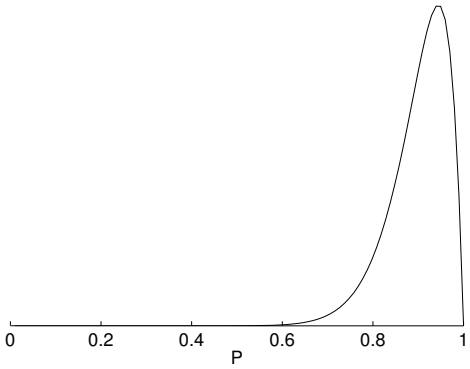


Illustration of a 90% probability interval.



My beta(18, 2) curve for the proportion of adult women in the U.S. that are employed.

```
P( X <= x)          x
   0.9500        0.5496
```

Activity 17-3: Employment Rate for Women (cont.)

We have redisplayed the beta curve which models my belief about the proportion of American adult women who are employed.

(a) Draw a vertical line inside the beta curve which approximately divides the area under the curve into half. From your line, what value of p divides the lower half and the upper half?

(b) Use Minitab to find the 50th percentile of the beta(18, 2). Contrast this value with the value of p that you guessed in part (a).

(c) Use Minitab to find the 10th percentile of the curve.

(d) Use Minitab to find the 90th percentile.

(e) From your work above, find a 80% probability interval for p. This will be an interval which contains the proportion p with probability .8.

(f) By finding appropriate percentiles, find a 90% probability interval for p.

Constructing a Beta Curve

We have seen how a beta curve represents one's opinions about the location of a proportion. Also, we've gotten some practice in summarizing this curve by computing probabilities (areas) under the curve and by computing percentiles. But how does one in practice construct his or her beta curve? We now discuss one simple method of finding the numbers for the beta curve which will approximately match one's prior beliefs about the proportion.

Let's consider a new example. Suppose I wish to learn about the proportion p of students at my university who have visited Europe. I want to construct a beta curve which models my knowledge about the value of this proportion. To find this curve, I first

- **guess at the value of the proportion**

After some thought, I guess that the proportion of students at my school that have been to Europe is .2.

Next, I have to decide how sure I am of this particular guess. Maybe I am very sure that the proportion is close to .2, since I have asked a large number of students if they have been to Europe.

On the other hand, perhaps I view .2 as a wild guess at the value of p since I don't know anything about the students' travel habits. To indicate the precision of my guess, I

- **state how many observations this guess is worth**

Here we think of our prior information in terms of data that we might have collected. Suppose, for example, that we think that 20% of the students have been to Europe. One way of expressing this prior belief is to imagine that we surveyed five students, where one (20%) had been to Europe and the remaining four (80%) had not (see the following table).

Student	Response
1	has been to Europe
2	has not been to Europe
3	has not been to Europe
4	has not been to Europe
5	has not been to Europe

In this case, our prior belief is equivalent to an imaginary sample of size 5. If we are *more sure* about the guess of 20%, we might think that this belief is equivalent to an imaginary sample of size 20, where 4 (20%) have been to Europe and the rest (80%) have not.

Student	Response	Student	Response	Student	Response	Student	Response
1	Europe	6	no Europe	11	Europe	16	Europe
2	no Europe	7	no Europe	12	no Europe	17	no Europe
3	no Europe	8	no Europe	13	no Europe	18	no Europe
4	no Europe	9	no Europe	14	no Europe	19	no Europe
5	no Europe	10	no Europe	15	no Europe	20	no Europe

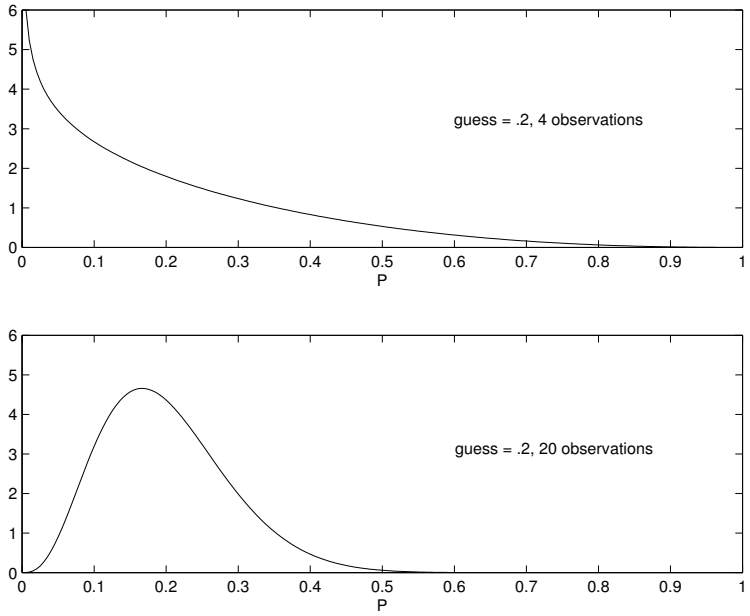
Let's let g represent our guess at the proportion p, and let n represent the number of observations that this guess is worth. Then the numbers of the beta curve which match this information are

$$a = n \times g \text{ and } b = n \times (1 - g).$$

For example, if our guess at the proportion of students who have been to Europe is $g = .2$ and this guess is worth $n = 5$ observations, then the beta numbers are

$$a = 5 \times .2 = 1, b = 5 \times (1 - .2) = 4.$$

The top graph of the figure on the next page shows the beta(1, 4) curve. Note that this curve has a pretty flat shape. This indicates that, although we think that the proportion is .2, values of the proportion between 0 and .4 are all pretty likely.



Two beta curves for the proportion of students who have visited Europe.

In contrast, suppose that our guess is $g = .2$ and this is worth (in our opinion) $n = 20$ observations. Then we compute

$$a = 20 \times .2 = 4, b = 20 \times (1 - .2) = 16.$$

The beta(4, 16) curve that matches this prior information is shown in the bottom graph of the figure. The shape of this curve is more peaked about .2 which indicates that we're pretty confident that the proportion of students who have visited Europe is between .1 and .3.

Important remark

Determining your beta density for a specific problem is a hard thing to do, since you haven't had much practice doing it. A good strategy is to start with reasonable choices of g and n, find the beta numbers, and then graph the corresponding beta curve. By looking at this curve, see if it appears to be consistent with your beliefs about the particular proportion. If it doesn't appear to match your opinion, then make changes to g and n. Redraw your new beta curve and see if it is a better match to your opinion. By this trial and error method, you'll eventually obtain a beta curve that you're satisfied with.

Activity 17-4: Does Frank have ESP?

Suppose one of your classmates, Frank, claims that he possesses ESP (extra-sensory perception). To test Frank's claim, the class plans to give Frank a variation of a classical ESP experiment. They make cards with four different faces ($\star$, $\triangle$, ∇, $\bigcirc$) and they create a deck containing a large number of cards of each face. A member of the class will hold up one of the cards and Frank will attempt to identify the face. This experiment will consist of 10 cards being shown and the class will record Frank's answer to each card.

Let p denote the proportion of times Frank will give the correct response when shown a large number of cards from the deck.

(a) Suppose that you think that Frank *does not* possess ESP and he is just guessing at the card's face. In this case, make a guess g at the proportion p.

(b) Suppose that you believe that the guess in (a) is worth 10 observations. Find the values of the beta numbers a and b which match this prior information.

(c) If you believed strongly that Frank did not possess ESP, would you assign your guess in (a) a large number of observations or a small number of observations? Explain.

(d) Suppose that Calvin believes that Frank does possess some ESP. Would Calvin's guess at the proportion p be larger or smaller than the value you chose in (a)? Explain.

Inference using a beta curve

We have discussed how one can use a beta curve to reflect his or her beliefs about a proportion p. Returning to my Europe example, suppose that my opinion about the proportion of students that have visited Europe is modeled by a beta(1, 4) curve. As in Topic 16, I will collect some data to learn more about p, and then use Bayes' rule to find my new, or updated, density for this proportion. It turns out that it is easy to perform Bayes' rule when one uses a beta curve as a prior density.

Suppose I take a random sample of 10 students from my school and ask each student if they have ever been to Europe. My responses are

No, Yes, No, No, No, No, No, Yes, No, Yes

We see that there were 3 students out of 10 that had been to Europe.

We follow the basic Bayes' rule recipe that we used in Topics 15 and 16.

- (**Prior**) We start with our prior density for the proportion p which is a beta(1, 4) curve. We write this as
$$PRIOR = Beta(1, 4) = p^{1-1}(1-p)^{4-1}$$

- (**Likelihood**) We find the likelihood of the proportion p. This is the probability of getting the above sample result (3 out of 10 going to Europe) if the proportion of all students going to Europe is indeed p. Assuming independent responses, the likelihood is given by
$$LIKE = p^3(1-p)^7.$$

- (**Posterior**) Using Bayes' rule, the posterior density for p is found by multiplying the likelihood by the prior:
$$POST = PRIOR \times LIKE = p^{1-1}(1-p)^{4-1} \times p^3(1-p)^7 = p^{4-1}(1-p)^{11-1}$$
Looking at the final form, we see that the posterior density for p is a beta curve with new numbers 4 and 11, that is, a beta(4, 11).

We have illustrated a general rule for updating probabilities for a proportion p when the prior is a beta curve:

> If our prior density for a proportion p is a beta(a, b) curve, and we take a random sample consisting of s successes and f failures, then the posterior density for the proportion will be a beta$(a + s, b + f)$ curve.

Activity 17-5: Does Frank have ESP? (cont.)

Suppose that you are interested in learning about Frank's ESP ability. We measure his ability by the proportion p, which represents the proportion of cards that he recognizes correctly in the ESP experiment. Suppose that your prior beliefs about this proportion are modeled by a beta(2.5, 7.5) curve. You observe the results of the experiment — 20 cards are shown to Frank and he recognizes 7 correctly.

(a) If we consider a "success" recognizing the card correctly and a "failure" making a mistake, how many successes and failures did you observe in this experiment? (That is, what are the values of s and f?)

(b) Find the numbers a and b for the beta curve which represents your beliefs about the proportion p *after* observing the above data.

(c) Suppose that you really believe that Frank has ESP if the proportion of cards p that he correctly recognizes, in the long run, is over .5. Using the beta density from (b), use Minitab to find the probability that p is over .5. From this calculation, would you say that it is likely that p exceeds .5?

(d) Find the probability that p is between .2 and .3.

(e) Using the beta density from (b), find a 90% probability interval for p.

Inference Using a Uniform Curve

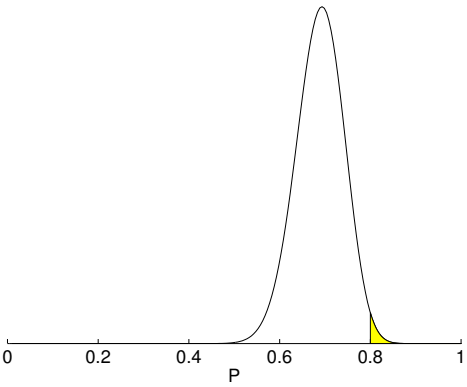
Suppose that I am interested in learning about the proportion p of patients of my dentist who pay their bills on time (say, within three months). I am not very knowledgable about the patients of my dentist and, in particular, know very little about the incomes of the patients. Thus it would be difficult to make an educated guess at the proportion p, and even if I could make a guess, I would not place much confidence in its accuracy. Thus, it would be helpful if I could construct a prior curve on p which reflects little knowledge about the location of this proportion.

In the cases where one has little prior knowledge about the proportion p, the **uniform curve** discussed earlier can be used. This curve says that all values of the proportion p between 0 and 1 are equally likely to occur. We can write the uniform curve as

$$PRIOR = 1$$

or equivalently as

$$PRIOR = p^{1-1}(1-p)^{1-1},$$



Posterior curve for the proportion of patients who pay their bill on time.

which we recognize as a beta curve with numbers $a = 1$ and $b = 1$.

Since the uniform curve is a special case of a beta curve, we can use our general recipe for computing the posterior density after data from a random sample are observed. If we observe s successes and f failures, then the posterior curve using a uniform prior will be beta($s + 1, f + 1$). In our example, suppose that 52 out of 75 randomly selected patients pay their bills on time. If we define a "success" as paying on time, then we observe

$$s = 52, \ f = 75 - 52 = 23,$$

and so the posterior curve using a uniform prior is

$$Beta(52 + 1, 23 + 1) = Beta(53, 24).$$

To illustrate using this posterior curve, suppose that the dentist is interested in the probability that over 80% of all his patients pay their bills on time. In other words, he is interested in the probability that $p > .8$ which is displayed as the shaded area below. Using Minitab, we find that the area to the left is .9886 (see output below), and so the probability of interest is

$$\text{Area to right of .8 = 1} - \text{ Area to left of .8 = 1} - .9886 = .0114.$$

This is a small probability, so it is unlikely that that over 80% of the patients pay on time.

```
Beta with first shape parameter = 53.0000 and second = 24.0000

          x      P( X <= x)
     0.8000         0.9886
```

Activity 17-6: Does Frank have ESP? (cont.)

Suppose that Dave, a friend of yours, knows little about extra sensory perception, and he is unwilling to even guess at the proportion of cards p that Frank will recognize correctly in the experiment. To reflect this lack of knowledge, Dave assigns a uniform curve as his prior on the proportion p. Recall the results of the experiment — Frank recognized 7 out of 20 cards correctly.

(a) Find the numbers for the beta posterior curve.

(b) Find the probability that Frank recognizes at least 50% of the cards in the long run. (That is, find the probability that p exceeds .5.)

(c) Use the posterior curve to find a 90% probability interval for p.

(d) Compare your answers to (b) and (c) with the answers in Activity 17-5 using an informative beta prior. Which probability interval has the shorter length? Can you explain why?

HOMEWORK ACTIVITIES

Activity 17-7: Probabilities of a Spinner

Many board games use a spinner like the one pictured below. One spins an arrow and it lands in the circle at a random point between 0 and 1. Any point between 0 and 1 has the same chance of being spun and so probabilities of different spinner results can be described using the uniform continuous probability curve shown below. By computing appropriate areas under this curve, find the following probabilities.

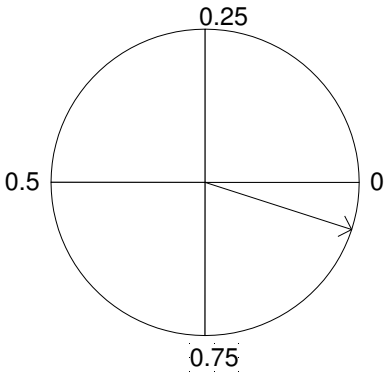
(a) Find the probability the spinner lands between 0 and .3.

(b) Find the probability that spinner lands between .15 and .75.

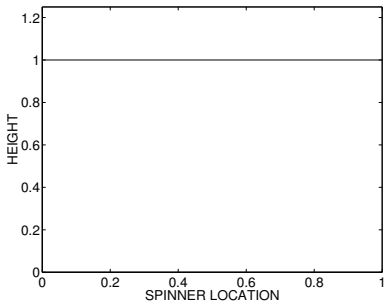
(c) Find the probability that a number greater than .63 is spun.

(d) Find the probability that the number .625 is spun.

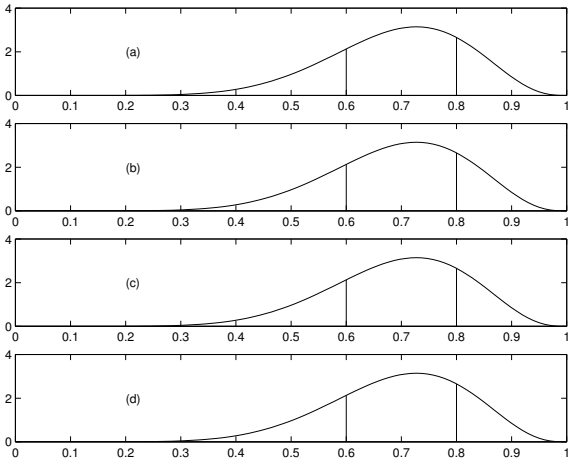
(e) Find the probability a number greater than 2 is spun.



A spinner.



Uniform curve for spinner location.



Activity 17-8: Finding Areas Under Beta Curves

What proportion of households in our town have an internet connection? My opinion about this proportion p can be represented by a beta curve with numbers $a = 9$ and $b = 4$. Some cumulative areas under this curve are shown in the Minitab output below. Find each of the probabilities below by (1) shading the correct area in the figure on the next page, and (2) computing the probability using the Minitab output.

```
Beta with first shape parameter = 9.00000 and second = 4.00000

          x      P( X <= x)
     0.6000         0.2253
     0.8000         0.7946
```

(a) The probability that the proportion is less than .6.

(b) The probability that the proportion is larger than .8.

(c) The probability that the proportion is between 60 and 80 percent.

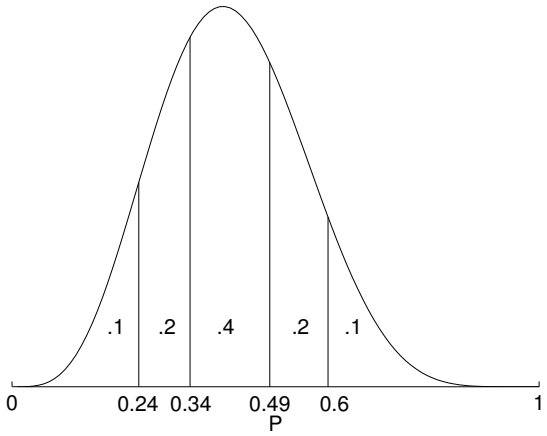
(d) The probability that the proportion is larger than .6.

Activity 17-9: Finding Percentiles Under a Beta Curve

The figure above divides a beta(5, 7) curve into areas and gives the values of p which bounds the given areas. Use this figure to answer the questions below.

(a) Find the 10th percentile.

(b) Find the 70th percentile.



Percentiles of a beta(5, 7) curve.

(c) Find the 90th percentile.

(d) Find a 80 percent probability interval for p. (This is an interval which bounds 80 percent of the probability.)

Activity 17-10: Matching Beta Curves with Prior Information

The figure below shows eight beta curves for a proportion p. Put the letter of the beta curve which matches the value of the prior guess g and the number of prior observations n given below.

(a) I guess that the proportion is .8 and my guess is worth 50 observations. ____

(b) I guess that the proportion is .9 and my guess is worth 10 observations. ____

(c) I guess that the proportion is .5 and my guess is worth 80 observations. ____

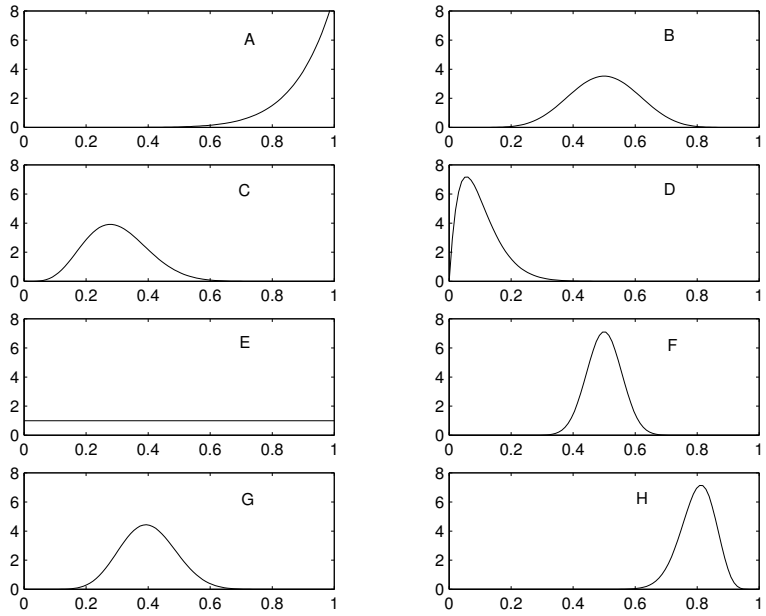
(d) I guess that the proportion is .5 and my guess is worth 2 observations. ____

(e) I guess that the proportion is .5 and my guess is worth 20 observations. ____

(f) I guess that the proportion is .3 and my guess is worth 20 observations. ____

(g) I guess that the proportion is .4 and my guess is worth 30 observations. ____

(h) I guess that the proportion is .1 and my guess is worth 20 observations. ____



Eight beta curves.

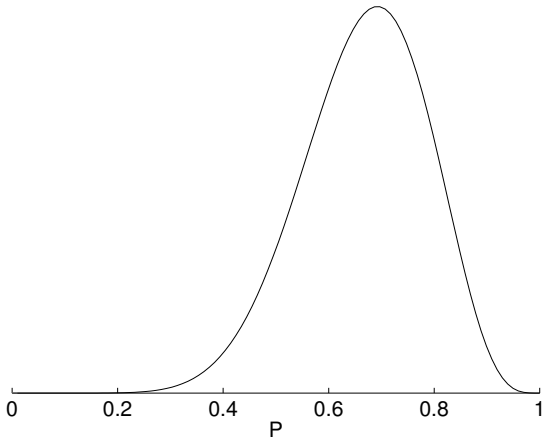
Activity 17-11: Marriage Ages (cont.)

Suppose you are interested in the proportion p of all marriages in your region where the bride is younger than the groom. After some thought, you decide that your beliefs about the location of this proportion are well-described by a beta(10, 5) curve that is shown below. Minitab was used to find cumulative areas under this curve and the output is placed below.

```
Beta with first shape parameter = 10.0000 and second = 5.00000

      x     P( X <= x)
 0.0000         0.0000
 0.1000         0.0000
 0.2000         0.0000
 0.3000         0.0017
 0.4000         0.0175
 0.5000         0.0898
 0.6000         0.2793
 0.7000         0.5842
 0.8000         0.8702
 0.9000         0.9908
 1.0000         1.0000
```

(a) Looking at the curve, what is the most likely value of the proportion p?

(b) Using the areas in the table, find the probability that p is less than .5.



Beta curve for proportion of marriages where bride is younger.

(c) Find the probability that the proportion is greater than .8.

(d) Find the probability that p is between .6 and .8.

(e) Find the probability that the proportion is less than .75. (Hint: The table doesn't give the exact probability for .75, but by looking at values close to .75, you can approximate the probability.)

Activity 17-12: Marriage Ages (cont.)

As in Activity 17-11, suppose that you are interested in the proportion of marriages in your area for which the bride is younger, and your information about this proportion is represented by a beta(10, 5) curve.

Suppose you inspect a random sample of 20 marriage records with the following results: the bride was younger in 12 marriages, the groom was younger for 4 marriages, and the bride and groom were the same age for 4 marriages.

(a) If we define a "success" a couple where the bride is younger than the groom, and a "failure" where the bride is *not* younger than the groom, find the number of successes and failures in the sample.

(b) Using the numbers of successes and failures from (a) and the prior beta curve, find the numbers for the posterior beta curve.

(c) Using the posterior beta curve, find the probability that the proportion p is smaller than .6. (You need Minitab to find the area under a particular beta curve.)

(d) By finding appropriate percentiles under the beta curve using Minitab, find a 90% probability interval for p.

(e) Suppose someone claims that the proportion of couples for which the bride is younger is at most 30%. Compute the probability that the proportion is at most 30%. From this computed probability, would you say that this claim is reasonable? Why?

Activity 17-13: Will the School Levy be Approved?

Suppose that your community is placing a school levy on the ballot and people are interested in the proportion p of the voters who will support the levy by voting "yes". For each of the following hypothetical people, give values of the guess g and the number of prior observations n which correspond to the statements.

(a) Joe believes that the election will be a dead heat — the number of voters that will support the levy will be equal to the number of voters against the levy. He thinks that his prior brief is equivalent to the information obtained by a sample of 20 voters.

(b) Helen agrees with Joe that the election will be a dead heat. But she is more sure than Joe about this result. (Give possible values for g and n.)

(c) Mark thinks that the school levy will be approved. The strength of his belief is equivalent to the strength of Joe's belief. (Give possible values for g and n.)

(d) Joy thinks that the school levy will be defeated and she is extremely confident of this result. (Give possible values for g and n.)

Activity 17-14: Students' Glasses in Class

Suppose you are interested in the proportion p of *all* students at your school that wear glasses or contacts.

(a) Before collecting any data, make an intelligent guess at the value of p.

(b) Make a guess on how many hypothetical observations your guess you made in part (a) is worth.

(c) From the information provided in parts (a) and (b), find the values of the beta numbers a and b which match this information.

(d) In the preliminary section, the class was polled regarding the use of glasses and/or contacts. Record in the table below the number of students who wear either glasses or contacts and the number of students who wear neither.

	Glasses or Contacts	Neither
Count		

(e) Assuming that the above data represents a random sample from the entire student body, find the values of the posterior beta curve.

(f) Using Minitab, find a 95% probability interval for the proportion of interest p.

(g) Suppose an eye doctor in town claims that over half of the students at your school wear glasses or contacts. Find the probability of this claim. Based on your calculation, would you agree with the doctor? Why?

Activity 17-15: What Proportion of Ball Games are Decided by One Run?

A British cricket fan is interested in the proportion of *all* current Major League baseball games that are decided by one run. Call this unknown proportion p. Since this fan is English, he knows little about baseball and so he assigns p a flat prior curve on the values 0 to 1. This is equivalent to a beta curve with numbers $a = 1$ and $b = 1$.

To get more information about p, the fan collects some data. The following table tallies the margin of victory in the 355 Major League baseball games played during June of 1992.

Margin	1	2	3	4	5	6	7
Tally	106	64	54	41	26	26	14
Margin	8	9	10	11	12	13	14
Tally	6	6	5	3	0	2	2

(a) How many games in the sample were decided by one run? How many games were *not* decided by one run?

(b) Using the above prior beta curve, find the numbers for the beta posterior curve.

(c) What proportion of games in the sample were decided by one run? (This sample proportion is approximately the value which is most likely in the beta posterior curve.)

(d) Use Minitab to find a 90% probability interval for the proportion p.

(e) Explain, using layman's terms, what it means for the interval in part (d) to have 90% probability.

(f) Is this particular sample a simple random sample (SRS) from the population of all Major League games? If not, is the sample likely to be biased in a certain direction with respect to margin of victory? Explain.

Activity 17-16: Does Your Magazine Have a High Proportion of Ads? (cont.)

(a) Recall that you guessed at the proportion p of all pages from your favorite magazine that contained at least one advertisement. Write down the name of your magazine and your guess at the proportion of ads.

(b) If your guess is worth about ten observations, find the numbers of the beta curve which match this prior information in part (a).

(c) Take a random sample of 20 pages from the most recent copy of your magazine by using a table of random digits. For each page selected, record the page number and whether or not the page contains at least one advertisement.

(d) What proportion of ads of this sample of pages contains at least one advertisement?

(e) Using the beta prior from part (b) and the observed data from (c), find the numbers of the beta posterior curve.

(f) Find a 95% probability interval for the proportion of all ads in your magazine that contain at least one ad.

(g) If someone thinks that a magazine has a high proportion of ads if at least 70% of its pages contain ads, would you conclude that your magazine has a high proportion of ads? Explain how you reached your conclusion.

Activity 17-17: Why Do People Vacation in Rotterdam? (cont.)

Activity 16-9 discussed a study carried out in Rotterdam to find out the reasons why tourists come to the city. Of 450 visitors to the Museum of Fine Arts, 243 indicated that they came to Rotterdam for the purpose of visiting the art museum. The people conducting the study were interested in learning about the proportion p of all Rotterdam tourists who came to the city to visit the art museum. Assume that this sample is representative of all tourists that visit this Dutch city.

(a) Assuming a uniform prior (with beta numbers $a = 1$ and $b = 1$), find the numbers of the posterior beta curve for the proportion p.

(b) Using Minitab, graph this posterior beta curve. From this graph, give a range of "likely" values for the proportion p.

(c) What is the probability that less than 50% of the visitors come to Rotterdam for the art museum?

(d) Find the probability that p is in the interval [.52, .56].

WRAP-UP

In this topic, we introduced the concept of a continuous model for a proportion. The proportion p was allowed to have a value anywhere between 0 and 1, and we used a continuous curve to model beliefs about this proportion. A beta curve described by two numbers a and b was used as our continuous model for p. When we have prior beliefs about the location of p, we find the numbers of the beta curve by making a guess g at the value of the proportion and then stating how many observations n this guess is worth.

If we use a beta prior curve, the posterior curve for p (after data are observed) also has a beta form with updated numbers. We make inferences about the proportion by summarizing the beta probability curve. We discussed two types of inferences: probabilities, found by computing areas under the beta curve, and percentiles, which are values of p such that a given probability is to the left.

In the next topic, we discuss the new situation where we are interested in learning about the mean of a population. The meaning of a model will change, but we again use Bayes' rule to learn about the model after observing data.

Topic 18: Learning About a Mean Using Discrete Models

Introduction

In the previous topic, we concentrated on learning about a proportion. In that situation, the population of interest was divided into two types, say smokers and non-smokers, and we learned about the proportion of smokers by taking a random sample from the population. Here we discuss the different situation where we wish to learn about the mean or average measurement in a population.

We first suppose that we take measurements, say pulse rates, from a sample of people. If this sample is selected randomly from some target group, say children of ages 7-12, then we imagine a distribution of pulse rates for all of the children in our target population. We suppose that this distribution is bell-shaped about some average value M. In this topic, we introduce the normal curve which is the most popular bell-shaped model for continuous measurements.

We focus on learning about the average measurement M. As in previous topics, we use Bayes' rule as our method for learning. One constructs a list of plausible values for the mean M and assigns probabilities to these values. Then posterior probabilities are computed for these values of M and inferences are based on this posterior distribution.

PRELIMINARIES

1. What do you think is the average daily high temperature in Bismarck, North Dakota in January?

2. Suppose you were able to measure the daily high temperature in Bismarck for 20 days in January. If you drew a stemplot of these 20 temperatures, what would be the shape of these data?

3. In an earlier topic, the 68-95-99.7 rule was used to approximate the fraction of data that would fall within one, two, and three standard deviations from the mean. This rule works well when

the data has a _____ shape.

4. What do you think is your usual pulse rate when you are sitting at your desk?

5. Measure your pulse rate by counting the number of heart beats in a 30 second interval. Record your measurement and the measurements for your classmates in the table below.

Student	Pulse	Student	Pulse	Student	Pulse
1		9		17	
2		10		18	
3		11		19	
4		12		20	
5		13		21	
6		14		22	
7		15		23	
8		16		24	

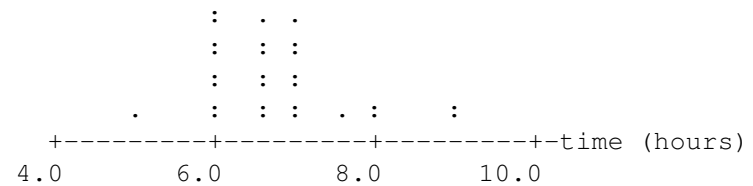
A Normal Curve Model for Measurement Data

Let us consider a simple example. Suppose we are interested in learning about the sleeping habits of college students in America. Perhaps we think that the students aren't performing well in their classes since they aren't getting enough sleep. We plan to take a sample of 28 students from the population of college students. For each student in the sample, we'll collect the number of hours that the student sleeps in a typical "school" night, and we wish to use this information to make some statement about the average amount of sleep for all college students.

Suppose that we observe the following sleep times (in hours):

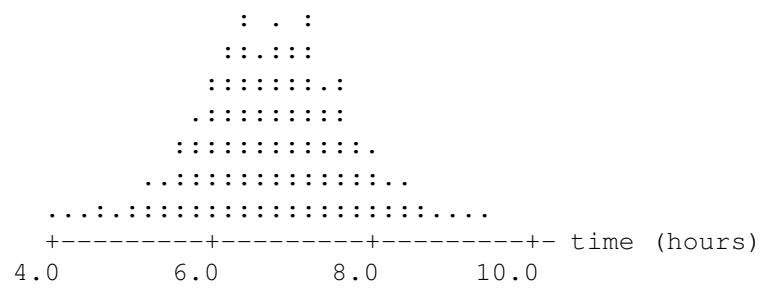
8.0 6.0 6.5 7.0 7.0 6.5 6.5 6.0 7.0 7.0
6.0 6.0 7.5 6.0 7.0 7.0 6.5 6.0 6.5 6.5
9.0 9.0 7.0 5.0 8.0 6.5 6.0 6.0

We graph the data using a dotplot.



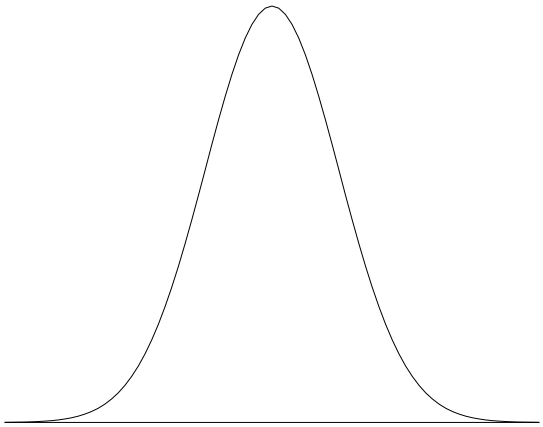
Note that the sleep times appear to clump in the middle — most of the times are in the 6-7 hour range with a few small and large times.

Next, suppose that we were able to sample 1000 college students (instead of only 28) and record their sleeping times. It is plausible that the dotplot of the times would look like the following:



The shape of these sleeping times is starting to take on a more distinctive shape. It appears mound-shaped and symmetric about the central value.

What would happen to this distribution of sleep times if we were able to sample one million students, instead of 28 or 1000? It likely would have the shape of the bell-shaped curve below, which we call a **normal** curve.



A normal curve.

Generally, measurement data, such as sleeping times for a group of individuals, will have a normal shape. Other examples of data whose shape will be approximately normal include

- heights of college men
- pulse rates of adults between the ages 30 and 40
- the exact fill amount of a number of 16 oz bottles of Coke
- student measurements of the diameter of a large ball

- student guesses at the age of a particular instructor
- scores of all students on the SAT math test

Mean and Standard Deviation of a Normal Curve

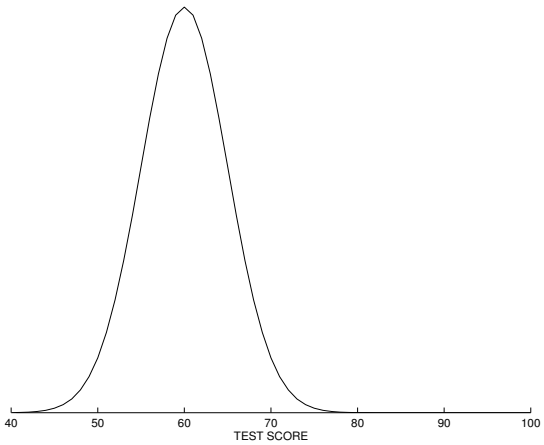
Although it may be reasonable to assume that our data has a normal shape, that won't be enough to explicitly define the data distribution. There are, in fact, many different normal curves that might be plausible for a given dataset. Each normal curve is described by two numbers, the **mean** M and the **standard deviation** h.

The mean M tells us about the **location** of the normal curve. To illustrate two normal curves with different means, consider the height distributions of U.S. women and men between ages 18 and 24. The women heights (measured in inches) have an approximate normal shape with mean $M = 64.5$ and standard deviation $h = 2.6$; the men heights (also in inches) are normal with mean $M = 70.0$ and standard deviation $h = 2.8$. The figure below plots the two normal curves using the same scale. Note that the two curves have basically the same shape — the difference is that the men heights curve is centered at 70 and the women heights curve is centered at 64.5.

Normal curves of men and women heights.

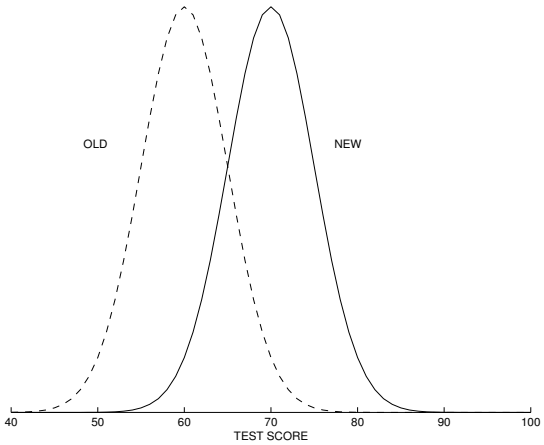
As a second example, suppose an instructor gives a test to a large group of students. The mean grade on the test is 60 percent with a standard deviation of 5. The distribution of test scores can be represented by the normal curve below.

Now suppose that the instructor believes that the grades on the test do not reflect the students' knowledge of the material and so she decides to curve the test scores by adding 10 percentage points to each grade. How does this grade adjustment change the normal curve of the test scores? The new test scores will have a mean of 70 percent with the same standard deviation of 5. The graph below



Normal curves of test scores.

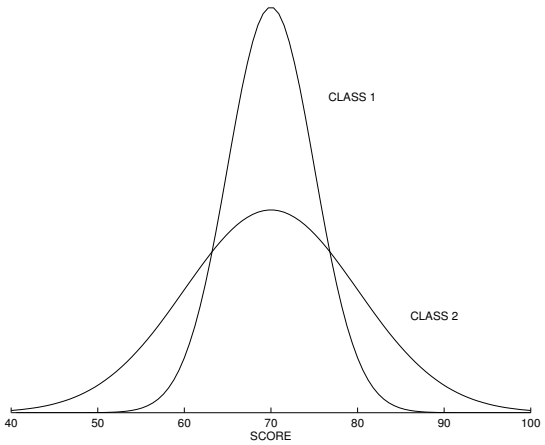
shows both normal curves where the old test scores are plotted using a dashed line and the new test scores using a solid line. We see that the shape of the normal curve doesn't change. But, by adjusting the scores by adding 10 points, we have moved or shifted the normal curve to the right by 10 points.



Normal curves of old and new test scores.

The second number associated with the normal curve, h, controls the **shape** of the curve. As the name standard deviation suggests, the number h tells us something about the spread or variability of the curve. Consider scores on a particular test for two history classes. We're told that both classes scored an average of 70 percent and the shapes of the distributions of the two sets of scores were normal shaped. Does this mean that the classes performed basically the same on the given test? To answer this question, let's plot the two normal curves below:

The test scores from the first class are normal shaped with mean $M = 70$ and standard deviation



Normal curves of test scores from two classes.

$h = 5$; the scores from the second class are normal with mean $M = 70$ and standard deviation $h = 10$. Although the classes have the same mean, the two curves are remarkably different. Looking at the scores from the first class, we see that practically all of the scores fell between 60 and 80 and it is rare for a student to get over 90 percent. In contrast, the scores from the second class have a broad range from 40 to 100. It is more common in this class to do really well or really poorly on the test.

The 68-95-99.7 rule (again)

Remember the 68-95-99.7 rule from Topic 4? That rule said that, if your data is approximately mound-shaped, then you can make an intelligent guess at the proportion of data that fall within one, two, and three standard deviations from the mean. This rule actually comes from properties of the normal curve which is the best-known mound-shaped distribution.

We can state this rule in terms of our normal curve with mean M and standard deviation h. For a normal curve,

- 68% of the data will fall between $M - h$ and $M + h$
- 95% of the data will fall between $M - 2h$ and $M + 2h$
- 99.7% of the data will fall between $M - 3h$ and $M + 3h$

What does this rule mean in practice? It means that, if our data is well-represented by a normal curve, then most of the data (that is, 95%) will fall within two standard deviations of the mean and it is pretty rare (5% chance) to see observations that fall outside 2 standard deviations of the mean.

Let's illustrate using this rule. Suppose someone tells me that daily high temperatures in Iowa in August are approximately normal shaped with a mean of $M = 90$ degrees and a standard deviation of $h = 5$ degrees. Using the 95 part of the 68-95-99.7 rule, I expect most of the high temps to fall between

$$90 - 2 \times 5 \text{ and } 90 + 2 \times 5 \text{ degrees}$$

which equals

$$80 \text{ and } 100 \text{ degrees}$$

So, if I visit Iowa in August, I would be pretty sure that I'm going to see a high temperature between 80 and 100 degrees. It would be unusual to see a high temperature lower than 80 degrees or higher than 100 degrees since this only happens 5% of the time.

Formula for a Normal Curve

Like the beta curve discussed in Topic 18, the normal curve has a relatively simple formula. Let x represent the data value. Then the normal curve is given by

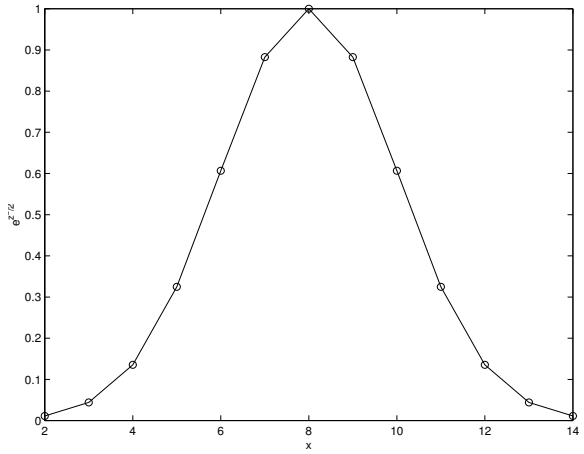
$$e^{-z^2/2},$$

where z is the standardized score

$$z = \frac{x - M}{h}.$$

We will illustrate computing this formula for one normal curve model for the sleeping data. Suppose that the mean of the curve is $M = 8$ and the standard deviation is $h = 2$. We want to compute the normal curve for data values x between 4 and 12. We illustrate the computations in the table below. The first column lists the data values. In the second column we change the data values to standardized scores z. In the last curve we compute $e^{-z^2/2}$. We plot the values of x and $e^{-z^2/2}$ on the figure below and connect them which displays the normal curve.

x	z	$e^{-z^2/2}$
2	-3.0	0.0111
3	-2.5	0.0439
4	-2.0	0.1353
5	-1.5	0.3247
6	-1.0	0.6065
7	-0.5	0.8825
8	0	1.0000
9	0.5	0.8825
10	1.0	0.6065
11	1.5	0.3247
12	2.0	0.1353
13	2.5	0.0439
14	3.0	0.0111

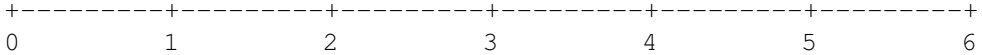


Plotting values of a normal curve.

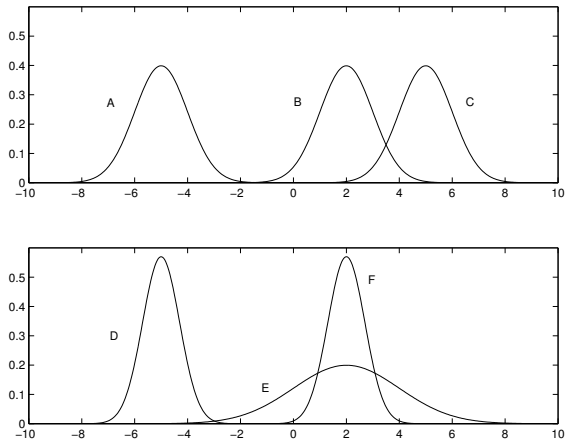
IN-CLASS ACTIVITIES

Activity 18-1: Distinguishing Normal Curves

(a) On the number line below, draw a normal curve with mean equal to 4.



(b) There are six normal curves in the figure below. Match the letter of the curve with the descriptions below.



(1) normal with $M = 5, h = 1$ ___

(2) normal with $M = 2, h = .7$ ____

(3) normal with $M = -5, h = .7$ ____

(4) normal with $M = 2, h = 1$ ___

(5) normal with $M = 2, h = 2$ ___

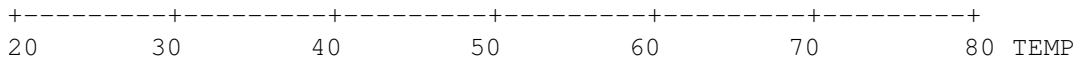
(6) normal with $M = -5, h = 1$ ___

(c) Consider a normal curve with mean $M = 3$ and standard deviation $h = 2$. Compute values of the normal curve by first computing the standardized score $z = \frac{x-M}{h}$ and then computing the curve $e^{-z^2/2}$. (The first row has been done for you.)

x	z	$e^{-z^2/2}$
0	-1.5	0.3247
1		
2		
3		

(d) Suppose that daily high temperatures (in degrees Fahrenheit) in Atlanta in January are well described by a normal curve with mean 50 and standard deviation 10.

(1) Draw the normal curve on the number line below.



(2) Find an interval where you expect 95% of the high temperatures to fall.

(3) Suppose that you travel to Atlanta in January and the high temperature is 75 degrees. Should you be surprised? Why or why not?

The Unknown Normal Curve

Let us return to our study of the sleeping habits of college students. We are interested in learning about the sleeping times of *all* college students, not just the times that we will collect in our sample. Recall that in statistical inference, we are interested about a large body of individuals that we refer to as the **population**, and we gain information about the population by collecting data from a subset that we call the **sample**.

Based on our earlier discussion, it seems reasonable to assume that the distribution of sleeping times can be represented by a normal curve that is described by two numbers, the mean M and the standard deviation h. For sake of simplicity, we will assume in this topic that we actually know the value of the standard deviation h. In this sleeping time example we'll assume that $h = .5$. (In typical practice, we won't know the value of h, and we'll discuss in the next topic what to do in this situation.)

So the sleeping times of all college students are normal (that is, have a normal shape) with mean M and standard deviation $h = .5$. So what is the value of M?

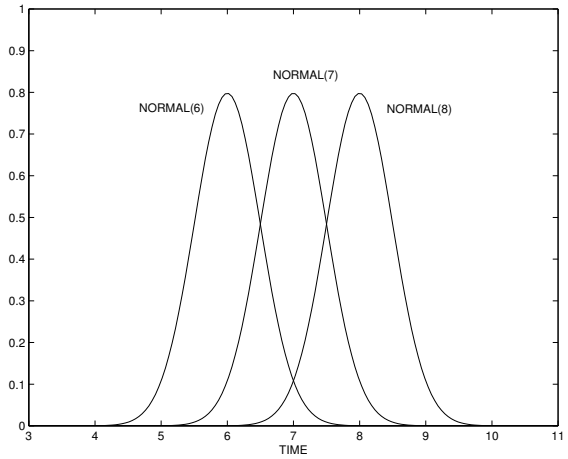
We don't know M.

Remember M is the mean of the normal curve which represents the distribution of sleeping times for all college students. If we were able to find the typical sleeping time for *every* college student in America, then we would be able to compute M by averaging all of the times. But of course we can't collect data from the entire population. So we never will know the exact value of the mean M.

However, you typically will have some opinion, before you observe any data, about likely and unlikely values of M. Remember, this number represents the average sleeping time (on a school

night) for all college students. Since we all sleep (or at least I hope we all do!), you should be able to make an intelligent guess at the "most likely" value of M. Also, since you are not sure if your guess is correct, you could list some plausible alternative values for the mean M.

After some thought, suppose that you think that M, the mean sleeping time for all college students, will be either 6 hours, 7 hours, or 8 hours. Then there are three possible measurement models for sleeping times of college students. Either the times are normal with a mean of 6 hours, the times are normal with a mean of 7 hours, or the times are normal with a mean of 8 hours. To emphasize this point, we'll let "Normal(M)" denote a normal measurement model with mean M. We display the three possible models in the figure below.



Three normal curve models.

After you list some plausible measurement models, you assign probabilities to these models which reflect your beliefs about which ones are more or less likely. Suppose that you think that each of the three models Normal(6), Normal(7), and Normal(8) are equally likely descriptions of the sleeping times of students. In this case, you would assign the equal probabilities of 1/3, 1/3, 1/3 to the three models. The table below lists the three models and their probabilities.

MODEL	PRIOR
Normal(6)	1/3
Normal(7)	1/3
Normal(8)	1/3

Activity 18-2: What is the Average Temperature in Bismarck?

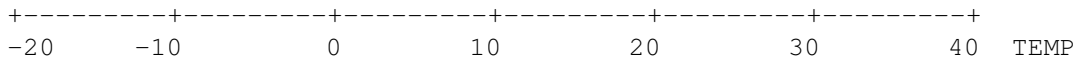
Suppose you are interested in the January temperatures in Bismarck, North Dakota. Specifically, you are interested in the long-term average daily high temperature M (measured in degrees Fahrenheit).

Now the actual January high temperatures in Bismarck can vary from day to day. However, it is reasonable to assume that the distribution of daily high temperatures resemble a normal curve with mean M and standard deviation $h = 15$ degrees. (I made a reasonable guess of the standard deviation based on my knowledge about the variation of temperatures across days.)

(a) Make an intelligent guess at the value of the long-term average daily high temperature M. (It might help to look at the average high temperatures of different cities displayed in a table in Activity 7-11.)

(b) Suggest two other plausible values for the average high temperature in Bismarck in January. These values will likely be close to the best guess that you made in part (a).

(c) On the number line below, plot the three normal measurement models that you specified in parts (a) and (b).



(d) Based on your beliefs, construct a probability distribution for your three models. In the table below, list your models and assign probabilities to each.

MODEL	PROBABILITY

Computing Posterior Probabilities Using One Observation

After we list some measurement models and assign probabilities to the models, we take a random sample from the population to get more information about the mean M. As in Topic 16, we use Bayes' rule to find our new probabilities for the measurement models. We illustrate the computation of the posterior probabilities first in the simple case where we take a **single** measurement randomly selected from the population.

In our example, we had three possible measurement models and each was assigned the same prior probability. Now suppose that we ask one student how often he or she sleeps on a typical school night and the student says 6.5 hours. What have we learned about our three models?

To find the new model probabilities by Bayes' rule, we first compute the **likelihoods** for the three models. Recall that the likelihood is the probability of observing our sample result for each of the possible models. Letting x denote our single measurement, we have observed

$$x = 6.5.$$

The probability of getting this measurement, if the mean is M, is equal to the value of the normal curve

$$LIKELIHOOD = e^{-z^2/2},$$

where

$$z = \frac{x - M}{h} = \frac{6.5 - M}{h}.$$

We illustrate the calculation of the likelihoods in the table below. First we compute the standardized score of the measurement $x = 6.5$ for each model. That is, we find the standardized score of 6.5 if the mean is $M = 6$, we compute it again when the mean is $M = 7$, and one last time for the mean value $M = 8$. These standardized scores are placed in the z column of the table. Then we compute the likelihood values from these z values.

MODEL	z	LIKELIHOOD $e^{-z^2/2}$
Normal(6)	$(6.5 - 6)/.5 = 1$	0.6065
Normal(7)	$(6.5 - 7)/.5 = -1$	0.6065
Normal(8)	$(6.5 - 8)/.5 = -3$	0.0111

Now that we have assigned prior probabilities to all models and have computed the likelihoods, the posterior probabilities for the models are found using the familiar "multiply, sum, divide" recipe that we used for learning about a proportion in Topic 16.

- We **multiply** the prior probabilities by the likelihoods to get products.
- We compute the **sum** of the products.
- We **divide** each product by its sum to get the final or posterior probabilities.

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POST
Normal(6)	1/3	0.6065	0.2022	0.4955
Normal(7)	1/3	0.6065	0.2022	0.4955
Normal(8)	1/3	0.0111	0.0037	0.0091
SUM			0.4081	

What have we learned in this calculation? Initially, we didn't favor any of the three measurement models — the normal curves with $M = 6$, $M = 7$ and $M = 8$ had equal prior probabilities. But after observing the measurement $x = 6.5$, the probabilities have significantly changed — the $M = 6$ and $M = 7$ models each have probabilities close to .5 and model $M = 8$ has a probability close to zero. Thus it is very unlikely that the mean sleeping time of all college students is equal to 8 hours. We are unsure if the mean time for all students is 6 or 7 hours — but after all, we can't learn that much about the population mean M based on a single observation.

Activity 18-3: What is the Average Temperature in Bismarck? (cont.)

(a) Copy your three models and probabilities from the table in Activity 18-2.

MODEL	PROBABILITY

(b) Suppose you sample one January day at random from the past year and the high temperature that day was 23 degrees. Compute the likelihoods of your three models from this single observation. Put your answers in the table below. (To compute the likelihood for a particular model, first compute the standardized score and then compute the value of the normal curve. Remember from Activity 18-2 that we are assuming that the population standard deviation $h = 15$.)

MODEL	z	LIKELIHOOD $e^{-z^2/2}$

(c) Find the posterior probabilities of the three models. Use the "multiply, sum, divide" recipe and put your calculations in the table below.

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POST
SUM				

(d) Look at your posterior probabilities. What was the most likely model? What can you say about the long-term average temperature in Bismarck?

Computing Posterior Probabilities Using a Sample of Observations

We have illustrated computing posterior probabilities for measurement models in the special case where we observe a single observation randomly selected from the population. But this is not the usual situation. One typically takes a random sample of measurements; in the college student study, we observed a sample of 28 sleeping times that was selected at random from the population of college students.

The only change in the computation is how we compute the likelihoods of the various models. Suppose we start with the same three possible models — $M = 6$, $M = 7$, $M = 8$ — and each is assigned a prior probability of 1/3. We observe the sleeping times (in hours)

8.0	6.0	6.5	7.0	7.0	6.5	6.5	6.0	7.0	7.0
6.0	6.0	7.5	6.0	7.0	7.0	6.5	6.0	6.5	6.5
9.0	9.0	7.0	5.0	8.0	6.5	6.0	6.0		

To find the likelihood of the model $M = 6$, we find the probability of getting this sample data if the population mean was really $M = 6$. If this data comes from a random sample, then we find the likelihood by multiplying the normal densities of all 28 observations. Remember the normal curve is given by $e^{-z^2/2}$ where z is the standardized score $\frac{x-M}{h}$. To compute the likelihood, we

- change each of the 28 observations to standardized scores using the mean value $M = 6$
- compute the normal densities from the standardized scores
- multiply all of the normal densities

This is certainly a tedious calculation and is best left to a computer. Fortunately, it turns out that there is an equivalent way of computing the likelihood that is much simpler. The likelihood can be expressed as the single normal curve

$$LIKELIHOOD = e^{-z^2/2},$$

where z is the standardized score

$$z = \frac{\sqrt{n}(\bar{x} - M)}{h},$$

where $\bar{x}$ is the mean of the sample of measurements and n is the size of the sample.

The table below illustrates the computation of the likelihoods for our example. We first compute the sample mean of the 28 measurements:

$$\bar{x} = \frac{8.0 + 6.0 + 6.5 + 7.0 + ... + 6.0 + 6.0}{28} = 6.75.$$

We next compute the standardized score z for each of the three measurement models ($M = 6$, $M = 7$, $M = 8$) using this value of the sample mean and the sample size $n = 28$. Then we compute the normal curve values $e^{-z^2/2}$ for all models.

MODEL	z	LIKELIHOOD $e^{-z^2/2}$
Normal(6)	$\sqrt{28}(6.75 - 6)/.5 = 7.9$	0.0000
Normal(7)	$\sqrt{28}(6.75 - 7)/.5 = -2.6$	0.0302
Normal(8)	$\sqrt{28}(6.75 - 8)/.5 = -13.2$	0.0000

We use these likelihood values to compute the posterior probabilities in the Bayes' table.

MODEL	PRIOR	LIKELIHOOD	PRODUCT	POST
Normal(6)	1/3	0.0000	0.0000	0.0000
Normal(7)	1/3	0.0302	0.0101	1.0000
Normal(8)	1/3	0.0000	0.0000	0.0000
SUM			0.0101	

Note that the Normal(7) model, the measurement model with a mean of 7 hours, gets all of the posterior probability. This tells us that, based on this random sample of 28 measurements, it is much more likely that the average sleeping time of all college students is 7 hours compared to 6 or 8 hours.

Actually, in this example, we probably would like to get a more accurate estimate at the average sleeping time. To get a more precise estimate at the population mean, we can choose a larger collection of models. Suppose we consider 16 measurement models, where M is equally spaced between 6.0 and 7.5 hours. In addition, suppose that we believe that all of the models are equally likely and so we assign each model a prior probability of 1/16. We have Minitab perform the calculation of the posterior probabilities which are displayed in the table below. Note that we are using briefer notation for the models in this table. The column M lists the different models — $M = 6.0$ corresponds to a normal measurement model with mean 6.0, and so on. These probabilities are displayed using a line graph below. We see from the table and the figure that the most probable values of the mean sleeping time are 6.7 and 6.8. If we select the most likely values of M, we see that the set of values

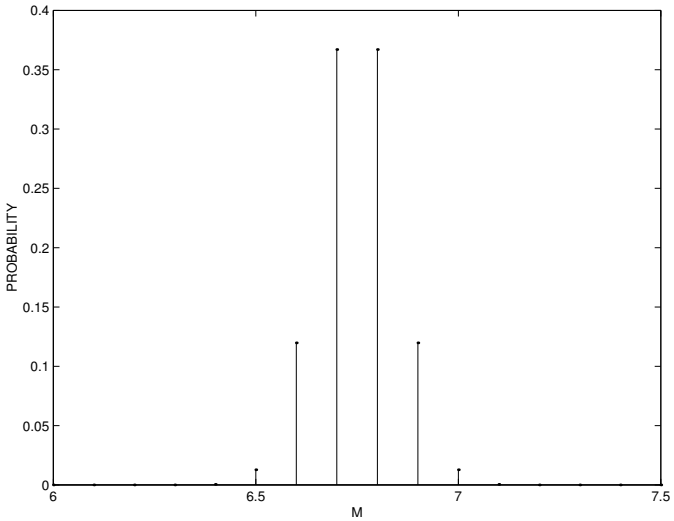
$$\{6.6, 6.7, 6.8, 6.9\}$$

has total probability

$$0.1198 + 0.3670 + 0.3670 + 0.1198 = 0.9736.$$

Thus, after taking these 28 measurements, we can say that the probability that the mean sleeping time is between 6.6 and 6.9 hours is approximately 97%.

M	POST	M	POST
6.0	0.0000	6.8	0.3670
6.1	0.0000	6.9	0.1198
6.2	0.0000	7.0	0.0127
6.3	0.0000	7.1	0.0004
6.4	0.0004	7.2	0.0000
6.5	0.0127	7.3	0.0000
6.6	0.1198	7.4	0.0000
6.7	0.3670	7.5	0.0000



Posterior probabilities for mean sleeping time.

Activity 18-4: What is the Average Temperature in Bismarck? (cont.)

To learn more about the average January temperature in Bismarck, we collect the high temperature for 10 randomly selected days in January, obtaining the values (in degrees Fahrenheit)

$$-5 \ 23 \ 21 \ 17 \ 23 \ 27 \ 22 \ 26 \ 6 \ 20$$

(a) Find the values of the sample mean $\bar{x}$ and the sample size n from this data.

(b) Compute the likelihood of the model $M = 20$. (First compute the standardized score z and then compute the normal curve. Remember from Activity 18-2 that we are assuming that the population standard deviation $h = 15$.)

(c) Suppose that one believes, before looking at any data, that M could be any one of the values 0, 1, 2, .., 30, and he or she believes that each value is equally likely. Minitab was used to compute the posterior probabilities tabulated below. The figure displays the posterior probabilities using a line graph.

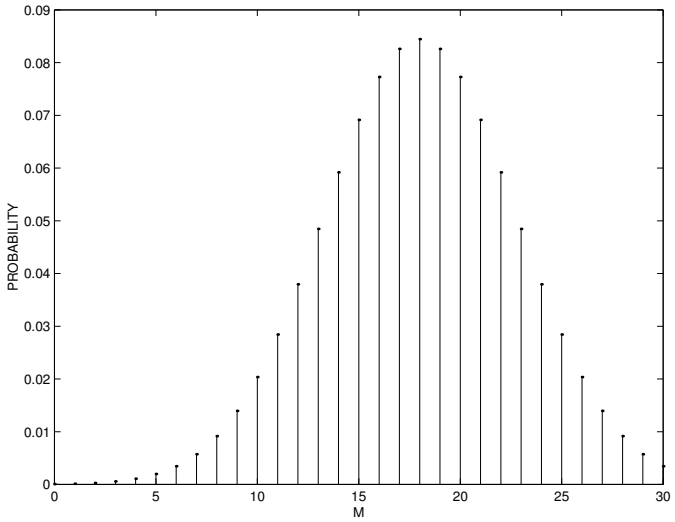
M	POST	M	POST
0	.0001	16	.0773
1	.0001	17	.0826
2	.0003	18	.0845
3	.0006	19	.0826
4	.0011	20	.0773
5	.0020	21	.0691
6	.0034	22	.0592
7	.0057	23	.0485
8	.0092	24	.0379
9	.0140	25	.0284
10	.0204	26	.0204
11	.0284	27	.0140
12	.0379	28	.0092
13	.0485	29	.0057
14	.0592	30	.0034
15	.0691		

(i) What is the most probable value of the long-term average M. What is its probability?

(ii) Find the probability that the long-term average is smaller than 10 degrees.

(iii) Find the 10 most likely values of M. Write these values and the associated probabilities below.

(iv) Find the probability that the average is one of the most likely values that you found in (iii).



Posterior probabilities for mean high January temperature in Bismarck.

HOMEWORK ACTIVITIES

Activity 18-5: Recognizing Normal Curve Data

For each of the following data, say if it can or cannot be modeled using a normal curve. If you don't think a normal curve model is appropriate for the data, explain why.

(a) The amount of gasoline (in gallons) that you purchase in 50 visits to the gas station this year.

(b) The number of points Michael Jordan scores in 20 consecutive basketball games.

(c) Suppose a fair die is rolled 100 times and the data consist of the individual rolls.

(d) The prices of 50 houses that are sold in your community this month.

(e) You record the ages of 80 college students, where a younger student (age 18 or younger) is recorded as a "1" and an older student (age over 18 years) is recorded as a "2".

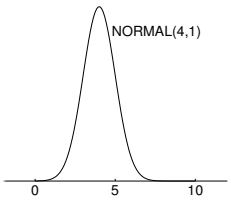
(f) The number of music cd's owned by 50 randomly selected students at your school.

(g) The amount of money spent by 40 families who visit Cedar Point on a particular summer day.

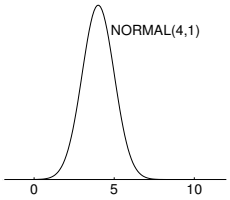
Activity 18-6: Distinguishing and Drawing Normal Curves

(a) On each graph below a normal curve with mean 4 and standard deviationi 1 has been plotted. On the same graph, draw a second normal curve with the indicated mean and standard deviation.

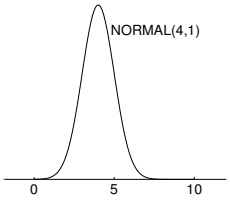
(1.) Draw a normal curve with mean $M = 7$ and standard deviation $h = 1$.



(2.) Draw a normal curve with mean $M = 4$ and standard deviation $h = 2$.



(3.) Draw a normal curve with mean $M = 0$ and standard deviation $h = 0.5$.

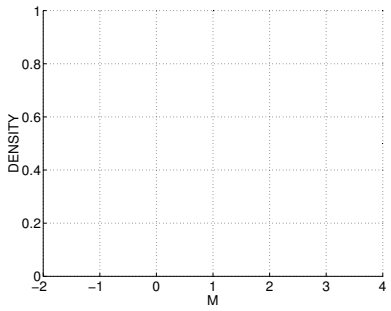


(b) In the table below, compute values of the normal curve with mean $M = 1$ and standard deviation $h = 2$. (First compute the standardized score $z = \frac{x-M}{h}$ and then compute the density $e^{-z^2/2}$.) Plot the values of the normal curve on the graph below.

x	z	$e^{-z^2/2}$
-1		
0		
1		
2		
3		

Activity 18-7: Lengths of *Wheel of Fortune* Phrases

I measured the length (in letters) of 96 phrases used in the *Wheel of Fortune* game show. The distribution of phrase lengths is approximately normal shaped with a mean of 16 letters and a standard deviation of 4 letters.



(a) Draw this normal curve on the number line below.

```
+---------+---------+---------+---------+---------+---------+
0         5         10        15        20        25        30      PHRASE LENGTH
```

(b) Find an interval of values that you think will contain about 68% of all of the phrase lengths.

(d) Find an interval of values that you think will contain about 95% of all of the phrase lengths.

(e) Suppose you are watching *Wheel of Fortune* and one of the phrases contains only 7 words. Would you be surprised at this small number? Why?

Activity 18-8: How Many States Do Students Visit?

In Topic 1, we collected data on the number of states that students had visited. Consider the number of states visited by all students that attend your school. Assume that the shape of this data is approximately normal with mean M and standard deviation h, where we assume that $h = 4$.

(a) Using your knowledge about the touring and moving habits of your fellow students, make a list of three plausible values of the mean M. Put your values of M and the associated probabilities in the table below, where you assume that the models are equally likely.

M	Probability

(b) Suppose that you sample one student and she says that she has visited $x = 10$ states. For each of the three values of the population mean M that you chose in part (a), compute the

standardized score

$$z = \frac{x - M}{h}$$

and the likelihood

$$e^{-z^2/2}.$$

Put your values of z and the likelihood in the table below.

M	z	LIKELIHOOD $e^{-z^2/2}$

(c) Using the prior probabilities from (a) and the likelihoods from (b), compute the posterior probabilities of the three models. Put your work in the table below.

M	PRIOR	LIKELIHOOD	PRODUCT	POST
SUM				

(d) Graph the posterior probabilities of the three models using a line graph. Describe how your probabilities for the values of M have changed after observing this single data value.

Activity 18-9: What is an Average Pulse Rate?

Suppose you are interested in learning about the mean 30-second pulse rate M for all students at your school. Note that it is a bit nonstandard to consider a 30-second pulse rate — usually it is measured as the number of heart beats for 60, rather than 30, seconds.

(a) Make a list of three possible values for the 30-second pulse rate M and put the values in the table below. Also put probabilities in the table, assuming the three values of M are equally likely.

M	Probability

(b) From the data collected in the preliminaries, compute the sample mean $\bar{x}$ and the sample size n. Assume that the standard deviation of the population of 30-second pulse rates is equal to $h = 5$.

(c) Compute the posterior probabilities of the three models.

(d) Based on your analysis, what is the most likely 30-second pulse rate for all students? What is its probability?

Activity 18-10: What is a Typical Marriage Age of a Bride?

In Activity 2-8, we obtained ages of a sample of 24 couples taken from marriage licenses filed in Cumberland County, Pennsylvania in June and July of 1993. The ages of the brides (in years) are given below.

22	32	50	25	33	27	45	47
30	44	23	39	24	22	16	73
27	36	24	60	26	23	28	36

Consider the population of bride ages of all marriages in Cumberland County during these two months in 1993. Suppose that we're interested in the mean bride age M of this population.

(a) Can we find the value of M by simply computing the mean age of the 24 brides in the sample above? Explain.

Suppose that we think that M conceivably could be any integer value from 25 to 40 and our prior assigns the same probability to each value of M. Using the above data and assuming that the population standard deviation $h = 12$, Minitab was used to compute the posterior probabilities.

M	POST
25	0.000
26	0.001
27	0.003
28	0.010
29	0.023
30	0.048
31	0.084
32	0.123
33	0.154
34	0.163
35	0.146
36	0.110
37	0.071
38	0.038
39	0.018
40	0.007

(b) To the area to the right of the table, construct a graph of the posterior probabilities.

(c) What is the most likely value of the mean M of the population of bride marriage ages?

(d) Find the probability that the mean age exceeds 30.

(e) Find the probability that the mean age is between 35 and 40.

(f) Find a 90% probability interval for the mean age M.

Activity 18-11: What is the Average Weight of a Newborn?

In Activity 2-4, we looked at the weights (in ounces) of newborn babies born in the local hospital. The observed weights are shown below.

147	102	107	90	126	105	143	123	117	126
132	110	121	87	110	125	129	114	124	102
97	123	126	118	136	121	133			

To learn about the mean weight M of *all* babies born at this hospital, we suppose that M can take on the values 101, 102, .., 130, and the 31 different values are equally likely. The data above was used together with the prior distribution to obtain the following posterior probabilities. (The value $h = 15$ was assumed for the standard deviation of all baby weights.)

M	POST	M	POST	M	POST
101	0.000	111	0.006	121	0.089
102	0.000	112	0.013	122	0.061
103	0.000	113	0.026	123	0.037
104	0.000	114	0.046	124	0.020
105	0.000	115	0.072	125	0.009
106	0.000	116	0.101	126	0.004
107	0.000	117	0.125	127	0.001
108	0.000	118	0.137	128	0.000
109	0.001	119	0.134	129	0.000
110	0.002	120	0.116	130	0.000

(a) Give the four "most likely" values of the mean weight M.

(b) Find the probability that M is one of the four values specified in part (a).

(c) Find a 90% probability interval for M.

(d) Suppose that a hospital spokesman claims that the mean weight of all babies born at this hospital exceeds 120. By computing an appropriate probability, check if this a reasonable statement.

WRAP-UP

In this topic, we were introduced to a normal curve model for a population of continuous measurements. This normal curve is described by a mean and a standard deviation, and we are interested in learning about the normal mean parameter M. We showed how the mean and standard deviation change the location and shape of the normal curve. Also, we were introduced to the formula for the normal curve which is used in computing a likelihood. Our method for learning is similar to the methodology for learning about a proportion in Topic 16. We construct a prior probability distribution which consists of a list of possible values of M and associated probabilities. After data are observed, we compute posterior probabilities using normal likelihoods and the usual multiply, sum, divide recipe. We find likely values for M and probability intervals based on the posterior distribution. In the next topic, we discuss learning about a mean when M is continuous and can take on values in an interval.

Topic 19: Learning About a Mean Using Continuous Models

Introduction

In the previous topic, we were interested in learning about a population of continuous measurements. We assumed that measurements from a population could sometimes be represented using a normal curve with an unknown mean M and a known standard deviation h. We performed inferences about the mean by specifying a set of plausible values for M, assigning prior probabilities to these values, and using Bayes' rule to compute posterior probabilities of the models

As an example, suppose that we're interested in learning about the average age that men marry in the state of Pennsylvania. In this case, our population consists of all of the ages of men in Pennsylvania that get married this year. We suppose that the distribution of all marriage ages is roughly normal shaped with a mean of M years and a standard deviation of h years. To make this discussion simple, we will assume that the standard deviation h is known to be 6 years.

Before we take a sample survey of recently married men, we construct a prior probability distribution on possible values of the mean M. Remember what M represents — in this example, M represents the mean age of *all* men who are currently getting married in Pennsylvania. To perform inference using the methods described in the previous topic, we need to first list a few plausible average marriage ages, say 20, 25, and 30 years, and then assign probabilities to these average values.

There can be problems actually doing this task. First, it is a little artificial to just state a few values of this average marriage age. In the above paragraph we said that the average age M could only be 20, 25, or 30 years. Why can't the average marriage age be 23, or 22.5, or even 24.106 years? If we think about this problem, it becomes clear that many different values of M are possible. It might be more realistic to say that the average age is located anywhere between 20 and 30 years. In other words, it may be better to assume that the unknown mean M takes on continuous values on an interval that extends from 20 to 30 years. Also, it can be difficult specifying prior probabilities

on the different measurement models. If one is specifying, say, ten different values of M, then he or she will have to assign probabilities to each of the ten models. This can be a difficult job, especially when one has informative prior beliefs about the population mean M.

In this topic, we will assume that the mean is continuous-valued and discuss two types of methods of representing prior beliefs. By use of a **uniform prior curve**, we say that we have little prior knowledge about the location of M, and this choice leads to a normal posterior distribution for the mean. In the case where we have some knowledge about the mean, a **normal prior curve** can be used to represent prior information and we will show that this also results in a normal posterior distribution.

PRELIMINARIES

1. Guess at an average January temperature in Bismarck, North Dakota.
2. Suppose you were able to measure the high temperature in Bismarck for the 31 days in January this year. Draw on the number line below what you think the distribution of 31 temperatures would look like.

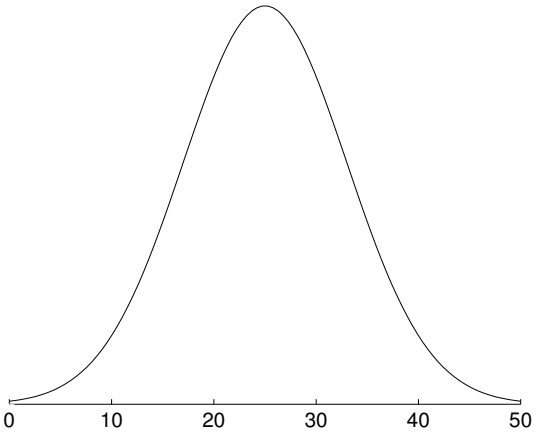
```
---+---------+---------+---------+---------+---------+--- TEMPERATURE
  -20       -10        0        10        20        30 (deg. Fahrenheit)
```

3. How many keys do you carry with you?
4. How many keys do you think a typical student carries at your school?
5. Explain why the answers to parts 3. and 4. are different.

Normal curves

Before we start talking about learning about a population mean, it will be helpful to discuss normal curves. Like the beta curve discussed in Topic 17, a normal curve is a probability curve for a continuous variable. Its shape is the popular bell or mound shape that has been discussed in previous topics. It is described by two numbers, a mean M and a standard deviation h. The mean M tells one where the curve is centered and the number h reflects the spread of the curve.

The figure below displays the normal curve with mean 25 and standard deviation 8. (This curve will be used later in this topic to reflect one's opinions about the average marriage age of men in a particular state.) We see that the curve reaches its highest point at the value 25 and the curve is symmetric about this value.



A normal curve with mean 25 and standard deviation 8.

Finding areas under a normal curve

Since the normal curve is a probability curve, the total area under the curve is equal to one. To find probabilities of intervals, we find areas under this curve. As in the case of a beta density, we use Minitab to compute **cumulative areas** under a normal curve, and using these values, we can find any area under the curve.

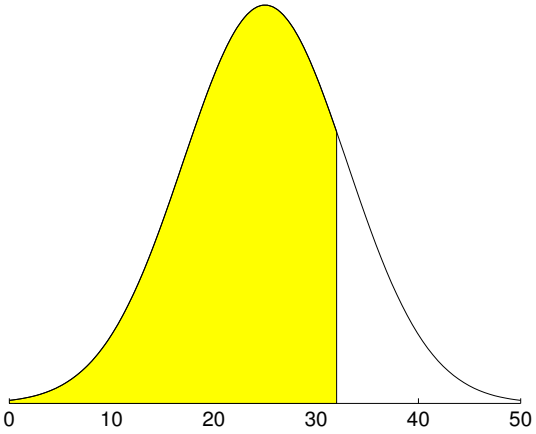
"Less than" areas

Suppose we wish to find the probability that the **variable is less than 32**. This probability is represented by the area under the curve to the left of 32, as shown in the figure below. We find this "less than" area by use of the Minitab "cdf" command. In the output below, note that we put the value that we're interested in (32) on the first line and the mean and standard deviation of the normal curve on the second line. The Minitab output gives the probability that the variable is less than or equal to the value. Here we see that the probability that the variable is less than 32 is equal to .8092.

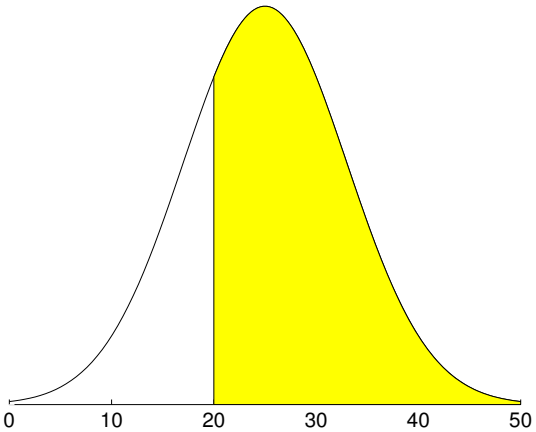
```
MTB > cdf 32;
SUBC> normal 25 8.
Normal with mean = 25.0000 and standard deviation = 8.00000

          x     P( X <= x)
   32.0000         0.8092
```

"Greater than" areas

Next, suppose that we wish to find the chance that the **variable is larger than 20**. This probability is represented by the area to the right of 20. Minitab won't find this probability directly, but it can find the area to the left — from the output below, we see that the area to the left is .2660.



Area under the normal curve to the left of 32.



Area under the normal curve to the right of 20.

```
MTB > cdf 20;
SUBC> normal 25 8.
Normal with mean = 25.0000 and standard deviation = 8.00000

          x      P( X <= x)
   20.0000         0.2660
```

Remember that the total area under the curve is 1. So the probability that the variable is greater than 20 is equal to the

$$\text{(Area to the right of 20)} = 1 - \text{(Area to the left of 20)} = 1 - .2660 = .7340.$$

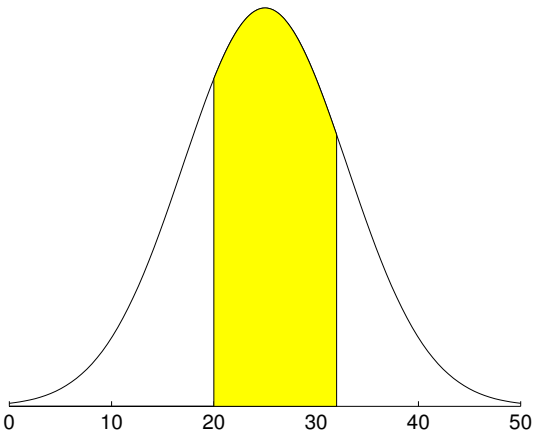
Areas between two values

The third type of area we might want to find is the area between two values. For example, suppose we want to find the probability that the variable is between 20 and 32, which is pictured in the figure below. Note that this area is the difference of two "less than" areas.

$$\text{(Area between 20 and 32)} = \text{(Area to the left of 32)} - \text{(Area to the left of 20)}$$

From the Minitab output, we see that the area to the left of 32 is .8092 and the area to the left of 20 is .2660, so the area between the two values is

$$\text{(Area between 20 and 32)} = .8092 - .2660 = .5432.$$

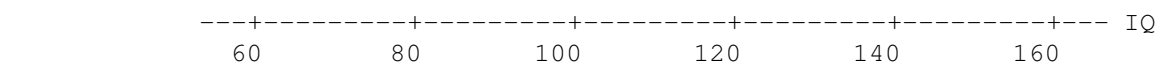


Area under the normal curve between 20 and 32.

Activity 19-1: Finding Areas Under a Normal Curve

Consider a normal curve with mean 100 and standard deviation 15. This curve represents the distribution of intelligence quotients (IQ's) for a large group of people.

(a) On the number line below, graph this normal curve.



The Minitab output below displays "less than" areas under this normal curve for a range of IQ values.

```
MTB > CDF C1;
SUBC>   Normal 100 15.

Normal with mean = 100.000 and standard deviation = 15.0000

          x     P( X <= x)
    80.0000         0.0912
    85.0000         0.1587
    90.0000         0.2525
    95.0000         0.3694
   100.0000         0.5000
   105.0000         0.6306
   110.0000         0.7475
   120.0000         0.9088
```

For each of the following parts, (1) draw the normal curve, (2) shade the area that you want to find, and (3) find the area from the Minitab output.

(b) Find the probability that an IQ score is less than 90.

(c) Find the probability that an IQ score is less than 110.

(d) Find the probability that an IQ score is between 100 and 110.

(e) Find the probability that an IQ score is larger than 120.

(f) Find the probability that an IQ score is between 85 and 120.

Finding percentiles for a normal curve

In the previous activity, we focused on obtaining areas or probabilities under the normal curve. We're also interested in computing percentiles for this curve. Recall our definition of percentile from our discussion of beta curves in Topic 17. A 25th percentile is the value of the variable, call it M_{25}, such that the area to the left of M_{25} is equal to .25 or 25 per cent. We illustrate this value in the figure below. This value is computed using the Minitab "invcdf" command. In the output shown below, the given area to the left (.25) is placed on the first line, and the mean and standard deviation of the normal curve is placed on the second line. The output gives the variable value to be 19.6041. This means that the area to the left of 19.6041 is equal to .25.

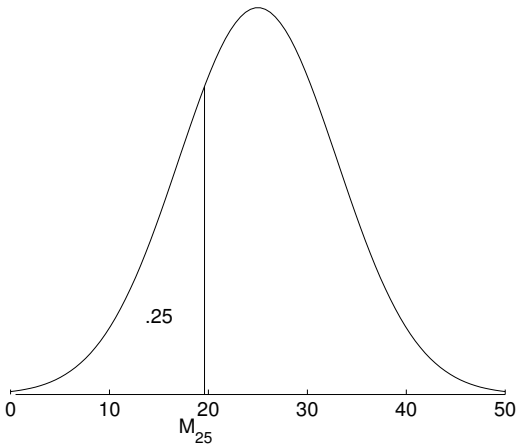


Illustration of 25th percentile of a normal curve.

```
MTB > invcdf .25;
SUBC> normal 25 8.
Normal with mean = 25.0000 and standard deviation = 8.00000

 P( X <= x)           x
    0.2500       19.6041
```

As a second example, suppose we wish to find two values which cover the middle 80% of the area under this normal curve. From the figure below, observe that, if we are interested in the middle 80%, then 10% of the area will fall to the left of the smaller value, and 10% will fall to the right of the larger value. We find these values by computing two percentiles for this normal curve — the left value is the 10th percentile and the right value is the 90th percentile.

We use the Minitab "invcdf" command twice to compute these two percentiles. From the output, we note that the 80% middle area lies between the values 14.7476 and 35.2524.

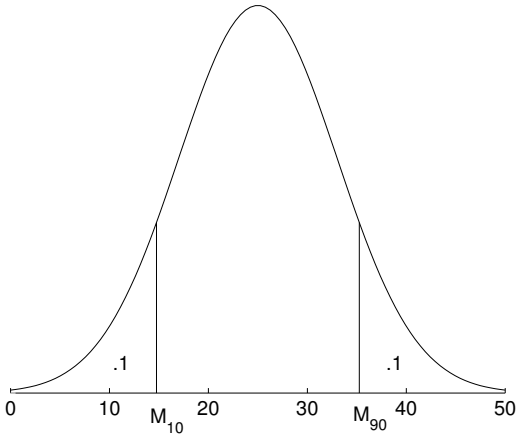


Illustration of 10th and 90th percentiles of a normal curve.

```
MTB > invcdf .1;
SUBC> normal 25 8.
Normal with mean = 25.0000 and standard deviation = 8.00000

 P( X <= x)          x
    0.1000       14.7476

MTB > invcdf .9;
SUBC> normal 25 8.
Normal with mean = 25.0000 and standard deviation = 8.00000

 P( X <= x)          x
    0.9000       35.2524
```

Activity 19-2: Finding Percentiles Under a Normal Curve

Again consider the population of IQ scores which is represented by a normal curve with mean 100 and standard deviation 15.

(a) Suppose that you wish to find the 75% percentile of IQ scores — call this percentile P. The proportion of IQ scores smaller than P is equal to ____ .

(b) What is the proportion of IQ scores greater than the value P?

The Minitab output below gives the percentiles of this normal curve for different probabilities or areas to the left. All of the questions below can be answered using this table.

```
MTB > InvCDF C1;
SUBC>   Normal 100 15.
```

```
Normal with mean = 100.000 and standard deviation = 15.0000

 P( X <= x)          x
    0.1000       80.7767
    0.2000       87.3757
    0.3000       92.1340
    0.4000       96.1998
    0.5000      100.0000
    0.6000      103.8002
    0.7000      107.8660
    0.8000      112.6243
    0.9000      119.2233
```

(c) Find the IQ scores such that 70% of all of the IQ's are smaller than that value.

(d) Find the IQ score such that the area to the *right* of that value is .7.

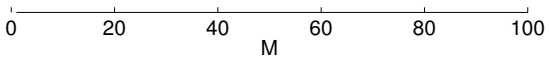
(e) Find an interval which contains 80% of the area under the IQ curve.

(f) Find an interval which contains 50% of the IQ scores. (More than one answer is possible.)

Learning About a Mean Using a Uniform Prior

In our example, we're interested in the average age M of all men who are currently getting married in Pennsylvania. Suppose that we think that M could be any number in a wide range. If we believe that the youngest man to be married is Pennsylvania is 15 years old and the oldest is 80 years old, then any value of M between 15 and 80 is plausible. Moreover, we have no reason to think that any values of M in the range 15-80 are any less or more likely than other values in the range. In this case, we represent our prior opinion by means of a flat curve, called a **uniform prior**, which is drawn in the figure below.

It is important to distinguish this prior from the one discussed in Topic 18. In that topic, we only assumed that the average marriage age was among particular whole numbers like 20, 21, and 30. That prior was called a **discrete prior** since the mean M was allowed only to take discrete values. Here we are using a **continuous prior** where **all** numbers between 15 and 80 are possible values for the average marriage age. A mean value of 20.12 is possible, as well as values of 34.15435 or 45.9.



Uniform prior for mean age of grooms in Pennsylvania.

Each model is a normal curve measurement model with mean M, but there are an infinite number of models of this type.

The posterior curve for the mean M is found using the familiar Bayes' rule recipe. Suppose we take a random sample of n measurements and compute the sample mean $\bar{x}$. We showed in Topic 18 that the likelihood is given by the normal density

$$LIKE = e^{-z^2/2},$$

where z is the standardized score

$$z = \frac{\sqrt{n}(\bar{x} - M)}{h}.$$

We can represent our uniform prior curve by the constant equation

$$PRIOR = 1.$$

Then the posterior curve for the mean M is obtained by multiplying the prior and likelihood curves:

$$POST = PRIOR \times LIKE = 1 \times e^{-z^2/2} = e^{-z^2/2}.$$

By inspecting this formula, we see that the posterior curve for M, when M is assigned a uniform prior, is a normal curve with mean $\bar{x}$ and standard deviation $h/\sqrt{n}$.

In our example, suppose we observe the following ages of the groom for a random sample of 12 married couples in Pennsylvania (in years):

22	38	31	42	23	55
26	24	19	42	34	31

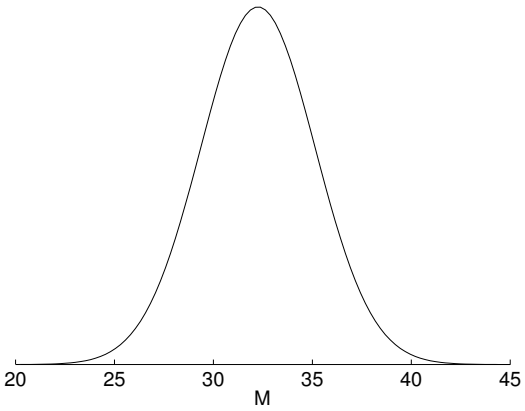
We compute the sample mean

$$\bar{x} = \frac{22 + 38 + 31 + 42 + 23 + 55 + 26 + 24 + 19 + 42 + 34 + 31}{12} = 32.25.$$

Suppose we know, from past experience dealing with marriage records, that the standard deviation h of the measurement model is equal to $h = 10$ years. If we assign a uniform prior to the average marriage age M, then the posterior curve will be normal with

$$\text{mean} = 32.25, \text{ standard deviation} = \frac{10}{\sqrt{12}} = 2.89.$$

A graph of this normal curve is displayed below.



Posterior curve for mean age of grooms in Pennsylvania.

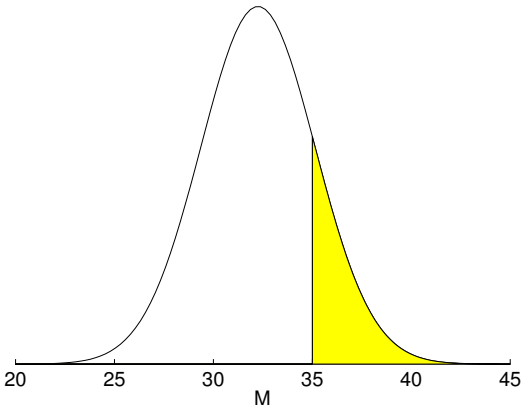
This posterior curve reflects our current beliefs about the location of the average marriage age M. We can answer specific questions regarding the mean M by calculating probabilities or percentiles for this normal curve. Specifically, let's consider the following questions.

1. **What is my best guess at the mean age of grooms in Pennsylvania?**

   A reasonable "best guess" at M is the value where the posterior normal curve is the highest. For a normal curve, the highest point occurs at the mean which is equal to $\bar{x} = 32.25$. So, based on this sample, I think that the mean marriage age is close to 32 years.

2. **Is it likely that the mean age of grooms in Pennsylvania is over 35 years?**

   We can answer this question by computing the probability that M is larger than 35. This probability is the area under the posterior normal curve for values of M larger than 35, as shown in the figure below. We find this area using Minitab. In the output below, see that



Probability that the mean age of grooms in Pennsylvania is over 35.

we use the Minitab "cdf" command — the value of interest (35) and the mean (32.25) and standard deviation (2.89) of the normal curve are put on the second line.

```
MTB > cdf 35;
SUBC> normal 32.25 2.89.
Normal with mean = 32.2500 and standard deviation = 2.89000

          x     P( X <= x)
   35.0000         0.8293
```

From the output, we see that the probability that the mean age M is smaller than 35 is equal to

$$\text{(Area to left of 35) = .8293.}$$

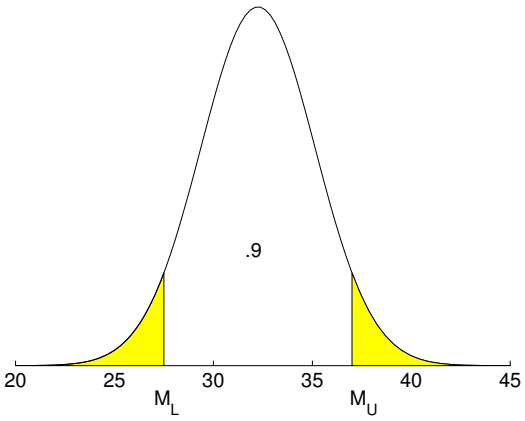
Thus the probability that the mean age is larger than 35 is equal to

$$\text{(Area to right of 35)} = 1 - .8293 = .1707.$$

Remember we were asked if it was "likely" for the mean age M to be larger than 35 years. The probability that M is larger than 35 years is about 17%. Seventeen percent is not likely, but since this probability is not tiny, it is still possible for the mean age to be this large.

3. **Can I construct an interval of values such that I am pretty confident that the mean M is in the interval?**

We can construct this interval by computing percentiles for the normal curve. Say, for example, that you want to construct a 90% probability interval. We find two values of M, call them M_L and M_U, such that the probability that M is smaller than M_L is .05 and the probability that M is larger than M_U is .05. These values, shown in the figure below, correspond to the 5th and 95th percentiles of the normal curve. The probability that M falls between M_L and M_U is .90 — thus (M_L, M_U) is a 90% probability interval for this average marriage age.



90% probability interval for the mean age of grooms in Pennsylvania.

Here we find these two percentiles using Minitab. To find a single percentile using the Minitab command "invcdf", we put the "left area" value on the first line and the normal mean and standard deviation on the second line. For example, to find the 5th percentile, we put the left area number .05 on the first line and the mean 32.35 and standard deviation 2.89 on the second line.

```
MTB > invcdf .05;
SUBC> normal 32.25 2.89.
Normal with mean = 32.2500 and standard deviation = 2.89000

 P( X <= x)          x
    0.0500       27.4964

MTB > invcdf .95;
SUBC> normal 32.25 2.89.
Normal with mean = 32.2500 and standard deviation = 2.89000

 P( X <= x)          x
   0.9500        37.0036
```

We see that the 5th and 95th percentiles of the posterior probability curve are 27.4964 and

37.0036, respectively. Thus the probability that the average marriage age M falls between 27.4964 and 37.0036 is 90%.

Activity 19-3: What is the Average Temperature in Bismarck? (cont.)

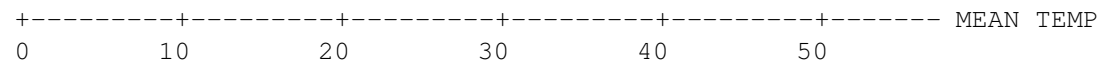
In Topic 19, we were interested in learning about the average January temperature M in Bismarck. We collected the following high temperatures (in degrees Fahrenheit) for 10 days in January:

$$-5 \ 23 \ 21 \ 17 \ 23 \ 27 \ 22 \ 26 \ 6 \ 20$$

Suppose that we have little prior knowledge about the average temperature in this city, and so we use a uniform prior to describe our beliefs about M. As in Activity 18-2, assume that the standard deviation h of the distribution of January temperatures is known to be equal to 15 degrees.

(a) Find the mean and standard deviation of the posterior curve for the average January temperature M.

(b) On the number line below, plot the posterior curve for M.



(c) Compute (without using Minitab) the probability that the mean January temperature is larger than 18 degrees.

(d) Using Minitab, compute the probability that the mean January temperature is smaller than 10 degrees.

(e) Using Minitab, compute the probability that the mean January temperature is between 15 and 25 degrees.

(f) Using Minitab, find a 90% probability interval for M.

What to Do if the Standard Deviation is Unknown?

In our computations of the posterior curve in this topic, we have assumed that one knows the value of the population standard deviation h. Remember we are learning about the normal measurement model — the mean of this measurement model is M and the standard deviation is h. In typical problems, if we don't know the mean value M, we also won't know the value of the standard deviation h. In our example, the assumption that we knew the standard deviation of the population of marriage ages of the Pennsylvania men is not very realistic.

What can we do when we don't know the standard deviation h of our population of measurements? A simple solution is to use the value of the standard deviation s from our sample of measurements as our best guess at h. If we just replace the value of h in our formulas by the sample value s, then we can compute the posterior density of the mean M that we are interested in.

However, if we just substitute the sample standard deviation s for h, the posterior density for M that we obtain is not very accurate. The expression for the posterior density doesn't account for the fact that we really don't know the value of the standard deviation h and are estimating its value by a number computed from the sample. To make the posterior density more accurate, we use the following adjusted formula for h when we estimate this standard deviation by the sample value s:

$$h = s\left(1 + \frac{20}{n^2}\right).$$

Let's revisit our marriage age example. Suppose that we don't know the value of the standard deviation h which describes the spread of the population of marriage ages. We observed the following marriage ages of 12 men:

22	38	31	42	23	55
26	24	19	42	34	31

By using the Minitab describe command, we compute the mean and standard deviation of these 12 ages:

```
MTB > desc 'ages'

Variable            N       Mean     Median     TrMean      StDev    SE Mean
C2                 12      32.25      31.00      31.30      10.53       3.04
```

From the output, we see that $\bar{x} = 32.25$ and $s = 10.53$.

If we assume that our prior beliefs about the mean age M are described by a uniform prior, then the posterior curve for M will have a normal shape with mean $\bar{x} = 32.25$ and standard deviation

$$\frac{h}{\sqrt{n}},$$

where h is now computed using the formula

$$h = s\left(1 + \frac{20}{n^2}\right) = 10.53\left(1 + \frac{20}{12^2}\right) = 11.99.$$

So the standard deviation of the normal curve is

$$\frac{11.99}{\sqrt{12}} = 3.46.$$

Let's compare this posterior curve with the one computed earlier. In the first case, when we knew the value of the population standard deviation h to be equal to 10 years, the standard deviation of our posterior normal curve was 2.89 years. In this case, when we are estimating h from our sample, the posterior curve has a standard deviation of 3.46 years. We know less about the average marriage M in this case; we know less since we are unsure about the values of both the mean M and the standard deviation h.

Activity 19-4: What is the Average Temperature in Bismarck? (cont.)

In Activity 19-3, the posterior density for the average January temperature M was computed assuming that the standard deviation h of the daily temperatures was equal to 15. Suppose now that you don't know the value of this population standard deviation (which is typically the case).

(a) Using Minitab, compute the standard deviation s of the sample of 10 temperatures listed in Activity 19-3.

(b) Find the mean and standard deviation of the posterior curve of the mean January temperature M.

(c) Using Minitab, construct a 90% probability interval for the mean temperature.

(d) Compare the probability interval computed in part (c) with the interval computed in Activity 19-3. Which interval is longer? Can you explain why one interval is longer than the other?

A Normal Prior

We have illustrated learning about a mean M in the case where we have little prior information about the mean and we use a uniform prior curve to reflect this lack of knowledge. What if we have some information about the location of the mean M? Here we explain how a normal curve can be used to model our beliefs about this unknown mean.

In our marriage ages example, the mean M represents the mean marriage age for all of the men in Pennsylvania. Before any data is observed, it is very possible that you might have some opinion about the value of M. Perhaps you have friends or relatives that have been married recently. These friends or relatives may not have been from Pennsylvania, but it is reasonable to think that the male marriage ages among your friends and relatives might be similar to the male marriage ages in Pennsylvania. In any event, based on your experiences, you likely have same opinion about plausible values for M.

We represent our prior information about a mean M by use of a normal curve with mean m_0 and standard deviation h_0. The center of this normal curve, m_0, represent your "best guess" at the mean M. In our example, we obtain the value of m_0 by making an intelligent guess at the average marriage age of men in Pennsylvania.

The standard deviation h_0 of this prior normal curve indicates how sure you are about your guess at the mean M. If you are extremely sure that M is close to your guess m_0, then a very small value of h_0 would be chosen. In contrast, a large value of the standard deviation h_0 means that you are relatively uncertain about your guess.

In many situations, it can be difficult to directly guess the value of the standard deviation h_0. It can be easier to state percentiles of the normal curve. If the user can make an intelligent guess at two percentiles of the curve, then one can find the mean and standard deviation of the normal curve which matches this information.

Let's illustrate choosing a normal curve for the marriage example. Suppose you say that

(A) Your best guess at the mean age M is 28 years. You think that it is equally likely that the mean M is smaller than 28 and that the mean is larger than 28.

(B) You are pretty confident (with probability .9) that the mean age is smaller than 35.

Statement (A) is a statement that the 50th percentile is equal to 28 and statement (B) says that the 90th percentile is equal to 35. In general, suppose that your guess at the 50th percentile is m_0 and your guess at the pth percentile is m_p. Then the normal curve with mean m_0 and standard deviation h_0 will match this information, where h_0 is computed using the formula

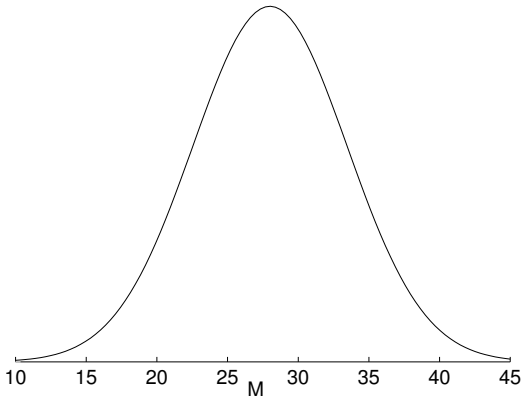
$$h_0 = \frac{m_p - m_0}{z_p}.$$

In the formula for h_0, z_p is the pth percentile of a normal curve with mean 0 and standard deviation 1; this the value of this special normal curve where the area to the left is equal to $p\%$.

For this example, $m_0 = 28$, and, using $p = 90$, we find using a computer program that $z_{90} = 1.2816$. So the standard deviation of the matching normal curve is given by

$$h_0 = \frac{35 - 28}{1.2816} = 5.46.$$

This normal curve with mean 28 and standard deviation 5.46 is displayed below. Looking at this curve, we see that most of the area under the curve falls between 20 and 35. Thus it appears that you are pretty confident that the average marriage age for these men in Pennsylvania falls between 20 and 35.



Normal prior for the mean age of grooms in Pennsylvania.

Activity 19-5: How Many Keys?

Suppose you are interested in learning about the mean number of keys carried by students at your school. Imagine that you were able to find out the number of keys carried by each student at school, and M is the mean of the entire collection of key numbers.

(a) Make a guess, m_0, at the value of the mean M. This number should be chosen so that M is equally likely to be smaller or larger than m_0. (This guess will be the 50th percentile of the prior on M.)

(b) Find a second number, call it M_{90}, such that your probability that M is smaller than M_{90} is equal to 90%.

(c) Use Minitab to find the values of the mean and standard deviation of the normal curve which match the two percentiles that you found in parts (a) and (b).

(d) Use Minitab to find the 10th percentile, M_{10}, of the normal curve found in part (c).

(e) If the normal curve accurately represents your prior beliefs, then you should feel pretty confident (with probability .9) that the mean number of keys is larger than M_{10}. Is this reasonable according to your prior opinion? If not, you should repeat parts (a), (b), (c), (d) to try to construct a normal curve which better approximates your prior beliefs.

Learning About a Mean Using a Normal Prior

Suppose that our prior beliefs about a mean M can be modeled using a normal curve with mean m_0 and standard deviation h_0. In our marriage example, we have a normal curve with mean 28 years and standard deviation 5.46 years which represents our opinion about the mean marriage age of men in Pennsylvania. Now we observe a random sample of marriages and record the ages of the grooms. How do we combine our prior beliefs with the information contained in the sample?

Our updated beliefs about the mean M are contained in the posterior curve. We obtain this curve by our basic recipe

$$POST = PRIOR \times LIKE.$$

Here the prior density has the form

$$e^{-z_P^2/2},$$

where z_P is the prior standardized score

$$z_P = \frac{M - m_0}{h_0}.$$

As before, the likelihood has the similar form $e^{-z^2/2}$, where z is the standardized score

$$z = \frac{\sqrt{n}(\bar{x} - M)}{h}.$$

When one multiplies the prior curve by the likelihood, one can show that the posterior curve also has the normal form where the mean and standard deviation are given by

$$m_1 = \frac{n\bar{x}/h^2 + m_0/h_0^2}{n/h^2 + 1/h_0^2},\ h_1 = \frac{1}{n/h^2 + 1/h_0^2}.$$

Since the above formulas look a bit complicated, we'll back up and show how the posterior curve combines the prior information and the information given in the data. First, it is helpful to define the term **precision**. The precision, denoted by c, is the reciprocal of the square of the standard deviation:

$$c = \frac{1}{h^2}.$$

The precision, as you might guess, measures the precision of the data or your prior beliefs. For example, if you are really sure you know the value of the mean M, you would give the normal curve a very small standard deviation which would result in a large value of the precision. Likewise, if you are unsure about the location of M, the normal prior curve would have a large spread and therefore a low precision.

The **data precision** measures how much information you have in the data. Recall that the standard deviation of the posterior curve when we had little prior information was

$$\frac{h}{\sqrt{n}}.$$

If we take the reciprocal of this standard deviation and square it, we obtain the data precision

$$c_D = \frac{n}{h^2}.$$

Likewise, the **prior precision** measures the amount of prior information. The prior standard deviation is h_0 and so the prior precision is

$$c_0 = \frac{1}{h_0^2}.$$

Using the notion of precision, here is a recipe for computing the mean and the standard deviation of the posterior curve. The table below contains the calculations.

1. First put the values of the prior mean and prior standard deviation in the "prior" row of the table. Also put the sample mean and the data standard deviation $h/\sqrt{n}$ in the "data" row of the table.

2. Compute the precision of the data c_D and the precision of the prior c_0. Put these two numbers in the "precision" column of the table.

3. The precision of the posterior, c_1, is found by **adding** the data and prior precisions:

   $$c_1 = c_0 + c_D.$$

   If we think of precision as information about the mean M, we are combining the information in the data and prior by adding the precisions. We put this value of the posterior precision in the table.

4. Compute the standard deviation of the posterior, h_1, by taking the reciprocal of this precision and then taking the square root:
$$h_1 = \frac{1}{\sqrt{c_1}}.$$

5. The posterior mean, m_1, can be expressed as a weighted average of the sample mean $\bar{x}$ and the prior mean m_0, where the weights depend on the precisions:
$$m_1 = \frac{c_0 m_0 + c_D \bar{x}}{c_1}.$$

	Mean	Standard Deviation	Precision
Prior	m_0	h_0	$c_0 = 1/h_0^2$
Data	$\bar{x}$	$h/\sqrt{n}$	$c_D = n/h^2$
Posterior	m_1	$h_1 = 1/\sqrt{c_1}$	$c_1 = c_0 + c_D$

We illustrate these calculations for the marriage age example.

1. Recall that our normal prior is normal with mean $m_0 = 28$ and standard deviation $h_0 = 5.46$ — we put these numbers in the first row of the table. From the data, we observe the sample mean $\bar{x} = 32.25$ and the data standard deviation is $h/\sqrt{n} = 3.46$ — these numbers are placed in the second row.

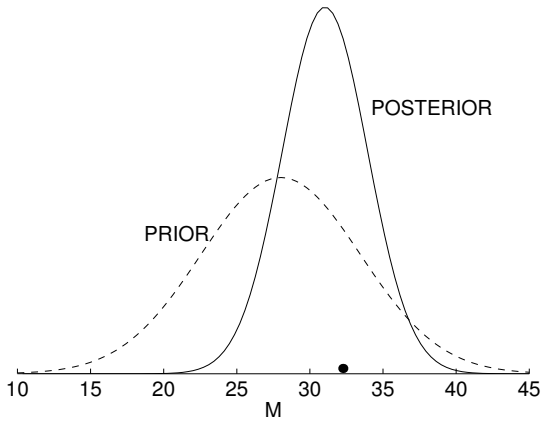
2. We next compute the precisions by taking the reciprocals and squaring the standard deviations:
$$c_0 = \frac{1}{5.46^2} = .0335 \ , c_D = \frac{1}{3.46^2} = .0835.$$

3. We sum the two precisions above to get the posterior precision:
$$c_1 = .0335 + .0835 = .1170.$$

4. We compute the posterior standard deviation by flipping and taking the square root of the posterior precision:
$$h_1 = \frac{1}{\sqrt{.1170}} = 2.923.$$

5. We compute the posterior mean:
$$m_1 = \frac{.0335 \times 28 + .0835 \times 32.35}{.1170} = 31.03.$$

	Mean	Standard Deviation	Precision
Prior	28	5.46	.0335
Data	32.25	3.46	.0835
Posterior	31.03	2.923	.1170



Prior and posterior curves for the mean age of grooms in Pennsylvania.

So the posterior curve for the mean marriage age M is normal with mean 31.03 years and standard deviation 2.923 years. On the figure below, the prior and posterior curves for the mean age are plotted. In addition, the value of the sample mean is plotted using a large dot on the horizontal axis.

We notice a few things from the graph.

1. The posterior curve for the mean M falls to the right of the prior curve towards the value of the sample mean $\bar{x}$. The posterior is a compromise between the information contained in the prior and the data. Note that the posterior mean, 31.03, falls between the prior mean, 28, and the data mean, 32.25.

2. The posterior curve has less spread than the prior curve. This is true because the posterior curve is based on more information than the prior curve. As one gains more information about a parameter, such as a population mean M, you are more sure of its value, and therefore the curve has a smaller standard deviation.

Activity 19-6: How Many Keys? (cont.)

In Activity 19-5, you constructed a normal curve prior for the mean number of keys carried by students at your school.

(a) Write down (from Activity 19-5) the mean and standard deviation of the normal curve prior.

Suppose that 20 students are surveyed and you find out the number of keys carried by each student. The data are given below:

6	1	9	9	3	13	13	8	9	9
7	11	6	17	7	8	12	8	8	5

The sample mean and sample standard deviation are given by

$$\bar{x} = 8.45, \;\; s = 3.62.$$

(b) Using the information supplied above, complete the table below.

	Mean	Standard Deviation	Precision
Prior			
Data			
Posterior			

(c) Compute the posterior mean of the mean number of keys M.

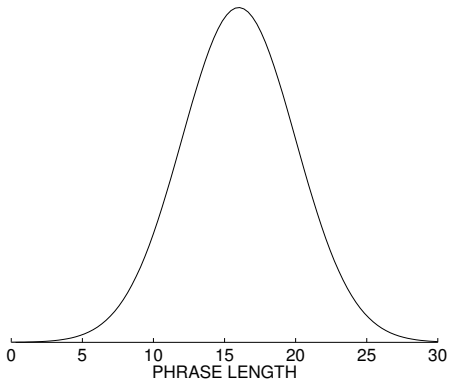
(d) The posterior curve for M has a normal form with mean _____ and standard deviation _____

(e) Using the posterior curve, find the probability that the mean number of keys exceeds 10.

HOMEWORK ACTIVITIES

Activity 19-7: Lengths of *Wheel of Fortune* Phrases (Continued)

Consider again the length (in letters) of 96 phrases used in the *Wheel of Fortune* game show. The distribution of phrase lengths is approximately normal shaped with a mean of 16 letters and a standard deviation of 4 letters. This normal curve is shown below.



Find each of the following probabilities by drawing the normal curve and shading the area that you need to find. Then use the Minitab table of cumulative probabilities to find the probability of interest.

```
Normal with mean = 16.0000 and standard deviation = 4.00000

          x     P( X <= x)
    12.0000         0.1587
    13.0000         0.2266
    14.0000         0.3085
    15.0000         0.4013
    16.0000         0.5000
    17.0000         0.5987
    18.0000         0.6915
    19.0000         0.7734
    20.0000         0.8413
```

(a) The probability a *Wheel of Fortune* phrase is shorter than 13 letters.

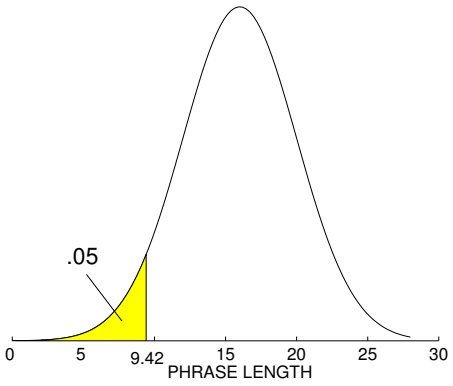
(b) The probability a phrase is between 12 and 18 letters.

(c) The probability a phrase is longer than 20 letters.

(d) The probability a phrase is shorter than 20 letters.

Activity 19-8: Lengths of *Wheel of Fortune* Phrases (Continued)

The table below gives some percentiles (computed using Minitab) of the normal curve which describes the distribution of *Wheel of Fortune* phrases. Reading the first line of the table, we see that the probability that a phrase is less than 9.4206 words (have you ever seen a phrase with 9.4206 words?) is .05 or 5 per cent. (See the figure below.) So it's relatively unlikely for a phrase to be nine words or shorter. Using this table, find




```
Normal with mean = 16.0000 and standard deviation = 4.00000

 P( X <= x)           x
```

```
    0.0500        9.4206
    0.2500       13.3020
    0.5000       16.0000
    0.7500       18.6980
    0.9500       22.5794
```

(a) The number such that 95% of the phrases are smaller than that number.

(b) A number such that 75% of the phrases are *larger* than that number.

(c) Two numbers which contain 90% of all of the *Wheel of Fortune* phrases.

(d) A number such that half of the phrases are smaller than that number and half are longer than that number.

Activity 19-9: Times to Commute to Work

Everyday I drive from home to work and I keep track of the commuting time. It doesn't always take the same amount of time. My trip will be shorter or longer depending on the weather, traffic, traffic lights, and so on. I record my commuting time for many days and the distribution of driving times is well approximated by a normal curve with mean 30 minutes and standard deviation 5 minutes.

(a) Draw this normal curve on the number line below.

```
    ----+---------+---------+---------+---------+---------+-- TIME
       10        20        30        40        50        60
```

Use the cumulative probabilities given below to answer the following questions:

```
MTB > CDF 'time';
SUBC>   Normal 30 5.
Normal with mean = 30.0000 and standard deviation = 5.00000

          x     P( X <= x)
   20.0000        0.0228
   22.0000        0.0548
   24.0000        0.1151
   26.0000        0.2119
   28.0000        0.3446
   30.0000        0.5000
   32.0000        0.6554
   34.0000        0.7881
   36.0000        0.8849
   38.0000        0.9452
   40.0000        0.9772
```

(b) What is the probability that it takes me under 28 minutes to drive to school?

(c) What is the chance that it takes between 28 and 32 minutes to get to school?

(d) What is the chance that it takes me over 40 minutes to get to school?

(e) Suppose I have a meeting at work scheduled for 10 am. When should I leave home so I am about 95% sure that I'll arrive on time for the meeting?

Activity 19-10: Marriage Ages (Continued)

In Activity 2-8, the ages of the bride and groom for some marriage licenses in Cumberland County, Pennsylvania was recorded. Suppose that one is interested in the age difference

AGE OF GROOM− AGE OF BRIDE

and we let M denote the mean age difference for *all* marriages in in Cumberland County, Pennsylvania in 1993.

(a) Make a best guess at the mean age difference M.

(b) Make a guess at the 10th percentile M_{10}. (This is a value such that you think the chance that the mean age difference M is smaller than this value is 10 percent.)

(c) Using the Minitab program "normal_s" to find the normal curve which matches your two prior statements in (a) and (b).

(d) Use Minitab to find the 90th percentile, M_{90}, of the normal curve you found in part (c). (This is the value that you're pretty confident that M is smaller than this value.) Does it seem to match your prior beliefs?

(e) If the value found in part (d) does not seem right, redo parts (a), (b), (c), (d) until you find a normal curve which seems to be a reasonable match to your opinion about the mean age difference.

Activity 19-11: Average Sentence Length of Shakespeare's Plays

Suppose that you are doing a study of the plays written by William Shakespeare and you're interested in his writing style. Specifically, suppose you are interested in estimating the mean sentence length M (measured by number of words) of all of the sentences in all of the plays written by Shakespeare.

(a) Make an intelligent guess, call it m_0, at the average sentence length of all the sentences in the plays of Shakespeare.

(b) Make a guess at M_{90}, the value such that you are 90 percent confident that the average sentence length is smaller than that value.

(c) Find the normal curve which matches your prior statements in parts (a) and (b).

(d) Use your normal curve to find the probability that the average sentence length exceeds 10 letters.

Activity 19-12: Marriage Ages (Continued)

In Activity 2-8, the ages of the bride and groom for some marriage licenses in Cumberland County, Pennsylvania was recorded.

(a) For each couple, compute the difference between the groom's age and the bride's age (AGE OF GROOM− AGE OF BRIDE).

(b) Compute the mean and standard deviation of the sample of differences found in (a).

(c) In Activity 19-10, you constructed a normal prior which reflected your opinion about the mean difference in marriage ages M. Use this normal prior and the information in your sample from part (b) to find the posterior normal curve of M.

(d) Use the posterior curve to find the probability that the age of the groom is typically greater than the age of the bride. (That is, find the probability that M is positive.)

Activity 19-13: Average Sentence Length of Shakespeare (Continued.)

To learn more about the average sentence length M of Shakespeare plays, twenty randomly selected pages were selected from *The Taming of the Shrew*. The length (in words) of the first complete sentence on each page was recorded — the lengths of the 20 sentences are displayed below.

15	6	8	13	34	23	6	23	18	4
2	15	6	4	5	17	38	17	9	7

The mean and standard deviation of these lengths are $\bar{x} = 13.50$ and $s = 9.99$, respectively.

(a) Use your normal prior found in Activity 19-11 and the data above to find the posterior normal curve for the mean sentence length M.

(b) Compare the prior and posterior normal curves. Specifically, is the mean of the posterior curve smaller or larger than the mean of the prior curve? Is the standard deviation of the posterior curve smaller or larger than the standard deviation of the prior curve? How has the data modified your prior opinion about the mean sentence length of Shakespeare plays.

(c) Find a 90% probability interval for the mean M.

Activity 19-14: Foot Lengths

Suppose you are interested in estimating the average foot length M (in centimeters) of all women at your school

(a) In Topic 2, you collected the foot lengths for the students in your class. Graph this data using a dotplot or a stemplot. Is it reasonable to assume that the distribution of all foot lengths for the women at your school can be approximated by a normal curve?

(b) Assume that the answer to part (a) is "yes". If your prior opinion about the mean foot length M is represented by a uniform curve, find the posterior normal curve for the mean foot length.

(c) Find a 90% probability interval for the mean length.

(d) Find the probability that the mean foot length is between 20 and 25 centimeters. [CHECK IF THIS IS REASONABLE]

Activity 19-15: What is the Average Age of a Newborn? (Continued.)

In Activity 19-11, we considered the problem of learning about the mean weight M of all babies born in a local hospital. The observed weights (in ounces) for a sample of 31 babies are shown below.

147	102	107	90	126	105	143	123	117	126
132	110	121	87	110	125	129	114	124	102
97	123	126	118	136	121	133			

(a) Find the mean and standard deviation of the sample of weights.

(b) Suppose that you have little prior knowledge about the average weight of a newborn and so you assign the mean M a uniform prior. Find the mean and standard deviation of the posterior normal curve for the mean weight M.

(c) Find a 90% probability interval for the mean weight.

(d) Find the probability that the mean newborn weight exceeds 120 ounces.

Activity 19-16: Boston Marathon Times

How long does it take to run the Boston Marathon (a famous cross-country footrace of 26 miles, 385 yards)? Below are the times (rounded to the nearest minute) for a random sample of 25 women runners who participated in the 1998 race.

229	233	221	211	213	241	228	278	200	243
240	256	233	238	236	200	250	291	218	217
260	228	236	225	221					

Suppose we're interested in using this data to estimate the mean running time for all women participants in the 1998 Boston Marathon — call this mean time M.

(a) Assuming a uniform prior curve for M, find the posterior normal curve.

(b) Find the probability that the average running time for the women participants exceeds 4 hours.

(c) Using the data above, find the proportion of women whose running time exceeded 4 hours. This proportion is approximately the probability that a randomly selected women will complete the race in over 4 hours.

(d) Explain why the answers to parts (b) and (c) are different.

WRAP-UP

In this topic, we discussed learning about a mean M of a population of measurements using continuous models. We first discussed computing probabilities and percentiles for a normal curve. Then we described two methods for learning about the mean M. One possibility is that we have little prior information about the location of M and a uniform curve is used as the prior. In that case, we showed that the posterior curve for M has is normal-shaped with mean $\bar{x}$ and standard deviation $h/\sqrt{n}$. We learn about the location of the population mean by summarizing this posterior curve. In other situations, we may have some knowledge about the location of the population mean. In that case, a normal prior was used to represent our prior knowledge, and the posterior curve for M again has a normal form with mean and standard deviation which depend on our prior information and the observed data.

Topic 20: Designing Experiments

Introduction

You have studied the idea of random sampling as a fundamental principle by which to gather information about a population. Since then you have discovered formal techniques of statistical inference which enable you to draw conclusions about a population based on analysis of a sample. With this topic you will begin to study another technique of data collection for use when the goal is not to describe a population but to investigate the effects that a variable has on other variables. You will investigate the need for controlled experiments and discover some fundamental principles which govern the design and implementation of such experiments.

PRELIMINARIES

1. If 95% of the participants in a large SAT coaching program improve their SAT scores after attending the program, would you conclude that the coaching was responsible for the improvement?

2. If recovering heart attack patients who own a pet tend to survive longer than recovering heart attack patients who don't own a pet, would you conclude that owning a pet has therapeutic benefits for heart attack patients?

3. Do you think that eating SmartFood popcorn really makes a person smarter?

4. Take a guess as to the mean IQ of a sample of people who claim to have had intense experiences with unidentified flying objects (UFO's).

5. Do you suspect that people who claim to have had intense experiences with UFO's tend to have higher or lower IQ's than other people, or do you think that there is no difference in their IQ's on the average?

6. If high school students who study a foreign language tend to perform better on the verbal portion of the SAT than students who do not study a foreign language, does that establish that studying a foreign language improves one's verbal skills?

7. If states which have capital punishment tend to have lower homicide rates than states without capital punishment, would you attribute that difference to the deterrence effect of the death penalty?

8. If states which have capital punishment have similar homicide rates to states without capital punishment, would you conclude that the death penalty has no deterrence effect?

IN-CLASS ACTIVITIES

Activity 20-1: SAT Coaching

Suppose that you want to study whether an SAT coaching program actually helps students to score higher on the SAT's, so you gather data on a random sample of students who have attended the program. Suppose you find that 95% of the sample scored higher on the SAT's after attending the program than before attending the program. Moreover, suppose you calculate that the sample mean of the improvements in SAT scores was a substantial 120 points.

(a) Explain why you can not legitimately conclude that the SAT coaching program caused these students to improve on the test. Suggest some other explanations for their improvement.

The SAT coaching study illustrates the need for a controlled experiment to allow one to draw meaningful conclusions about one variable causing another to respond in a certain way. The fundamental principle of experimental design is control. An experimenter tries to control for possible effects of extraneous variables so that any differences observed in one variable of interest can be attributed directly to the other variable of interest.

The variable whose effect one wants to study is called the explanatory variable (or independent variable), while the variable which one suspects to be affected is known as the response variable (or dependent variable).

The counterpart to a controlled experiment is an observational study in which one passively records information without actively intervening in the process. As you found when you discovered the distinction between correlation and causation in Topic 7 (think of televisions and life expectancies), one can not infer causal relationships from an observational study since the possible effects of confounding variables are not controlled.

(b) Identify the explanatory and response variables in the SAT coaching study.

(c) Is the SAT coaching study as described above a controlled experiment or an observational study? Explain.

Experimenters use many techniques in their efforts to establish control. One fundamental principle of control is comparison. One important flaw in that SAT coaching study is that it lacks a control group with which to compare the results of the group that attends the coaching program.

Activity 20-2: Pet Therapy

Suppose that you want to study whether or not pets provide a therapeutic benefit for their owners. Specifically, you decide to investigate whether heart attack patients who own a pet tend to recover more often than those who do not. You randomly select a sample of heart attack patients from a large hospital and follow them for one year. You then compare the sample proportions who have survived and find that 92% of those with pets are still alive while only 64% of those without pets have survived.

(a) Identify the explanatory and response variables in this study.

(b) Is this study a controlled experiment or an observational study? Explain.

This pet therapy study points out that analyzing a comparison group does not guarantee that one will be able to draw cause-and-effect conclusions. A critical flaw in the design of the pet therapy study is that subjects naturally decide for themselves whether or not to own a pet. Thus, those who opt to own a pet may do so because they are in better health and therefore more likely to survive even if the pet is of no benefit at all.

Experimenters try to assign subjects to groups in such a way that confounding variables tend to balance out between the two groups. A second principle of control which provides a simple but effective way to achieve this is randomization. By randomly assigning subjects to different treatment groups, experimenters ensure that hidden confounding variables will balance out between/among the groups in the long run.

Activity 20-3: Vitamin C and Cold Resistance

Suppose that you want to design a study to examine whether taking regular doses of vitamin C increases one's resistance to catching the common cold. You form two groups — subjects in one group receive regular doses of vitamin C while those in the other (control) group are not given vitamin C. You assign subjects at random to one of these two groups. Suppose that you then monitor the health of these subjects over the course of a winter and find that 46% of the vitamin C group resisted a cold while only 14% of the control group resisted a cold.

(a) Identify the explanatory and response variables in this study.

(b) Is this vitamin C study a controlled experiment or an observational study? Explain.

(c) Identify some of the confounding variables which affect individuals' resistance to a cold that the random assignment of subjects should balance out.

The design of the vitamin C study could be improved in one respect. There is one subtle confounding variable that the randomization can not balance out because this variable is a direct result of the group to which a subject is assigned. The very fact that subjects in the vitamin C group realize that they are being given something that researchers expect to improve their health may cause them to remain healthier than subjects who are not given any treatment. This phenomenon has been detected in many circumstances and is known as the placebo effect.

Experimenters control for this confounding variable by administering a placebo ("sugar pill") to those subjects in the control group. This third principle of control is called blindness since the subjects are naturally not told whether they receive vitamin C or the placebo. When possible, experiments should be double-blind in that the person responsible for evaluating the subjects should also be unaware of which subjects receive which treatment. In this way the evaluator's judgment is not influenced (consciously or sub-consciously) by any hidden biases.

Activity 20-4: AZT and Pregnancy (cont.)

Recall from Activity 9-3 that a study was performed to investigate whether the drug AZT reduces the risk of an HIV-positive pregnant woman giving birth to an HIV-positive baby.

(a) Identify the explanatory and response variables in this study. Are these measurement or categorical variables? If they are categorical, are they also binary?

(b) Explain how the study could be designed to make use of the principle of comparison.

(c) Explain how the study could be designed to incorporate the principle of randomization.

(d) Explain how the study could be designed to take into account the principle of blindness.

In the context of medical studies, controlled experiments such as the AZT study which randomly assign subjects to a treatment are called clinical trials. Many medical issues do not lend themselves to such studies, however, so researchers must resort to observational studies. Three types of observational studies are often used in medical contexts:

- case-control studies, in which one starts with samples of subjects who do and who do not have the disease and then looks back into their histories to see which have used and which have not used a certain treatment or condition
- cohort studies, in which one starts with samples of subjects who do and who do not use the treatment or condition and then follows them into the future to see which do and which do not develop the disease
- cross-sectional studies, in which one simply takes a sample of subjects and classifies them according to both variables (have disease or not, use treatment or not)

While controlled experiments are the only way to establish a causal relationship between variables, observational studies can also provide researchers with important information.

Activity 20-5: Smoking and Lung Cancer

(a) Explain how one could in principle design a controlled experiment to assess whether smoking causes lung cancer. Also indicate why conducting such an experiment would be morally unethical.

(b) What kind of observational study are researchers conducting if they take a random sample of 500 people who suffer from lung cancer and a random sample of 500 people who do not suffer from lung cancer and then compare the proportions of smokers in each group?

(c) What kind of observational study are researchers conducting if they take a random sample of 500 smokers and a random sample of 500 non-smokers, follow them for 20 years, and compare the proportions who develop lung cancer in each group?

Activity 20-6: SmartFood Popcorn

Describe (in as much detail as possible) how you might design an experiment to assess whether eating SmartFood brand popcorn really makes people smarter. You will have to describe how you would measure "smartness" as part of the experimental design. Try to adhere to the three principles of control that we listed above.

HOMEWORK ACTIVITIES

Activity 20-7: UFO Sighters' Personalities

In a 1993 study researchers took a sample of people who claim to have had an intense experience with an unidentified flying object (UFO) and a sample of people who do not claim to have had such an experience. They then compared the two groups on a wide variety of variables, including IQ. The sample mean IQ of the UFO group was 101.6 and that of the control group was 100.6. Is this study a controlled experiment or an observational study? If it is an observational study, what kind is it? Explain your answers.

Activity 20-8: Mozart Music

Researchers in a 1993 study investigated the effect of listening to Mozart music before taking an IQ test. Subjects were randomly assigned to one of three groups and would either listen to Mozart music, be told to relax, or be given no instructions. The sample mean IQ in the Mozart group was 119, in the relax group was 111, and in the silent group was 110.

(a) Is this study a controlled experiment or an observational study? If it is an observational study, what kind is it? Explain your answers.

(b) Identify the explanatory variable and the response variable.

(c) Classify each variable as a categorical or measurement variable.

Activity 20-9: Language Skills

Students who study a foreign language in high school tend to perform better on the verbal portion of the SAT than students who do not study a foreign language. Can you conclude from these studies that the study of a foreign language causes students to improve their verbal skills? Explain.

Activity 20-10: Capital Punishment

Suppose that you want to study whether the death penalty acts as a deterrent against homicide, so you compare the homicide rates between states that have the death penalty and states that do not.

(a) Is this a controlled experiment? Explain.

(b) If you find a large difference in the homicide rates between these two types of states, can you attribute that difference to the deterrence effect of the death penalty? Explain.

(c) If you find no difference in the homicide rates between the two types of states, can you conclude that the death penalty has no deterrence effect? Explain.

Activity 20-11: Literature for Parolees

In a recent study 32 convicts were given a course in great works of literature. To be accepted for the program the convicts had to be literate and to convince a judge of their intention to reform. After thirty months of parole only six of these 32 had committed another crime. This group's performance was compared against a similar group of 40 parolees who were not given the literature course; 18 of these 40 had committed a new crime after thirty months.

(a) What proportion of the literature group committed a crime within thirty months of release? What proportion of this group did not commit a crime?

(b) What proportion of the control group committed a crime within thirty months of release? What proportion of this group did not commit a crime?

(c) Construct a segmented bar graph to compare these conditional distributions of crime commission between the literature and control groups.

(d) Which fundamental principles of control does this experiment lack? Comment on how this lack hinders the conclusion of a cause-and-effect relationship in this case.

Activity 20-12: Gun Control Legislation

(a) Suppose that a nation passes a strict gun control measure and finds five years later that the national homicide rate has increased. Can you conclude that the passage of the gun control measure caused the homicide rate to increase? Explain.

(b) Would your answer to (a) differ if the homicide rate had actually decreased and you were asked to conclude that the passage of the gun control measure caused the homicide rate to decrease? Explain.

Activity 20-13: Baldness and Heart Disease (cont.)

Reconsider the study mentioned in Activity 9-14, where researchers took a sample of male heart attack patients and a sample of men who have not suffered a heart attack and compared baldness ratings between the two groups. Their goal was to determine if baldness has an effect on one's likelihood of experiencing a heart attack.

(a) Identify the explanatory and response variables.

(b) Is this study a controlled experiment or an observational study? If it is an observational study, what kind is it? Explain your answers.

Activity 20-14: Pet Therapy (cont.)

Reconsider the study described in Activity 20-2. Explain how you could (in principle) design a controlled experiment to investigate the proposition that owning a pet has therapeutic benefits for heart attack patients.

Activity 20-15: Assessing Calculus Reform

Many colleges and universities in the 1990's have developed "calculus reform" courses which substantially alter the way that calculus is taught. The goal is that the reform courses help students to understand fundamental calculus concepts better than traditionally taught courses do.

(a) If you simply compare scores on a standardized calculus test between students in colleges that teach a reform course and students in colleges that teach a traditional course, would you be able to conclude that any differences you might find are attributable to the teaching style (reform or traditional)?

(b) Describe how you might design an experiment to assess whether this goal is being met. Be sure that your experimental design incorporates aspects of comparison, randomization, blindness, and double blindness. Also explain the need for each of these aspects in your design.

Activity 20-16: Subliminal Messages

Explain in as much detail as possible how you might design and conduct an experiment to assess whether listening to audiotapes with recorded subliminal messages actually helps people to lose weight.

Activity 20-17: Experiments of Personal Interest

Think of a situation in which you would be interested in determining whether a cause-and-effect relationship exists between two variables. Describe in as much detail as possible how you might design and conduct a controlled experiment to investigate the situation.

WRAP-UP

This topic has introduced you to principles of designing controlled experiments. You have explored the limitations of observational studies with regard to establishing causal relationships between variables. You have also learned that control is the guiding principle of experimental design and discovered the principles of comparison, randomization, and blindness as specific techniques for achieving control.

In the next topic you will investigate how one can use inferential procedures based on Bayes' rule to draw conclusions about the results of experiments. We focus on inferential methods for comparing two proportions.

Topic 21: Learning About Two Proportions

Introduction

In this topic, we consider the problem of comparing two population proportions. Suppose that one is interested in comparing the drinking habits of undergraduates who belong to fraternities or sororities (the Greeks) with the habits of students who don't belong in these organizations. There are two populations of interest — the undergraduates at the college who belong to fraternities or sororities and the undergraduates who are non-Greeks. Suppose that a "drinker" is defined as one who drinks an alcoholic beverage at least four days in every week. Let p_G denote the proportion of Greeks who drink and p_N denote the proportion of non-Greeks who drink. It is of interest to compare the population proportions p_G and p_N.

To learn about these proportions, a survey is taken. Suppose that a random sample of 20 Greeks are sampled and 12 are drinkers; a second sample of 20 non-Greeks are taken and only 8 are drinkers. Is this sufficient evidence that the population proportions are nonequal?

As in Topics 15-19, we will learn about these two proportions by means of probability and Bayes' rule. The situation becomes a bit more complicated, however, since there are two unknown parameters.

- A **model** in this case consists of values for both proportions p_G and p_N. We can represent a collection of models by means of a two-way table. Rows of the table correspond to values of p_G, columns correspond to values of p_N; each cell of the table corresponds to a single combination of values of p_G and p_N.

- A **prior** is hard to specify since one has to specify a probability for each combination of values of the two proportions. In this topic, we'll focus on uniform sets of probabilities that are easy to specify.

- **Posterior probabilities** are computed using the same "multiply, sum, divide" recipe that we

used in Topics 16 and 17. The calculations are tedious in this case primarily since there are many terms to multiply in the likelihood.

- **Inference.** To compare the two proportions, we will focus on the **difference** $d = p_G - p_N$. If the value of $d = 0$, then Greeks and non-Greeks have the same drinking habits; if $d > 0$, Greeks generally drink more than non-Greeks. We will focus on two issues: (1) is there support for the statement that the proportions are equal and (2) if the proportions are indeed different, what is the size of the difference $d = p_G - p_N$.

PRELIMINARIES

1. Suppose a drug, called drug A, will cure 50% of patients who have a specific illness. Suppose that a second drug, drug B, is equally effective. What percentage of patients with the same illness will be cured using drug B?

2. If drug B is more effective than drug A in curing patients with the specific illness, what would be a plausible cure rate for drug B?

3. Suppose that 10 patients are randomly assigned to receive drug A while another 10 are randomly assigned to receive drug B. If 7 patients in the drug A group and 5 patients in the drug B group recover, would you be fairly convinced that drug A is superior to drug B?

4. Consider the population of college students who regularly exercise. Make a guess at the proportion of these students who are on a diet.

5. Consider the population of college students who *don't* regularly exercise. Make a guess at the proportion of these students who are on a diet.

6. Take a guess concerning the proportion of Americans who would respond in the affirmative to the question, "Do you think the United States should forbid public speeches in favor of communism?"

7. Suppose that one group of subjects is asked the question in 6, while another group is asked, "Do you think the United States should allow public speeches in favor of communism?". Would you expect these groups to differ in terms of their opposition to communist speeches? If so, which group do you think would more likely oppose communist speeches?

IN-CLASS ACTIVITIES

Activity 21-1: Is the New Drug Better?

Suppose a researcher at a pharmaceutical company is interested in determining the relative effectiveness of two drugs to treat a specific illness. We will refer to the two drugs as drug A and drug B. We measure the effectiveness of a drug by the proportion of people with the illness who will be cured within a specific time interval. Let p_A denote the proportion of all people who will be cured using drug A and p_B the corresponding proportion for drug B.

Some questions of interest to the researcher are

- Are the two drugs equally effective?
- What's the probability that drug A is more effective than drug B?
- If the two drugs are not equally effective, how much better is the more effective drug?

We will answer these questions using the same general framework that was used to learn about a single proportion. First, we define a *model* which gives the cure rates for both drugs. Next, we will describe how one constructs a prior distribution on a collection of models. Last, given data, we will use Bayes' rule to find posterior probabilities for these models. The questions of the researcher can be answered using the posterior distribution.

A model

There are two proportions that are unknown — the cure rate for drug A and the cure rate for drug B. These numbers represent the proportions of all people with the illness who are cured by the two drugs. A model gives values for both of these proportions. For example, one model says that $p_A = .3$ and $p_B = .5$; another model says that both p_A and p_B are equal to .4. To make the following discussion as easy as possible, we will assume that each proportion can take the three values .3, .5, .7. This means that the cure rate for each drug can be either 30%, 50% or 70%. If each proportion has three values, there are nine models — for each cure rate for drug A, say .3, there are three possible cure rates for drug B.

A convenient way of listing all possible models is by the following two-way table.

		Cure rate 2 (p_B)		
		.3	.5	.7
Cure rate 1 (p_A)	.3			
	.5			
	.7			

Each cell of the table corresponds to a model. For example, the top-left cell corresponds to the model where $p_A = .3$ and $p_B = .3$ — this is the model which says that both drugs have cure rates of 30%.

The prior

Next we assign probabilities to the different pairs of cure rates (p_A, p_B). By now you understand that it can be difficult to specify a prior distribution for a single proportion p. It is even more difficult to construct a set of prior probabilities for two proportions which reflects your beliefs. So we will focus on the use of two sets of prior probabilities which reflect vague beliefs about the values of the cure rates for the two drugs.

A flat prior

Suppose that you think that the nine models, represented by the nine cells of the table, are equally likely. This means that you are uncertain about the cure rates for the two drugs. Then you would assign each model a prior probability of 1/9. These prior probabilities are displayed in the below table. We call this a *flat prior*, since the probabilities are spread out flatly over the possible models.

The flat prior

		Cure rate 2 (p_B)		
		.3	.5	.7
Cure rate 1 (p_A)	.3	1/9	1/9	1/9
	.5	1/9	1/9	1/9
	.7	1/9	1/9	1/9

We are interested in comparing the two cure rates. If we use a flat prior, what is the probability the two rates are equal? We find this by adding up the probabilities over the models along the diagonal where $p_A = p_B$. (These are marked in bold in the table below.) So our prior probability that the rates are equal is $1/9 + 1/9 + 1/9 = 1/3$.

		Cure rate 2 (p_B)		
		.3	.5	.7
Cure rate 1 (p_A)	.3	**1/9**	1/9	1/9
	.5	1/9	**1/9**	1/9
	.7	1/9	1/9	**1/9**

What is the probability that drug A has a larger cure rate? To answer this, we focus on the models where the cure rate p_A is larger than p_B (marked in bold below). If we sum the probabilities of the models in this lower diagonal region, we obtain that the probability that drug A is better is $1/9 + 1/9 + 1/9 = 1/3$.

		Cure rate 2 (p_B)		
		.3	.5	.7
Cure rate 1 (p_A)	.3	1/9	1/9	1/9
	.5	**1/9**	1/9	1/9
	.7	**1/9**	**1/9**	1/9

Let's summarize what this prior distribution says. Either the two cure rates are equal, drug A has a larger cure rate, or drug B has a larger rate. These three possibilities have the following prior probabilities:

Model	Probability
the two cure rates are equal	1/3
drug A has a higher cure rate	1/3
drug B has a higher cure rate	1/3

We computed the probabilities of "equal cure rates" and "drug A is better" above. The probability that drug B has a higher cure rate is 1/3 since all three probabilities in the table must add up to one.

Likelihoods

To learn about the two drugs, patients with the illness will be selected. Some patients will be assigned to drug A and the remainder to drug B. We will observe if each patient is cured within the time interval.

To update our model probabilities, we compute likelihoods, and compute posterior probabilities by Bayes' rule. We'll illustrate this computation first for a very small dataset.

Suppose that two people with the illness are in the experiment — one person takes drug A and the other person takes drug B. The result of this experiment is

$$\{\text{person with drug A is cured, person with drug B is not cured}\}$$

For this data, we compute likelihoods. For each one of the nine models (p_A, p_B), we compute the probability of this outcome:

$$\begin{aligned} \text{LIKELIHOOD} &= \text{Prob}(\{\text{person with drug A is cured, person with drug B is not cured}\}) \\ &= \text{Prob}(\text{person with drug A is cured}) \times \text{Prob}(\text{person with drug B is not cured}). \end{aligned}$$

For example, suppose that $p_A = .3$ and $p_B = .3$ (each drug has a 30% cure rate). The probability that the person using drug A is cured is given by the proportion $p_A = .3$ and the probability that a person using drug B is not cured is given by one minus the proportion $1 - p_B = 1 - .3$. So the likelihood is given by

$$\begin{aligned} \text{LIKELIHOOD} &= .3 \times (1 - .3) \\ &= .21. \end{aligned}$$

We put this value in the likelihood table below — we put it in the cell corresponding to the row $p_A = .3$ and the column $p_B = .3$.

The likelihoods

		Cure rate 2 (p_B)		
		.3	.5	.7
Cure rate 1 (p_A)	.3	.21		
	.5			
	.7			

(a) Compute the likelihoods for the remaining eight cells of the table.

The posterior

Suppose we use a flat prior for the models. We compute the posterior probabilities in the usual way:

- for each model (cell of the table), we multiply the prior probability by the likelihood
- we add up all of the products
- we divide each product by the sum

After our data {person with drug A is cured, person with drug B is not cured}, we obtain the following model probabilities:

The posterior assuming a flat prior

		Cure rate 2 (p_B)		
		.3	.5	.7
Cure rate 1 (p_A)	.3	.093	.067	.040
	.5	.156	.111	.067
	.7	.218	.156	.093

These probabilities reflect our new opinion about the cure rates after seeing this data. How have our probabilities changed? Using the above table of posterior probabilities, find

(b) the probability the cure rates are equal

(c) the probability drug A is better

(d) the probability drug B is better

(e) Put the values you computed in (b), (c), (d) in the below table.

Model	Prior Prob.	Post. Prob
the two cure rates are equal	.33	
drug A has a higher cure rate	.33	
drug B has a higher cure rate	.33	

Look at the table and describe how your opinion about the effectiveness of the two drugs has changed.

Activity 21-2: Is the New Drug Better? (cont)

Let's consider a different set of prior probabilities that reflects a different opinion about the relative effectiveness of the two drugs. Suppose that that you think it is equally likely that the drugs have the same cure rates or that the drugs have different rates. In other words, your probability that the proportions p_A and p_B are equal is 1/2, and so the probability that the proportions is different is $1 - 1/2 = 1/2$. Otherwise, you have no knowledge about the actual effectiveness of the two drugs. This prior information can be represented by the following probabilities:

The testing prior

		Cure rate 2 (p_B)		
		.3	.5	.7
Cure rate 1 (p_A)	.3	2/12	1/12	1/12
	.5	1/12	2/12	1/12
	.7	1/12	1/12	2/12

Note that each model where the proportions p_A and p_B are equal is given a probability of 2/12. The models off of the diagonal where the proportions are unequal are each assigned a probability of 1/12. We refer to this prior as a "testing prior" since it is useful for testing the statement that the proportions are equal.

(a) As a check, use the prior distribution above to fill in the following table:

Model	Probability
the two cure rates are equal	
drug A has a higher cure rate	
drug B has a higher cure rate	

Suppose the researcher is interested if the two proportions are equal. He uses the above testing prior on the nine models where the probability that $p_A = p_B$ is equal to .5. After collecting some data, he will decide that the proportions are not equal if the posterior probability that $p_A = p_B$ is sufficiently small, say under .2.

Below I have listed different datasets and the posterior probabilities using the testing prior. For each part,

- find the posterior probability that the proportions are equal
- decide if $p_A = p_B$

(b) Data: 4 of 8 cured using drug A; 2 of 8 cured using drug B

		Cure rate 2 (p_B)		
		.3	.5	.7
Cure rate 1 (p_A)	.3	.270	.050	.005
	.5	.272	.200	.009
	.7	.135	.050	.009

(c) Data: 8 of 16 cured using drug A; 4 of 16 cured using drug B

		Cure rate 2 (p_B)		
		.3	.5	.7
Cure rate 1 (p_A)	.3	.248	.016	.000
	.5	.480	.131	.001
	.7	.119	.016	.000

(d) Data: 16 of 32 cured using drug A; 8 of 32 cured using drug B

		Cure rate 2 (p_B)		
		.3	.5	.7
Cure rate 1 (p_A)	.3	.100	.001	.000
	.5	.817	.030	.067
	.7	.050	.001	.000

(d) Note that the first dataset had 16 patients, the second dataset had 32, and the third dataset had 64. What happens to the posterior probability of equality as you take a larger sample?

Activity 21-3: Exercising and Dieting.

A group of students were interested in studying the relationship between exercising and dieting. They hypothesized that college students who exercise were more likely to diet than students who didn't exercise. A phone survey of 114 randomly selected students was performed. Each student was asked two questions:

- How many times per week do you exercise?
- Do you consider yourself on a diet?

Students who exercised three or more times a week were classified as "exercisers". The remaining students were called "nonexercisers". For each group of students, the number who were dieting and who were not dieting was recorded. The data is summarized in the table below.

	Dieting? Yes	No
Exercisers	44	29
Nonexercisers	19	22

This is a problem in comparing two proportions. Suppose that the group of all college students is divided into two groups — those who exercise and those who don't exercise. Let

p_E = the proportion of exercisers who diet

p_N = the proportion of nonexercisers who diet

The students in the study believe that the proportion of dieters among exercisers is greater than the proportion of dieters among nonexercisers — that is, $p_E > p_N$. They would like to learn about the size of the difference in proportions

$$d = \text{difference in proportions} = p_E - p_N.$$

We will use the same approach as in our two drugs activity. We will construct a collection of models for the two proportions — these models will be displayed using a two-way table. Our prior probabilities will be flat over all possible models. Posterior probabilities will be computed from the above data and we will learn about the size of the difference in proportions by interpreting this table of probabilities.

First, what does our prior look like? To make this problem more realistic, we'll let each proportion take on the nine values .1, .2, ..., .9. A model is a pair of values for the two proportions p_E and p_N and we display all of the models by the following table with nine rows and nine columns.

p_E \ p_N	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1									
.2									
.3									
.4									
.5									
.6									
.7									
.8									
.9									

The students want to learn about the difference in proportions $d = p_E - p_N$. Each model in the above two-way table corresponds to a particular value of d. For example, if $p_E = .4$ and $p_N = .2$, then $d = .4 - .2 = .2$. Conversely, for a specified value of d there are many models.

In the grid below I have put an 'X' in the cells where the two proportions are equal. These correspond to the cells where the difference $d = 0$.

p_E \ p_N	.1	.2	.3	.4	.5	.6	.7	.8	.9
.1	X								
.2		X							
.3			X						
.4				X					
.5					X				
.6						X			
.7							X		
.8								X	
.9									X

In the above table

(a) Put a "+" in the cells where the difference in probabilities $d = .1$. (To get you started, $p_E = .2$ and $p_N = .1$ is one of these cells.)

(b) Put a "#" in all cells where the difference $d = .2$.

(c) Put a "$\otimes$" in all cells where the difference $d = .3$.

The prior

We assume that each of the 81 models is equally likely before you see any data and so each cell of the table is assigned a probability of 1/81.

The posterior

Using the Minitab program 'pp_disc', I compute posterior probabilities. When I run this program by typing

```
exec 'pp_disc'
```

I have to input

- the *lo* and *hi* values for each proportion (.1 and .9)
- the *number of models* for each proportion (9)

- the *data*: 44 successes and 29 failures in the first sample; 19 successes and 22 failures in the second sample (we're calling a "success" someone who is dieting)

The program p_disc outputs the following table of posterior probabilities for all models. For ease of reading, I have rounded each probability to three decimal places — a probability of 0 actually is .000 rounded to three decimal places.

		p_N								
		.1	.2	.3	.4	.5	.6	.7	.8	.9
p_E	.1	0	0	0	0	0	0	0	0	0
	.2	0	0	0	0	0	0	0	0	0
	.3	0	0	0	0	0	0	0	0	0
	.4	0	0	0	.001	.001	0	0	0	0
	.5	0	0	.007	.055	.070	.016	.001	0	0
	.6	0	0	.033	.261	.328	.077	.003	0	0
	.7	0	0	.007	.055	.069	.016	.001	0	0
	.8	0	0	0	0	0	0	0	0	0
	.9	0	0	0	0	0	0	0	0	0

(d) What is the most likely model (values of p_E and p_N)?

(e) What is the probability that the proportions are equal? (You sum the probabilities over the diagonal cells where $p_E = p_N$.)

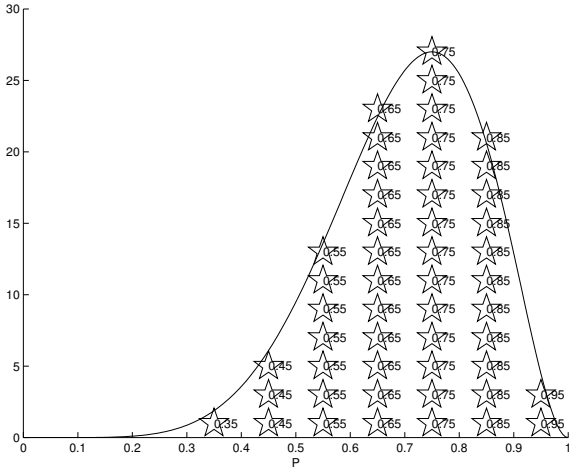
(f) What is the probability that the proportion p_N is larger?

(g) Find the probability that the difference in probabilities $d = p_E - p_N = -.8, -.7, ..., .8$. Put your answers in the table below. (To find the probability that $d = .2$, say, you have to find all of the cells in the probability table where $d = .2$, and then add up the probabilities of all these cells to get Prob($d = .2$). You repeat this process for other values of d.)

$d = p_E - p_N$	PROBABILITY
$-.8$	
$-.7$	
$-.6$	
$-.5$	
$-.4$	
$-.3$	
$-.2$	
$-.1$	
0	
.1	
.2	
.3	
.4	
.5	
.6	
.7	
.8	

(h) Find the three most likely values of the difference d. Put the values and the probabilities below.

The probability that d is between _____ and _____ is _____.



A beta (7, 3) curve and some representative stars.

Summarizing a beta curve using simulated values.

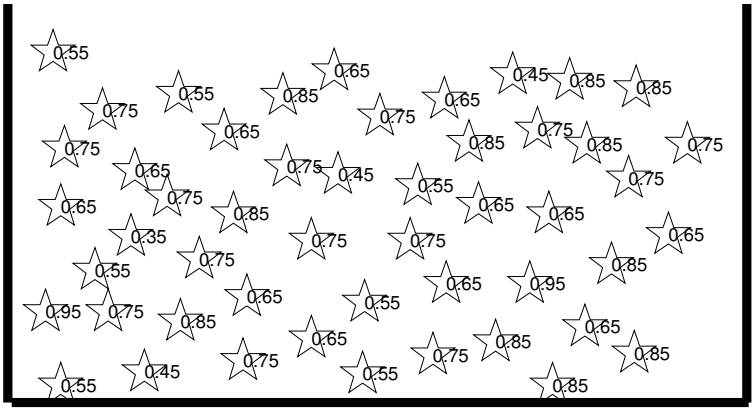
Before we discuss comparing two proportions using continuous models, we will introduce the **simulation** method of summarizing a beta curve for a proportion. Suppose that I am interested in the proportion p of households in my community that has access to the internet. My beliefs about this proportion are modeled using a beta(7, 3) curve. I am interested in summarizing this curve. What is my best guess at the proportion? What is my probability that p is smaller than .5? What is a 50% probability interval for this proportion?

To introduce the use of simulation, suppose we divide the range of values of the proportion into 10 subintervals. For each subinterval, we represent the height of the beta curve by a stack of stars. (See the figure below.) The fraction of stars at a given location is approximately equal to the probability that the proportion p is in the given subinterval. Note from the figure that each star is labeled by the proportion value it represents. So we can think of the beta curve as a collection of proportion values (stars), with one .35, three .45's, seven .55's, and so on.

Now suppose that we collect all of the stars that are pictured and place them into a box shown below. We select random values of the proportion p by randomly selecting stars with replacement from the box. Here "with replacement" means that after we select a star, say a one that is labeled .45, we place it back in the box before selecting the next one.

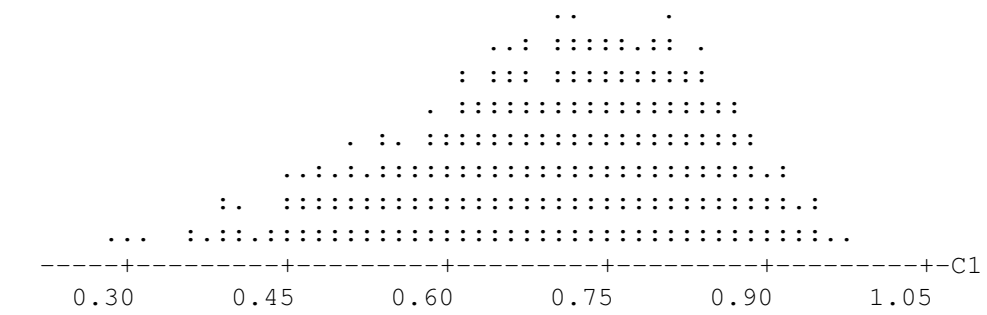
Suppose that we select a large number of stars from the box. The group of stars that are selected represents a **simulated random sample** from the beta(7, 3) curve. We use these random values of the proportion p to learn about the beta curve.

Actually, in practice, we use many more intervals than 10 to perform this simulation, but the basic idea behind the sampling is the same as described above.



A box of 50 stars from the beta(7, 3) curve.

I had Minitab simulate 1000 values from the beta(7, 3) curve. A dotplot of the simulated values is displayed below. Also, the output below shows some summary statistics from these 1000 simulated proportion values.



Variable	N	Mean	Median	TrMean	StDev	SE Mean
C1	1000	0.70003	0.71363	0.70375	0.13525	0.00428

Variable	Minimum	Maximum	Q1	Q3
C1	0.28537	0.97591	0.61203	0.80489

One can think of this simulated sample as 1000 typical proportion values from a beta(7, 3) curve. Note that the shape of the dotplot of the simulated proportions is approximately the same as the beta curve that was pictured earlier.

If we are interested in an **average proportion value** of the beta curve, we can simply compute an average of the simulated sample. For example, we see from the output that the sample mean of the simulated proportions is equal to .70003. This is approximately equal to the mean of the beta(7, 3) curve.

Suppose that we wish to find the probability that p is smaller than .5.

$$\mathrm{P}(p < .5).$$

We found this probability earlier by finding an area under the beta curve. Using the simulated sample, we find this probability by computing the fraction of proportion values that are smaller than .5. Using Minitab, we find that 96 out of the 1000 simulated values are smaller than .5. So the probability of interest is given by

$$\mathrm{P}(p < .5) = \frac{96}{1000} = .096.$$

Next, suppose we wish to find an interval which contains the middle 50% of the area under the beta(7, 3) curve. In other words, we want to find the 25% and 75% percentiles of the beta curve, or the quartiles of the curve. Using the simulation output, we compute this interval by finding the lower and upper quartiles of the sample of 1000. From the output, we see that the quartiles are given by .61203 and .80489. So the probability that the proportion is between .61203 and .80489 is 50%.

Activity 21-4: Employment Rate for Women (cont.)

In Activity 17-2, we were interested in the proportion p of adult women in the U.S. that are currently employed. A beta(18, 2) curve was used to represent the prior beliefs of one of the authors about the location of this proportion. Minitab was used to simulate the following 20 proportion values from this beta curve. A stemplot of the simulated values is also shown.

```
0.906   0.828   0.914   0.868   0.824   0.801   0.772   0.912   0.987
0.837   0.902   0.975   0.803   0.964   0.929   0.966   0.961   0.787
0.882   0.909

7 7
7 8
8 00
8 223
8
8 6
8 8
9 00011
9 2
9
9 6667
9 8
```

(a) Using these simulated values, approximate the probability that the proportion p is less than .9.

(b) Approximate the probability that the proportion is between .85 and .95.

(c) Use the simulated values to find the mean of the beta(18, 2) curve.

(d) Find an interval which covers the middle 50% of the probability of the beta(18, 2) curve.

Note: The above activity is a simple example to illustrate working with simulated proportions. In practice, we will simulate a large number, say 1000, of proportion values to get accurate summaries of the beta curve.

Using Beta Curves to Learn About Two Proportions.

Let us return to Activity 21-2 where we were interested in comparing the effectiveness of two drugs to treat a specific illness. There were two proportions of interest: p_A is the proportion of all people who will be cured using drug A, and p_B is the cure proportion for the people using drug B. In Activity 21-2, we assumed that these proportions could each take on the three possible values .3, .5, .7. Here we generalize this problem to the scenario where each proportion is continuous and can take on all possible values between 0 and 1.

Suppose that we know little about the effectiveness of either drug and so have no knowledge about the size of either curing proportion. Recall from Topic 17 that we represent little prior beliefs about a single proportion p by means of a uniform curve. This curve says that we believe, before looking at any data, that all proportion values between 0 and 1 are equally likely to be true.

We extend this idea to constructing a model for two proportions. If we know little about the curing proportion for drug A, then it is reasonable to assign a uniform curve for our prior for this proportion:

$$PRIOR(p_A) = 1.$$

Likewise, if we are unsure about the size of the drug B curing proportion, we assign p_B a uniform prior.

$$PRIOR(p_B) = 1.$$

Last, if our beliefs about the first proportion p_A are independent or unrelated to our beliefs about the second proportion p_B, we can multiply the above priors to get our prior for the two proportions:

$$PRIOR(p_A, p_B) = 1.$$

After our prior for the two proportions has been determined, we use Bayes' rule and our familiar POST = PRIOR $\times$ LIKELIHOOD recipe to determine the posterior curve. Suppose, as in Activity 21-2, that 4 out of 8 patients are cured that use drug A, and 2 of 8 are cured using drug B. The likelihood is the probability of getting this data if the true proportion values are p_A and p_B. Using the same thinking as in Activity 21-2, this probability is given by

$$\text{LIKELIHOOD} = p_A^4(1-p_A)^4 p_B^2(1-p_B)^6.$$

If we multiply the likelihood by the uniform curve prior, we obtain the posterior curve for the two proportions.

$$\text{POST} = \text{PRIOR} \times \text{LIKELIHOOD} = p_A^4(1-p_A)^4 p_B^2(1-p_B)^6.$$

Looking at the form of the posterior curve, we see that the posterior curve for the first proportion p_A has the form

$$p_A^4(1-p_A)^4,$$

which we recognize (from Topic 17) as a beta(5, 5) curve. Similarly, the second proportion has a posterior curve

$$p_B^2(1-p_B)^6,$$

which is a beta(3, 7) curve.

Let us state this result in general:

- We are interested in two proportions p_1, p_2, and our prior beliefs are described by means of a uniform curve.
- We take independent random samples from the two populations — we observe s_1 successes and f_1 failures in the first sample, and s_2 successes and f_2 failures in the second sample. (In this example, $s_1 = 4, f_1 = 4$ and $s_2 = 2, f_2 = 6$.)
- Then the posterior curve for p_1 will be beta($s_1 + 1, f_1 + 1$) and the posterior curve for p_2 will be beta($s_2 + 1, f_2 + 1$).

Activity 21-5: Pregnancy, AZT, and HIV (cont.)

Recall from Activity 9-3 that medical experimenters randomly assigned 164 pregnant, HIV-positive women to receive the drug AZT during pregnancy, while another 160 such women were randomly assigned to a control group which received a placebo. Of those in the AZT group, 13 had babies who tested HIV-positive, compared to 40 HIV-positive babies in the placebo group.

(a) Let p_{AZT} denote the proportion of all potential AZT-takers who would have HIV-positive babies and p_{PLAC} denote the proportion of all potential placebo-takers who would have HIV-positive babies.

The researchers are making the conjecture that AZT would prove beneficial for these patients. In terms of the population proportions, this means that (circle one)

(i) $p_{AZT} < p_{PLAC}$

(ii) $p_{AZT} = p_{PLAC}$

(iii) $p_{AZT} > p_{PLAC}$

(b) If we define a "success" as having a baby who is tested HIV-positive, what are the values of s_1, f_1, s_2, f_2? (These are the number of successes and failures in the two samples.)

(c) Suppose that our prior beliefs about each proportion are represented by a uniform curve. Then

(i) the posterior distribution for p_{AZT} is a beta curve with numbers ____ and ____ .

(ii) the posterior distribution for p_{PLAC} is a beta curve with numbers ____ and ____ .

Using Simulated Values to Learn About the Difference of Two Proportions.

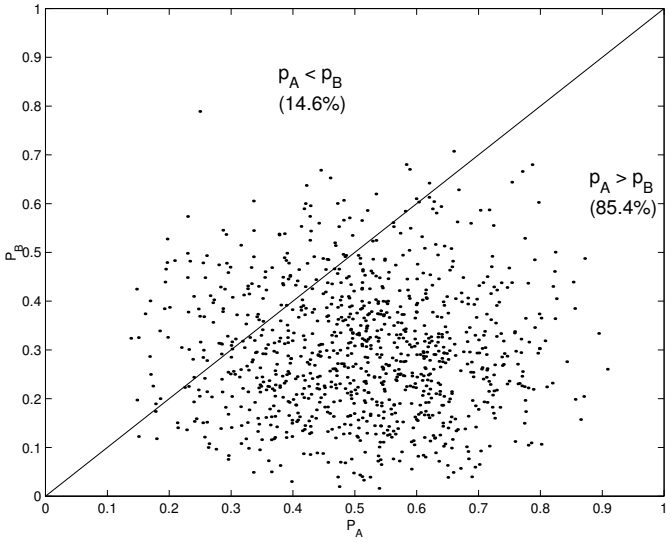
To summarize the posterior probabilities for the two proportions, we use the simulation method that was used earlier to describe a single beta curve. In our example, recall that p_A has a beta(5, 5) posterior curve and p_B has a beta(3, 7) curve. To simulate a pair of proportion values (p_A, p_B), we simulate the first proportion value from a beta(5, 5) curve and then the second proportion from a beta(3, 7) curve. I did this on Minitab — I got the values

$$0.4124, \;\; 0.5808$$

So (.4124, .5808) is one "typical" pair of proportion values (p_A, p_B).

Actually, we need to simulate many pairs of proportion values to get a good understanding about the location of these two proportions. I used Minitab to generate 1000 pairs of proportion values. In the figure below, the simulated pairs (p_A, p_B) have been graphed using a scatterplot.

In Topic 6, when we discussed scatterplots, we were looking for some pattern in the plot, which indicated that there was an increasing or decreasing relationship between the two variables. Here we don't see any increasing or decreasing pattern in the plot of p_A and p_B, which is a reflection of the fact that the posterior probabilities for p_A are independent of the probabilities for p_B. Actually,



Scatterplot of 1000 simulated pairs (p_A, p_B) from the posterior distribution.

in this situation, we're interested if one proportion is larger or smaller than the other. To make this judgment, the line $p_A = p_B$ has been drawn on top of the scatterplot. The points in the scatterplot *above* the line correspond to proportion pairs (p_A, p_B) where the first proportion p_A is *smaller* than the second proportion p_B. The points *below* the line correspond to pairs where p_A is *larger* than p_B.

The figure indicates that 85.4% of the points are below the line, which means that the probability that $p_A > p_B$ is .854. Equivalently, the probability that $p_A < p_B$, which is the fraction of points above the line, is .146. So there is some evidence from our data that the cure rate from drug A is larger than the cure rate for drug B. But, since .854 is not that large a probability, there is not strong evidence for the superiority of drug A.

As discussed earlier, we can measure the superiority of drug A by looking at the difference between the probabilities

$$d = p_A - p_B.$$

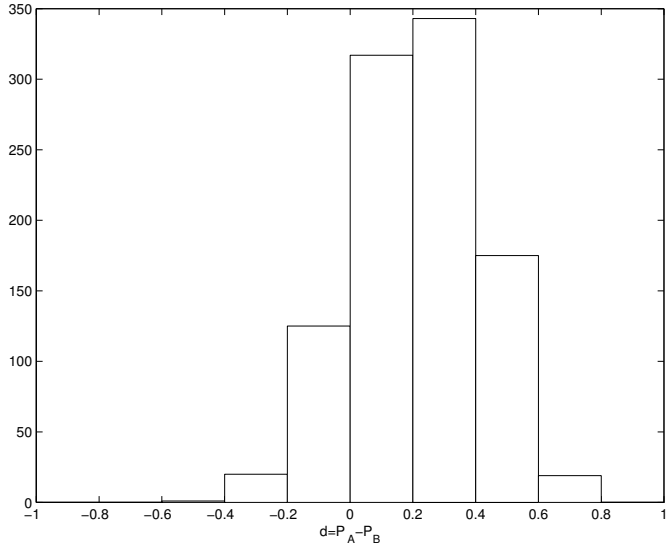
From our simulated pairs (p_A, p_B), we can obtain simulated values of the difference in proportions d. For example, if the first pair of simulated proportion values is

$$p_A = 0.4124, p_B = 0.5808,$$

then the corresponding simulated value of d is

$$d = 0.4124 - 0.5808 = -0.1684.$$

In our example, we simulated 1000 pairs of proportion values. If we compute the difference d for each pair, we obtain a simulated sample from the posterior curve of d. The table below shows a



Histogram of 1000 simulated values from the posterior distribution of $d = p_A - p_B$.

grouped count table for the simulated values of d and the figure below displays the corresponding histogram.

d	Count	Probability
(-1.0, -0.8)	0	0.000
(-0.8, -0.6)	0	0.000
(-0.6, -0.4)	1	0.001
(-0.4, -0.2)	20	0.020
(-0.2, 0.0)	125	0.125
(0.0, 0.2)	317	0.317
(0.2, 0.4)	343	0.343
(0.4, 0.6)	175	0.175
(0.6, 0.8)	19	0.019
(0.8, 1.0)	0	0.000

We learn about the relative effectiveness of the two drugs by looking at the posterior probabilities of the difference in proportions $d = p_A - p_B$. First, remember that 4 of the 8 patients were cured using drug A, and 2 of the 8 patients using drug B were cured. The *sample proportions* of cured using the two drugs are

$$\hat{p}_A = \frac{4}{8} = .5, \;\; \hat{p}_B = \frac{2}{8} = .25.$$

So a good guess at the relative effectiveness of the two drugs is the difference in sample proportions

$$\hat{p}_A - \hat{p}_B = .5 - .25 = .25.$$

Note that the center of the histogram of posterior probabilities for d is about .25. So we think that the proportion of *all* patients cured using drug A, p_A, is about .25 larger than the proportion of patients cured who use drug B, p_B.

Although a good guess at the difference in cure proportions is .25, the posterior histogram shows that there is much uncertainty about the true value of d. What's the probability that drug A has a higher cure rate? This is same as finding the probability that the difference in proportions $d > 0$. Looking at the table of probabilities, we find P($d > 0$) by adding the probabilities of d being in the intervals (0, .2), (.2, 4), (.4, .6), (.6, .8), .(8, 1). So the probability is equal to

$$P(d > 0) = .317 + .343 + .175 + .019 + .000 = .854.$$

(We found this probability earlier using a different method.)

To learn about the size of the difference of cure proportions d, it is useful to construct a probability interval. This is an interval of values which contain d with a large probability. Looking at the table of probabilities, we find four intervals which are most likely to contain d. In this case, the intervals are given by

$$(.2, .4), (0, .2), (.4, .6), (-.2, 0),$$

which contain the difference in proportions $d = p_A - p_B$ with probability

$$.343 + .317 + .175 + .125 = .960.$$

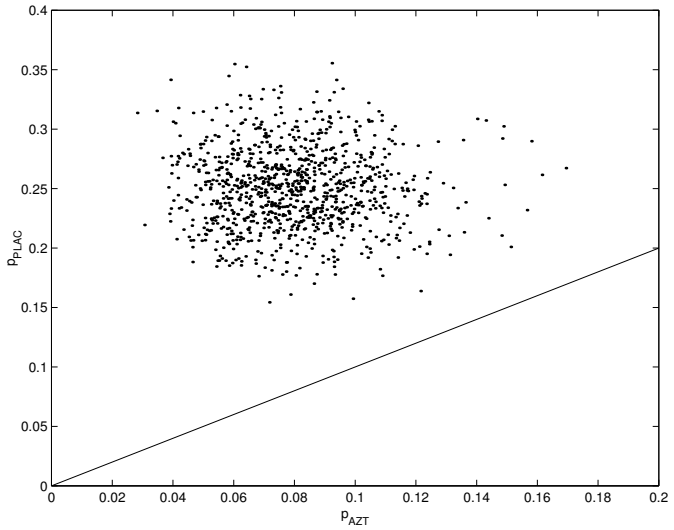
So the interval $(-.2, .6)$ is a 96% probability interval for d. Since this interval contains negative and positive values, we can't really say with any confidence that one drug is better than the other.

Activity 21-6: Pregnancy, AZT, and HIV (cont.)

Recall from Activity 9-3 and Activity 21-5 that medical experimenters randomly assigned 164 pregnant, HIV-positive women to receive the drug AZT during pregnancy, while another 160 such women were randomly assigned to a control group which received a placebo. Of those in the AZT group, 13 had babies who tested HIV-positive, compared to 40 HIV-positive babies in the placebo group.

(a) Find the proportion of babies who tested HIV-positive from the women who received the drug AZT.

(b) Find the proportion of babies who tested HIV-positive from the women who received the placebo.

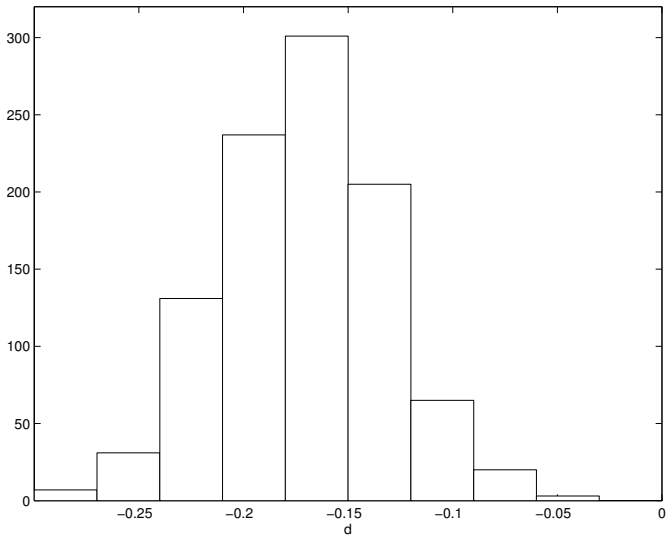


Scatterplot of 1000 simulated pairs (p_{AZT}, p_{PLAC}) from the posterior distribution.

(c) Suppose we are interested in $p_{AZT} - p_{PLAC}$, the difference in proportions of HIV-positive babies from *all* women who take the drug AZT and the placebo, respectively. Use your answers from (a) and (b) to make a good guess at the value of $p_{AZT} - p_{PLAC}$.

(d) Assuming uniform curve prior probabilities for both proportions, 1000 pairs of proportion values (p_{AZT}, p_{PLAC}) were simulated from the posterior distribution. A scatterplot of these simulated proportions is shown in the figure below. The line is drawn at the values where the two proportions are equal. Based on the graph, can you say that one proportion is larger than the second proportion? Why?

(e) A table of counts and the corresponding histogram are shown below for the difference in probabilities $d = p_{AZT} - p_{PLAC}$. Find the probability that the difference d is smaller than $-.18$.



Histogram of 1000 simulated values from the posterior distribution of $d = p_{AZT} - p_{PLAC}$.

Interval	Count
(-0.30 -0.27)	7
(-0.27 -0.24)	31
(-0.24 -0.21)	131
(-0.21 -0.18)	237
(-0.18 -0.15)	301
(-0.15 -0.12)	205
(-0.12 -0.09)	65
(-0.09 -0.06)	20
(-0.06 -0.03)	3
(-0.03 0.00)	0

(f) Find a 90% probability interval for the difference d.

HOMEWORK ACTIVITIES

Activity 21-7: Literature for Parolees? (cont.)

Activity 20-11 describes a study which investigated the possible beneficial aspects of learning literature in the reform of prisoners. 32 convicts were given a course in great works of literature — of this group, 6 committed another crime within 30 months of parole. This group was compared with a control group of 40 parolees who were not given the literature course; 18 of these 40 committed a crime within 30 months.

Let p_L denote the proportion of all "literature" prisoners who would commit a crime and p_N denote the proportion of "non-literature" prisoners who would commit a crime. The researchers were interested if the literature prisoners were less likely to commit a crime. Equivalently, they were interested in the probability that $p_L < p_N$.

(a) Suppose that each proportion can take on each of the equally spaced values .1, .2, ..., .9. There are 81 possible models for p_L and p_N and we assign a flat prior over all models. The models and initial probabilities are displayed in the following two-way table:

		p_N								
		.1	.2	.3	.4	.5	.6	.7	.8	.9
p_L	.1	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81
	.2	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81
	.3	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81
	.4	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81
	.5	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81
	.6	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81
	.7	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81
	.8	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81
	.9	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81	1/81

(b) In the table above, circle all of the cells where p_L is smaller than p_N. Find the prior probability that p_L is smaller than p_N.

(c) I use program 'pp_disc' to compute the posterior probabilities. Here a "success" is defined as a parolee committing a crime. (This is a strange definition of success.) The literature parolees had 6 successes and 26 failures; the non-literature parolees had 18 successes and 22 failures. The posterior probabilities are presented below.

		p_N								
		.1	.2	.3	.4	.5	.6	.7	.8	.9
p_L	.1	0	0	.014	.081	.082	.016	0	0	0
	.2	0	.001	.041	.244	.245	.048	.001	0	0
	.3	0	0	.014	.086	.087	.017	0	0	0
	.4	0	0	.001	.009	.009	.002	0	0	0
	.5	0	0	0	0	0	0	0	0	0
	.6	0	0	0	0	0	0	0	0	0
	.7	0	0	0	0	0	0	0	0	0
	.8	0	0	0	0	0	0	0	0	0
	.9	0	0	0	0	0	0	0	0	0

(d) Circle the cell which has the largest probability. This cell corresponds to what values of the two proportions?

(e) Find the posterior probability that p_L is smaller than p_N.

(f) Assuming a valid experimental design, is there good evidence that the literature program is beneficial in reducing crime of parolees? Explain.

Activity 21-8: Do Cellular Phones Cause Accidents?

In New York, researchers were interested in studying the relationship between driving habits and accidents. 60 resident drivers with a record of a recent accident returned a mail survey. Of these drivers, 8 (13%) used a cellular phone while driving. 77 drivers that were accident-free also returned the survey — 7 (9.1%) of these people reported using a cellular phone while driving.

Let p_A denote the proportion of all New York drivers with history of accidents who use a cellular phone. Let p_N represent the proportion of "accident-free" drivers who use a cellular phone. The researchers were interested in learning about the size of the difference in proportions $d = p_A - p_N$.

The researchers believe that the size of each proportion is small. So, for each proportion, they use the six model values .05, .1, .15, ..., .3. A flat prior which assigns equal probabilities to the 36 model pairs (p_A, p_N) is used.

The posterior probabilities (found using the program 'pp_disc') are displayed below.

		p_N					
		.05	.10	.15	.20	.25	.30
p_A	.05	.004	.013	.004	0	0	0
	.10	.066	.193	.060	.006	0	0
	.15	.087	.253	.079	.008	0	0
	.20	.037	.108	.034	.004	0	0
	.25	.008	.022	.007	.001	0	0
	.30	.001	.003	.001	0	0	0

(a) Circle all of the cells where the two proportions are equal — that is, where the difference $d = 0$. Find the probability that $d = 0$.

(b) Similarly, find the probabilities that $d = p_A - p_N$ is equal to $-.25$, $-.2$, $-.15$, $-.1$, $-.05$, .05, .1, .15, .2, .25. Put you answers in the below table.

$d = p_A - p_N$	PROBABILITY
$-.25$	
$-.20$	
$-.15$	
$-.10$	
$-.05$	
0	
.05	
.10	
.15	
.20	
.25	

(c) What is the most likely value of the difference in probabilities d? What is its probability?

(d) The four most likely values of d are _____.

The probability that d is one of these values is _____.

(e) Can the researchers conclude that accident drivers are more likely to use cellular phones than non-accident drivers? Explain.

Activity 21-9: Wording of Surveys

Much research goes into identifying factors which can unduly influence people's responses to survey questions. In a 1974 study researchers conjectured that people are prone to acquiesce, to agree with attitude statements presented to them. They investigated their claim by asking subjects whether they agree or disagree with the following statement. Some subjects were presented with form A and others with form B:

- form A: Individuals are more to blame than social conditions for crime and lawlessness in this country.
- form B: Social conditions are more to blame than individuals for crime and lawlessness in this country.

The responses are summarized in the table:

	Blame individuals	Blame social conditions
Form A	282	191
Form B	204	268

(a) Let p_A represent the population proportion of all potential form A subjects who would contend that individuals are more to blame, and let p_B denote the population proportion of all potential form B subjects who would contend that individuals are more to blame. If the researchers' claim about acquiescence is valid, should p_A be greater than p_B or vice versa?

(b) Calculate the sample proportion of form A subjects who contended that individuals are more to blame. Then calculate the sample proportion of form B subjects who contended that individuals are more to blame.

(c) If uniform curves are used as prior distributions, find the numbers of the posterior beta curves for p_A and p_B.

(d) One thousand pairs of proportions (p_A, p_B) were simulated from the appropriate beta curves. A frequency table of the 1000 simulated values of the difference in proportions $d = p_A - p_B$ is displayed below. Use this table to find the probability that the researchers' claim is true. Write a one-sentence conclusion.

d	(0, .05)	(.05, .1)	(.1, .15)	(.15, .20)	(.20, .25)	(.25, .30)	(.30, .35)
Count	0	26	314	528	127	5	0

Activity 21-10: Wording of Surveys (cont.)

Researchers have conjectured that the use of the words "forbid" and "allow" can affect people's responses to survey questions. In a 1976 study one group of subjects was asked, "Do you think the United States should forbid public speeches in favor of communism?", while another group was asked, "Do you think the United States should allow public speeches in favor of communism?". Of the 409 subjects asked the "forbid" version, 161 favored the forbidding of communist speeches. Of the 432 subjects asked the "allow" version, 189 favored allowing the speeches.

(a) Calculate the sample proportion of "forbid" subjects who oppose communist speeches (i.e., favor forbidding them) and the sample proportion of "allow" subjects who oppose communist speeches (i.e., do not favor allowing them).

(b) By using the computer to simulate the relevant beta posterior distributions, compute the probability that the researchers' conjecture is true.

(c) A 1977 study asked 547 people, "Do you think the government should forbid the showing of X-rated movies?" 224 answered in the affirmative (i.e., to forbid). At the same time a group of 576 people was asked, "Do you think the government should allow the showing of X-rated

movies?" 309 answered in the affirmative (i.e., to allow). Repeat (a) and (b) for the results of this experiment. (Be very careful to calculate and compare relevant sample proportions. Do not just calculate proportions of "affirmative" responses.)

(d) A 1979 study asked 607 people, "Do you think the government should forbid cigarette advertisements on television?" 307 answered in the affirmative (i.e., to forbid). At the same time a group of 576 people was asked, "Do you think the government should allow cigarette advertisements on television?" 134 answered in the affirmative (i.e., to allow). Repeat (a) and (b) for the results of this experiment; heed the same caution as in (c).

(e) Write a paragraph summarizing your findings about the impact of forbid/allow distinctions on survey questions.

Activity 21-11: Questioning Smoking Policies

An undergraduate researcher at Dickinson College examined the role of social fibbing (the tendency of subjects to give responses that they think the interviewer wants to hear) with the following experiment. Students were asked, "Would you favor a policy to eliminate smoking from all buildings on campus?" For half of the subjects, the interviewer was smoking a cigarette when asking the question; the other half were interviewed by a non-smoker. Prior to conducting the experiment, the researcher suspected that students interviewed by a smoker would be less inclined to indicate that they favored the ban. It turned out that 43 of the 100 students interviewed by a smoker favored the ban, compared to 79 of the 100 interviewed by a non-smoker. Let p_S denote the proportion of students favoring the ban who are interviewed by a smoker and p_N denote the corresponding proportion of students who are interviewed by a non-smoker. Assuming uniform priors, 1000 values of the difference in proportions $d = p_S - p_N$ were simulated and summarized using the count table below. Estimate d using a 90% probability interval and write a brief conclusion to the researcher.

d	Count
(-0.55, -0.50)	11
(-0.50, -0.45)	48
(-0.45, -0.40)	186
(-0.40, -0.35)	300
(-0.35, -0.30)	258
(-0.30, -0.25)	140
(-0.25, -0.20)	40
(-0.20, -0.15)	13
(-0.15, -0.10)	4

Activity 21-12: Age and Political Ideology (cont.)

Reconsider the data tabulated in Activity 9-2 concerning age and political ideology. Of the 296 people in the "under 30" age group, 83 identified themselves as politically liberal. Of the 586 people in the "over 50" age group, 88 regarded themselves as liberal.

(a) Use the posterior distribution of a difference in proportions to determine if these sample results provide strong evidence that the proportion of liberals among all "under 30" people is larger than the proportion from that among all "over 50" people.

(b) Estimate the difference in these population proportions with a probability interval.

(c) Write a few sentences describing your findings from (a) and (b) regarding the question of whether the proportion of liberals differs between "under 30" and "over 50" people and, if so, by about how much.

Activity 21-13: BAP Study

Researchers investigating the disease Bacillary Angiomatosis and Peliosis (BAP) took a study of 48 BAP patients and a control group of 94 subjects. The following table lists the numbers of people in each group who had the indicated characteristics.

	Case patients ($n = 48$)	Control patients ($n = 94$)
# Male	42	84
# White	43	89
# Non-Hispanic	38	75
# With AIDS	24	44
# Who own a cat	32	37
# Scratched by a cat	30	29
# Bitten by a cat	21	14

(a) Is this study a controlled experiment or an observational study? If it is an observational study, what kind is it? Explain.

(b) Use the computer to see if the proportion of case patients exceeds the proportion of control patients for any of the variables listed in the table.

(c) In light of the type of study involved, can you conclude that these variables cause the disease? Explain.

Activity 21-14: Baldness and Heart Disease (cont.)

Reconsider the data presented in Activity 9-14 and mentioned in Activity 20-13 concerning baldness and heart disease.

(a) Calculate the sample proportion of the heart disease patients who had some or more baldness (i.e., some, much, or extreme baldness). Calculate the same for the control group.

(b) Assuming uniform priors, 1000 values from the posterior distribution of the difference in proportions $d = p_D - p_C$ were simulated, where p_D is the proportion of heart disease patients who had some or more baldness and p_C is the proportion of control group patients showing baldness. From this table, compute the probability that heart disease patients tend to have more baldness than control patients.

d	Count
(0, 0.02)	2
(0.02, 0.04)	30
(0.04, 0.06)	99
(0.06, 0.08)	231
(0.08, 0.10)	347
(0.10, 0.12)	199
(0.12, 0.14)	77
(0.14, 0.16)	15
(0.16, 0.18)	0

(c) Construct a 90% probabililty interval for the difference in population proportions d. Write a sentence or two describing what the interval reveals.

(d) What conclusion can you draw about a possible association between baldness and heart disease? Be sure to keep in mind the design of the study as you address this question.

Activity 21-15: Sex on Television

In his book *Hollywood vs. America: Popular Culture and the War on Popular Values*, Michael Medved cites a pair of studies that examined instances of and references to sexual activity on prime-time television programs. A 1981 study examined a sample of 47 references to sex on prime-time television shows and found that 6 of the references were to sexual intercourse between partners who were married to each other. A similar (but larger) study in 1991 examined a sample of 615 references to sex and found that 44 of the references were to sex between married partners.

(a) What proportion of the sexual references in the 1981 study were to sex between married partners? In the 1991 study?

(b) Use the computer to compute the probability that the proportion of all sexual references which describe married sex has decreased from 1981 to 1991.

(c) Use the computer to find a 95% confidence interval for the difference in the proportions of references to married sex among all such television references between 1981 and 1991; record the interval below.

Activity 21-16: Employment Discrimination

In the legal case of Teal vs. Connecticut (1982), a company was charged with discrimination in that blacks passed its employment eligibility exam in smaller proportions than did whites. Of the 48 black applicants to take the test during the year in question, 26 passed; of the 259 white applicants to take the test, 206 passed.

(a) What is the sample proportion of black applicants who passed the test? What is the sample proportion of white applicants who passed?

(b) Compute the appropriate probability to see if the data provide strong evidence that the proportion of blacks who pass the test is less than that of whites. Write a one- or two-sentence conclusion (as if reporting to the jurors in the case).

Activity 21-17: Kids' Smoking

A newspaper account of a medical study claimed that the daughters of women who smoked during pregnancy are more likely to smoke themselves. The study surveyed children, asking them if they had smoked in the last year and then asking the mother if she had smoked during pregnancy. Only 4% of the daughters of mothers who did not smoke during pregnancy had smoked in the past year, compared to 26% of girls whose mothers had smoked during pregnancy.

(a) What further information do you need in the description above to determine if daughters of mothers who smoke during pregancy are more likely to smoke than daughters of mothers who do not smoke?

(b) Suppose that there had been 50 girls in each group. Let p_N denote the proportion of smokers among all children of mothers who did not smoke during pregnancy, and let p_S denote the corresponding proportion among children of mothers who did smoke. Use the computer to compute the probability that p_S exceeds p_N.

(c) Repeat (b) supposing that there had been 50 girls whose mothers had smoked and 200 whose mothers had not.

(d) Repeat (b) supposing that there had been 200 girls in each group.

(e) Is this study a controlled experiment or an observational study?

(f) Even if the probability that p_S exceeds p_N is close to one, does the study establish that the pregnant mother's smoking caused the daughter's tendency to smoke? Explain.

WRAP-UP

In this topic, we discussed the general problem of comparing two population proportions. A model in this case is a pair of proportion values, say (p_1, p_2). In the case where there are only a few proportion values of interest, we represented all possible models by means of a two-way table. Bayes' rule was used to compute the posterior probabilities in the table, and we focused on posterior probabilities for the difference in proportions $d = p_1 - p_2$. In the case where each proportion is continuous-valued, then we used uniform curves to represent our prior beliefs, and the posterior distribution for each proportion were described by a beta curve. Simulation was used as our method for obtaining a representative sample of proportion values from the posterior probability distribution. We decided which proportion is larger by inspection of a count table of the simulated values of d, and we learned about the size of the difference in proportions by means of a probability interval.

Appendix: Sample Survey Project

Introduction

In this project, you perform your own statistical inference using methods described in this book. Specifically, you will take a sample of students from your school to learn about one or more proportions of interest. After you take your sample, you will use inferential methods to see how the observations have modified your beliefs about each proportion. In this appendix, we outline the different steps of the project and discuss constructing a prior, selecting a random sample, summarizing the sample information, and computing the posterior probability distribution.

Getting Started

You begin by thinking of one proportion of your student body that you wish to learn about. This proportion might be the fraction of students in agreement with a particular issue, the fraction who prefer one flavor of ice cream to another, or the fraction who participate in a special activity. Suppose, for the sake of illustration, that you are interested in the proportion of the student body who regard their political philosophy as conservative. Then p would denote the proportion of conservative students in the entire student body.

Once you decide on a proportion of interest, you construct a question that will be asked to each student in your sample. If, for example, you are interested in the proportion of students who think of themselves as conservative, you might ask the question

"Do you think your political philosophy is conservative?"

The possible responses to this question would be "yes", "no", or "I don't know" and you learn about p by counting the number of yes's in your sample.

Constructing a Prior

Before a sample is taken, you likely have some opinions about the location of the proportion value p. Your opinions about this proportion are represented by means of a prior probability distribution. For simplicity, suppose that p can be one of the eleven values 0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1. By following the worksheet on page 393, you can construct a prior distribution which approximately reflects your knowledge about p.

Taking a Random Sample

In Topic 10, we discussed how to take a Simple Random Sample (SRS) which is one type of random sample. However, the procedure for taking a SRS (labeling each member of the population and using the random digit table) is impractical when the size of the population is large. This will be the case if you have many students at your school.

A **Stratified Random Sample** is another type of random sample which is easier to take when you have a listing of the population, such as a phonebook containing the names and phone numbers of all students at your school.

Here is how you take a Stratified Random Sample:

1. Decide on a step size – here we'll use 50 but other values can be used.
2. Decide on a starting place to sample in the population listing. (This should be done in some random fashion.) Say we decide to start on the 17th listing on page 50 of the phonebook.
3. Add the step size (50) to each listing to get the next one to sample. So we'll sample the

   17th, 67th, 117th, 167th, 217th ... listings
4. Continue until you've got a large enough sample.
5. What if a person is not home when you call? I would just forget this person and keep sampling. However, this procedure might introduce a bias in your sampling procedure. Why?

WORKSHEET TO CONSTRUCT PRIOR PROBABILITY DISTRIBUTION FOR PROJECT

Your name: _____

Question you will ask (with a yes or no response):

_____

In my problem, p is the proportion of all students who

_____

Make a guess at what YOU think the value of p is: _____

Constructing your prior:

1. First fill out the **Whole Number** column.

   (a) Assign the number 10 to the value of p that you think is most likely. Put this number in the table.

   (b) Consider other values of p that are not the most likely. If you think another value of p is half a likely as the most likely value, give it a 5. If you think a value of p is 1/10th as likely as the most likely value, give it a 1. Continue until you've assigned whole numbers to all 11 values of p.

   (c) If you give a value of p a 0, this means that you think that it is impossible that the proportion is this value.

2. Convert the whole numbers to probabilities by (1) finding the sum of the whole numbers (put the sum in the **Sum** row) and (2) dividing each whole number by the sum to get the probabilities (put in the **Probability** column).

p	Whole Number	Probability
0		
.1		
.2		
.3		
.4		
.5		
.6		
.7		
.8		
.9		
1		
Sum		

Data Analysis

Suppose that you have taken your random sample and you have a list of responses of the type "yes" or "no" which are the responses to your question. You can summarize these data using the basic techniques described in Topic 1. A count table is helpful for finding the number and proportion of yes's and no's and a bar graph can be used to display these data.

Statistical Inference

After the prior distribution for the proportion has been constructed and the data are taken, we use the methodology of Topic 16 to compute the posterior probability distribution. There is a Minitab program called p_disc that can be used in this computation. You enter the prior distribution into two columns of the spreadsheet, input the number of yes's and no's, and the program computes the posterior distribution. You use this probability distribution to construct a probability interval for p as described in Topic 16.

The Project Report

It is helpful to write a report which describes all parts of this scientific study. This report can be divided into three stages.

- **The first stage** of the report describes the choice of a survey question and the construction of the prior probability distributions. How did you decide on your particular inference problem? Is there any personal experience or things you heard or things you read in the newspaper that motivated you to choose your question? Before you took your sample, what did you think you would find out? Include the prior worksheet that you used to construct your prior distribution.

- **The second stage** of the report describes the "data phase" of the investigation. Describe in detail how you took your random sample. Describe the method you used, how many students were contacted, and any difficulties you experienced in collecting these data.

  Give the results of your survey, including the number who said yes, no, or something else. Include any graphs that you made to summarize these data. If you have the raw data for the individual students available, include these in the report.

- **The third stage** of the report describes the statistical inference. Write down the posterior probability distribution and graph the probabilities. If you are constructing a probability interval for the proportion, describe the methodology you are using and give all the details of

your computation. Attach Minitab output if this computer package was used in the computations.

It is important to explain, using language that a layman would understand, what your interval estimate means. In particular, if you are computing a 95% interval, explain what 95% means. Interpret this interval in the context of your example.

To conclude this part of the report, you should explain what you learned from this project. Were you surprised by the results? How different were the prior and posterior distributions? What problems did you experience in doing the project? If you had to do the project over again, what would you do differently?