DATA ANALYSIS AND PROBABILITY FOR TEACHERS

JIM ALBERT

BRIEF TABLE OF CONTENTS

JUNE 2008

EXPLORING DATA

Topic D1: Statistics, data, and variables

Topic D2: Graphing data

Topic D3: Summaries of data

Topic D4: Comparing batches and relative standing

Topic D5: Relationships between categorical variables

Topic D6: Relationships between quantitative variables

Topic D7: Relationships – summarizing by a line

COLLECTING DATA

Topic C1: Obtaining Data by Sampling

Topic C2: Obtaining Data by Experiments

PROBABILITY

Topic P1: Probability – a measure of uncertainty

Topic P2: Sample space and assigning probabilities

Topic P3: Let me count the ways

Topic P4: Computing probabilities by simulation

Topic P5: Conditional Probability

Topic P6: Probability distributions

Topic P7: Coin tossing distributions

Topic P8: Continuous distributions

Topic P9: The Normal distribution

INTRODUCTION TO STATISTICAL INFERENCE

TOPIC I1: Introduction to Inference: Estimating a Proportion

TOPIC I2: Introduction to Hypothesis Testing

TOPIC I3: Learning about a Population Mean

DATA ANALYSIS AND PROBABILITY FOR TEACHERS

JIM ALBERT

EXTENDED TABLE OF CONTENTS

JULY 2008

EXPLORING DATA

Topic D1: Statistics, data, and variables

- SPOTLIGHT: The U.S. Census Bureau
- WARM-UP ACTIVITY: Getting to know you
- Statistics and data
- Organizing and summarizing data: some initial thoughts
- ACTIVITY: Meet the states data
- Reading articles in the media
- Graphs in the media
- TECHNOLOGY ACTIVITY: Introduction to *Tinkerplots*
- CLASSROOM CAPSULE: Children's Well-Being
- EXERCISES

Topic D2: Graphing data

- SPOTLIGHT: Who are the baseball players?
- Graphing categorical data
- Graphing quantitative data – distribution and shape
- Stemplot
- Histogram
- Experiment with different graphs
- TECHNOLOGY ACTIVITY (Fathom): Choosing the bins of a histogram
- ACTIVITY: The shape of the data
- Classes of data and shapes
- ACTIVITY: Matching variables and shapes
- CLASSROOM CAPSULE: Graphing health and liking school
- EXERCISES

Topic D3: Summaries of data

- SPOTLIGHT: Nutrition value of ice cream?
- Summarizing categorical data
- How many calories are there in an "average" scoop of ice cream? (Introducing the median)
- Deviations and the mean
- Geometrical interpretation of the mean
- Comparing the median and the mean

Measures of spread: quartiles and the IQR

Measures of spread using deviations

Interpreting s: the 68/95/99.7 rule for bell-shaped data

ACTIVITY: Collecting some data on cities

ACTIVITY: V is for Variation

TECHNOLOGY ACTIVITY (Fathom): Deviations, the Mean, and Measures of Spread

ACTIVITY: Measurement bias

ACTIVITY: Matching statistics to histograms

CLASROOM CAPSULE: Summarizing risky behavior

EXERCISES

Topic D4: Comparing batches and relative standing

SPOTLIGHT: Where's the best place to live?

Comparing batches of categorical data

Comparing batches of quantitative data

Relative standing

Flagging possible outliers

ACTIVITY: Comparing men and women in the class dataset

ACTIVITY: Matching statistics with boxplots

ACTIVITY: Counting pasta

CLASSROOM CAPSULE: Comparing achievement scores of rich and poor countries

EXERCISES

Topic D5: Relationships between categorical variables

SPOTLIGHT: The Titanic: women and children first?

Titanic survival data

ACTIVITY: Predictable pairs

Simpson's Paradox [TO BE DONE]

CLASSROOM CAPSULE: Exploring two-way tables

EXERCISES

Topic D6: Relationships between quantitative variables

SPOTLIGHT: Measuring climate

Introduction: Looking at weather data

Relationships – scatterplots

A simple correlation formula – the QCR

A correlation coefficient

Interpreting the correlation coefficient

ACTIVITY: Matching correlations with scatterplots

TECHNOLOGY ACTIVITY (Fathom): Guessing correlations

EXERCISES

Topic D7: Relationships – summarizing by a line

SPOTLIGHT: Measuring a car

Relationships – summarizing by a least-squares line

Making appropriate and inappropriate predictions

ACTIVITY: Fitting a line by eye to Galton's data

The median-median line – a robust alternative method of fitting a line

TECHNOLOGY ACTIVITY (Fathom): Fitting a "best" line

Plotting residuals

TECHNOLOGY ACTIVITY (Fathom): Exploring some Olympics data

ACTIVITY: Regression to the mean

Different ways of looking at relationships

TECHNOLOGY ACTIVITY: Using *Tinkerplots* to study relationships

CLASSROOM CAPSULE: Summarizing association between two measures of family health

EXERCISES

COLLECTING DATA

Topic C1: Obtaining Data by Sampling

SPOTLIGHT [TO BE DONE]

Population and sample

Examples of popular sampling methods: Elvis Presley and Alf Landon

Simple random sampling

ACTIVITY: Random rectangles

TECHNOLOGY ACTIVITY (Fathom): Biased sampling of rectangles

Practical concerns in sampling

EXERCISES

Topic C2: Obtaining Data by Experiments

SPOTLIGHT [TO BE DONE]

An apple a day: different ways of collecting data

Music and math achievement

An experiment to detect the Mozart effect

Basic principles of experiments

ACTIVITY: Jumping frogs

EXERCISES

PROBABILITY

Topic P1: Probability – a measure of uncertainty

SPOTLIGHT: How Risky is …?

WARM-UP ACTIVITY: Some questions on probability

The classical view of a probability

The frequency view of a probability

ACTIVITY: Tossing and spinning a poker chip

The subjective view of a probability

A calibration experiment

CLASSROOM CAPSULE: Thinking about probability

EXERCISES

Topic P2: Sample space and assigning probabilities

SPOTLIGHT: The casino game of Roulette

WARM-UP ACTIVITY: Writing down some sample spaces

Different representations of a sample space

Assigning probabilities

A more formal look at probability

The three probability axioms

The complement and addition rules

CLASSROOM CAPSURE: What can happen?

EXERCISES

Topic P3: Let me count the ways

SPOTLIGHT: Rolling dice and Yahtzee

Equally likely outcomes

The multiplication rule

Permutations

Combinations

Arrangements of non-distinct objects

Which rule?

Playing Yahtzee

ACTIVITY: Mothers and babies

ACTIVITY: Sampling from a bag

CLASSROOM CAPSULE: Playing scrabble

EXERCISES

Topic P4: Computing probabilities by simulation

SPOTLIGHT: Buffon's needle simulation

Simulating a lottery game

Basic components of a simulation experiment

The collector's problem

TECHNOLOGY ACTIVITY (Fathom): Mixed-up letters

ACTIVITY (TI -84 Plus Calculator, Fathom): The longest run

ACTIVITY (TI -84 Plus Calculator, Fathom):   Sampling people from a room

ACTIVITY (TI -84 Plus Calculator, Fathom): : The birthday problem

CLASSROOM CAPSULE: Random ties

CLASSROOM CAPSULE: Waiting in line

EXERCISES

Topic P5: Conditional Probability

SPOTLIGHT: The three-card problem

New information, reduced sample space, and conditional probability

Conditional probability in everyday life

Conditional probability in a two-way table

Definition of conditional probability and multiplying probabilities

The multiplication rule under independence

Learning using Bayes' rule

TECHNOLOGY ACTIVITY (Fathom): Rolling two dice

TECHNOLOGY ACTIVITY (Fathom): How many defectives?

CLASSROOM CAPSULE: Spinning a random spinner

EXERCISES

Topic P6: Probability distributions

SPOTLIGHT: The hat check problem

A random variable

Summarizing a probability distribution

Standard deviation of a probability distribution

Interpreting the standard deviation for a bell-shaped distribution

TECHNOLOGY ACTIVITY (Fathom): Constructing a probability distribution by simulation

TECHNOLOGY ACTIVITY (Fathom): Playing roulette

ACTIVITY: How many keys?

ACTIVITY: Investing money: comparing safe and risky investments

CLASSROOM CAPSULE: Playing the lottery

EXERCISES

Topic P7: Coin tossing distributions

SPOTLIGHT: Galton's board

Probabilities of a coin-tossing experiment

Binomial experiments

ACTIVITY: Coin flipping: Is it Real or Fake?

TECHNOLOGY ACTIVITY (Fathom): Simulated coin flipping

ACTIVITY: Is a professional athlete streaky?

Binomial computations

Mean and standard deviation of a binomial

Negative binomial experiments

ACTIVITY: Graphing binomial and negative binomial experiments

CLASSROOM CAPSULE: Probabilities of the Galton board

EXERCISES

Topic P8: Continuous distributions

SPOTLIGHT: A spinner baseball game

The uniform distribution
A probability density/waiting for a bus
The probability function (the cdf F(x))
Finding probabilities using F
Summarizing a continuous random variable
Percentiles
TECHNOLOGY ACTIVITY (Fathom): Spinning away
TECHNOLOGY ACTIVITY (Fathom): Waiting for the shuttle
TECHNOLOGY ACTIVITY(TI-84 Plus or Fathom): A test with a bimodal distribution
TECHNOLOGY ACTIVITY: Estimating areas by simulation [TO BE DONE]
CLASSROOM CAPSULE: Census ages
EXERCISES
Topic P9: The Normal distribution
SPOTLIGHT: Early use of the normal curve
Modeling data by a normal curve
Computing normal probabilities
Computing normal percentiles
Binomial probabilities and the normal curve
Sampling distribution of the mean: the central limit theorem
The central limit theorem works for any population
TECHNOLOGY ACTIVITY (Fathom): Sampling heights
TECHNOLOGY ACTIVITY (Fathom): Rolling biased dice
CLASSROOM CAPSULE: Sampling from a U-shaped population
EXERCISES

INTRODUCTION TO STATISTICAL INFERENCE

TOPIC I1: Introduction to Inference: Estimating a Proportion
SPOTLIGHT [TO BE DONE]
A classroom survey
Population, parameter, sample, and statistic
Sample estimates: bias and variance
TECHNOLOGY ACTIVITY (Fathom): The taxi problem
Construction of a confidence interval
A large sample confidence interval for a proportion
Understanding a confidence interval
Choosing a sample size
Some cautions
TECHNOLOGY ACTIVITY (Fathom): Penny ages
EXERCISES
TOPIC I2: Introduction to Hypothesis Testing

SPOTLIGHT [TO BE DONE]

Introduction: A taste test

Stating the hypotheses

A statistical test

A large-sample test for a proportion

Statistical significance

Two-sided tests and confidence intervals

Decisions, two-errors, and confidence

TECHNOLOGY ACTIVITY (Fathom): Is the machine working?

EXERCISES

TOPIC I3: Learning about a Population Mean

SPOTLIGHT [TO BE DONE]

How long is a cell phone call?

Review: the pattern of sample means

A confidence interval for a mean

Understanding a confidence interval for a mean

Testing about a mean

More about a test for a mean

ACTIVITY: Estimating family size

ACTIVITY: Estimating the total of a restaurant bill

TECHNOLOGY ACTIVITY (Fathom): Is the neighborhood expanding?

EXERCISES

TOPIC D1: STATISTICS, DATA, AND VARIABLES

SPOTLIGHT: THE U.S. CENSUS BUREAU

The U.S. Census Bureau is the leading provider of data regarding the people and economy of the United States. Following independence in 1776, there was an immediate need to count the number of people in the entire country. Secretary of State Thomas Jefferson was the overseer of the first census taken in 1790 that counted 3.9 million inhabitants. It was clear in the early history of the United States that there was a need to collect statistics to help people understand the current status of the country and its growth. The content of the census has changed over time. In 1810, information was collected regarding the manufacturing and quality of products, in 1840 questions were added on fisheries, and in 1850 data were collected on taxation, churches and crime.

The Census Bureau oversees a census of the U.S. population every ten years. In addition, censuses are conducted on economic activity and state and local governments every five years, and every year the Census Bureau conducts over 100 other surveys.

Much of the data collected by the Census Bureau is publicly available on its website at www.census.gov. On the *American Factfinder* page of this Census website, one can find a wide variety of data regarding the population, economic activity, and geography of the United States. There is information about age, education, income and race of different states. There are tables and graphs regarding home ownership, including home values and mortgages, and a wealth of information regarding different types of businesses and industries.

In this topic, we will access data available in the Education section of the Census website. There are extensive data on school enrollment at different levels for all states. Data have been collected to learn about children's academic achievement, differences in achievement due to gender or race, the relative number of children attending private and public schools, and the availability of classes for gifted students. In addition, data are available documenting the educational attainment of adults and the relationship between educational attainment and job earnings. The site, through its data tables, describes the differences in educational attainment between age groups and between the different regions of the country. The material posted on this website provides a great opportunity to learn about the education of Americans.

PREVIEW

In this opening topic, we get introduced to the science of statistics. Briefly, statistics is the process of formulating one or more questions about our world, collecting relevant data, and organizing and summarizing the data to answer the questions of interest. When we collect data, we will see that there are different ways that variables can be measured, and these different measurement types affect how we work with the data to draw conclusions. In this topic, we will get experience reading statistical studies and graphs in the media, and see how these reports reflect the different parts of a statistical investigation.

In this topic your learning objectives are:

- To understand statistics as a science of learning about our world.
- To understand that the science of statistics has four basic components common to any investigation that collects data to learn about our world.
- To understand the different measurement types of variables.
- To begin to have a critical view towards statistical studies reported in the media. Topic D1

NCTM Standards

✓In Grades 9-12, all students should understand the meaning of measurement data and categorical data, and of the term variable.

✓In Grades 9-12, all students should evaluate published reports that are based on data by examining the design of the study, the appropriateness of the data analysis, and the validity of conclusions.

WARM-UP ACTIVITY: GETTING TO KNOW YOU

What do you know about your fellow students? Suppose you are interested in learning about the group of students that attend your school. In particular, you might be interested in learning about various physical traits of students such as gender and hair length and about social characteristics such as students' interest in movies, their diets, and their sleeping habits. We can obtain a convenient and interesting data set by asking questions from ourselves. Below is a list of questions that can be answered by every student in the class. In addition, through a class discussion, other questions can be proposed that will be answered by all of the students.

Your instructor will prepare a data set containing the student responses to all of the questions. This data set will be used to illustrate various methods for graphing and summarizing data. Using these descriptive methods, we will be able to draw some general conclusions about the students that attend our school.

QUESTIONS:

1. What is your height in inches? __________________

2. What is your gender? __________________

3. How many pairs of shoes do you own? __________________

4. Choose a number between 1 and 10. __________________

5. How many movie DVDs do you own? __________________

6. What time (to the nearest half-hour) did you go to bed last night?

   ____

7. What time (to the nearest half-hour) did you wake up this morning?

    ____

8. How much did you spend on your last haircut (including the tip)?

    ____

9. How many hours do you plan to work on a job per week this semester?

    ____

10. For an evening meal, do you prefer water, soda (pop), or milk?

    ____

QUESTIONS GENERATED FROM CLASS DISCUSSION

11. ____

12. ____

13. ____

14. ____

15. ____

Statistics and Data

There is a general confusion about the meaning of "statistics." If we look at the *American Heritage Dictionary of the English Language*, we find two very different definitions of statistics.

1. *Statistics* is a collection of numerical data.
2. *Statistics* is the mathematics of the collection, organization, and interpretation of numerical data.

Let's focus first on the first definition. To many people, statistics are simply numbers like

61, 73, .406

But these are not statistics, they are simply numbers. Statistics are ***numbers with a context*** or underlying meaning. The author is a baseball fan and the above three numbers have interesting contexts:

- 61 is the number of home runs hit by Roger Maris in 1961
- 73 is the number of home runs hit by Barry Bonds in 2001
- .406 is the batting of Ted Williams in 1941 (the last player to have a batting average of over .400 for a single season)

A baseball fan cares about these statistics since they make him or her think about outstanding baseball players and their accomplishments.

We will refer to information collected from people or objects as ***data***. Actually, data is the plural form and a single piece of information is called ***datum***, although data is now commonly used to represent one or more pieces of information.

Why Do We Collect Data?

Why do we care about data? We collect data to help us learn about our world. We start off with some questions about something we wish to learn about, and we find appropriate data to help us answer these questions.

The Second Definition of Statistics

We've seen that statistics are numerical data that we work with. But statistics has a second deeper meaning -- it's the science of using these data to learn about our world and make conclusions.

The science of statistics has four basic components:

- **FORMULATING QUESTIONS**: First, state some questions or problems that we would like to address by collecting relevant data.
- **COLLECTING DATA:** Second, specify effective ways of collecting data that are useful in answering the questions of interest.
- **ORGANIZING & SUMMARIZING**: Next, organize and summarize the collected data to learn about its general features.
- **MAKING CONCLUSIONS**: Last, use the data to make conclusions. (It turns out that probability or chance plays an important role in decision-making.)

Any statistical study reported in the media will have these four components. At the beginning, there will be some questions that motivated the researcher to study a problem. If there were no questions, then there would be no reason to proceed further into a statistical study. Second, the researcher will collect data that he or she believes will be useful in answering the question. We will see that data can be collected or found from many sources. Next the researcher organizes the data in some useful way and make graphs and or calculations that are helpful in answering the main questions. Finally, the researcher has to use the graphs and calculations to address the questions of interest. It is possible that the data are insufficient or inconclusive on answering the questions and perhaps a new statistical study will be undertaken.

In this topic, we illustrate the four components of a statistical study for several examples. We begin with a description of our own study on learning about the educational achievement of Americans and then we look at statistical studies reported in the media.

Example: Educational Attainment

Formulating questions

Let's get started by looking at some educational data. All of us completed high school and most of us will be completing college. We take these accomplishments for granted, but we realize that not everyone in the United States completes high school and certainly many students don't go to college. There is a current budget crisis in the state of Ohio and the legislature has to prioritize its spending. Should the money go to four-year state universities, or should it go to community colleges that specialize in two-year technical programs? In the discussion about this budget crisis, some people have said that one problem is that a relatively low number of Ohio adults are college educated. These adults are more familiar with technical education and may not appreciate the added value of a four-year college degree in preparing students for a variety of careers.

This brief discussion raises the following questions:

1. Is it common for a person in the United States to complete high school?
2. Are there differences in high school graduation rates between states? If so, which states have higher rates, and which states have lower rates?
3. How likely is it for a United States adult to complete college? How does the college completion rate vary between states?
4. How does Ohio compare with other states with respect to the college completion rate?
5. Is there a connection between a state's high school graduation rate and its college graduation rate?

Collecting Data

Where do we find data? Many books, such as almanacs, have data on population and demographic information on states in the United States. A quick way to get data is through the Internet.

Use your favorite search engine on the Internet and do a search for

educational attainment state

One of the sites it might pick up is from the U.S. Census Bureau:

www.census.gov/population/socdemo/education/p20-536/tab13.pdf

This file contains a table "Educational Attainment of the Population 25 Years and Over, By State, March 2000." The source for the data in the table is the U.S. Census Bureau and these data were put on the Internet on December 19, 2000.

Here's part of the table:

(Numbers in thousands)

			Bachelors
	Total	Completed	degree
	25 years	High School	or more
State	and over	Percent	Percent
Alabama	2,790	77.5	20.4
Alaska	359	90.4	28.1
Arizona	2,996	85.1	24.6

For each of the 50 states plus the District of Columbia, this table gives

- the name of the state or district
- the total number of residents 25 years or over (we'll call these adults)
- the percent of these adults that completed high school
- the percent that obtained a bachelor's degree

How did the U.S. Census Bureau get these data? One possibility is that the bureau asked every single adult in the United States their educational status. Actually, the government does attempt to collect data from all adults every 10 years -- this is called the census. But a census is very expensive and rarely done. These reported high school and college percentages are actually numbers (or statistics) computed based on a sample of adults taken from each state.

How did the Census Bureau take their samples? That's a good question and it should be asked whenever there is sampling reported in an article in the media. We'll talk about basic principles for taking "good" samples in a later topic.

Organizing and summarizing; drawing conclusions

Now that we have listed some questions of interest and have found relevant data, we next have to organize and summarize the data and then use our work to answer our questions. In topics D2 and D3, we will revisit this example and introduce methods of graphing and summarizing data that will help us learn about the graduation rates of Americans.

Variables and Variable Types

In this example, we have collected different types of information from each state. The object that we are collecting information from is called the ***observational unit***. In this case, the observational unit is the state. The different types of information we collect for each state are called ***variables***. Here some variables are the name of the state, the percent of that state's adults that completed high school and the percent of adults that have a bachelor's degree. There are two distinct types of variables depending on how the variable is recorded. The name of the state is an example of a ***categorical variable*** -- this is in which its values can be grouped into different categories. The percentage of adults that graduate from high school is a ***quantitative variable*** -- this is a variable where the values are numerical and refer to the quantity or size of something.

As a second example, suppose we record the current grade point average, the hair color, and the number of music cds owned for 30 students in a class. Here the student would be the observational unit. Hair color would be a categorical variable, and the grade point average and the number of cds owned would be quantitative variables.

SPECIAL NOTE: Sometimes it can be difficult to tell if a collected “number” is a categorical or a quantitative variable. For example, is a person’s social security number quantitative or categorical? For a variable to be of the quantitative type, it must make sense to add, subtract, multiply or divide these values. Is it meaningful to subtract two social security numbers, say 123 55 005 and 222 44 2121? The answer is “no” and that indicates that a person’s social security number is an example of a categorical variable.

When we collect data, it is important to recognize if a given data value represents a categorical or quantitative variable. Our exploration of data will depend on its type. The way we explore categorical data will be fundamentally different from our treatment of quantitative data.

PRACTICE: GETTING TO KNOW YOU

In the "Getting to Know You" activity, the students in your class were asked several questions. For each question,

- explain why you (or someone else) might be interested in the answers to this question for a group of students
- state if the answer is a quantitative or a categorical variable

(The first two questions have been filled in for you.)

(a) Question: What is your height in inches?

Why? If a particular student was 70 inches tall, she might want to know how her height compares to the heights of other women at her school.

Type of variable: quantitative

(b) What is your gender?

Why? You might be interested in the relative numbers of men and women at your school.

Type of variable: categorical

(c) How much money do you have with you right now, in change only?

Why?:

Type of variable:

(d) Choose a number between 1 and 10. ___________________

Why?:

Type of variable:

(e) How many audio CDs do you own? ___________________

Why?:

Type of variable:

(f) What time did you go to bed last night? ___________________

Why?:

Type of variable:

(g) What time did you wake up this morning? ________________

Why?:

Type of variable:

(h) How much did you spend on your last haircut (including the tip)?

Why?

Type of variable:

(i) How many hours do you plan to work on a job per week this semester?

Why?

Type of variable:

(j) How many cups of coffee did you drink yesterday?

Why?

Type of variable:

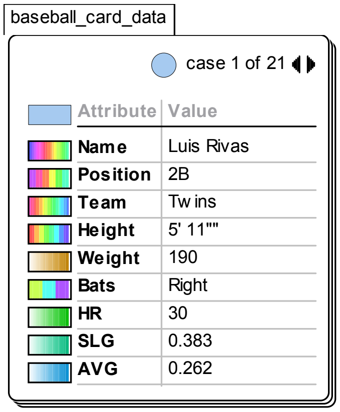
REFLECTION. Think of three additional pieces of information that you would like to learn about your fellow students. For each piece of information, explain why you are interested in this information, state the question you would ask, and describe if the answer to the question would be a quantitative or categorical variable.

Organizing and summarizing data: some initial thoughts

How does one organize and summarize data? To give some initial thoughts about how one gets started on this task, it is helpful to recall an episode from the author's youth. He enjoyed baseball and liked to collect baseball cards. He would buy a number of packs of cards and then spread all of the cards on the rug in my room. Each baseball card contained some data about a particular player, such as his age, height, weight, and statistics of his batting or pitching performance for recent seasons.

When your author looked at this collection of cards on the rug, he wanted to manipulate them in some way to get a better understanding of the statistics on the cards. In a similar manner, we wish to perform different operations on our dataset to learn about its basic features. We will discuss more formally in Topics D2 and D3 how to graph and summarize a single batch of data. But here we describe some basic operations that might be helpful in organizing data.

It can be fun to revisit one's youth, so the author recently purchased several packs of baseball cards for the 2006 season. There is a variety of data printed on each baseball – the data card below for Luis Rivas illustrates some of the variables collected.



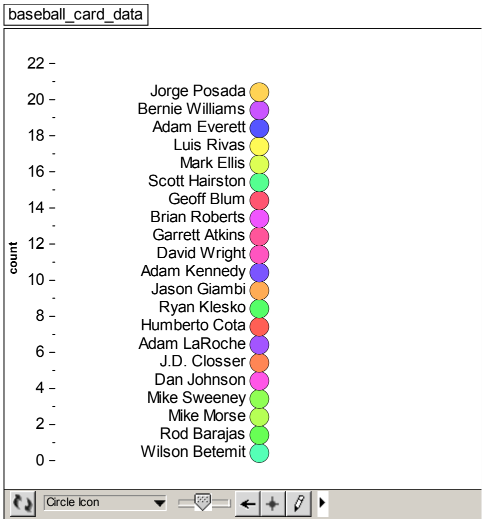
baseball_card_data

case 1 of 21

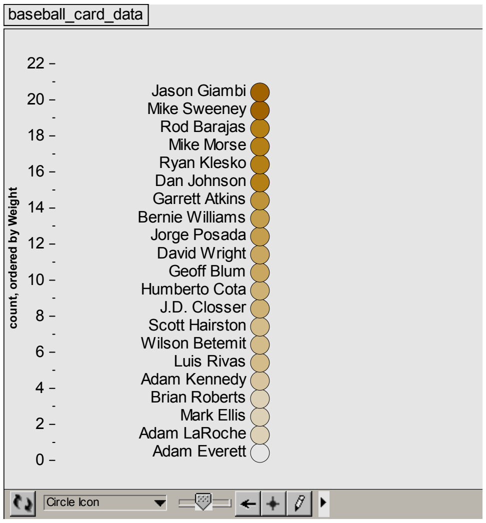
Attribute	Value
Name	Luis Rivas
Position	2B
Team	Tw ins
Height	5' 11""
Weight	190
Bats	Right
HR	30
SLG	0.383
AVG	0.262

For this card, there are several categorical variables measured such as the player's Name, his fielding Position, the Team that he plays for, and the side of the plate that he bats (variable Bats). There are also quantitative variables available such as the player's height, his weight, the number of home runs (HR) hit in his major league career, and his career slugging percentage (SLG) and his career batting average (AVG).

Suppose one collects all of the 21 baseball cards of players who are not pitchers and have played at least one season in the Major League. We list the names of the players on the 21 cards as shown below.

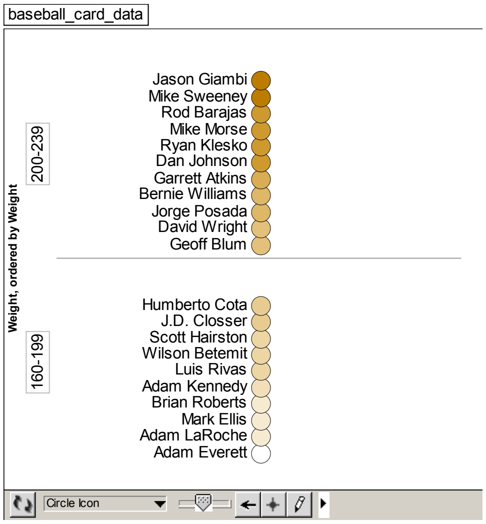


One way of organizing data is to *arrange* or sort the values on the basis of one variable. For example, suppose the author is interested in the weights of the baseball players and so he sorts the cards from the heaviest to the lighter player.



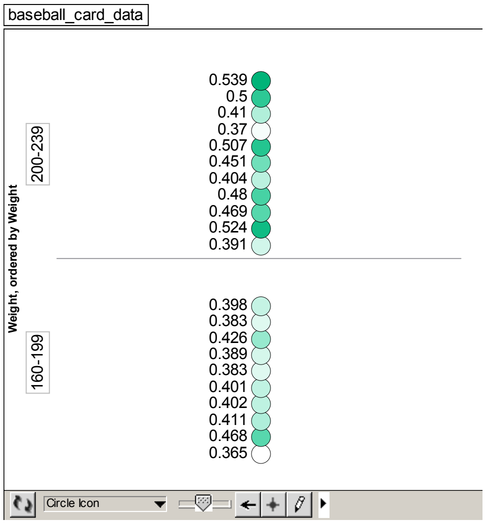
By doing this, we can identify the players who are unusually heavy (such as Jason Giambi) or light (such as Adam Everett) in this group. David Wright is in the middle of this sorted list, so he appears to have an "average" weight in this player group.

Another way to organize data is to *divide* them into two or more groups by the values of one variable. In this baseball example, we might be interested in breaking the players into the "light" players and the "heavy" players, where "heavy" is defined to be a player who weighs at least 200 pounds. In the below figure, we use a horizontal line to break the data into the two groups.



Once the data has been divided into groups by one variable, we might look at the relationship between the variable and a second variable. In our example, we might wonder if the light players and the heavy players differ with respect to another variable. One might think that a bigger player is more likely to hit extra-base hits such as doubles, triples, or homeruns. One measure of the ability of a baseball hitter to get extra-base hits is the slugging percentage (SLG). Do the light and heavy players differ with respect to slugging percentage?

To answer this question, the slugging percentage for each of our 21 players is recorded next to each player.



Looking at values of SLG for the two groups, it does appear that the heavy players have some of the largest values. To see if this first impression is correct, we summarize the slugging percentages in each group. A basic summary is an average such as a mean that is found by summing the SLG for each group and dividing by the number of players. If we do this, we get the following table.

Group	Average SLG
Heavy players	.458
Light players	.403

It does appear that heavy players tend to have a higher slugging percentage – by the table, it seems that heavy players, on average, have a .458 - .403 = .055 higher SLG than light players.

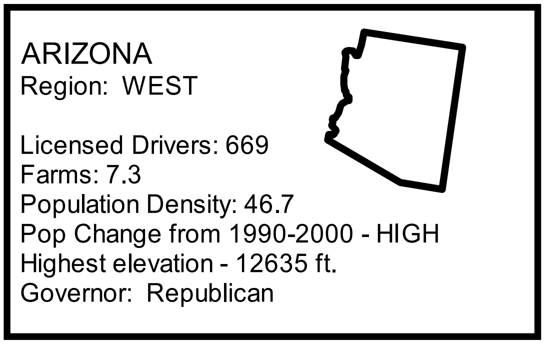
In the following activity, we will use the "arrange, divide, and summarize" data operations to learn about characteristics of different states in the U.S.

ACTIVITY: MEET THE STATES DATA

The United States is a diverse country in many ways. One will find significant differences between the states with respect to population density, geography, climate, employment, and cost of living. By exploring data collected from various states in this activity, we will begin to appreciate the diversity of the U.S. and start thinking about ways of effectively organizing, graphing, and summarizing these data to draw conclusions about this diversity. It is not important that you use a particular type of graph or summarize the data by the "right" statistic. Instead, you should choose methods that seem helpful in answering your questions.

This activity will be done in pairs. Each pair of you will be given a pack of State Cards, where one sample State Card is shown below. You are supposed to pose questions about several variables and then work with the data in various ways to answer these questions. (If these State Cards are not available, this activity can be done using packs of baseball or other sports cards that readily available in stores.)

MATERIALS NEEDED: One pack of special State Cards.



ARIZONA
Region: WEST

Licensed Drivers: 669
Farms: 7.3
Population Density: 46.7
Pop Change from 1990-2000 - HIGH
Highest elevation - 12635 ft.
Governor: Republican

On a particular card, there are recorded:

- STATE: The name of the state.
- REGION: The location of the state in the United States.
- DRIVERS: The number of licensed drivers per 1000 residents in August 2002.
- FARMS: The number of farms (in thousands) in 2001.
- DENSITY: The population per square mile of land area (2000).
- POP CHANGE: The percentage change in population between 1990 and 2000 (HIGH OR LOW).
- ELEVATION: The highest point in the state (measured as feet above sea level).
- GOVERNOR: The political party of the state governor in 2002.

Your assignment is:

(1) Choose one quantitative variable that you are interested in and formulate a few questions about the variable.

My variable is ___________________

Questions about my variable:

(a)

(b)

(c)

(2) Arrange the cards from low to high with respect to the variable you chose in (1).

(3) Find the state with the lowest value, the state with the highest value, and the state with the "middle" value with respect to your variable.

	State	Value
Lowest value		
Highest value		
"Middle" value		

(4) Think of a second quantitative variable that you believe is related or associated with your first variable.

My second variable is ____________________

(5) Break the states into two groups of approximately equal size – the states that are low with respect to your first variable and the states that are high with respect to the first variable.

(6) For each group, find the "average" value of the second quantitative variable. (A simple average is the mean found by summing the values of the variable and dividing by the number of states.) Summarize your work below.

For states that were low in____________, the average of ____________ was _________

For states that were high in ___________, the average of ____________ was _________

(7) Based on your work above, do you believe there is a relationship between your two variables? Explain.

EXTENSIONS: An almanac is a good source of data for different states. Make a list of 20 states and use the almanac to find an interesting variable for each state. Examples of interesting variables might be (1) birth rate, (2) percentage of population not covered by health insurance, (3) average temperature in July, (4) land area, (5) the mean SAT score, and (6) the percentage of people that voted for the Republican candidate in the most recent Presidential election. Answer the seven questions above using your collected data.

Reading Articles in the Media

Everyday we can read articles describing the results of statistical studies in the newspaper or the Internet. To illustrate, I looked at articles published in *USA Today* for a particular week in April, 2005. Here are some headlines of some relevant articles:

- "Alcohol's role in health not clear." An earlier study had suggested that there was evidence that a few alcoholic drinks a day was good for the heart. But this article discusses a more recent study that is inconclusive about the impact of drinking in reducing the risk of heart disease.
- "Perchance, to dream – of a good night's sleep." This article investigates the sleeping patterns of business travelers. By the use of several surveys, the article concludes that these "road warriors" are getting insufficient sleep.
- "Study: Fewer than expected dying from obesity." There is a general concern about the impact of obesity on death rates among Americans. But it is difficult to precisely estimate the number of deaths that are due to this health risk and this article describes how the conclusions from different studies vary.
- "2030 Forecast: Mostly gray." The Census Bureau recently projects, on the basis of current data, how the elderly population will grow faster than the total population.
- "Fewer high schoolers use Ecstasy, study finds." The drug Ecstasy was recently a popular drug among certain teens and young adults. But this article describes statistical evidence that this drug is losing popularity among this particular group of Americans.

Whenever we read an article such as these from *USA Today*, we should ask ourselves what questions were asked, how the data were collected, and if the conclusions drawn from the data appear valid based on the information that is provided. Unfortunately, many of the articles are brief, and it can be difficult to determine if the conclusions make sense due to the incomplete description of the study. But even if the article seems incomplete, you can think about what needs to be described to make the article more complete.

To illustrate this critical view, consider the following article about the possible benefits in eating grapefruit.

Grapefruit Lowers Weight, Fights Cancer
Studies find benefits to eating the citrus

By Kathleen Doheny

WEDNESDAY, Aug. 25 (HealthDayNews) -- A grapefruit or two a day, along with a healthy diet, could help shrink widening waistlines.

This finding comes from one of several studies on the benefits of citrus fruits presented Wednesday at the annual meeting of the American Chemical Society in Philadelphia.

The so-called grapefruit diet -- which advocates mostly eating grapefruit with some protein -- has been popular on and off for weight loss for years, said Dr. Ken Fujioka, director of nutrition and metabolism research at the Scripps Clinic in San Diego and lead author of a study evaluating grapefruit for weight loss. Most nutrition experts have deemed the grapefruit-and-protein regimen unhealthy, and Fujioka is not advocating any return to such a strict diet. However, his findings do suggest that a grapefruit or two each day, added to a balanced diet, might help the weight-conscious stay svelte.

In the study, Fujioka and his colleagues assigned 100 men and women who were obese to one of four groups. One group received grapefruit extract, another drank grapefruit juice with each meal, another ate half a grapefruit with each meal, while the fourth group received a placebo. "They weren't trying to diet," he said. "To make everyone even [on activity], all were asked to walk 30 minutes three times a week."

At the end of 12 weeks the placebo group lost on average just under half a pound, the extract group 2.4 pounds, the grapefruit juice group 3.3 pounds, and the fresh grapefruit group 3.5 pounds.

"In this study they had one and a half grapefruits a day," he noted. "That's not easy to do." And participants ate the fruit more like an orange: "They cut it in half, then into four sections, then separated the fruit from the skin." Eating grapefruit this way is thought to yield more beneficial compounds, he explained.

Exactly how grapefruit might spur weight loss isn't known, Fujioka said, but "it appears to help insulin resistance," which develops as people become obese.

The weight loss associated with eating grapefruit isn't surprising to another expert familiar with the study. "Eat fruit before any meal and you will lose weight," said Julie Upton, an American Dietetic Association spokeswoman. "The fiber fills you up, and fruit has fewer calories than other foods."

One half of a grapefruit has 60 calories, no fat, and six grams of fiber.

In analyzing this study, we ask the following questions:

1. What were the main questions addressed by the statistical study?

Here the investigators were trying to learn about the benefit of eating grapefruit towards the goal of losing weight. There was some support for a grapefruit-only diet in the past, and the scientists wished to learn if a moderate consumption of grapefruit would help to lose weight.

2. What data were collected to answer the questions?

The scientists measured the number of pounds lost in a 12-week period for each person in the experiment.

3. How did they collect the data?

Initially 100 obese people were selected to participate in the study. These people were placed into four groups where each group had a different type of grapefruit diet.

4. What variables were measured in the data? Label each variable collected as quantitative or categorical.

For each person there are two relevant variables: the group or diet plan and the number of pounds lost in 12 weeks. Group would be a categorical variable and pounds lost would be a quantitative variable.

5. What were the conclusions drawn from this statistical study? Do you believe that the conclusions are valid based on the information provided?

The groups that had grapefruit in their diet lost more weight, on average, than the group that didn't have any grapefruit in their diet. Based on the limited information in the article, it is difficult to say that eating grapefruit will help any obese person to lose weight. The last comment by the American Dietetic Association spokeswoman indicates that eating any fruit before a meal may help a person lose weight.

Graphs in the Media

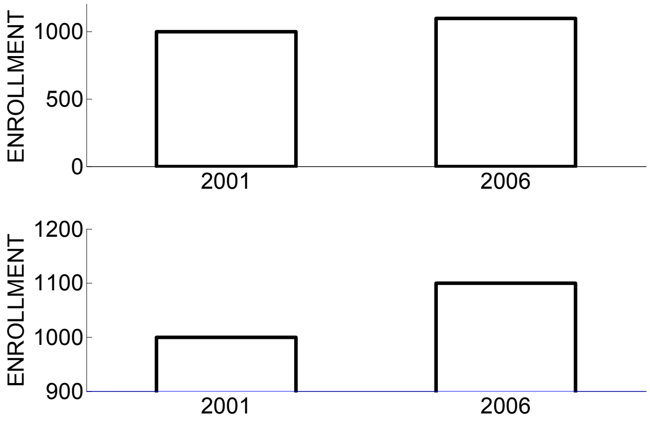
In newspapers, we will often see graphical displays that are used to convey numerical information. In some cases, the primary purpose of a graphical display is not to communicate information, but rather to entertain the reader. When we read a graph, we should think about the information that is being displayed and decide if the picture is an accurate representation of the information. Here are some specific questions to ask.

1. What is the main message of the graph?

2. What is the source of the data from which the numerical information has been computed?

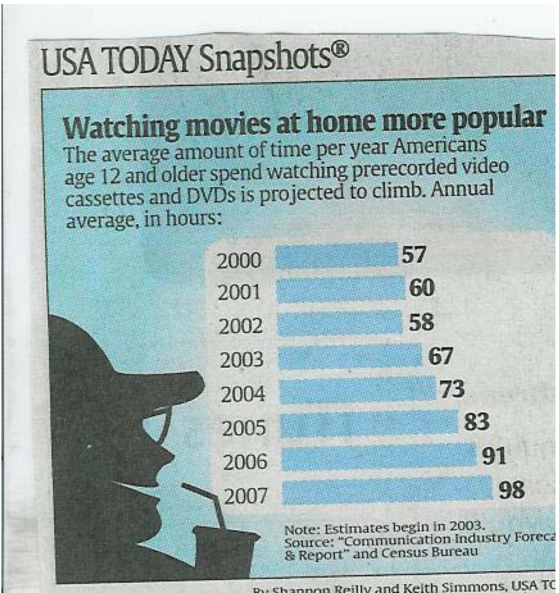
3. In a typical graph, there will be a primary display that shows the numerical information and other material (sometimes called chart junk) that is included to make the graph more attractive. Does the extra material distract the reader from seeing the primary graph?

4. Does the graph accurately represent the data? It is important that a graph follows the *area principle* where the area of the bar or figure should correspond to the number that one wishes to represent. For a simple illustration of this principle, suppose a university president wishes to construct a graph to show how the enrollment at the university has climbed in the last five years. The enrollment has increased from 1000 students in the year 2001 to 1100 students in the year 2006, a 10% increase. The figure below shows two sets of bar charts that could be used to display the enrollment numbers. The top graph obeys the area principle – the area of the bar for the year 2005 is 10% larger than the area of the bar for the year 2001 that does correspond to the actual increase in enrollment. In contrast, the bottom graph does not obey the area principle. Since the area of the 2006 bar is approximately twice the area of the 2001 bar, one gets the misleading impression that enrollment has doubled in the five-year period. The problem with this bottom graph is that the baseline enrollment in this graph has been set to 900. To give an accurate representation for bar graphs such as these, one always should use a baseline of zero instead of some arbitrary positive number.



Let's illustrate this critical perspective through several graphs that recently appeared in USA Today.

"Watching movies at home more popular"


USA TODAY Snapshots®

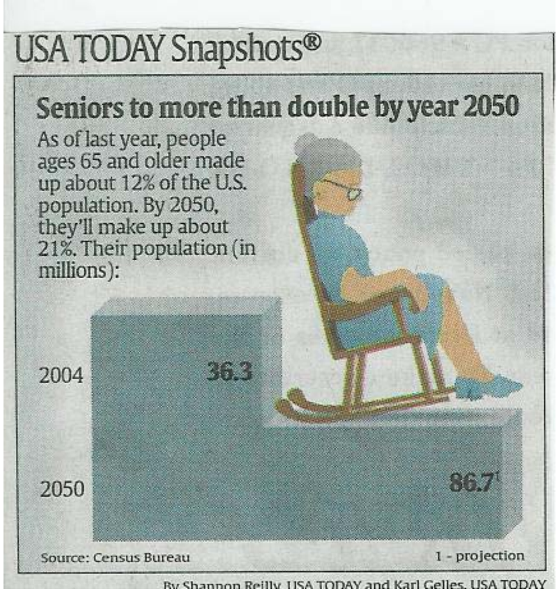
Watching movies at home more popular

The average amount of time per year Americans age 12 and older spend watching prerecorded video cassettes and DVDs is projected to climb. Annual average, in hours:

Note: Estimates begin in 2003.
Source: "Communication Industry Foreca
& Report" and Census Bureau

By Shannon Reilly and Keith Simmons, USA TC

This graph shows how Americans' watching of video cassettes and DVDs has increased over time. For each year from 2000 to 2007, this graph shows the average number of hours watching movies per year by Americans age 12 or older. Blue bars are used to display these averages. To be an accurate representation of these data, the lengths of the bars should be proportional to the data values. For example, since 91 hours is roughly 50% longer than 60 hours, the length of the bar for 91 hours should be approximately 50% longer than the length of the bar for 60 hours. This seems to be generally true, so this display seems to be an accurate picture of the information.

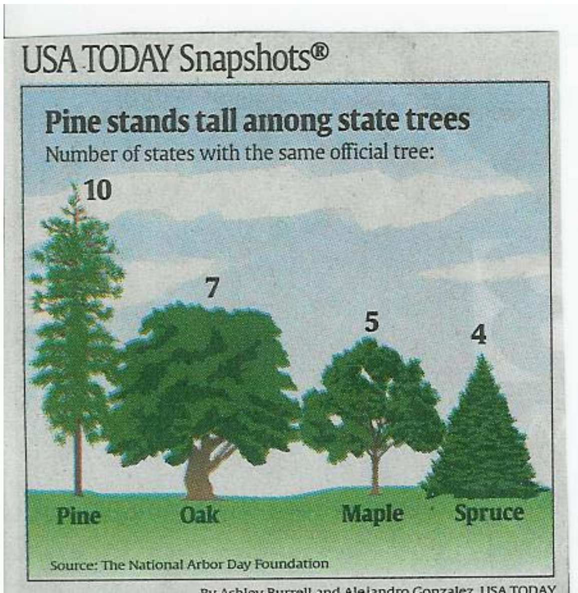


"Seniors to more than double by year 2050"

This graph is used to show visually how the percentage of seniors (those ages 65 or older) will dramatically increase in the coming years. The population sizes (in millions) of seniors in the years 2004 and 2050 are displayed using bars. The length of the top horizontal bar corresponds to 36.3 million, the population of seniors in 2004 and the length of the bottom horizontal bar corresponds to 86.7 million, the population of seniors that is projected in 2050. The lengths of the bars look reasonable – the length of the bottom bar is over twice as long as the length of the top bar that reflects the data. But this display is hard to read. One focuses on the graphic of the woman in the rocking year that is irrelevant to the message of the graphic. Also the two horizontal bars have been merged into a single grey step and one might be tempted to look at the vertical heights rather than the horizontal heights.

"Pine stands tall among state trees"

Every state in the United States has an official "state tree" and the point of this graphic is to show what trees are most popular among the state



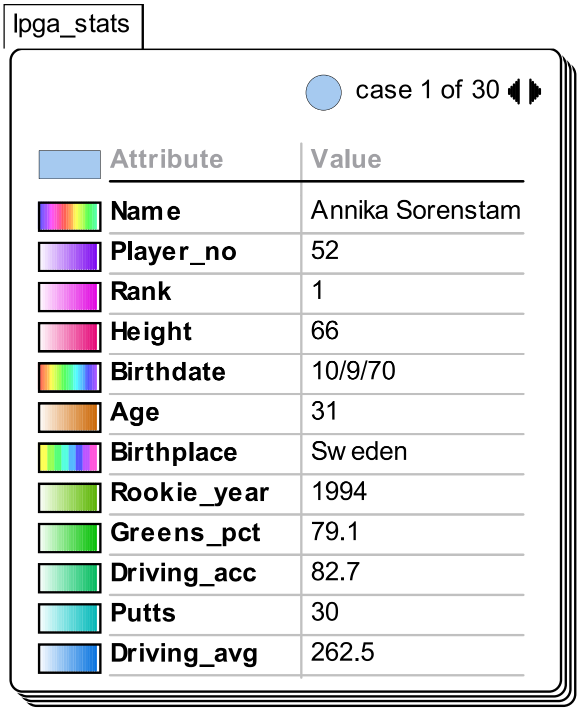
trees. We see that the pine tree is the most popular state tree (for 10 states), followed by the oak (7 states), maple (5 states), and spruce (4 states). Is this a good graphic of these data? The heights of the four tree diagrams do appear to correspond to the numbers – the height of the pine tree (10 states) is twice the height of the maple tree (5 states). But note that there are substantial differences in the widths of the diagrams. The area of the oak display is actually larger than the area of the pine display, which gives the impression the number of oak-tree states exceeds the number of pine-tree states. This graph does not obey the area principle.

TECHNOLOGY ACTIVITY: INTRODUCTION TO *TINKERPLOTS*

DESCRIPTION: This activity provides an introduction to the graphing package *Tinkerplots.* The author has collected data from the top 30 players in the Ladies Professional Golf Association in a recent year. For each golfer, the dataset contains

- NAME: the golfer's name
- PLAYER_NO: the number of the player as recorded on the LPGA database
- RANK: her rank in terms of money won (a rank of 1 corresponds to the golfer who has won the most)
- HEIGHT: her height in inches
- BIRTHDATE: her birth date
- AGE: her age in years when this data were collected
- COUNTRY: country of her birthplace
- ROOKIE: the year she was a rookie on the LPGA tour
- GREEN_PCT: green accuracy (the percentage of greens hit in regulation)
- DRIVING_ACC: driving accuracy (the percentage of drives that landed in the fairway)
- PUTTS: the average number of putts per round
- DRIVING_AVG: the average length of a drive (in yards)

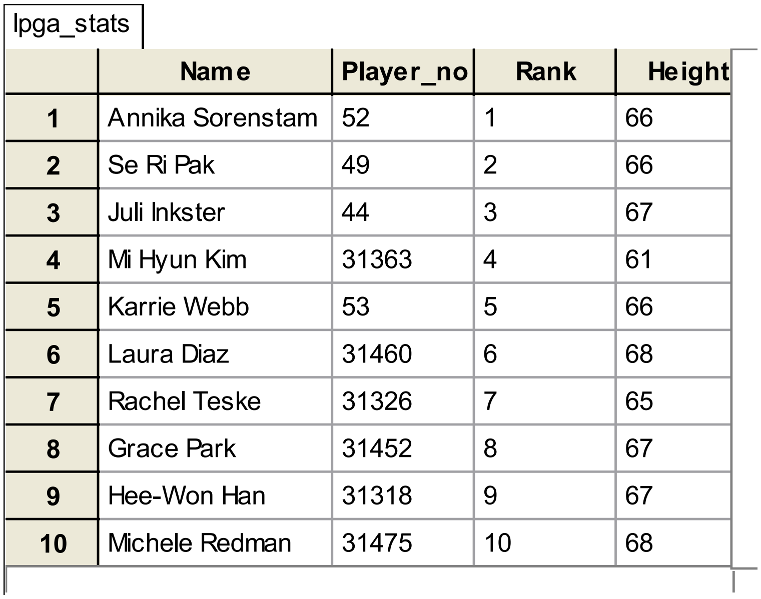
Start the program *Tinkerplots*. Import the data into *Tinkerplots* from the file lpga_stats.txt. You should see the Data Card for the first woman golfer, Annika Sorenstam:



Attribute	Value
Name	Annika Sorenstam
Player_no	52
Rank	1
Height	66
Birthdate	10/9/70
Age	31
Birthplace	Sweden
Rookie_year	1994
Greens_pct	79.1
Driving_acc	82.7
Putts	30
Driving_avg	262.5

You can look at other Data Cards by clicking on the arrows at the top of the card.

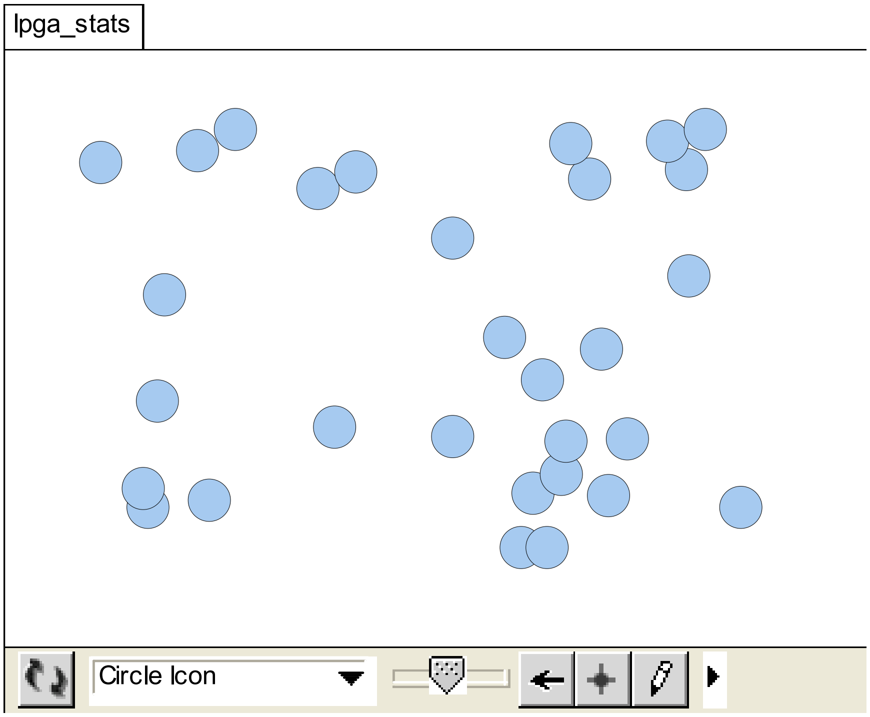
An alternative representation of the data is by a Case Table, where the players appear as rows, and the variables appear as columns in the table. You can get this representation in *Tinkerplots* by dragging down the Table icon.



lpga_stats

	Name	Player_no	Rank	Height
1	Annika Sorenstam	52	1	66
2	Se Ri Pak	49	2	66
3	Juli Inkster	44	3	67
4	Mi Hyun Kim	31363	4	61
5	Karrie Webb	53	5	66
6	Laura Diaz	31460	6	68
7	Rachel Teske	31326	7	65
8	Grace Park	31452	8	67
9	Hee-Won Han	31318	9	67
10	Michele Redman	31475	10	68

If you drag down a Graph object, you see the 30 golfers represented as a random collection of case icons or dots. This representation is analogous to taking the stack of "golf cards" and scattering them on the floor.



Tinkerplots allows one to organize these case icons in different ways.

- One can *separate* the icons in two or more groups by use of any of the variables.
- One can *order* the icons from highest to lowest on some variable.
- When the icons are separated into groups, one can *stack* the icons in each group.

Use one or more *Tinkerplots* graphs to answer the following questions. For each question, you should show your graph by either printing and pasting it into your book, or electronically copying the graph in a word processing document.

1. Look at the Birthplace variable. What countries are represented by these women golfers? Are these golfers predominately from particular countries? Are Americans well-represented in this group?

2. Look at the Age variable. Who are the youngest and oldest golfers in this group? What is a typical age among these golfers? Explain how you found this typical golfer.

3. Look at the Rookie Year variable. (This is the first year that the golfer played as a professional.) Which golfer has played the longest on the tour? Which golfer has played the shortest? On average, how many years have these golfers played on the LPGA tour. (Explain how you computed the average.)

4. Look at one of the golf statistic variables (GREEN_PCT, DRIVING_ACC: PUTTS: or DRIVING_AVG). It is desirable for a golfer to have …

- a *high* percentage of greens hit in regulation (GREEN_PCT)
- a *high* percentage of drives hit in the fairway (DRIVING_ACC)
- a *low* average number of putts (PUTTS)
- a *high* average driving length (DRIVING_ACG)

Can you pick out a couple of the more successful golfers and the less successful golfers with respect to this variable? Are the best golfers (the ones ranked from 1 to 5) good with respect to this particular golf statistic?

WRAP-UP

In this opening topic, we have been interested to the science of *statistics*. In a statistical investigation, we begin with some questions of interest, and then we collect relevant data that we think will be helpful in answering the questions. *Data* is the general term for information that we collect about individuals and *statistics* is a term used to describe numerical and categorical data. Generally, several variables are collected from each individual, and these variables differ by how they are *measured*. It is important to distinguish *quantitative* variables from *categorical* variables – the distinction is important in how we graph and summarize data. Whenever a statistical study is reported in the media, it is important to judge if the collected data are useful in addressing the main questions of the study. We were introduced to the use of graphical displays in the media. When we view a graph in a newspaper, we should think if the graphical display is a clear and accurate representation of the quantitative or categorical information.

Classroom Capsule: "Children's Well-Being"

Overview: The students will be introduced to an interesting UNICEF study on the well-being of children from 21 industrialized countries.

Objectives: The students will start to think about the meaning of "well-being of a child". After they list several attributes of a child's well-being, they read an article on a UNICEF student from the Baltimore Sun. From the article, they identify the relevant data that are collected and the variables that are studied.

Description: Ask the students the following questions. (They can answer these questions orally or by writing.)

1. Consider an average child who is brought up in the United States and an average child brought up in France. Who do you think would be better off? Why?
2. What do you think are the positive aspects of a child brought up in the U.S.?
3. What do you think are the negative aspects of a child brought up in the U.S.?
4. How would you measure a child's material well-being?

5. How would you measure a child's health and safety?
6. How would you measure a child's educational well-being?
7. How would you quantify a child's family and peer relationships?
8. How would you measure the health and risk behavior of a child?
9. How would you measure a child's subjective sense of well-being?

Next, read the following article from the *Baltimore Sun* that compares the well-being of children from the United States and other industrialized countries.

From the Baltimore Sun (May 4, 2007)

U.S. scores low in study on children's well-being By Julie Deardorff

America is one of the richest countries in the world. It's also one of the worst industrialized places for kids to grow up and has a greater percentage of depressed people than impoverished, war-torn nations do, according to two major studies.

The first unflattering finding comes from a recent UNICEF child-welfare study that measured everything from the number of books in the home to infant-mortality rates, drinking and drug use and the percentage of children who eat meals with their families.

Of 21 wealthy nations surveyed, the United States ranked second to last. Only Britain was worse. Child well-being was highest in the Netherlands, Sweden, Denmark and Finland, places that invest heavily in their children.

The problem isn't just that, compared with the European countries, the United States lacks day-care services and has poorer health and preventive-care coverage, which has left 9 million children without health insurance.

America finished dead last in terms of infant-mortality rates, vaccinations, the percentage of newborns with low birth weights and deaths from accidental injuries. We finished second to last when the researchers assessed a child's diet, physical activity and weight, exposure to violence and bullying and the number of 15-year-olds who smoke, drink and have sex.

And, in what could explain why we're among the most depressed people on Earth, according to a study of 14 nations conducted jointly by the World Health Organization and Harvard Medical School, we finished second to last when researchers examined relationships with family members and friends and family structure.

American children often don't eat the main meal of the day with their parents. Children say they don't spend time "just talking" to their parents. And they generally don't find their peers "kind and helpful," according to the study.

It shouldn't really be a surprise, then, that 9.6 percent of Americans suffer from depression or bipolar disorder, according to the WHO/Harvard study; that binge eating or drinking is up; or that children are heavily medicated for depression and attention-deficit disorder.

In material goods, American children have it all. But to make them feel loved, cherished and supported, they need family, community, a higher sense of purpose and meaningful cultural traditions - all things money can't buy.

Questions from reading the article:

1. What were the main questions addressed by the statistical studies described in the article?
2. What data were collected to answer the questions?
3. How did they collect the data?
4. What variables were measured in the data? Label each variable collected as quantitative or categorical.
5. What were the conclusions drawn from this statistical study? Do you believe the conclusions were valid based on the information provided?

Share and Summarize: Here are some important items that should be discussed for this example.

(a) There are many ways to describe a child’s well-being. A child will have a good upbringing if he or she has sufficient material possessions, is healthy and safe, has sufficient education, good family and peer relationships, does not engage in bad behavior or risky situations, and has a positive self-image.

(b) Although we can agree that all six aspects are important, it can be difficult to measure the corresponding attributes. For example, how do you measure the quality of a child’s family relationships?

(c) If we can find a suitable way of measuring a particular attribute, say self-worth, the next problem is how to collect data from a representative subset of the population of interest. For example, if we wish to compare the self-worth of American and French children, how do we take samples of the two populations?
(d) A newspaper article will typically focus on the conclusions of a statistical study and say little about the details of a study such as the manner in which the data were collected.
Application or Extension: Each of the students could be asked to find another newspaper article that compares the well-being of children of the United States and other countries. The student could discuss the article using the same questions given above. What information is given in your article that was not contained in the *Baltimore Sun* article?

EXERCISES

1. **Reading Articles**

Two articles that recently appeared in the media are copied below. Each article discusses the results of a statistical study. For each article, write a paragraph describing the different parts of the study. In particular, answer the following questions:

1. What were the main questions addressed by the statistical study?
2. What data were collected to answer the questions?
3. How did they collect the data?
4. What variables were measured in the data? Label each variable collected as quantitative or categorical.
5. What were the conclusions drawn from this statistical study? Do you believe that the conclusions are valid based on the information provided?

ARTICLE 1

Americans' Costly Health Care Not Better

By Jennifer Warner WebMD Medical News

May 5, 2004 -- Americans may pay more for health care than other countries, but they may not necessarily be getting the best medical care.

A new study shows the U.S. health care system ranks near the top on some but not all major health indicators and could take a lesson from the superior performance of other countries on several key areas, such as asthma and transplant surgery survival.

Researchers say the findings show there may be little evidence to back up the mantra often cited by policymakers, "Americans have the best medical care in the world."

"It is well known that the United States spends much more on health care per capita than other countries, and it is commonly assumed that we have the best health care system in the world, " says researcher Peter S. Hussey of the John Hopkins Bloomberg School of Public Health in Baltimore, Md., in a news release. "However, the results of our study show that the United States performs better than other countries in only a few areas, while performing worse in others."

"This raises the question of what Americans receive for all of the money devoted to health care," says Hussey.

Comparing Health Care Systems

Researchers say the study, which appears in the May/June issue of *Health Affairs*, is the first to use a universal set of standards to compare the quality of health care in five countries: Australia, Canada, England, New Zealand, and the U.S.

An international group of researchers collected and examined data on 21 health indicators that reflect the quality of medical care, including:

- 5-year cancer survival rates
- Cancer screening rates
- Avoidable events, such as suicide, asthma deaths, and smoking prevalence
- Vaccination rates
- Transplant survival rates

The study showed that no one country consistently scored among the best or worst overall.

For example, the U.S. had the highest breast cancer survival rate, cervical cancer screening rate, and lowest smoking (tied with Canada). But the U.S. performed among the worst in other areas, such as asthma-related deaths and survival after kidney and liver transplants. In fact the U.S. was the only country in which asthma-related deaths were increasing rather than decreasing.

The U.K. scored best in five of the eight avoidable event indicators, including pertussis and hepatitis B, but scored the lowest in five of the nine survival rate indictors and had the lowest cancer survival rate of the five countries studied.

"Each country in our study has areas of care where it can learn from the other countries and areas where it could teach others," says Hussey. "That tells us that there are opportunities for improvement in the quality of health care in all five countries."

SOURCES: Hussey, P.*Health Affairs*, May/June 2004; vol 23: pp 89-99. News release, John Hopkins Bloomberg School of Public Health.

ARTICLE 2:

Mass. to track all traffic stops after study finds profiling

The Northeastern University study, released Tuesday, was commissioned four years ago by the Legislature and included 366 departments — from cities and towns and the state police, to university, state transit and Amtrak police agencies. Just 92 got a passing grade.

Public Safety Secretary Edward A. Flynn warned against condemning departments that failed until more information can be gathered. The study caused the state to order 249 departments to collect a year's worth of data on all traffic stops.

"We are not today finding any agency guilty of having engaged in racial profiling," he said. "Data collection is not punishment."

Flynn said requiring agencies to collect more data will provide a clearer picture of racial profiling in Massachusetts.

"Every community deserves an explanation from its police department on how it uses its authority," Flynn said.

Northeastern used four statistical tests in analyzing 1.6 million traffic citations issued between April 1, 2001, and June 30, 2003: Ticketing resident minorities disproportionately more than whites; ticketing all minorities disproportionately more than whites; searching minorities more often than whites; and issuing warnings to whites more often than minorities.

According to the study, 15 police departments failed all four tests, 42 failed three tests, 87 failed two tests and 105 failed one. Among those that failed all four were Boston, Springfield and Worcester.

The Executive Office of Public Safety will use $1 million in grant money over the next six months to set up a uniform system for all police departments to report traffic stops, including those that do not result in any citations or written warnings. That information will be gathered over another year, then analyzed again.

Jack Collins, general counsel for the Massachusetts Chiefs of Police Association, said the added paperwork is a "witch hunt" and unnecessary.

Bishop Filipe Teixeira, a Catholic bishop from Brockton, said he's heard complaints in his community about minorities being targeted.

"We do have bad apples in the police departments," Teixeira said. "We have enough data. Let's get into action."

2. In a local or national newspaper, find three articles that describe conclusions from a statistical study. Summarize the information in each article by answering the questions given in Exercise 1.

3. **Graphical Displays in the Media**

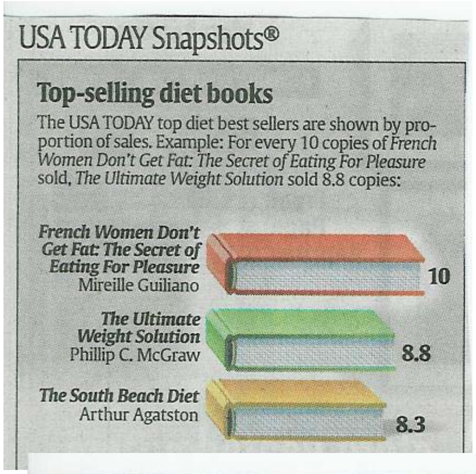
Four graphical displays from USA Today are shown below. For each display, answer the following questions:

a. What is the source of the information? (That is, where did the information come from?)

b. How is the information portrayed in the graphical display?

c. What is the basic message communicated in the graphic?

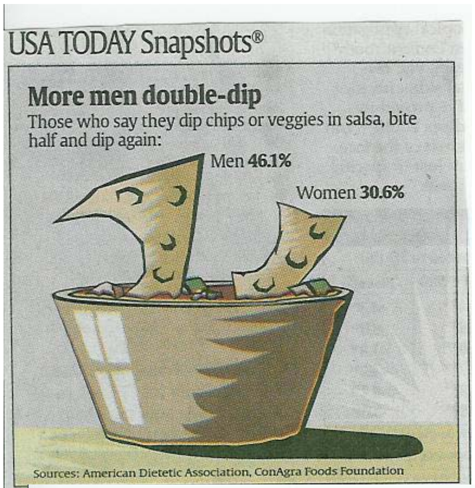
d. Is the graphical display an accurate representation of the data? In particular, does the graphical display follow the area principle, where the areas of the bars or shape are proportional to the data values?



USA TODAY Snapshots®

Top-selling diet books

The USA TODAY top diet best sellers are shown by proportion of sales. Example: For every 10 copies of *French Women Don't Get Fat: The Secret of Eating For Pleasure* sold, *The Ultimate Weight Solution* sold 8.8 copies:

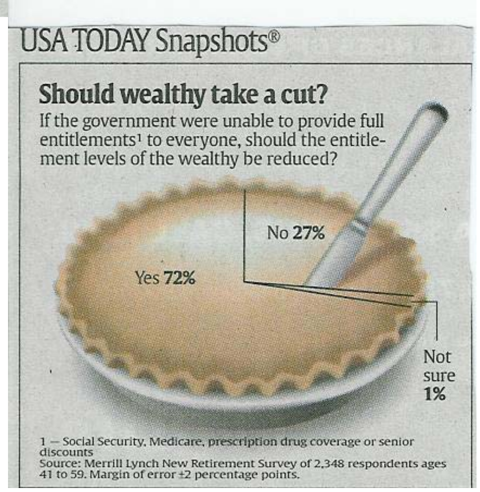


USA TODAY Snapshots®

More men double-dip

Those who say they dip chips or veggies in salsa, bite half and dip again:

Sources: American Dietetic Association, ConAgra Foods Foundation



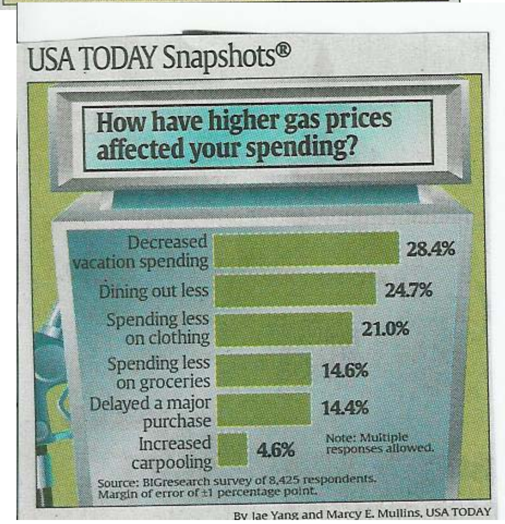
USA TODAY Snapshots®

Should wealthy take a cut?

If the government were unable to provide full entitlements[1] to everyone, should the entitlement levels of the wealthy be reduced?

[1] — Social Security, Medicare, prescription drug coverage or senior discounts

Source: Merrill Lynch New Retirement Survey of 2,348 respondents ages 41 to 59. Margin of error ±2 percentage points.



USA TODAY Snapshots®

How have higher gas prices affected your spending?

Note: Multiple responses allowed.

Source: BIGresearch survey of 8,425 respondents. Margin of error of ±1 percentage point.

By Jae Yang and Marcy E. Mullins, USA TODAY

TOPIC D2: GRAPHING DATA

SPOTLIGHT: WHO ARE THE BASEBALL PLAYERS?

Baseball has been traditionally called the “great American game,” and has been played at a professional level in the United States since 1871. Although the rules of the game have remained virtually unchanged since then, there have been dramatic changes in the types of men who play the game. These changes are described in *The New Bill James Historical Baseball Abstract*. In the beginning years, professional baseball was played primarily by U.S. immigrants and residents of eastern U.S. cities such as Brooklyn, Philadelphia, and Baltimore. One prominent group of immigrants who played baseball was the Irish. Toward the end of the century, many players were men who had played baseball in college. The creation of the American League in 1901 introduced many players from the Midwest region of the U.S. At this time, the baseball rosters had players from many different vocations. In the 1920’s, more men from rural parts of the country joined professional baseball, and the number of players with college educations declined. The 1930’s introduced more players from Southern U.S. and California.

Before 1947, only white Caucasians played baseball in the Major Leagues, and African-American baseball players played in their own Negro League. Although Jackie Robinson broke the color barrier in 1947 when he was signed by the Brooklyn Dodgers, professional baseball was slow to adopt African-Americans and other minorities to their teams. But by the 1960’s, a large number of African-American and Latin-Americans played the game. The 1970’s saw the introduction of a suburban generation of American players who learned baseball at an early age through organized leagues; during this time the number of African-American players reached its peak. In recent years, professional baseball has seen a rise in the number of Latin-American players and American players of Latino ancestry, and a decline in the number of African-American players. Professional baseball has also recently seen the introduction of players from the Far East, especially Japan.

PREVIEW

In this topic, you begin your exploration of data. You saw in Topic D1 that variables differ by how they are measured. Now you'll come to see that the appropriate graph for a particular variable depends upon the measurement type. By graphing quantitative data, you will be introduced to the notion of a dataset's distribution, which illustrates the variability in the data values. You will learn about a distribution's shape and begin to make informal judgments about the center and spread of the values. You'll also gain experience interpreting different graphs.

In this topic your learning objectives are to:

- Understand how to construct and interpret different graphs for a single collection of data.
- Understand there are choices in constructing a graph and the appearance and the usefulness of a graph can be dependent of these choices.
- Use the graph to write a descriptive paragraph about the distribution of a dataset that contains information about the shape, center, spread, and any unusual characteristics of the distribution.
- Compare the usefulness of different types of graphs in representing a distribution.

NCTM Standards

✓In Grades 6-8, students should select, create, and use appropriate graphical representations of data.

✓In Grades 6-8, students should discuss and understand the correspondence between data sets and their graphical representations.

✓In Grades 9-12, all students should for, univariate measurement data, be able to display the distribution and describe its shape.

GRAPHING CATEGORICAL DATA

The opening spotlight describes the dramatic changes in the backgrounds of professional baseball players over the years. You may be asking yourself, "Who are the professional baseball players today?" To look more carefully at the backgrounds of current players, here are the countries of birth for each Major League Baseball player born in the year 1975. (These are the players that would be 30 years old in the 2005 baseball season. D.R. is the Dominican Republic and P.R. is Puerto Rico):

U.S.
D.R.
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
Venezuela
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
P.R.
U.S.
U.S.
U.S.
P.R.
U.S.
U.S.
D.R.
U.S.
U.S.
South Korea
U.S.
U.S.
U.S.
P.R.
U.S.
U.S.
U.S.
U.S.
P.R.
D.R.
U.S.
U.S.
U.S.
U.S.
U.S.
P.R.
U.S.
U.S.
U.S.
U.S.
Australia
U.S.
U.S.
D.R.
U.S.
U.S.
Venezuela
D.R.
U.S.
D.R.
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
D.R.
D.R.
Mexico
Mexico
U.S.
U.S.
Mexico
Venezuela
Canada
U.S.
Venezuela
D.R.
U.S.
D.R.
Cuba
Cuba
Venezuela
Venezuela
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
Canada
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
Canada
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
D.R.
Panama
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
D.R.
Mexico
D.R.
Nicaragua
U.S.
D.R.
Japan
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
Panama
D.R.
U.S.
U.S.
U.S.
U.S.
P.R.
U.S.
U.S.
Venezuela
D.R.
U.S.
U.S.
U.S.
Venezuela
Cuba
U.S.
Mexico
Venezuela
Cuba
D.R.
U.S.
U.S.
U.S.
U.S.
D.R.
D.R.
D.R.
D.R.
U.S.
U.S.
U.S.
D.R.
U.S.
U.S.
U.S.
Venezuela
U.S.
Colombia
U.S.
U.S.
D.R.
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
Venezuela
U.S.
U.S.
Panama
U.S.
Curacao
U.S.
D.R.
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
Japan
D.R.
U.S.
Cuba
U.S.
P.R.
Venezuela
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
U.S.
D.R.
U.S.
Venezuela
Canada

Display D2.1: Countries of birth for all Major League Baseball players born in 1975. (Source: Lahman baseball database, baseball1.com.)

A first step in organizing these categorical data is to construct a **frequency table** where you list the possible countries of origin and the corresponding counts or **frequencies** of each country. Looking at the frequency table below (Display D2.2), you see that a large number of players (135) are from the United States, but many other countries are represented in Major League Baseball. You may recognize that many of these countries are from the Latin America region. Because there are many countries with small counts, it is helpful to collapse the countries into the three categories "U.S.," "Latin

America" and "Other." A frequency table using these new categories is also shown below (Display D2.3).

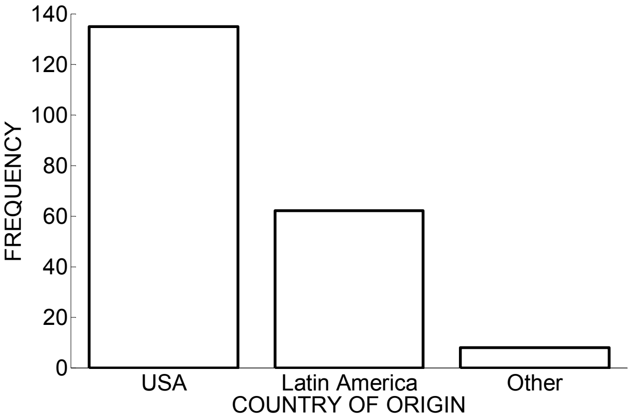
Country of Birth	**Frequency**
Australia	1
Canada	4
Columbia	1
Cuba	5
Curacao	1
D.R.	26
Japan	2
Mexico	5
Nicaragua	1
P.R.	7
Panama	3
South Korea	1
U.S.	135
Venezuela	13

Display D2.2: Frequencies of countries of birth for all Major League Baseball players born in 1975.

Country of Birth	**Frequency**
U.S.	135
Latin America	62
Other	8

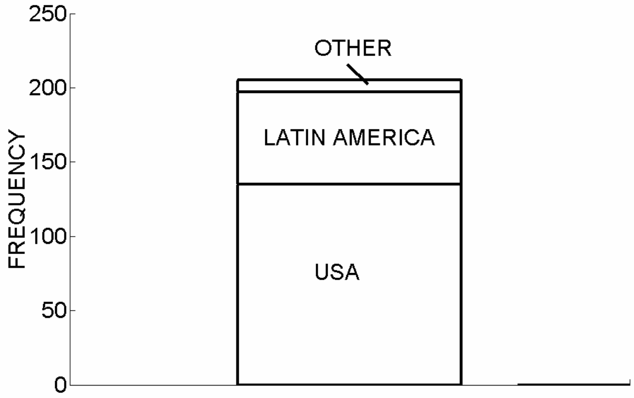
Display D2.3: Frequencies of countries of birth, by categories, for all Major League Baseball players born in 1975.

After you construct a frequency table, you are typically interested in describing the relative sizes of the category frequencies, and you can do this by constructing a suitable graph. There are several types of "good" graphs for representing categorical data –a *bar chart*, a *segmented bar chart*, and a *pie chart* are illustrated in this section. To construct a **bar chart,** you list the possible categories on one axis and then construct a bar for each category along the second axis, where the height (or length) of the bar corresponds to the frequency.



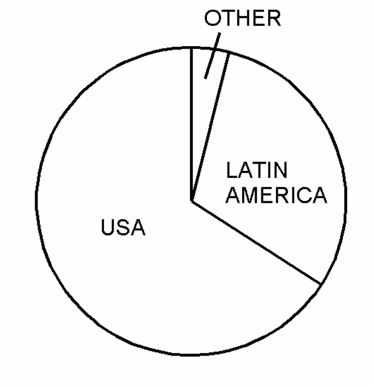
Display D2.4: Bar chart of countries of birth for all Major League Baseball players born in 1975.

A **segmented bar chart** is similar to a bar chart, but you stack the bars for the different categories into a single bar. The height (or length) of each section of the combined bar is proportional to the corresponding category frequency. You will see in Topic D4 that segmented bar charts are useful for comparing two or more batches of categorical data. A **pie chart** represents the entire group of ballplayers as a circle, and each country is represented by a slice of the pie. For a pie chart, the areas of the slices are proportional to the frequencies of the categories.



[ART EDITS: change "USA" to "U.S."; insert "Country of Birth" as horizontal axis label; use title capitalization for the categories "Other" and "Latin America," and for the vertical axis label "Frequency"]

Display D2.5: Segmented bar chart of countries of birth for all Major League Baseball players born in 1975.



Display D2.6: Pie chart of countries of birth for all Major League Baseball players born in 1975.

This data exploration can be put in the context of the four basic components of statistics described in Topic D1. You began by asking "Who are the current baseball players?" To help answer this question, the places of birth of current major league baseball players born in the year 1975 were collected. These places of birth were grouped by country and the relative frequencies of countries were displayed using different types of graphs. By looking at any one of these three graphs, it is possible to draw a general conclusion about the nationality of current ballplayers. It is clear from the graphs that a majority of the ballplayers are born in the U.S. and most of the non-American-born players are from Latin America.

What is the best graph for representing categorical data? It is important that a graph present an accurate representation of the data in the frequency table. To be an accurate display, a graph should obey the **area principle,** where the area of the bar or object corresponding to a particular category is proportional to its frequency. All of the graphs above obey the area principle. (Recall that you saw several misleading graphs for categorical data that did *not* obey the area principle in Topic D1.)

Although all three graphs are accurate representations of the data, pie charts will not be featured in this book. A single pie chart helps you see the relative sizes of the counts for a single batch of categorical data. However, when several pie charts are used, it becomes difficult to compare batches of data because you have to visually compare the sizes of angles of the slices of the pie chart. It is generally easier for people to make visual comparisons of lengths of lines, and so bar charts and stacked bar charts are more effective for graphical comparison of batches. (The use of these graphs will be revisited in topic D4.)

PRACTICE: GRAPHING CATEGORICAL DATA

Suppose you are interested in buying a used sedan car. To help understand what types of cars are available under a cost of $10,000, you visit the website *usedcars.com* and it gives you a list of 50 available sedans that are located within 30 miles of your hometown. The table below gives the details of each car. (For mileage, under 70,000 miles is "low," 70,000–100,000 miles is "medium," and over 100,000 miles is "high.")

Model	**Year**	**Color**	**Mileage**
Chevrolet	1996	blue	high
Mercury	1999	white	high
Chevrolet	1997	blue	medium
Ford	1997	tan	high
Plymouth	1997	burgundy	medium
Mercury	1997	red	high
Plymouth	1998	green	medium
Dodge	1999	red	low
Ford	1999	white	medium
Mercury	1997	blue	high
Ford	1996	copper	medium
Mitsubishi	2001	green	low
Ford	2000	white	low
Buick	1997	white	high
Pontiac	1998	green	medium
Mercury	1999	blue	medium
Buick	1999	green	medium
Ford	2000	green	low
Cadillac	1997	green	medium
Buick	2000	green	low
Kia	2002	silver	low
Chrysler	1999	white	medium
Dodge	2000	blue	medium
Oldsmobile	1999	green	low
Buick	2000	maroon	medium
Pontiac	2000	silver	low
Ford	2001	blue	medium
Ford	2002	gray	medium
Chevrolet	2001	black	low
Oldsmobile	2000	green	medium
Chevrolet	2000	beige	low
Dodge	2002	brown	low
Ford	2002	red	low
Chevrolet	2000	beige	medium
Honda	1999	green	medium
Mercury	1999	tan	low
Mercury	1998	silver	medium
Mitsubishi	2002	green	low
Ford	2001	maroon	low
Saturn	2002	black	low

Chevrolet	2000	brown	low
Oldsmobile	2000	green	medium
Pontiac	2000	white	medium
Kia	2002	maroon	low
Dodge	2001	tan	medium
Pontiac	2002	burgundy	low
Nissan	2001	white	low
Ford	2002	gold	low
Honda	1999	silver	medium
Ford	2003	tan	low

Display D2.7: Details of 50 used sedan cars. (Source: *usedcars.com*)

1. Suppose you are interested in the manufacturers of these “cheap” used cars. So you classify the car models into four groups: those manufactured by General Motors (Cadillac, Chevrolet, Oldsmobile, Pontiac, Saturn and Buick), Ford (Ford and Mercury), Chrysler (Chrysler, Dodge, and Plymouth), and foreign (all others). Construct a frequency table and bar chart for the car model.

2. REFLECTION Based on your bar chart, make some comments about what you have learned about the models of these 50 cars. (For example, what are the popular and unpopular car manufacturers among these used cars?) Can you explain why there are so few foreign cars listed?

3. Suppose you consider the model year to be a categorical variable. Construct a segmented bar chart and a pie chart for this variable.

4. Construct a graph of the colors of the cars and use this graph to describe the popular and unpopular colors in this group of used cars.

5. REFLECTION If you were to purchase a used car, what other variables would you collect besides the ones listed above? Explain why these additional variables would be important to you.

GRAPHING QUANTITATIVE DATA – DISTRIBUTION AND SHAPE

In topic D1, you were introduced to a dataset from the U.S. Census Bureau that contained the percentages of adults completing high school and college for all states and the District of Columbia. (See Display D1.1 on page XXX.) Here are all of the data for the high school completion rates.

State	**Completed High School (percent)**	**State**	**Completed High School (percent)**	**State**	**Completed High School (percent)**
Alabama	77.5	Kentucky	78.7	North	85.5

				Dakota	
Alaska	90.4	Louisiana	80.8	Ohio	87.0
Arizona	85.1	Maine	89.3	Oklahoma	86.1
Arkansas	81.7	Maryland	85.7	Oregon	88.1
California	81.2	Massachusetts	85.1	Pennsylvania	85.7
Colorado	89.7	Michigan	86.2	Rhode Island	81.3
Connecticut	88.2	Minnesota	90.8	South Carolina	83.0
Delaware	86.1	Mississippi	80.3	South Dakota	91.8
District of Columbia	83.2	Missouri	86.6	Tennessee	79.9
Florida	84.0	Montana	89.6	Texas	79.2
Georgia	82.6	Nebraska	90.4	Utah	90.7
Hawaii	87.4	Nevada	82.8	Vermont	90.0
Idaho	86.2	New Hampshire	88.1	Virginia	86.6
Illinois	85.5	New Jersey	87.3	Washington	91.8
Indiana	84.6	New Mexico	82.2	West Virginia	77.1
Iowa	89.7	New York	82.5	Wisconsin	86.7
Kansas	88.1	North Carolina	79.2	Wyoming	90.0

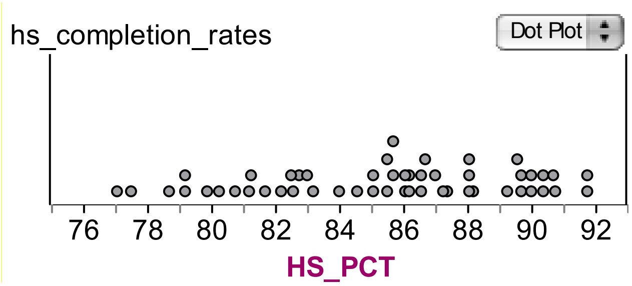
Display D2.8: The percentages of adults (25 years or over) that have have completed high school, by state. (Source: U.S. Census Bureau, *Educational Attainment of the Population 25 Years and Over, By State, Including Confidence Intervals of Estimates: March 2000,* released December 19, 2000.)

You may recall from Topic D1 that these completion rates were collected in order to answer questions similar to the following:

- What is a typical high school completion rate for the states?
- Are there sizable differences in the completion rates for states? Can states with high and low rates be identified?
- Are there particular states that stand out (either in the high end or in the low end) in terms of getting their high school students to graduate?

The first step in exploring these completion rates and answering the questions is to construct a graph. For quantitative data, a graph can help illuminate patterns—such as frequently occurring values, clusters, and gaps—that are not easily visible from a table of values. In a sense, a graph gives you a way to see how the variable actually varies.

A simple graph that is easy to construct by hand is a **dotplot** (often called a *line plot* in the school curriculum). You draw a number line covering the smallest and largest data values and then you place a dot on the number line for each data value. Here is a dotplot of the high school completion rates. This dotplot was created with *Fathom,* a software package designed to collect, explore, and analyze data. Note that the dots are placed on a higher level when they start to overlap. For example the second dot at the value 79.2 is placed over the first dot at 79.2, and the three dots at the value 88.1 are placed on three levels.



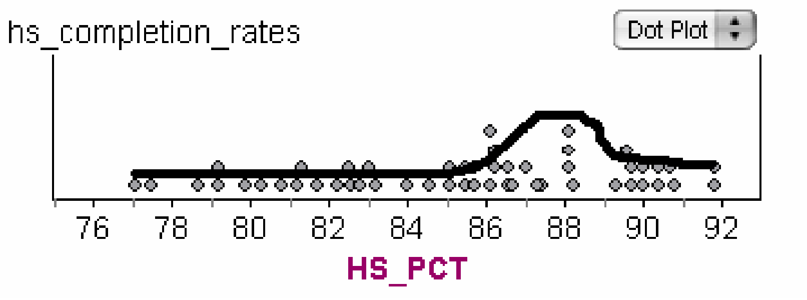
Display D2.9: Dotplot of high school completion rates.

From the graph alone, you may begin to see answers to the questions at hand. For example, there are fairly sizeable differences—ranging between 77 and 92 percent—in the high school completion rates. And because so many values are clustered between 85 and 91 percent, a typical completion rate might be somewhere in this interval.

What you see from the dotplot are the values that the variable takes and how often each occurs; this is called the **distribution** of high school completion rates. What do you look for in this distribution?

SHAPE

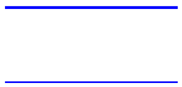
For one thing, you look at the general **shape** of the data. To help understand the shape of the data, you can draw a smooth curve over the dots in the dotplot. The curve helps you focus on the general pattern of the data rather than small gaps and clusters amongst the individual values.



Display D2.10: Dotplot of high school completion rates with a smooth curve sketched.

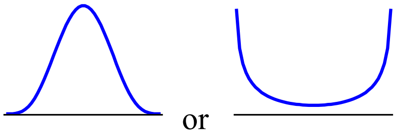
There are several common shapes that you may see in the smooth curve that you draw over the dotplot.

- *Uniform*



In a **uniform** distribution, each data value occurs at roughly the same frequency. The overall shape of the distribution is flat.

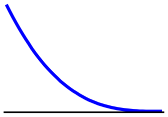
- *Symmetric*



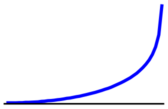
Symmetric describes a distribution in which the data values drop off (or increase) at the same rate at the left and right ends. You can imagine dividing the data into two halves, where the left half is approximately a mirror image of the right.

The symmetric distribution illustrated on the left—when the majority of the data values are in the middle—is frequently described as *bell-shaped* in school textbooks.

- *Skewed*

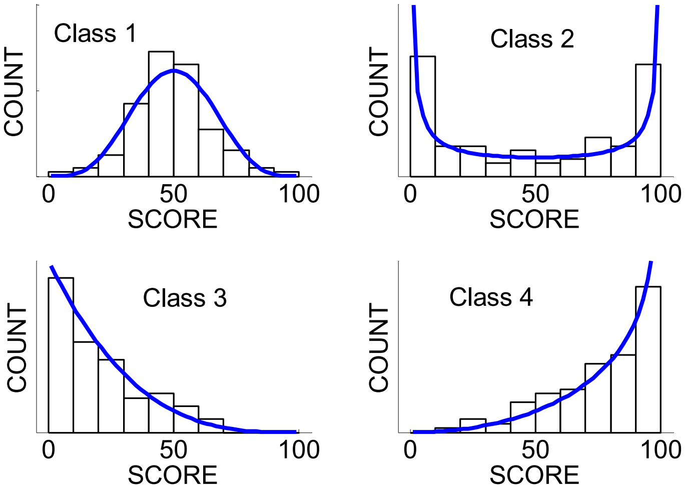


A distibution is called **skewed right** when the data values pile up at the low end and the frequency of data values decrease slowly as you move right toward larger values.



Conversely, **skewed left** is when the data values pile up for large values and the frequencies decrease slowly as you move left toward smaller values.

To help understand different shapes, consider these graphs of four batches of test scores corresponding to some hypothetical classes (Display D2.11). The shapes of the datasets are indicated by smooth curves drawn over the graphs. The scores of Class 1 have a symmetric shape with the frequencies of students scoring less than 50 dropping off in the same way as the frequencies of students scoring more than 50. Class 2's scores have a symmetric "u-shaped" distribution. For this class it was common to score close to 100 or close to 0. The scores of Class 3 are skewed right where low scores were the most common, and Class 4 scores are skewed left with a large number of high scores. You would probably be happiest with the test scores of Class 4, although the shape of the test scores for Class 1 is customary for standardized tests.



Display D2.11: Distributions of test scores for four hypothetical classes.

For the high school completion data, you see from the dotplot in Display D2.10 that there is a tight clump of states with high completion rates and there is a wide interval of states with low completion rates between 77 and 84 percent. So, the shape of the high school completion rates is somewhat skewed left.

CENTER

When examining a distribution, you also look for a **center,** or representative data value. You will learn more precise definitions in later topics, but for now look for a representative value that is located in the center of the distribution.

By looking at the dotplot in Display D2.10, you see that the center of the entire distribution is about 85. So, it might be reasonable to say that 85 percent is a representative high school completion rate for a state.

SPREAD

Along with noticing a central value, you want to say something about the **spread** of the percentages. The spread helps illuminate characteristics of the data that the center alone cannot. For example, consider two hypothetical classes of three students that take the same test. One class scores 40, 70, and 100, and the other class scores 69, 70, and 71. While both classes have a center of 70—and would appear to be identical if you were told only the center—describing the spread of the values would emphasize how drastically different the two classes performed.

Like the center, there are different ways of defining spread, and you will learn precise definitions later. One quick way to describe the spread is to name the interval that contains all of the values. From inspecting the data table in Display D2.8, you see that the minimum and maximum rates are 77.1 and 91.8 percent, respectively. So, the high school completion rates for all states and the District of Columbia fall in the interval [77.1, 91.8].

The interval that contains all of the data might be wide if there are one or more unusually small or large values. As an alternative, you can give an interval that contains a particular high percentage of the data. Looking at the dotplot in Display D2.9, you see that most of the values fall in the 80's. Actually, 37 states have a high school completion rate greater than or equal to 80 percent but less than 90 percent. Because there are 51 states and $\frac{37}{51} \approx 0.73$, you may further describe the spread by saying, "About 73% of the states have high school completion rates between 80 and 90 percent."

INTERESTING FEATURES

In addition to talking about the shape, center, and spread of the data, you also want to point out any interesting features of the distribution. These could include

- unusually small or large data values that stand out from the majority of the data; these are commonly called **outliers**
- gaps in the data
- clusters of several data values

For the high school completion rate data, no unusually small or large rates stand out. But it is interesting to note that the two smallest rates (77.1 and 77.5) correspond to West Virginia and Alabama, two southern states, and the two highest rates (91.8 and 91.8) correspond to South Dakota and Washington, two northern states. This suggests that there might be a relationship between a state's high school completion rate and its location. You'll explore this relationship further in a later topic.

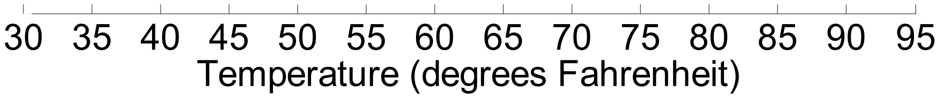
PRACTICE: GRAPHING QUANTITATIVE DATA

Here are the daily high temperatures (in degrees Fahrenheit) for the city of Atlanta, Georgia, in the month of March 2005.

March 2005						
		41	48	55	61	66
61	67	55	47	54	60	75
76	62	59	50	40	55	58
65	66	58	67	70	79	80
67	64	74	77	65		

Display D2.12: Daily high temperatures for Atlanta, Georgia. (Source: Georgia Automated Environmental Monitoring Network http://www.griffin.uga.edu/aemn/cgi-bin/AEMN.pl?site=GAAA&report=hi)

1. Construct a dotplot of the temperatures on the scale below.



2. Draw a smooth curve over the dotplot and describe the shape of the distribution of temperatures.

3. Give a center value for the temperatures and describe the spread by giving an interval that contains a high percentage of the data.

4. Are there any interesting features about these data, such as outliers, gaps, or clusters?

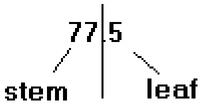
5. REFLECTION Is there any possible explanation for the large spread in temperatures in the data? (Think about how the weather in Atlanta changes in March.)

6. REFLECTION If you collected and graphed the daily high temperatures for Miami, Florida, in March 2005 and compared the distribution to the one above for Atlanta, would you expect to see any differences in the shape, center, and spread? How do you think the distribution of daily high temperatures in March 2005 for Minneapolis, Minnesota, would compare? Explain.

STEMPLOT

The dotplot is one way of graphing a single batch of quantitative data. There are alternative graphs that may also be useful. A **stemplot** is a clever way of quickly organizing a batch of data by hand using the digits of the data values. (Some school textbooks use the longer term *stem-and-leaf plot*.)

To construct a stemplot of the high school completion rates, you take each data value and divide it into two parts, called the **stem** and the **leaf**. For example, Alabama's rate of 77.5 percent could be written as a whole-number stem and a decimal leaf. A vertical line separates the two parts.



To create the entire stemplot, you first write down a chronological list of all possible stems on the left (here 77 to 91); you draw a vertical line to the right of the stems; and then you record each leaf after the appropriate stem. Doing this alphabetically for all of the states in Display D2.8, you get the stemplot shown below left (Display D2.13). To better summarize this data, it is helpful if you rewrite the leaf values in ascending ordered after each stem. This creates the stemplot with ordered leaves shown below right (Display D2.14). Notice that each stemplot includes a key at the bottom that helps the reader interpret the data values.

```
77 | 51
78 | 7
79 | 292
80 | 83
81 | 723
82 | 6825
83 | 20
84 | 06
85 | 157157
86 | 1226167
87 | 430
88 | 2111
89 | 7736
90 | 484700
91 | 88

77|5 means 77.5 percent
```

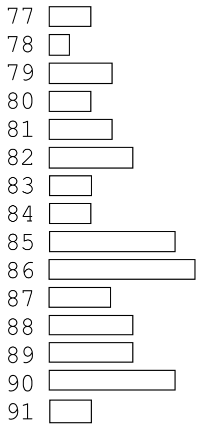
Display D2.13: Basic stemplot of high school completion rates.

```
77 | 15
78 | 7
79 | 229
80 | 38
```

```
81 | 237
82 | 2568
83 | 02
84 | 06
85 | 115577
86 | 1122667
87 | 034
88 | 1112
89 | 3677
90 | 004478
91 | 88

77|5 means 77.5 percent
```

Display D2.14: Stemplot of high school completion rates with ordered leaves.

A stemplot effectively groups the data into equal-sized intervals, or **bins**. For example, you see from the leaves of the first line of either stemplot that two high school completion rates are in the bin between 77.0 and 77.9 percent (specifically 77.1 and 77.5). Similarly, one rate falls in the bin between 78.0 and 78.9; three rates fall in the bin between 79.0 and 79.9; and so on. To emphasize the number of rates that fall into each bin, you can imagine replacing the leaves by bars where the length of each bar is proportional to the number of leaves.



Display D2.15: Stemplot of high school completion rates with leaves replaced by bars.

Note that, like the dotplots in Displays D2.9 and D2.10, the stemplots above still show the shape of the distribution of high school completion rates. The overall shape is still skewed left, with the frequency of values decreasing as you move toward smaller values. (If you have trouble visualizing this, imagine turning the stemplot 90° counterclockwise so that the stems form a horizontal number line.) And you still see that the most of the values fall between 80 and 90 percent.

However, unlike a dotplot, a stemplot allows you to see the actual data values. For example, you see from the stemplot in Display D2.14 that the two largest high school completion rates are 91.8 and 91.8 percent (corresponding to Washington and South Dakota). The dotplot in Display D2.9, in contrast, shows that the two highest rates are very close to 92 percent, but you would have to guess the decimal precision. So, one of the advantages of a stemplot is that you can see the shape and characteristics of the distribution without losing sight of the data values. For this reason, a stemplot is a useful and very detailed graph for relatively small data sets, say up to 50 values.

PRACTICE: CONSTRUCTING STEMPLOTS

When you construct a stemplot, a fundamental decision is how to divide each data value into the stem and leaf. Your decision may or may not lead to a "good" stemplot that makes it easy to see the data distribution.

The table below gives a variety of data about the states in the Midwest and Northeast sections of the U.S. Driver Rate is the number of licensed drivers per 1000 residents in August 2002; Farms is the number of farms (in thousands) in 2001; Density is the population per square mile of land area in 2000; and Elevation is the highest point in the state, measured as feet above sea level.

State	**Driver Rate (per 1000)**	**Farms (thousands)**	**Density (persons per square mile)**	**Elevation (feet above sea level)**
Illinois	641	76.0	224.6	1255
Indiana	654	63.0	170.5	1257
Iowa	667	93.5	52.3	1670
Kansas	710	63.0	52.9	4039
Michigan	697	52.0	175.9	1979
Minnesota	598	79.0	62.5	2301
Missouri	689	108.0	81.7	1772
North Dakota	715	30.3	9.2	3506
Ohio	723	78.0	277.8	1549
South Dakota	720	32.5	10.0	7242
Wisconsin	703	77.0	95.5	1951
Connecticut	779	3.9	706.9	2380
Maine	722	6.7	41.7	5267
Massachusetts	707	6.0	813.7	3487

New Hampshire	752	3.1	140.4	6288
New Jersey	672	9.6	1143.9	1803
New York	573	37.5	402.7	5344
Pennsylvania	670	59.0	274.2	3213
Rhode Island	624	0.7	1013.3	812
Vermont	831	6.6	66.3	4393

Display D2.16: Driver rates, number of farms, population density, and highest elevation for Northeast and Midwest states. (Source: *The World Almanac and Book of Facts, 2004*.)

1. Construct two stemplots for Driver Rate, as described below. Provide a key for each stemplot.

a. One stemplot where the stems are the hundreds and tens places and the leaves are the ones places. (For example, Illinois' rate of 641 would have a stem of 64 and a leaf of 1).

b. A second stemplot where the stems are the hundreds places and the leaves are the tens. (Illinois would have a stem of 6 and a leaf of 4.)

SPECIAL NOTE: It is common practice to record only a single digit for each leaf. Hence, for the stemplot in 1b, you drop the units places. Although you can either truncate the superfluous digits or round, it is usually quicker to truncate. Truncating also makes it easier to find a particular data value within the stemplot.

2. Which stemplot, 1a or 1b, gives a better display of the distribution of the Driver Rate data? Explain.

3. For each of the remaining variables (Farms, Density, and Elevation), construct a stemplot using the "best" division between stem and leaf. If the choice is not obvious, construct two stemplots using different divisions and choose the better graph.

HISTOGRAM

The **histogram** is a traditional method of graphing quantitative data, especially useful when you have a dataset with a large number of observational units. Similar to the modified stemplot in Display D2.15, a histogram uses bars to show the frequency of data values that fall into classes. To construct a histogram, you first create equal-sized bins that cover all of the data values; you classify all of the data values into these bins; and then you graph the bin frequencies using connected bars. One sticky point is what to do when a data value falls on the boundary between two bins; in this case, it is customary to place the data value into the bin on the right.

Recall that all of the high school completion rates in Display D2.8 fall between 77.1 and 91.8 percent. So, you could create these bins of width 2 starting from 76:

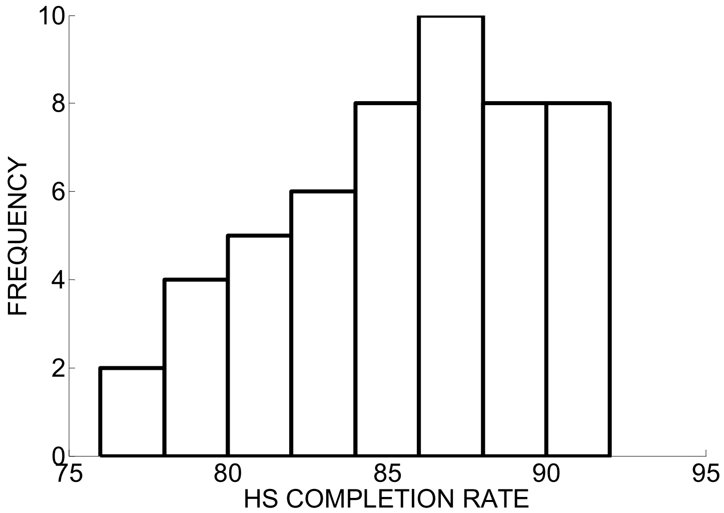
[76, 78), [78, 80), [80, 82), [82, 84), [84, 86), [86, 88), [88, 90), [90, 92)

Counting the number of completion rates that fall in [76, 78), [78, 80), and so on you obtain this frequency table.

Bin	[76,78)	[78,80)	[80,82)	[82,84)	[84,86)	[86,88)	[88,90)	[90,92)
Frequency	2	4	5	6	8	10	8	8

Display D2.17: Frequency table for high school completion rates.

This table tells us that two states have high school completion rates between 76 and 78 percent, four states have rates between 78 and 80 percent, and so on. To construct the histogram, you graph the frequencies (2, 4, 5, ...) against the bin boundaries using a bar chart. Notice, however, that the bars touch each other, thereby implying that the data is a continuous quantitative variable rather than a discrete categorical variable.



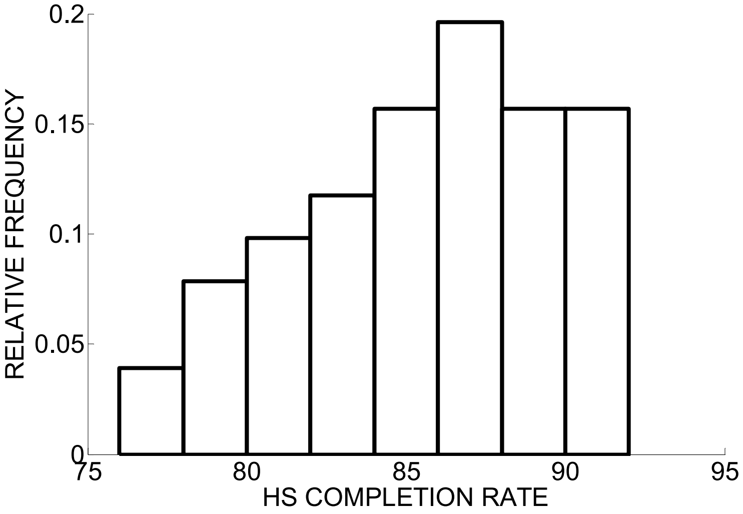
Display D2.18: Histogram of high school completion rates.

Sometimes it is helpful to further convert the bin frequencies to proportions, or **relative frequencies,** by dividing each frequency by the total number of values (51). Doing this for each bin, you get this **relative frequency table**:

Bin	[76,78)	[78,80)	[80,82)	[82,84)	[84,86)	[86,88)	[88,90)	[90,92)
Frequency	2	4	5	6	8	10	8	8
Proportion	0.039	0.078	0.098	0.118	0.157	0.196	0.157	0.157

Display D2.19: Relative frequency table of high school completion rates.

To construct a **relative frequency histogram,** you graph the proportions (0.039, 0.078, …) against the bin boundaries. Notice that the relative frequency histogram obeys the area principle; in fact, it looks identical to the histogram in Display D2.18 except for a change in the vertical axis labels. Also notice that the sum of all of the bars' heights is 1.



Display D2.20: Relative frequency histogram of high school completion rates.

These histograms provide a third view of the distribution of high school completion rates. Similar to what you previously saw from both the dotplot and stemplot, the histogram shows that the shape of the data is skewed left and most of the rates fall between 80 and 90 percent. One disadvantage of the histogram is that you lose sight of the actual data values within the bins. For example, the histogram in Display D2.18 shows that there are eight states that have high school completion rates between 90 and 92 percent, but you don't know the values in this particular bin—all eight data values could be exactly the same or wildly different. In contrast, the dotplot in Display D2.9 at least shows you that there are three pairs of values that are the same and two that are unique; and the stemplot in Display D2.14 shows you that the actual values are 90.0, 90.0, 90.4, 90.4, 90.7, 90.8, 91.8, and 91.8. Nonetheless, the histogram remains a popular type of graph because it is a suitable and compact way to show the distribution of a large amount of quantitative data.

PRACTICE: INTERPRETING HISTOGRAMS

Look again at the relative frequency table for the high school completion rates with bins of width 2:

Bin	[76,78)	[78,80)	[80,82)	[82,84)	[84,86)	[86,88)	[88,90)	[90,92)
Frequency	2	4	5	6	8	10	8	8
Proportion	0.039	0.078	0.098	0.118	0.157	0.196	0.157	0.157

Display D2.21: Relative frequency table of high school completion rates.

1. What proportion of high school completion rates are less than 82 percent?

2. What proportion of rates are between 84 and 90 percent?

3. Suppose that you decide instead to use bins of width 4 starting the 75: [75, 79), [79, 83), [83, 87), [87, 91), [91, 95). Find the bin frequencies and bin proportions and complete this relative frequency table.

Bin	[75, 79)	[79, 83)	[83, 87)	[87, 91)	[91, 95)
Frequency					
Proportion					

4. Construct a relative frequency histogram using bins in question 3.

5. Compare your four-bin relative frequency histogram in question 4 with the eight-bin relative frequency histogram in Display D2.20. Which is a “better” graph of the data? Explain.

6. REFLECTION You have seen that you can graph the high school completion rates using a dotplot, a stemplot, and a histogram.

(a) Which would you prefer if you had to make the graph by hand?

(b) If you could use a software package to make the graph, which would you prefer?

(c) If instead of exploring the high school completion rates of 51 rates you needed to explore the rates for the 254 counties in Texas, which graph might be better?

EXPERIMENT WITH DIFFERENT GRAPHS

When you construct a stemplot or a histogram, the data are grouped into bins, but you have to make choices about how that grouping is done. The choice of bin size can have a dramatic effect on the view of the data distribution, and some sizes are better than others for seeing the features of the distribution. It is good practice to experiment with several sizes for bins, choosing the one that seems to best represent the data.

To illustrate the use of different class sizes for stemplots, consider this table that gives the average price per gallon of regular gasoline for each U.S. state and the District of Columbia in May 2004.

State	**Average Gasoline Price ($)**	**State**	**Average Gasoline Price ($)**	**State**	**Average Gasoline Price ($)**
Alaska	1.919	Kentucky	1.767	New York	1.903
Alabama	1.726	Louisiana	1.730	Ohio	1.830
Arkansas	1.741	Massachusetts	1.775	Oklahoma	1.689
Arizona	1.978	Maryland	1.778	Oregon	2.062
California	2.159	Maine	1.792	Pennsylvania	1.789
Colorado	1.851	Michigan	1.852	Rhode Island	1.818
Connecticut	1.846	Minnesota	1.797	South Carolina	1.681
District of Columbia	1.845	Missouri	1.722	South Dakota	1.810
Delaware	1.762	Mississippi	1.737	Tennessee	1.735
Florida	1.821	Montana	1.887	Texas	1.699
Georgia	1.697	North Carolina	1.736	Utah	1.942
Hawaii	2.162	North Dakota	1.851	Virginia	1.712
Iowa	1.773	Nebraska	1.818	Vermont	1.763
Idaho	1.961	New Hampshire	1.732	Washington	2.032
Illinois	1.887	New Jersey	1.687	Wisconsin	1.886
Indiana	1.832	New Mexico	1.781	West Virginia	1.837
Kansas	1.795	Nevada	2.113	Wyoming	1.788

Display D2.22: Average gasoline prices, May 2004. (Source: http://www.fuelgaugereport.com)

Suppose you first construct a stemplot by dividing each price so that the units and tenths digits form the stem and the hundredths digit form the leaf; the thousandths digit is truncated. For example, Kansas' price of $1.795 becomes 17 | 9. In general, this stemplot uses classes of width 0.10—that is, any value in the interval [1.70, 1.80) would be placed on the 17 stem. Then you get this stemplot of average gasoline prices:

```
16 | 88899
17 | 122333334666777888999
18 | 111233344555888
19 | 01467
20 | 36
21 | 156

16 | 8 means $1.68 per gallon
```

Display D2.23: Stemplot of average gasoline prices.

From this graph, you see that the gasoline prices are skewed right with most of the prices clustered in the $1.70's and $1.80's. But because the 51 data values are placed into only six classes, you might be missing some features of the distribution that are hidden by so few classes. So, it may be worthwhile to try grouping the data using more classes.

In the stemplot in Display D2.23, all ten possible leaves (0, 1, 2, …, 9) were placed after a single stem. You can stretch the stemplot out by placing half of the leaves (0, 1, 2, 3, 4) after one stem, and the other half (5, 6, 7, 8, 9) after a duplicate stem. This is called a "5 leaves per stem" stemplot because there are five possible leaves after each stem. Using 5 leaves per stem changes the classes to width 0.05—that is, any value in the interval [1.70, 1.75) is placed on the first 17 stem, and any value in the interval [1.75, 1.80) is placed on the second 17 stem. If you use this alternative approach for these data, you get this new stemplot:

```
16 | 88899
17 | 122333334
17 | 666777888999
18 | 111233344
18 | 555888
19 | 014
19 | 67
20 | 3
20 | 6
21 | 1
21 | 56

16 | 8 means $1.68 per gallon
```

Display D2.24: Stemplot of average gasoline prices with 5 leaves per stem.

In Display D2.24 you see more structure in these prices than you saw in Display D2.23. You see that more prices fell in the interval $1.75–$1.79 than in any other interval; a majority of the prices (30 out of 51, or 59%) fall between $1.70 and $1.85;

there appears to be more detail and a smoother flow in the right-skewed shape; and you see that the two highest prices were $2.15 and $2.16 (corresponding to California and Hawaii.)

You can stretch the stemplot even further by placing the possible leaves after five stems: 0–1, 2–3, 4–5, 6–7, and 8–9. This stemplot variation is called "2 leaves per stem" (because there are two possible leaves after each stem) and changes the classes to width 0.02. Applying this approach, you get:

```
16 | 88899
17 | 1
17 | 2233333
17 | 4
17 | 666777
17 | 888999
18 | 111
18 | 2333
18 | 44555
18 |
18 | 888
19 | 01
19 |
19 | 4
19 | 67
19 |
20 |
20 | 3
20 |
20 | 6
20 |
21 | 1
21 |
21 | 5
21 | 6

16 | 8 means $1.68 per gallon
```

Display D2.25: Stemplot of average gasoline prices with 2 leaves per stem.

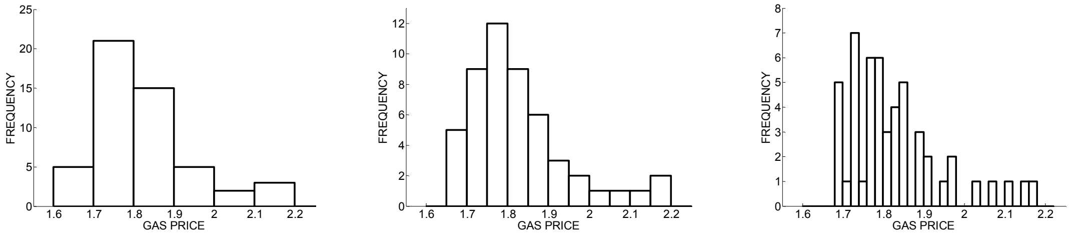
I think you'll agree that the stemplot in Display D2.25 is stretched too far—it is harder to see the basic right-skewed shape; there are now several gaps in the stemplot; and the two highest gasoline prices are no longer grouped into the same class.

What is the best stemplot for these data? To answer this question, think about how well the each stemplot illustrates shape, center, spread, and interesting features of the distribution. Considering these elements, the stemplot in Display D2.24 with 5 leaves per stem appears to be the best—you see the basic shape of the distribution; you can make some meaningful observations about the center and spread; and you can see a few extreme prices that are much higher than the rest.

SPECIAL NOTE: The previous examples illustrate stemplots with 10, 5, and 2 leaves per stem. Could you construct a stemplot with 3 or 4 leaves per stem? No, because you want to have the same possible number of leaves after each stem. (Conceptually, this is the same idea as creating bins that each has the same width in a histogram.) In order for the ten possible leaf digits (0 through 9) to divide evenly among duplicate stems, you have to use a factor of ten. Because 10, 5, 2, and 1 are the only factors of ten, 3 or 4 leaves per stem will not work. The factors of ten also explain why 10 leaves per stem results in one of each stem, 5 leaves per stem results in two duplicates of each stem, and 2 leaves per stem results in five duplicates for each stem.

Similar to choosing class sizes for a stemplot, the choice of bin sizes is important when you construct a histogram. Although a computer program like Fathom will automatically select the bins according to a rule, it is helpful to experiment with wider or narrower bins to get a better graph of the data.

Here are three relative frequency histograms that use the same intervals for bins as were used for classes in the three stemplots above.



Display D2.26: Histograms of average gasoline prices using bins of width 0.10, 0.05, and 0.02.

The middle histogram, which uses bins of width 0.05, seems to best represent the distribution of gasoline prices of the 51 states. The first histogram on the left uses too few bins and you lose information about the distribution of gasoline prices within individual bins.   The last histogram on the right, in contrast, uses too many bins and it is harder to see the basic shape of the distribution.

TECHNOLOGY LAB: CHOOSING THE BINS OF A HISTOGRAM

Fathom provides an easy way to modify the choice of bins in a histogram. By experimenting with the choice of bin width, you will understand the problems in constructing a histogram with bin widths that are either too small or too large.

PART A: Heights of college women

The datafile female_heights.txt contains the heights (recorded in inches) for 428 women taking introductory statistics at a Midwestern college one semester.

1. Import this data into Fathom.

2. Construct a histogram of the heights. Change the vertical scale to "density" by selecting the graph and choosing menu item Graph→Scale→Density.

(Note: By choosing the vertical scale of a histogram to be "density", the proportion of data values in a particular bin will be equal to the area of the corresponding bar of the histogram.)

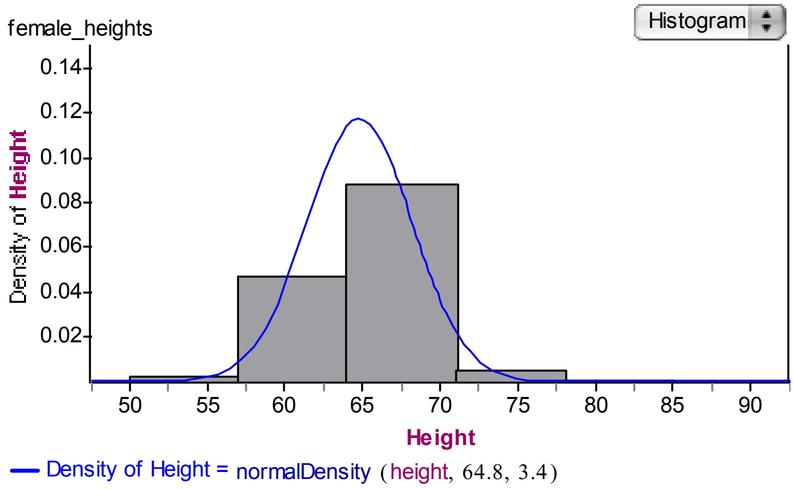
3. Note that the shape of these heights is symmetric. In a later topic, you will learn that the shape of this dataset can be well-described by a special bell-shaped curve called the normal density. With the histogram still selected, draw this normal curve on top of the histogram by selecting menu item Graph→Plot Function and typing `normaldensity(height,64.8,3.4)` in the formula editor window that appears.

Fathom give you two ways of changing the bins in a histogram:

- When you move the selection arrow between the bars of the histogram it changes to a double-headed arrow. When this happens, you can change the width of the bins by holding the mouse button and dragging. If you hold-and-drag on the left-most edge of the first bar, you can also change the starting value for the bins.
- Or, if you double-click inside the graph, a graph inspector window appears. Under the Properties tab, you can set the bin width by typing a new value for binWidth. You can also set the starting value for the bins by editing binAlignment. (Note: If your selections for binWidth and binAlignment result in the first few bins being empty, then Fathom will default to another value for binAlignment.)

The smooth curve represents the shape of the distribution of data if it contained infinitely many heights. The goal of this activity is to construct a histogram for the heights that is a good match for the smooth curve.

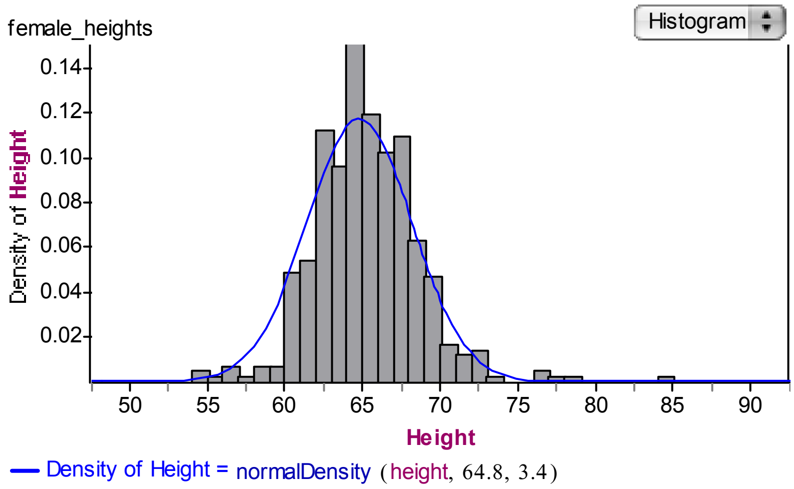
4. Using Fathom, choose only five bins of width 7 starting from 50; you should obtain a histogram such as the one pictured below.



Display D2.27: Fathom histogram of women heights with bin width 7 inches.

Notice that the histogram is not a good match to the curve. The curve gradually increases until a height of 65 inches and then gradually decreases for large values. In contrast, the histogram is not smooth as it takes big jumps at heights of 57 and 64 inches, and a big drop for the height of 71 inches. To make the histogram smoother, you need to choose a bin width smaller than 7 inches.

5. Redraw the histogram on Fathom using a small bin width of 1 inch. Your histogram should look similar to the histogram in Display D2.28 that uses approximately 30 bins of width 1 starting from 54 inches. The histogram appears to be a better match to the curve, but there is a new problem. We don't see the big jumps and drops that we saw in the first histogram. But because the number of scores in each bin is small, the bar heights appear to be more erratic and the heights of the bars don't follow the increasing and decreasing behavior of the curve.



Display D2.28: Fathom histogram of women heights with bin width of 1 inch.

To summarize the last two parts, your objective is to construct a histogram that

- is **a good match** to the underlying distribution curve (that is, the histogram doesn't have big jumps or big drops)
- is **relatively smooth** so you don't see much random fluctuation in the heights of the bars

6. Experiment with different bin widths and different starting values to find the settings that you think are "best." (*Hint:* Your bin width should be between 1 and 7 inches.)
7. Does the choice of the bin width and starting value depend on the number of data values in the dataset? To explore this question, the datafile female_heights_50.txt contains a smaller set of 50 women heights. Import these data into Fathom, create a histogram, and draw a bell-shape curve. Now find the "best" choices for bin width and starting value.
8. Based on your work, if you have more data, should you choose fewer (wider) bins or more (narrower) bins? Explain.

PART B: Histogram for skewed data.

Students in an introductory statistics class were asked the question: "How many movie dvds do you own." The datafile dvds.txt contains the response to this question for 636 students. These data can be modeled by a special right-skewed curve, an exponential density curve that you will learn more about in another topic.

1. Import these data into Fathom and construct a histogram. As before, select a density scale for the histogram.

2. Draw this distribution shape on top of the histogram by selecting menu item Graph→Plot Function and typing exponentialDensity(dvds, 27.3) in the formula editor.

3. Using the two criteria described above ("good match" and "smooth"), find a good choice for the bin width and starting value of the histogram.

PART C: Histogram for Old Faithful.

Old Faithful is a famous geyser in Yellowstone National Park, Wyoming. The waiting times between successive eruptions of Old Faithful follow a predictable pattern. The datafile oldfaithful.txt contains 106 times between eruptions. [SOURCE?]

1. Import these data into Fathom and construct a histogram. As before, when comparing histograms with different bin widths, it is helpful to choose the density scale.

2. Experiment with these bin widths and decide on the best choice. (Remember that on Fathom one can use exact values of the bin width using the Graph Inspector.)

10 9 8 7 6 5 4 3 2 1

3. Using your best choice of histogram, draw a smooth curve over the bar heights. Describe the shape of this smooth curve.

PRACTICE: EXPERIMENTING WITH DIFFERENT GRAPHS

Consider again the data on four variables for the states in the Midwest and Northeast regions of the U.S. (Display D2.16 on page XXX).

1. Construct three stemplots for Driver Rate data, as described below. For each data value, use the hundreds place as the stem and the tens place as the leaf. Don't forget to provide a key for each stemplot.

a. One stemplot that uses 10 leaves per stem. (This was the stemplot that was drawn in the previous practice activity.)

b. A second stemplot that uses 5 leaves per stem.

c. A third stemplot that uses 2 leaves per stem.

2. Which stemplot, 1a, 1b, or 1c, gives a better display of the distribution of the Driver Rate data? Explain.

3. Construct three histograms (or relative frequency histograms) for Elevation, as described below.

a. One histogram that uses bins of width 1000 starting from 0.

b. A second histogram that uses bins of width 500 starting from 0.

c. A third histogram that uses bins of width 250 starting from 0.

4. Which histogram, 3a, 3b, or 3c, gives a better display of the distribution of the Elevation data? Explain.

5. REFLECTION What does it mean for one graph to be a better display of the distribution of data than another? You can answer this question by drawing a “good” graph and a “bad” graph and explaining why the good graph is a better representation of the data than the bad graph. [too open-ended?]

ACTIVITY: THE SHAPE OF THE DATA

In this activity, you get some experience in collecting and graphing data, and studying the shapes of the corresponding distributions. You’ll see data distributions that have a variety of shapes, including uniform, symmetric, and skewed.

Each person needs: a tennis ball, a single die, a centimeter ruler

1. Individually, devise a way to use a ruler to measure the diameter of a tennis ball in centimeters. Combine everyone’s measurements into a class dataset.

2. Individually, count the number of rolls of a die until you observe all six possible outcomes (1, 2, 3, 4, 5, and 6) at least once. Combine everyone’s results into a class dataset.

3. Individually, devise a way to use a ruler to measure the thickness of a single page of your textbook in centimeters. Combine everyone’s measurements into a class dataset.

4. Refer to Display D2.29 on page XXX. Work individually or as a class to create a dataset of the first (left-most) digits of the counties’ populations.

5. Use Display D2.29 to create dataset of the last (right-most) digits of the counties’ populations.

Complete steps 6–9 for *each* of the five datasets that you collected above.

6. Construct a graph of the data.

7. Based on the graph, describe the shape of the distribution (uniform, symmetric, skewed right, skewed left, or other). Thinking about the type of data that was collected and the way it was collected, explain why the shape of the distribution "makes sense."

8. Find the *mean* and the *median.* (These measures of center will be fully defined in Topic D3. Briefly, the mean is the arithmetic average calculated by adding all of the data values and dividing by the number of values. When the data values are put in chronological order, the median is the data value, or the mean of two data values, that is in the middle of the dataset.)

a. Is either measure of center significantly greater than the other, or are the mean and median approximately equal?

b. If the two measures of center are different, can you find any reasons why? Does it appear to be related to the shape of the distribution?

9. Describe any other interesting features (unusually large or small values, gaps, or clusters).

County	Population
Autauga County	43,671
Baldwin County	140,415
Barbour County	29,038
Bibb County	20,826
Blount County	51,024
Bullock County	11,714
Butler County	21,399
Calhoun County	112,249
Chambers County	36,583
Cherokee County	23,988
Chilton County	39,593
Choctaw County	15,922
Clarke County	27,867
Clay County	14,254
Cleburne County	14,123

County	Population
Graham County	33,489
Greenlee County	8,547
La Paz County	19,715
Maricopa County	3072,149
Mohave County	155,032
Navajo County	97,470
Pima County	843746
Pinal County	179,727
Santa Cruz County	38,381
Yavapai County	167,517
Yuma County	160,026
Arkansas County	20,749
Ashley County	24,209
Baxter County	38,386
Benton County	153,406

County	Population
Kern County	661,645
Kings County	129,461
Lake County	58,309
Lassen County	33,828
Los Angeles County	9,519,338
Madera County	123,109
Marin County	247,289
Mariposa County	17,130
Mendocino County	86,265
Merced County	210,554
Modoc County	9,449
Mono County	12,853
Monterey County	401,762
Napa County	124,279
Nevada County	92,033

Coffee County	43,615
Colbert County	54,984
Conecuh County	14,089
Coosa County	12,202
Covington County	37,631
Crenshaw County	13,665
Cullman County	77,483
Dale County	49,129
Dallas County	46,365
DeKalb County	64,452
Elmore County	65,874
Escambia County	38,440
Etowah County	103,459
Fayette County	18,495
Franklin County	31,223
Geneva County	25,764
Greene County	9,974
Hale County	17,185
Henry County	16,310
Houston County	88,787
Jackson County	53,926
Jefferson County	662,047
Lamar County	15,904
Lauderdale County	87,966
Lawrence County	34,803
Lee County	115,092
Limestone County	65,676
Lowndes County	13,473
Macon County	24,105
Madison County	276,700
Marengo County	22,539
Marion County	31,214
Marshall County	82,231
Mobile County	399,843
Monroe County	24,324
Montgomery	223,510

Boone County	33,948
Bradley County	12,600
Calhoun County	5,744
Carroll County	25,357
Chicot County	14,117
Clark County	23,546
Clay County	17,609
Cleburne County	24,046
Cleveland County	8,571
Columbia County	25,603
Conway County	20,336
Craighead County	82,148
Crawford County	53,247
Crittenden County	50,866
Cross County	19,526
Dallas County	9,210
Desha County	15,341
Drew County	18,723
Faulkner County	86,014
Franklin County	17,771
Fulton County	11,642
Garland County	88,068
Grant County	16,464
Greene County	37,331
Hempstead County	23,587
Hot Spring County	30,353
Howard County	14,300
Independence	34,233
Izard County	13,249
Jackson County	18,418
Jefferson County	84,278
Johnson County	22,781
Lafayette County	8,559
Lawrence County	17,774
Lee County	12,580
Lincoln County	14,492

Orange County	2,846,289
Placer County	248,399
Plumas County	20,824
Riverside County	1,545,387
Sacramento County	1,223,499
San Benito County	53,234
San Bernardino	1,709,434
San Diego County	2,813,833
San Francisco	776,733
San Joaquin County	563,598
San Luis Obispo	246,681
San Mateo County	707,161
Santa Barbara	399,347
Santa Clara County	1,682,585
Santa Cruz County	255,602
Shasta County	163,256
Sierra County	3,555
Siskiyou County	44,301
Solano County	394,542
Sonoma County	458,614
Stanislaus County	446,997
Sutter County	78,930
Tehama County	56,039
Trinity County	13,022
Tulare County	368,021
Tuolumne County	54,501
Ventura County	753,197
Yolo County	168,660
Yuba County	60,219
Adams County	363,857
Alamosa County	14,966
Arapahoe County	487,967
Archuleta County	9,898
Baca County	4,517
Bent County	5,998
Boulder County	291,288

County	
Perry County	11,861
Pickens County	20,949
Pike County	29,605
Logan County	22,486
Lonoke County	52,828
Madison County	14,243
Chaffee County	16,242
Cheyenne County	2,231
Clear Creek County	9,322

Display D2.29: Populations of a number of counties from Alabama, Alaska, Arizona, Arkansas, California, and Colorado from the 2000 U.S. Census. (Source: *http://www.census.gov/popest/counties/files/CO-EST2005-ALLDATA.csv*)

CLASSES OF DATA AND SHAPE

There are two general classes of quantitative data, called Counts/Amounts, and Measurements. The famous statistician John Tukey in his EDA (Exploratory Data Analysis) book gave special names to these classes since they tend to have common shapes. Some familiarity with these classes will be helpful in our exploration and summarization of data.

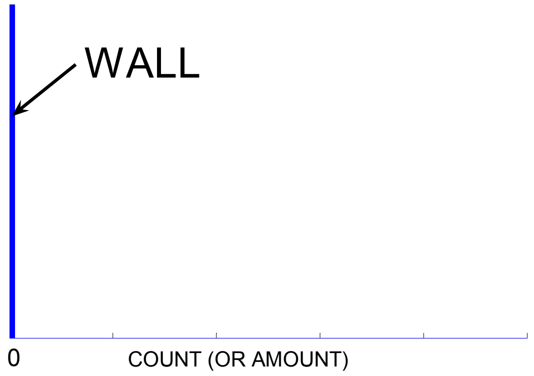
- COUNTS or AMOUNTS. You can collect the **count** of something, like the **number** of people in your family, the **number** of cookies you ate today, or the **number** of students who attend your school. Also you may be interested in the **amount** of money you spent on books this semester, the **amount** of time you spend studying for this course, or the **amount** of money you have in your saving account. Batches of counts or amounts are typically right-skewed.
- MEASUREMENTS. Data may be a collection of **measurements**, such as the **heights** of individuals of a single gender, the **time** it takes you to drive from home to school for many trips, or the **highest recorded temperature** for a city for many days in a particular month. Measurements are typically symmetric.

Why are counts and amounts typically right-skewed? As an example, the author collected the duration (in minutes) of 58 phone calls for a local company from the monthly bill:

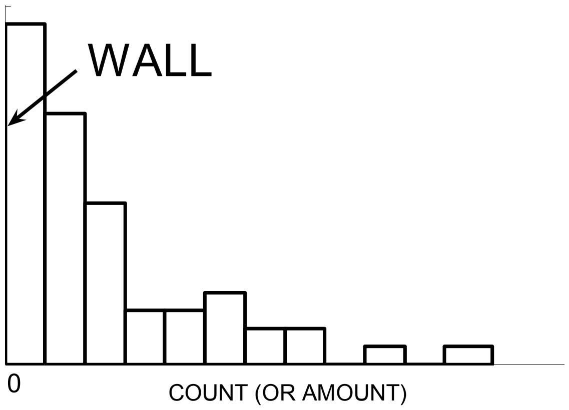
```
2.8    1.8    0.9    0.3    1.2    1.2    2.4   12.2    1.5
5.3    0.9    0.6    4.3    1.0    0.7    0.5    1.2    1.2
```

```
6.9    0.6    0.3    4.5    0.6    7.8    0.8    0.9    0.8
5.3    7.5    4.2    2.7    1.5    2.4    6.0    3.6    6.3
1.2    2.7    1.1    0.6    0.4    2.7    1.0    1.4    2.1
0.6    3.1    1.6    1.5    3.0    5.7    0.3    9.9    1.7
3.7    0.6    2.0    3.0
```

Here the duration would be an example of an Amount – the amount of time that a phone call lasts. Note that a phone call's duration can't be negative, so all of the data must fall to the right of a "wall" of zero minutes. Also in many situations small positive values of Amounts are likely. It follows that the Amounts will tend to crowd up against the wall, resulting in a right-skewed distribution shape.



Indeed, if we graph the phone durations using a histogram, we see the right-skewed shape.

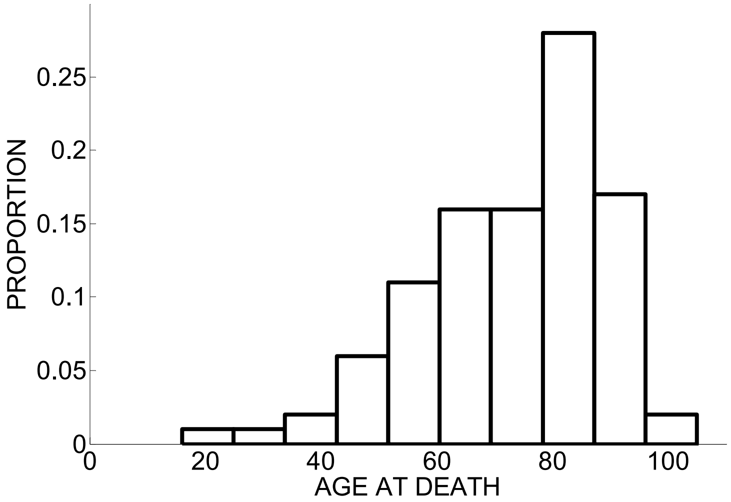


In contrast, consider a simple illustration of a batch of measurements. Suppose that each student in a class is asked to guess at an age of the instructor and suppose the instructor appears to be 35 years old. The error in a student's guess is defined to be

ERROR = GUESS – TRUE AGE = GUESS – 35.

Now some students will make guesses that are too small, resulting in negative errors, and other students will make guesses too high that result in positive errors. Most of the errors will be close to 0 and there will be (roughly) the same number of positive errors and negative errors. Here the distribution of measurements will be approximately symmetric. Unlike the first example, there is no wall that restricts the value of the error of the guess.

Is it possible for real-life data to be left-skewed? It is more likely to see right-skewed or symmetric data, but some variables have left-skewed distributions. This happens when data must fall to the left of a wall, where the wall represents the largest possible data value. One example of a left-skewed variable mentioned earlier would be test scores of an "easy" test. Here all scores must fall to the left of 100 percent and there will be many high grades and a trail of low grades. A second example would be age at death; a histogram of some ages of death collected from the obituary column of a local newspaper is displayed below. Note that most of the ages at death are between 70 and 90 years and the bins proportions decrease slowly for smaller values. This skewness is caused by a limit to the length of human life.



ACTIVITY: MATCHING VARIABLES AND SHAPES

DESCRIPTION: In this activity, you will get some experience matching histograms with different variables. Before doing this activity, it will be helpful to review the previous material on classes of data to understand some common distribution shapes.

For each group of variables, read the descriptions. Then write the variable name over the corresponding histogram.

GROUP 1 OF VARIABLES:

[HOME RUNS] The number of home runs hit for all baseball players who had at least 300 at-bats (opportunities) to hit. Home run numbers tend to be somewhat right-skewed – low home run counts are more common than large home run counts.

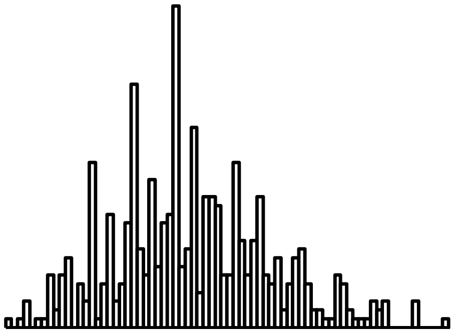
[FOOTBALL SCORES] The score of the winning team for a large number of (American) college football games. Particular football scores are popular, like 7, 14, etc.

[RUNNING TIMES] The times of women who ran in a marathon running race. Marathon running times are pretty symmetric, but it is more common to have a LARGE (slow) time than a SMALL (fast) time.

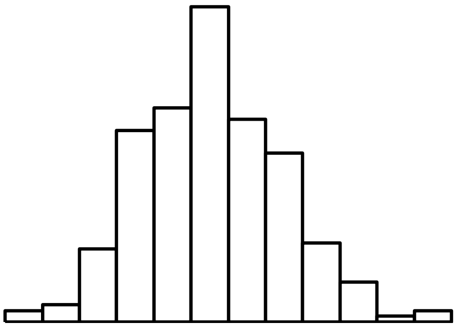
[BATTING AVERAGES] The batting averages for all baseball players who had at least 300 at-bats (opportunities) to hit. Batting averages tend to be very symmetric.

[SOCCER SCORES] The score of the winning team for a large number of soccer games played in club games in England. (If the game was tied, then the common score is recorded.) Soccer scores tend to be small.

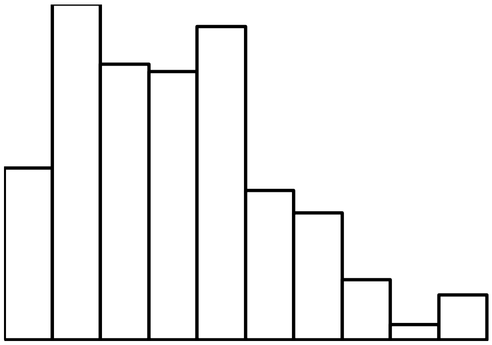
A



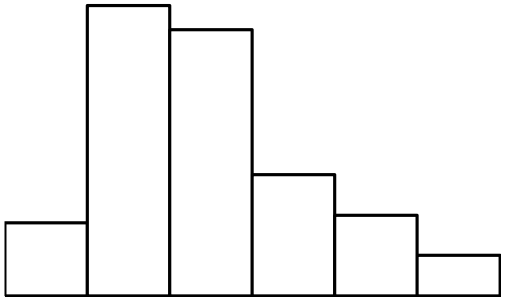
B



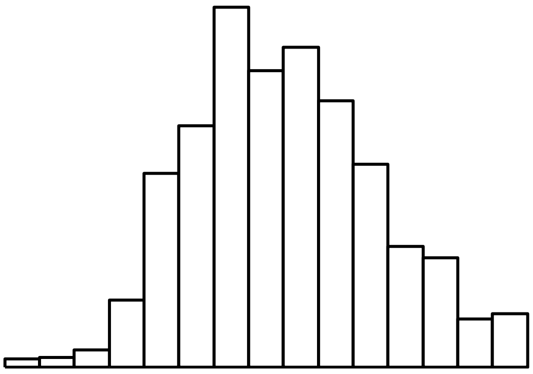
C



D



E



GROUP 2 OF VARIABLES:

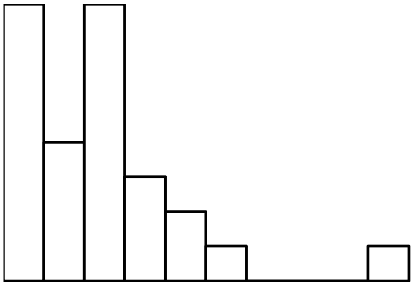
[DVDS] The number of movie DVDs owned by students in a college statistics class. Most students own few movie DVDs, but a couple of students own many DVDs.

[MILES FROM HOME] The miles from the school to the student's hometown for students in a college statistics class. Students from this school tend to come from three regions of the state – one region within 30 miles of the school, a second region about 100 miles from the school, and a third region about 150 miles from the school.

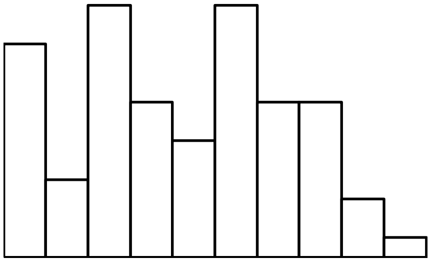
[TV] The average numbers of hours watching the television for students in a college statistics class. Most students watch only a little TV each day, but several students watch a lot of TV.

[SLEEP] The number of hours of sleep for students in a college statistics class. Students tend to get between 7 to 9 hours of sleep, but a few students got only a few hours of sleep.

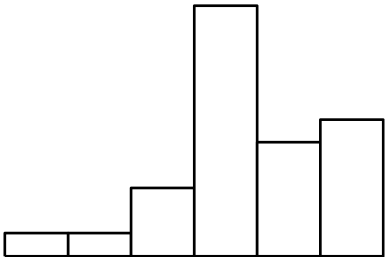
F



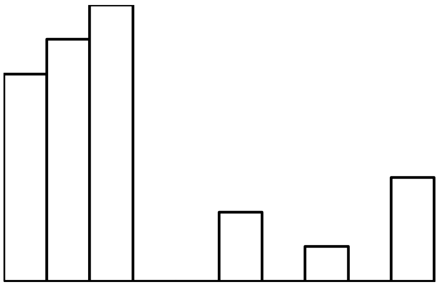
G



H



I



Classroom Capsule: Graphing Health and Liking School

Overview: The students will learn about different graphs for a single group of measurement data and understand principles for good graphs.

Objectives: The students will get experience drawing stemplots, dotplots, and histograms. They will understand that the construction of any graph involves choices, and the quality of the graphical display depends on these choices. For a given dataset, the students will learn that they can be several suitable graphs. The students will be able to criticize graphical displays presented in the media.

Description: In the UNICEF study, we focus on two variables that were measured for a group of 21 countries: “health”, the percentage of young people rating their health as "fair or poor", aged 11, 13, and 15, and “swell”, the percentage of young people 'liking school a lot', aged 11, 13, and 15. The table of the variables is shown below.

Country	health	swell
Austria	15.6	36.1
Belgium	13.1	17.9
Canada	13.7	21.9
Czech_Republic	11.8	11.6
Denmark	14.8	21.4
Finland	11.0	8.0
France	NA	21.7
Germany	14.9	29.5
Greece	10.1	29.5
Hungary	14.9	26.3
Ireland	12.9	22.3
Italy	12.5	13.0
Netherlands	17.2	34.4
Norway	18.5	38.9

Poland	14.4	17.3
Portugal	19.1	31.1
Spain	9.0	22.8
Sweden	13.2	21.6
Switzerland	9.1	22.3
United_Kingdom	22.6	19.0
United_States	19.8	23.4

1. For the health variable, construct three stemplots.

(a) First, construct a stemplot where you break between the units and tenths places and have ten leaves per stem

(b) Next, construct a stemplot where you break between the tens and units places and have two leaves per stem

(c) Last, construct a stemplot where you break between the tens and units places and have five leaves per stem

2. Among the three stemplots you drew in part 1, which is the best graph of the data? Why?

3. Using the best graph, write a descriptive paragraph of the data include statements about shape, center, spread and any unusual features.

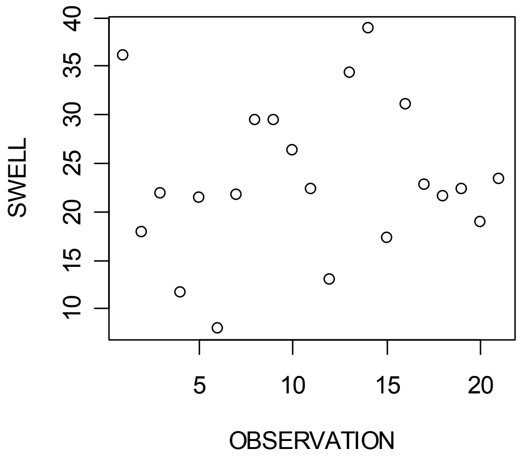
4. On the basis of the graph, how you would you describe American children's health in comparison with the 20 other countries in this study?

5. Four different graphs are drawn of the "swell" variable. The first graph is an index plot where the swell measurement is plotted as a function of the observation number and the remaining three graphs are histograms using different bin widths. Rank the four graphs in order (1, 2, 3, 4) from the "best" display and the "worst" display. In the spaces below give your rankings and explain why you gave these rankings.

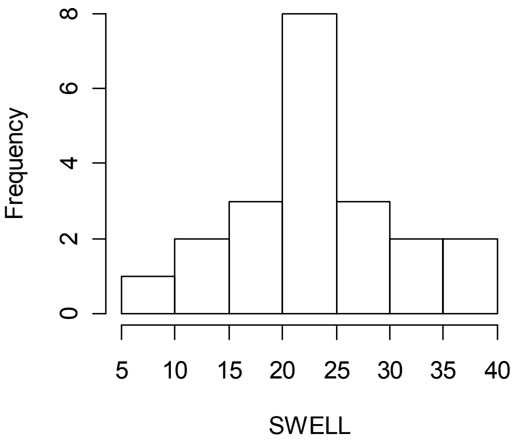
	GRAPH	WHY?
Best		
Next best		

Next best		
Worst		

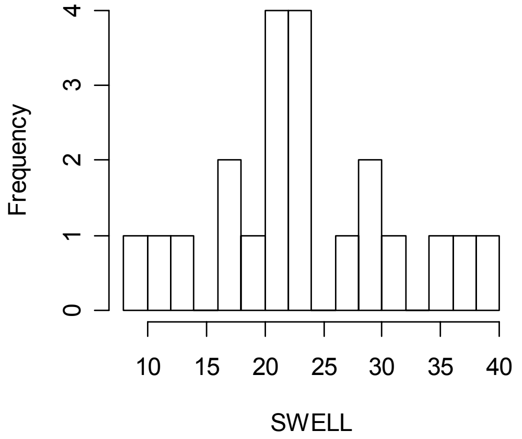
GRAPH A



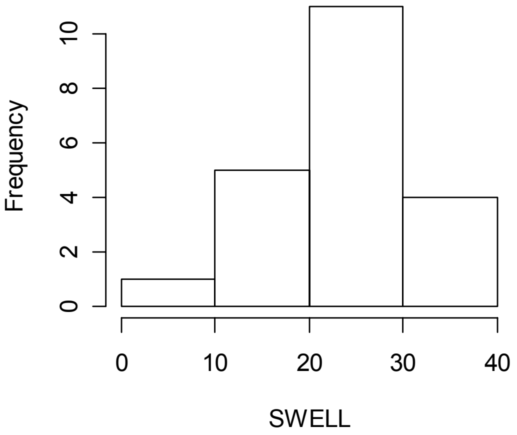
GRAPH B



GRAPH C



GRAPH D



Share and Summarize: One graphs data for a reason. Here the reason is to see the distribution of the data. If the data is graphed well, then it will be relatively easy to detect its shape, locate the center value, and make some statement about the spread of the values. It should be easy to summarize a dataset from looking at its graph.

Following are the three stemplots of the health variable. Which is the best graph?
1. Stemplot 1 has too many lines or bins. It is difficult to see the distributional shape and there are gaps in the display that may not be meaningful.
2. Stemplot 3 only uses four lines and it is difficult to see the distributional shape – over half of the values are found on a single line.
3. Stemplot 2 seems to be the best display among these three. The shape (roughly symmetric about 13%) is visible and there is one possible outlier (22%) at the high end.

STEMPLOT 1 STEMPLOT 2 STEMPLOT 3

```
   9 | 01        0. | 99        0. | 99
  10 | 1         1* | 011       1* | 011223334444
  11 | 08         t | 22333     1. | 57899
  12 | 59         f | 44445     2* | 2
  13 | 127        s | 7
  14 | 4899      1. | 899       1|0 means 10%
  15 | 6         2* |
  16 |            t | 2
  17 | 2
  18 | 5            1|0 means 10%
  19 | 18
  20 |
  21 |
  22 | 6

9|0 means 9.0%
```

After the class has decided on the best graphical display, then talk about describing the health variable, including statements about shape, average value, and spread. The health variable for the United States is at the right end of the graph, indicating that the health of children in the U.S. is worse than most of the countries in the study.

In the discussion about part 5, Graph A that constructs an index plot of the observations, is not a good graph. In some cases, this might be useful, but the purpose of this graph is not obvious and it doesn't display the distribution of the data well. The three histograms (graphs B, C, D) use different choices for the number of bins. Histogram B is the best choice since it best shows the distributional shape and it is easiest to find the average value. A stemplot actually is a type of histogram where one sees the data values – choosing the right number of bins is the same issue as choosing the right stemplot.

Application or Extension: Use a graphing calculator to construct a histogram for a dataset of interest. First construct the "automatic" histogram where the bins are determined by the calculator. Then modify the histogram by choosing half as many bins, and by choosing twice as many bins. Confirm that the automatic choice for bins results in the best graphical display.

WRAP-UP

A first step in exploring a batch of data is to construct a suitable graph. For categorical data, **bar charts, segmented bar charts,** and **pie charts** are useful for comparing the **frequencies** of different categories. For quantitative data, **dotplots** and **stemplots** are easy-to-construct graphs that help you visualize the **distribution** of the data. **Histograms** are particularly helpful when you explore the distribution of a large volume of quantitative data. All of these graphs follow the **area principle** where the area of a object representing a data value is proportional to its frequency. In practice, it is useful to experiment with several graphs—such as stemplots with different numbers of **classes** or histograms with different numbers of **bins**—in order to find the graph that "best" represents the distribution.

When you look at the distribution of a batch of quantitative data, you are interested in its basic **shape,** a **center** value, the **spread** of values, and interesting features such as **outliers,** gaps, or clusters. Some basic distribution shapes include **uniform, symmetric, skewed right,** and **skewed left.** You will see in the next topic that the shape of a distribution influences the way that you summarize a dataset.

EXERCISES

1. **Religions of Countries**

Suppose you are taking a college class on world religions and you are asked about the predominant religions in the world. *The World Almanac* gives the chief religion for a large group of countries. To save time, you decide to record the chief religion for a sample of countries randomly selected from the almanac list with the hope that the distribution of major religions in this sample will be similar to the distribution of religions for the entire list of countries. (In the below list, Indigenous refers to a religion that is unique to that particular country.)

Country	**Chief Religion**
Afghanistan	Muslim
Antigue and Barbuda	Protestant
Azerbaijan	Muslim
Belarus	Eastern Orthodox
Bosnia and Herzegovina	Muslim
Burkina Faso	Muslim
Cape Verde	Roman Catholic
Columbia	Roman Catholic
Croatia	Roman Catholic
Denmark	Lutheran
Ecuador	Roman Catholic
Estonia	Lutheran
Gabon	Christian
Greece	Greek Orthodox
Guyana	Christian
India	Hindu
Israel	Jewish
Kenya	Protestant
Kyrgyzstan	Muslim
Libya	Muslim
Madagascar	Indigenous
Malta	Roman Catholic
Micronesia	Roman Catholic
Mozambique	Indigenous
Netherlands	Roman Catholic
Norway	Lutheran
Papua New Guinea	Indigenous
Portugal	Roman Catholic
Rwanda	Roman Catholic
San Marino	Roman Catholic
Seychelles	Roman Catholic
Somalia	Muslim
Suriname	Hindu
Taiwan	Buddhist
Trinidad and Tobago	Roman Catholic
Uganda	Roman Catholic
Uraguay	Roman Catholic
Vietnam	Buddhist

Display D2.30: Chief religions of 38 countries. (Source: *The World Almanac and Book of Facts 2005,* World Almanac Education Group, Inc., 2005)

a. What type of variable is Chief Religion?

b. Construct a frequency table for Chief Religion.

c. Suppose you decide to combine the religions into the categories Christianity-based (Christian, Protestant, Roman Catholic, Greek Orthodox, Lutheran, East-Orthodox), Muslim, and Other (all remaining religions). Construct a frequency table of these categories.

d. Construct a bar chart for Chief Religion using the categories of part c.

e. Based on your work, which chief religions are shared by the most countries in this group?

2. Types of High Schools in Fort Worth

The high schools in the United States can be categorized into three distinct types: **public** schools are administered and financed by the local and state government, **private** schools are not administered by the government, and **charter** schools are organized and controlled by educators, parents, or private groups with an expressed purpose or philosophy. Suppose a family is moving to Fort Worth, Texas and they are interested in the distribution of the three types of schools in this city. Display D2.31 gives the school type for the 66 high schools in Fort Worth.

public	private	public	public	public	public
private	private	public	charter	public	private
public	public	public	public	public	charter
public	private	charter	private	public	public
private	private	public	charter	public	private
public	private	private	private	public	public
public	private	private	private	public	private
charter	charter	private	private	private	private
public	public	public	private	public	charter
public	public	public	public	private	public
public	charter	charter	private	public	public

Display D2.31: School type for 66 high schools in Fort Worth, Texas. (Source: http://www.greatschools.net)

a. Construct a frequency table for School Type.

b. Construct a bar chart for School Type.

c. What is the most common school type among the high schools in Fort Worth?

d. Would it be accurate to say that over half of the Fort Worth high schools are public?

3. **Educational Attainment of Americans**

To get some insight about the education background of American adults, this table gives the highest level of education for a group of adult Americans, age 21 or older.

Level of Education	**Frequency**
4 or more years of college	12
1 to 3 years of college	20
Grade 12	17
Grade 11	2
Grade 10	4
Grade 9	1
Grade 5, 6, 7, or 8	5
None or preschool	2

Display D2.32: Educational attainment of 63 adults. (Source: Collected from the 2000 U.S. Census.)

a. Suppose you are interested in categorizing the Level of Education into the three groups: "not completed high school," "completed high school," and "some college." Construct a frequency table using these three groups.

b. Construct a segmented bar chart for the frequencies you found in part 3a.

c. How many adults in this group had at least a high school education?

4. **Scores on the AP Calculus Exam**

Many high school students take the Advanced Placement (AP) Calculus course; this provides an opportunity for students to take a college-level calculus course while still in high school. At the end of the course, the student takes an AP Calculus Exam; if the student obtains a high grade on this exam, he or she can get credit for a college calculus course. There are two AP calculus exams that are given; the "AB exam" covers the content of the first calculus course in of and the "BC exam" covers the content of the first and second calculus courses. Display D2.33 gives the grade distribution of all students who took the AB exam and the BC exam in 2005.

AP Calculus AB Exam	
Exam Grade	Number of Examinees
Score of 5	38,539
Score of 4	36,347

AP Calculus BC Exam	
Exam Grade	Number of Examinees
Score of 5	23,877
Score of 4	9,237

Score of 3	33,006
Score of 2	31,141
Score of 1	46,959

Score of 3	10,929
Score of 2	3,695
Score of 1	6,677

Display D2.33: Grade distribution of all students who took the AB or BC forms of the AP calculus exam in 2005. (Source: The College Board, apcentral.collegeboard.com.)

a. Construct a histogram of the frequency of students obtaining the different grade levels for the AB exam.

b. What is the shape of this histogram?

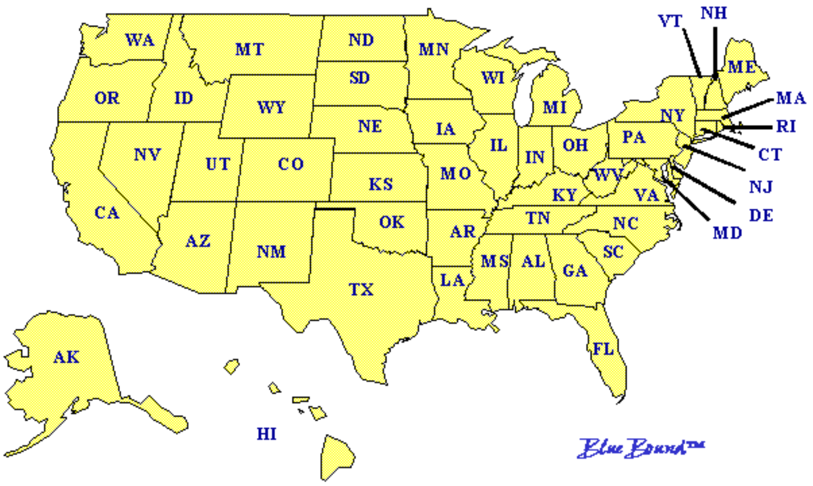
c. What is the most common score on the AB exam?

d. Answer parts a, b, and c for the grades on the BC exam.

e. Suppose that college credit is given if a student scores 3 or higher on the exam. From your work, estimate the proportion of students who would get college credit for each exam.

5. **How Many States Have You Visited?**

Look at the map below and count how many different U.S. states you have visited in your lifetime. Assume that "visited" means "stayed overnight."



a. What is the total number of states that you've visited?

Students in a statistics class were also asked, "How many different states have you visited in your lifetime?" Here are their answers:

46	10	15	14	8	10	13	11	18	13	13
12	10	11	5	30	32	11	25	14	12	11

Display D2.34: Number of states visited by 22 students.

b. Construct a dotplot of the data in Display D2.32.

c. Write a short paragraph that describes the distribution's shape, center, spread, and any interesting features.

d. Find an interval of values that approximately contains 50% of the data values.

e. Compare the number of states you have visited (from part 5a) with the distribution for the students in the statistics class. Can you consider yourself a "well-traveled person" relative to the students in this class? Explain.

6. **How Do You Sleep?**

Answer parts 6a, 6b, and 6c based on your own experience.

a. What time did you go to bed last night?

b. What time did you wake up this morning?

c. To the nearest quarter hour, how many hours of sleep did you get last night?

Students in a statistics class were also asked to calculate how many hours of sleep they each got last night. Here are the results:

7.00	6.50	10.00	6.25	9.50	8.50	6.00	8.00	8.75	5.50	9.00
9.38	5.75	5.50	8.25	8.50	8.00	8.50	7.80	8.75	7.75	6.00

Display D2.35: Hours of sleep for 22 students.

d. Construct a suitable graph of the data in Display D2.33.

e. Describe the distribution of this data, including shape, center, spread, and any interesting features.

f. How does your amount of sleep (from part 6c) compare with the distribution for the students in the statistics class? Would it be appropriate to say that you get an "average" amount of sleep? Explain.

Assume that a person's amount of sleep is categorized as "little" if it is less than 7 hours, "moderate" if it is greater than or equal to 7 hours but less than 9 hours, and "high" if it is greater than 9 hours.

g. Use the data in Display D2.33 and the categories "little," "moderate," and "high" to construct a frequency table and bar chart.

h. What proportion of students get a "little" amount of sleep?

i. In this exercise you first treated the amount of sleep as a quantitative variable, and then as a categorical variable. Which treatment did you prefer? Why?

7. **Gross Sales of Movies Starring Julia Roberts**

The table below lists the gross sales of 24 movies featuring Julia Roberts.

Movie	**Gross Sales (millions of dollars)**
America's Sweethearts (2001)	94
Confessions of a Dangerous Mind (2002)	16
Conspiracy Theory (1997)	76
Dying Young (1991)	34
Erin Brockovich (2000)	126
Everyone Says I Love You (1996)	10
Flatliners (1990)	61
Full Frontal (2002)	3
Hook (1991)	120
I Love Trouble (1994)	31
Mary Reilly (1996)	6
Mexican, The (2001)	67
Michael Collins (1996)	11
Mona Lisa Smile (2003)	64
My Best Friend's Wedding (1997)	12
Notting Hill (1999)	116
Ocean's Eleven (2001)	183
Pelican Brief, The (1993)	101
Prêt-à-Porter (1994)	6
Pretty Woman (1990)	178
Runaway Bride (1999)	152
Sleeping with the Enemy (1991)	101
Something to Talk About (1995)	51
Stepmom (1998)	91

Display D2.36: Gross sales of several Julia Roberts movies, as of March 2004. (Source: Internet Movie Database www.imdb.com.)

a. Construct a stemplot of the gross sales for these Julia Roberts movies by breaking between the tens and units digits, and using ten leaves per stem. (So for example, the first gross sales of 94 would have a stem of 9 and a leaf of 4. The gross sales of 6 or 06 would have a stem of 0 and a leaf of 6.)

b. Construct an alternative stemplot of the gross sales by again breaking between the tens and units digits and using five leaves per stem. Compare your two stemplots. Which one do you think is a better representation of the distribution of gross sales? Why?
c. Describe the features of the distribution of this data.

8. **How Random Are You?**

Students in a statistics class were asked to choose a number between 1 and 20. Here are the responses:

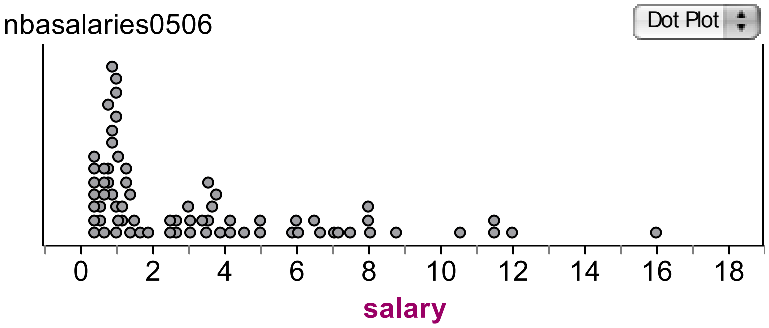
11	2	7	13	13	16	7	17	8	8	12
5	8	12	7	20	17	7	8	2	13	13

Display D2.37: Selection of numbers between 1 and 20 for 22 students.

a. Construct a stemplot of these numbers by breaking between the tens and units places and using five leaves per stem.
b. Can you find particular numbers that were popular among the students? Is there any explanation why these numbers might be popular?
c. Can you find numbers that were relatively unpopular? Is there any explanation why these numbers might be unpopular?
d. Do you think a distribution of random numbers (created by rolling a 20-sided die or by a computer algorithm) would look different from this distribution of student-chosen numbers? If so, explain how.

9. **Salaries of Basketball Players**

Only a small number of players have the opportunity to play for the NBA basketball league, but if they do get this opportunity, they receive very large salaries. The 2005-06 salaries (in millions of dollars) were collected for all players on five professional NBA teams (Hawks, Nuggets, Lakers, Hornets, Kings). There are a total of 72 players in this dataset. Display D2.37 shows a dotplot of these salaries.



Display D2.38: Dotplot of salaries (in millions of dollars) of players on five NBA teams for the 2005-06 season. (Source: USA Today, April 2006.)

a. Write a short paragraph describing the features of this distribution of salaries.

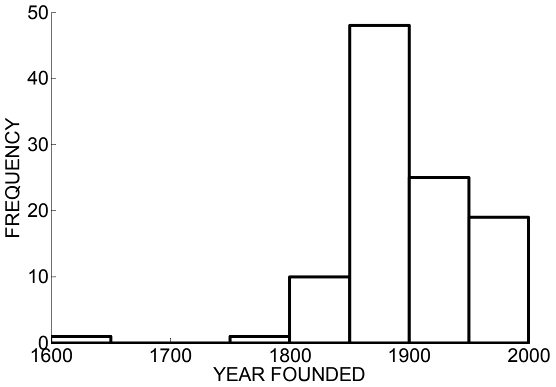
b. Estimate the number of players with a salary less than $2 million.

c. Estimate the most common salary among these players.

d. Suppose a "star" basketball player earns more than $10 million in this season. How many stars are there in this dataset? Given that this dataset includes the salaries for five teams, how many stars are there, on average, on each team?

10. **Year Founded for 104 Colleges**

It is interesting that Harvard University was founded in 1636, over one hundred years before the United States was founded. This motivates the question: "Generally, when were most of colleges in the United States founded?" To help answer this question, a histogram was constructed for the year founded for 104 U.S. colleges selected at random from a list of all colleges.



Display D2.39: Year founded for a selection of colleges. (Source: *U.S. News and World Report College*, 2004)

a. Describe the shape of these data.

b. Name a center value that represents a typical year founded for these schools.

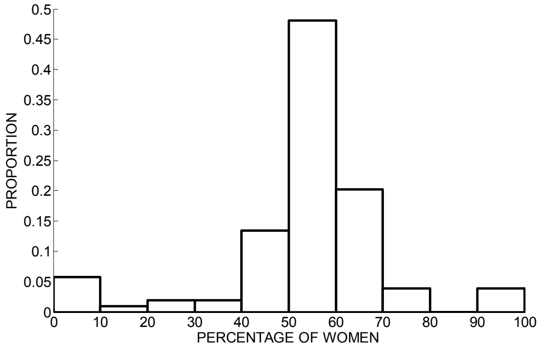
c. There is one outlier in the data corresponding to Harvard University that was founded in 1636. Can you guess at the locations of the colleges that were founded between 1750 and 1800?

d. Estimate the number of schools that were founded before 1900.

e. Estimate the number of schools that were founded in 1950 or later.

11. **Percentage of Women for 104 Colleges**

Many college freshmen are interested in the ratio of men to women at their school. This relative frequency histogram shows the percentage of women enrolled at the same 104 schools randomly selected for exercise 10:



Display D2.40: Percentage of women enrolled at a selection colleges. (Source: *U.S. News and World Report College*, 2004)

a. Describe the shape, center, spread, and any interesting features of this distribution.

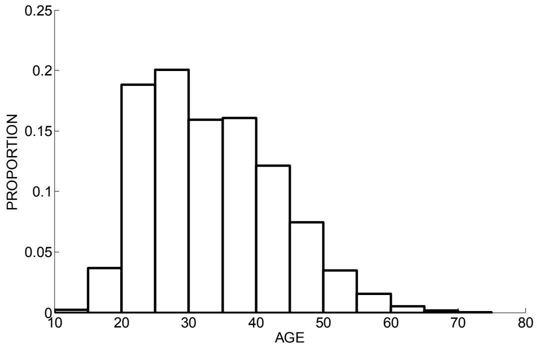
b. What proportion of schools have 50 percent or more women?

c. Based on this graph, do you believe that there are more women or men attending colleges in the United States? Explain.

d. What explanations can you provide for the colleges that have unusually small or large proportions of women?

12. **Ages of Women Participating in a Marathon**

The marathon race is unique in that it includes participants from a wide range of ages. This relative frequency histogram shows the ages of women participating in the Grandma's Marathon race in Duluth, Minnesota.



Display D2.41: Ages of women participating in the 2003 Grandma's Marathon, Duluth, Minnesota (Source: www.grandmamarathon.com)

a. Describe the shape of this distribution of ages.

b. What is a center value for the ages of women in this marathon?

c. What proportion of these women were age 20 or older, but less than age 30?

d. Find an interval that contains approximately the middle 50% of the runners' ages.

e. If a total of 10,000 women competed in this marathon, how many women were older than 50?

13. **Fares of Airplane Flights to Different Cities**

The National Council of Teachers of Mathematics is considering hosting a national conference in Detroit, Michigan. Assuming that participants may come from all of the United States, what is the fare for a round-trip airplane flight to the conference? To explore this question, consider this table of fares from Detroit, Michigan.

City	Fare ($)
Boston, MA	327
Chicago, IL	92
Denver, CO	198
Fairbanks, AK	696
Fargo, ND	369
Honolulu, HI	701
Houston, TX	219
Kansas City, MO	220

City	Fare ($)
New Orleans, LA	280
New York, NY	170
Orlando, FL	236
Philadelphia, PA	258
Phoenix, AZ	204
Portland, OR	352
Raleigh, NC	224
San Diego, CA	312

Las Vegas, NV	242	San Francisco, CA	310
Miami, FL	252	Sante Fe, NM	521

Display D2.42: Lowest-cost round-trip airplane fares from Detroit, Michigan, April 2004. (Source: *www.orbitz.com*)

a. Construct a dotplot, stemplot, and histogram for these fares. Which graph do you think best represents the data, and why?

b. Describe the distribution of fares, including shape, center, spread, and any interesting features.

c. Why is there so much spread in plane fares? Give several reasons for this spread.

14. **Nutritional Content of Ben & Jerry's Ice Cream**

Ben & Jerry's is one of the most famous makers of gourmet ice cream. The table below gives nutritional information for some of their flavors:

Flavor	**Calories**	**Sodium (grams)**
Brownie Batter	310	115
Butter Pecan	290	80
Cherry Garcia®	260	50
Chocolate	260	50
Chocolate Chip Cookie Dough	270	90
Chocolate Fudge Brownie™	270	80
Chubby Hubby®	330	160
Chunky Monkey®	300	45
Coffee	240	60
Coffee HEATH® Bar Crunch	290	115
Dublin Mudslide™	270	80
Everything But The... ®	320	80
Fudge Central®	300	60
Half Baked®	280	90

Flavor	**Calories**	**Sodium (grams)**
Karamel Sutra®	280	75
Mint Chocolate Cookie	270	100
New York Super Fudge Chunk®	310	55
Oatmeal Cookie Chunk	280	120
One Sweet Whirled	280	85
Peanut Butter Cup™	380	140
Phish Food®	280	90
Pistachio Pistachio®	280	125
Primary Berry Graham	270	110
Strawberry	240	50
Uncanny Cashew™	290	130
Vanilla	240	55
Vanilla HEATH® Bar Crunch	300	120
Vanilla Swiss Almond	280	65

Display D2.43: Calories and sodium in half-cup servings of select Ben & Jerry's ice cream flavors. (Source: *www.benjerry.com*)

a. Construct a histogram of Calories.

b. Write a short paragraph about the distribution of Calories, including shape, center, spread, and any interesting features.

c. What is your favorite flavor? How does the amount of calories in your flavor compare to the distribution of calories for the flavors in Display D2.41? (Is it average, below average, or above average?) (*Note:* If your favorite flavor and/or brand is not listed, try using the Internet to find its nutritional information.)

d. Construct a suitable graph for sodium.

e. Describe the distribution of sodium.

f. How does the amount of sodium in your favorite flavor compare to the distribution?

g. What explanations can you provide for the flavors that have unusually high amounts of sodium?

15. Data Shapes

Suppose you collect each of the following variables for each student in a high school class. Describe the expected shape of the distribution of the variable.

a. The number of first cousins.

b. The amount of money spent for the last haircut.

c. The pulse rate (number of beats in one minute).

d. The number of homework study hours in a week.

e. The composite score on the ACT exam.

16. Data Shapes

For each of the following data variables, construct a hypothetical graph of the distribution and describe its shape.

a. The month of the birthday (a number between 1 and 12) is recorded for a group of 1000 people.

b. The salaries for 50 people who work in a local company is collected.

c. A test with 100 possible points is administered to a large group of students. A majority of the students score more than 90 points.

d. The number of points scored by a basketball team is recorded for a large number of games.

TOPIC D3: SUMMARIES FOR DATA

SPOTLIGHT: NUTRITION VALUE OF ICE CREAM?

Ice cream is one of the favorite American desserts. Actually ice cream has a long history as it can be traced back to the 4th century B.C. The Roman emperor Nero (A.D. 36-68) had ice brought from the mountains and combined with fruit toppings and King Tang (A. D. 618-97) of Shang, China had a method of creating ice and milk concoctions. Europe was likely introduced to ice cream from China and "milk ices" were served to the Italian and French royal courts.

Ice cream was imported to the United States and was served by some famous Americans such as George Washington and Thomas Jefferson. The first ice cream parlor in America opened in New York City in 1776 and supposedly the term "ice cream" was first used by American colonists. There were great advances in the technology of making ice cream in the 19th century. Nancy Johnson in 1846 received a patent for a hand-cranked freezer that established the basic method for making ice cream still used today. The first large-scale commercial ice cream plant was started by Jacob Fussell in 1851. The ice cream cone was introduced at the 1904 St. Louis World's Fair. There is some debate about the origin of the ice cream sundae. One possible inventor of the sundae was Chester Platt who served ice cream with cherry syrup and candied cherry at his drug store in 1893 for Reverend John Scott on a Sunday and called the concoction "Cherry Sunday." Soft ice cream was introduced in the 20th century by a chemical research team in Britain who discovered a method of doubling the amount of air in ice cream.

The United States produces about 900 million gallons of ice cream annually. Almost one-tenth of the nation's milk supply is used to produce ice cream and other frozen desserts. Americans eat an average of about 20 quarts of ice cream annually. According to the web site http://www.sendicecream.com/, ice cream consumption is highest in July and August, most ice cream is purchased on Sunday, and Portland, Oregon purchases the most ice cream on a per capita basis.

In an article published by the Center of Science in the Public Interest (CSPI) in July/August 2003, there is a general concern about the excessive fat and calories contained in ice cream desserts served by major restaurants. The daily recommended number of calories for a male adult, age 25-50, is 2900 and a day's allowance for saturated fat is 20 grams. The CSPI article shows that many of the desserts served by Ben and Jerry's, TCBY, Baskin-Robbins, and Haagen-Dazs contain over 1000 calories and over 20 grams of saturated fat. One problem is that customers often don't know what they are getting, since these stores don't give nutrition labels to their products. "With ice cream portions like these, it is no wonder that two out of three Americans are overweight, diabetes rates are rising, and heart disease is the leading cause of death", says Marion Nestle, chair of the nutrition and food studies department at New York University.

PREVIEW

In the last topic, we were introduced to the notion of a data distribution and learned about different graphs useful for seeing particular characteristics of the distribution such as shape, average, and spread. In this topic, we first describe useful ways of summarizing categorical data and then we will be introduced to some basic methods for describing the average and spread for a group of quantitative data.

In this topic, the learning objectives are to:

- Understand how to summarize categorical data by computing percentages and a mode.
- Understand how to interpret two popular measures of center, the median and the mean, and understand when these two measures will be the same or different.
- Understand how to measure spread by the quartiles and by a typical size of a deviation from the mean.
- Relate a graph of a data distribution with the measures of center and spread.

NCTM Standards

✓In Grades 6-8, all students should find, use, and interpret measures of center and spread, including mean and interquartile range.

✓In Grades 9-12, all students should recognize how linear transformations of univariate data affect shape, center, and spread.

SUMMARIZING CATEGORICAL DATA

In this topic we focus on some numbers that are useful for summarizing a collection of data. First, let's consider data of the categorical type. In Topic D2, we looked at the country of birth of all professional baseball players who were born in the year 1975. We organized the data by use of the following frequency table.

Country of origin	Frequency
USA	135
Latin America	62
Other	8

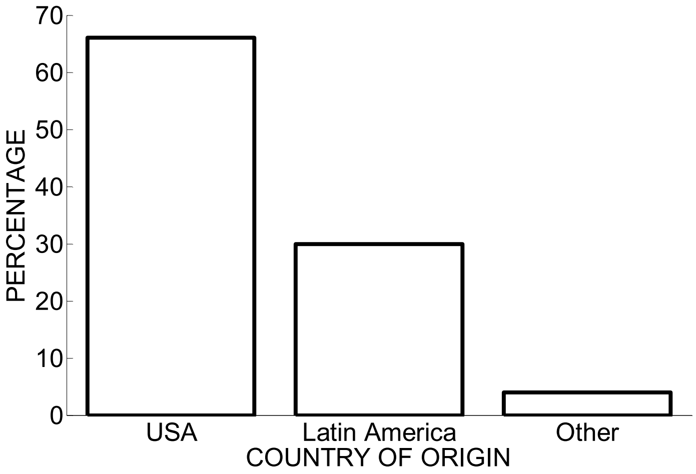
How do we summarize these data? First, to understand the relative sizes of the frequencies, it is helpful to compute the proportions of the categories – we find these by dividing each frequency by the total count (205). So, for example, the proportion of USA ballplayers is 135/205 = .659 and the proportion of Latin American players is 62/205 = .302. We convert these proportions to percentages in the table by multiplying by 100. If we multiply this proportion of American players (.659) by 100 and round to the nearest integer, we get the percentage of American ballplayers to be 66.

Country of origin	Frequency	Proportion	Percentage

USA	135	0.659	66
Latin America	62	0.302	30
Other	8	0.039	4
TOTAL	205	1.000	100

For categorical data, a useful summary is the *mode*, the category with the highest frequency or percentage. Here the mode of the country of origin is USA. Also it is helpful to describe other categories that have large percentages. For this example, we could say that approximately 66 percent of these baseball players are American, but a sizeable (30%) percentage of the players are from Latin America. The remaining countries make up only a small percentage of the players.

In topic D2, we introduced a bar chart where we graphed the category frequencies on the vertical scale against the category names on the horizontal scale. One variation of this display is to plot the category proportions or the category percentages on the vertical scale. The figure below shows a bar chart of the category percentages.



The frequency and percentage versions of the bar chart have the same visual appearance. But perhaps the percentage bar chart is more useful since one can read the category percentages directly from the graph.

PRACTICE: SUMMARIZING CATEGORICAL DATA

In the Practice Graphing Categorical Data section of Topic D2, we collected several variables for 50 used sedan cars.

1. We classified the cars into four groups by the manufacturer (General Motors, Ford, Chrysler, and Foreign). Copy the frequencies of the four groups from your work in Topic D2. Compute the proportion and percentage of each group.

Manufacturer	Frequency	Proportion	Percentage
General Motors			
Ford			
Chrysler			
Foreign			

2. Find the mode of the manufacturer for these 50 used cars.

3. Cars in this group were also classified by mileage (low, medium, and high). Construct a frequency table for the mileage below including the proportions and percentages.

Mileage	Frequency	Proportion	Percentage
Low			
Medium			
High			

4. Construct a bar chart for mileage using percentage as the variable on the vertical axis.
5. Find the mode of the mileage for these used cars.

HOW MANY CALORIES ARE IN AN "AVERAGE" SCOOP OF ICE CREAM? (INTRODUCING THE MEDIAN)

On the Ben and Jerry Ice Cream website http://www.benjerry.com/, one finds a list of many of their ice cream flavors and the number of calories contained in a single serving of each flavor. Here is a listing for 13 flavors:

Flavor	Calories/serving
Aloha Macadamia	330
Apple Crumble	280
Bovinity Divinity	290
Cherry Garcia™	260
Chocolate Chip Cookie Dough	300
Chocolate Fudge Brownie	280
Chubby Hubby™	350
Chunky Monkey™	310
Coffee Heath™ Bar Crunch	310
Concession Obsession	300
Festivus (limited edition)	300
Island Paradise	240
Kaberry Kaboom	240

A first step in understanding the variation in these calorie measurements is to construct a stemplot.

```
    24 | 00
    25 |
    26 | 0
    27 |
    28 | 00
    29 | 0
    30 | 000
    31 | 00
    32 |
    33 | 0
    34 |
    35 | 0

24|0 means 240 calories
```

We see a center cluster of calorie measurements in the 280 to 310 range and values ranging from 240 to 350.

We would like to summarize this distribution of calorie numbers with a single "average." There are a couple different averages that are commonly reported in the media. The first one we will discuss is the ***median*** (denoted by M) that is the *middle value* when the data values are arranged in ascending order.

We find the median in two steps:

- We find the position or location of the median in the list when the measurements are arranged in increasing order.
- The median is the data value that has the middle position.

What is the position of the median? We first arrange the calorie measurements in increasing order. We assign position 1 to the smallest value, position 2 to the next smallest value, etc.

Calories	240	240	260	280	280	290	300	300	300	310	310	330	350
Position	1	2	3	4	5	6	7	8	9	10	11	12	13

We can see that the middle position of the 13 measurements is 7. The median is the data value in the 7th position, which we see is M = 300.

Suppose, instead, that you have 10 measurements, like the first 10 calorie measurements, 330, 280, 290, 260, 300, 280, 350, 310, 310, 300. We list these measurements in ascending order:

Calories	240	240	260	280	280	290	300	300	300	310
Position	1	2	3	4	5	6	7	8	9	10

In this case (with an even number of measurements), there is not one middle measurement. In this case we define the position to be the average of the middle two positions (5+6)/2 = 5.5 and the median is defined to be the average of the measurements with these two middle positions. So M = (280 + 290)/2 = 280.5.

Generally if n is the number of values in the dataset, then the position of the median is given by

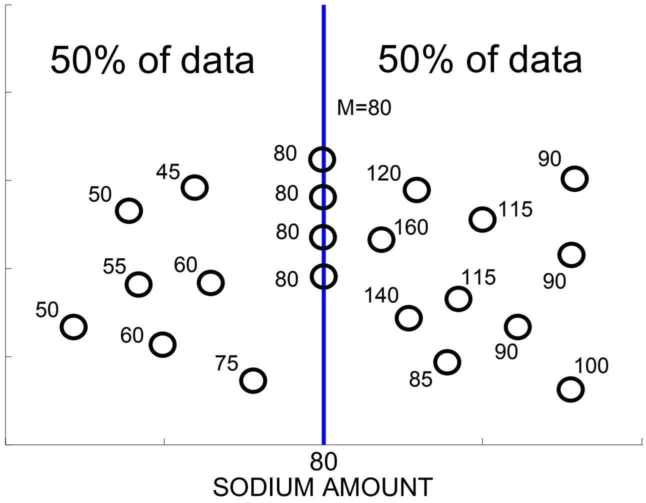
$$pos(M) = \frac{n+1}{2}.$$

- For our first example of 13 measurements, n = 13 and pos(M) = (13+1)/2 = 7.
- For our second example of 10 measurements, n = 10 and pos(M) = (10+1)/2 = 5.5

We illustrate the computation of the median for each of the data distributions (calories and sodium amounts for a collection of ice cream flavors) displayed using stemplots. For each distribution, we show the number of measurements n, the position of the median pos(M), and the value of the median M.

```
       Calories of                Sodium amounts
       28 flavors                 for 21 flavors

           24 | 000               4  | 5
           25 |                   5  | 005
           26 | 00                6  | 00
           27 | 00000             7  | 5
           28 | 0000000           8  | 00005
           29 | 000               9  | 000
           30 | 000               10 | 0
           31 | 00                11 | 55
           32 | 0                 12 | 0
           33 | 0                 13 |
           34 |                   14 | 0
           35 |                   15 |
           36 |                   16 | 0
           37 |
           38 | 0

24|0 means 240 calories           4|5 means 45 grams of sodium

   n = 28                         n = 21
   pos(M)=(28+1)/2=14.5           pos(M)=(21+1)/2= 11
   M=(280+280)/2=280 calories     M = 80 grams
```

The median has a simple interpretation as a center of a dataset. The value of M divides the data into two groups of the same size. So we can say (approximately) that half of the measurements are smaller than M and half are larger than M. For the sodium amounts, approximately half of the values are smaller than 80 grams. We illustrate the interpretation of the median by the graph below. The individual observations are represented by circles; the median M = 80 can be thought as a dividing line between the "low" sodium amounts and the "high" sodium amounts. When we calculate a median, the actual sodium measurements are not important – it only matters if the sodium amount is below or above the median M.



PRACTICE: COMPUTING A MEDIAN

Consider again the daily high temperatures (in degrees Fahrenheit) for Atlanta for the month of March 2005. A stemplot of these daily temperatures is displayed below.

```
    4 | 01
    4 | 78
    5 | 04
    5 | 555889
    6 | 01124
    6 | 5566777
    7 | 04
    7 | 5679
    8 | 0

4|0 means 40 degrees F
```

1. Find the sample size and the position of the median.

2. Find the median M.

3. Based on our work, we can say that (approximately) half of the daily high temperatures are larger than ____________.

4. Suppose by mistake the largest temperature 80 degrees really should have been 90 degrees. Calculate the median of the new dataset. Did the value of the median change? If so, by how much?

DEVIATIONS AND THE MEAN

The second measure of center, the ***mean*** (usually denoted by $\bar{x}$) is the value you get when you sum all of the values and divide by the number of values. For our collection of 13 calories of ice cream flavors, the mean is equal to

$$\bar{x} = \frac{330+280+290+260+300+280+350+310+310+300+300+240+240}{13} = 291.54$$

.

We saw that the median had a simple interpretation -- what's the interpretation of the mean?

To help understand the mean, we introduce a new idea – a *deviation*. The deviation of a data value from a given number, say c, is the difference of that data value from c:

deviation = data value – c.

For our original 13 ice cream calorie numbers, suppose that c = 300; this number can represent our guess at a typical calorie number. To obtain the deviations, we subtract 300 from each data value.

Calories	Deviation = Calories - 300
330	330 – 300 = 30
280	280 – 300 = -20
290	290 – 300 = -10
260	260 – 300 = -40
300	300 – 300 = 0

280	280 – 300 = -20
350	350 – 300 = 50
310	310 – 300 = 10
310	310 – 300 = 10
300	300 – 300 = 0
300	300 – 300 = 0
240	240 – 300 = -60
240	240 – 300 = -60

How good is our guess of 300 as a typical number of calories? We can judge the goodness of this guess by adding up all of the deviations – we call this the *sum of the deviations*.

SUM OF DEVIATIONS = 30 +(-20) + (-10) + … + (-60) + (-60) = -110.

Here the sum of the deviations is negative (-110) – this means that our guess of 300 as a typical number of calories is a bit high.

One way of defining a "best" measure of center is to find the value of c such that the sum of deviations about c is equal to zero. By trial and error, we could try different values of c, compute the sum of deviations about that value, and then find the value of c that makes the sum of deviations equal to zero. But fortunately we don't have to do this work -- one can show mathematically that the sum of deviations about the mean $\bar{x} = 291.54$ is equal to zero.

Another way of stating this property is that the sum of positive deviations about the mean (that is, the deviations of the data values above the mean) will be equal to the sum of negative deviations about the mean (corresponding to the data values below the mean).

PRACTICE: DEVIATIONS AND THE MEAN

1. The below table contains the high temperatures of Atlanta, Georgia in the first thirteen days in May 2005. In the below table, suppose that your guess at the "average" is 50

degrees. Find the deviations (in the first empty column of the table) and the sum of the deviations.

Temperature	Deviation = Temperature – 50	Deviation = Temperature - 55
41		
48		
55		
61		
66		
61		
67		
55		
47		
54		
41		
48		
55		

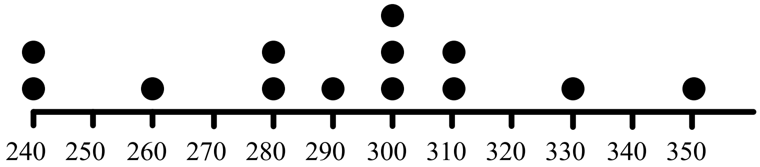
2. Now suppose that your guess at the average is 55 degrees. Find the deviations (in the last column of the table) and sum of deviations for this guess.

3. Based on your work from parts 1 and 2, do you think that 50 or 55 is closer to the mean $\bar{x}$?

4. Compute the mean $\bar{x}$. Was your choice in part 3 correct?

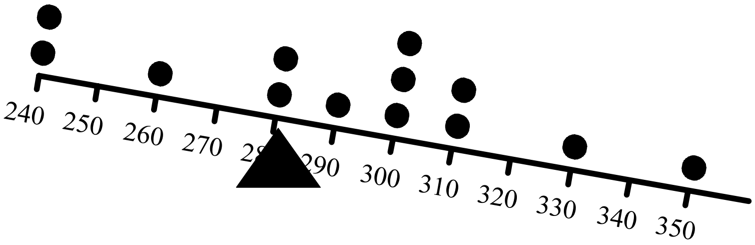
GEOMETRICAL INTERPRETATION OF THE MEAN

There is a geometrical interpretation of this result about deviations and the mean. First we draw a dotplot graph of the data.

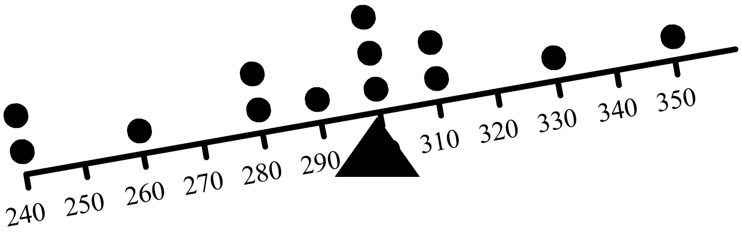


Suppose that we placed a stiff board under the graph and each dot on the graph represents a weight of a given amount. We place a fulcrum ▲ under the board and try to balance the weights.

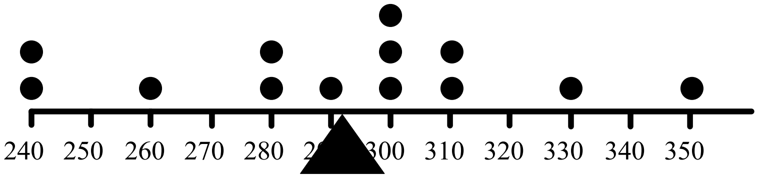
If we try to place the fulcrum at the value 280, the dotplot is unbalanced -- there is too much weight on the right. Here the positive deviations about 280 outweigh the negative deviations about 280.



We slide the fulcrum to the right at 300 -- now the dotplot is unbalanced with too much weight on the left. Here the negative deviations are greater than the positive deviations.



After some more moving of the fulcrum, we find a spot so that the board is balanced. The sum of the positive deviations is equal to the sum of the negative deviations.



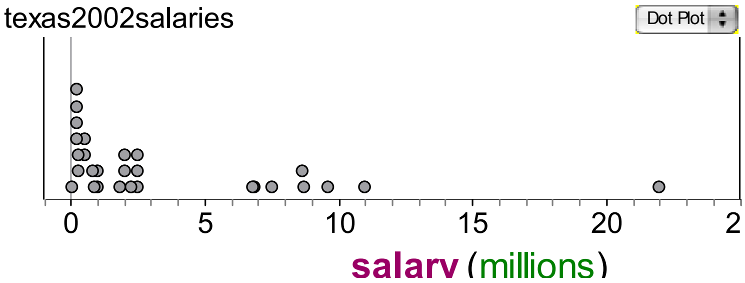
The location of the fulcrum will be at the mean value $\bar{x} = 291.54$.

COMPARING THE MEDIAN AND THE MEAN

In the first example, the median and mean of the group of 13 calorie measurements were approximately the same -- will that usually be the case?

Actually, no. In many situations, the mean and median will be different and one should understand why.

Suppose we look at the salaries of the players on the 2002 Texas Rangers baseball team. The salaries have been graphed below using a dotplot.



Note that there is a cluster of low values, a second cluster of salaries about 10,000,000 (10 million dollars) and a single large value (that corresponds to Alex Rodriguez who was making 22 million dollars that particular season).

Here the median and mean salaries are very different. The median salary is M = $2,000,000 and the mean is $\bar{x}$ = $3,634,728, so these two measures are over 1.5 million dollars apart.

Here are some comments about the interpretation of these measures and why the mean and median are different.

1. First, the median is easy to interpret. We can say that approximately half of the Rangers have salaries larger than 2 million dollars and half have salaries smaller than 2 million.
2. The mean is larger than the median since there are a number of large salaries that make the mean larger. When the dataset is skewed to the right, the mean will be larger than the median.
3. What is a better measure of average -- the mean or the median? The answer depends on what type of average you are interested in. If you are interested in a typical or representative salary, then the median is the better average, since many players have salaries close to 2 million dollars. The median is a reasonably typical salary. On the other hand, suppose the owner is concerned about the total amount of money that he or she is spending on payroll. The mean is a better measure for this owner since it includes all of the players' salaries and one can compute the total payroll from the mean. Here the mean is equal to $\bar{x}$ = $3,634,728 and there are 29 players – so the total payroll is

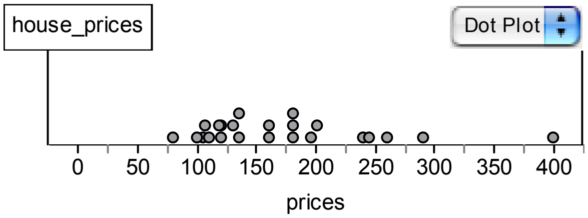
PAYROLL = 3,634,728 x 29 = $105,407,112.

The owner is paying over 105 million dollars for the salaries of his players.

PRACTICE: COMPARING THE MEDIAN AND THE MEAN

Recently the author collected the sale prices for some houses that were open for viewing on a particular weekend. Here are the prices (in thousands of dollars) and a graph of the data.

```
195   135   104   399   107   120   240   130   180   180   180   100   160
245   200   259   160    80   110   120   290   118   135
```



1. Write a short paragraph about these data.

2. Based on your description in part (a), do you believe that the mean and median would be approximately equal, or do you think that one measure would be larger than the other? Why?

3. Compute the mean and median for these data.

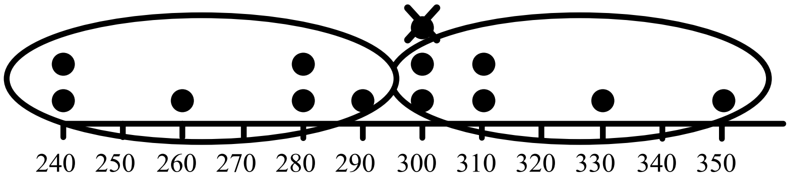
4. Would it be possible to change the value of a single observation, so that the mean and median for the new data would be approximately equal? Explain.

MEASURES OF SPREAD: QUARTILES AND IQR

In this first half of this topic, we talked about ways of measuring an average value in a dataset. Here we focus on ways of describing the spread or variation in a dataset of measurement data. To define our first measure of spread, the IQR, we need to define additional division points in our data.

Let's return to our Ben and Jerrys' ice cream example, where the number of calories in a single serving of 13 flavors of ice cream was recorded. We earlier found the median and the mean for these data. How can we measure the spread in the calorie values that we saw in the dotplot?

In this example we have 13 data values. To define a measure, we wish to divide the dataset into two halves. In a case such as this where the number of observations is odd, we discard the middle measurement (the median) indicated below with an X. Then the measurements can be divided equally into a lower half and an upper half that are circled below.



We find the median of the lower half and the median of the upper half of measurements. The median of the lower half of measurements is the median of {240, 240, 260, 280, 280, 290} which is equal to 270 calories. The median of the upper half is 310 calories.

These new dividing points are called *quartiles* -- the lower quartile Q_L is the value such that (approximately) one quarter of the data is smaller than Q_L, and the upper quartile Q_U is the value such that one quarter of the data is larger than Q_U. Here Q_L = 270 and Q_U = 310. By reporting the quartiles Q_L and Q_U, we get some idea about the spread in the data.

Unfortunately, there is not a universally accepted definition of a quartile -- you will find different definitions in different statistics books. Here is our definition of a quartile that is easy to learn.

1. After arranging the data in order, divide into two halves. If we have an even number of data values, we can make a clean break into halves. If there is an odd number of data values (as in the above example), we discard the median and divide the remaining observations into halves.
2. The lower and upper quartiles, Q_L and Q_U, are the medians of the lower half and the upper half of the data, respectively.

A useful summary of a group of measurement data are the five numbers

$$(\text{LO}, Q_L, \text{M}, Q_U, \text{HI}),$$

where LO and HI are respectively the lowest and highest values in the dataset -- we call this the ***five-number summary***.

To illustrate, we find the five-number summary for the following prices of houses (in thousands of dollars)

```
195   135   104   399   107   120   240   130   180   180   180   100   160
245   200   259   160    80   110   120   290   118   135
```

By graphing these data using a stemplot with ascending leaves, we have an ordered arrangement of the data. (Note that we have placed the unusually high price of 399 thousand dollars on a separate line so that the display isn't too long.) We show the calculation of the five-number summary below.

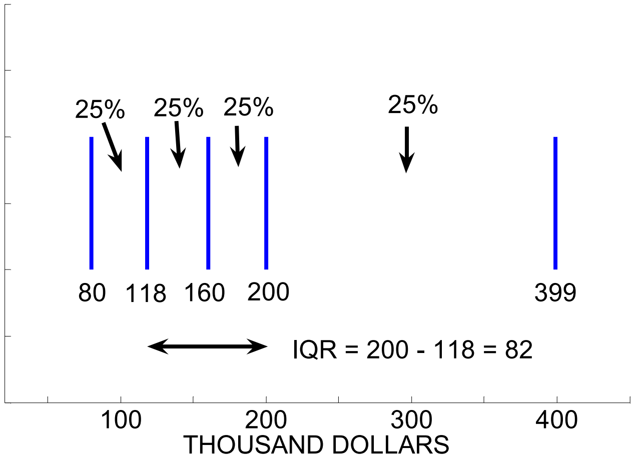
```
8  | 0            n = 23
9  |              pos(M) = (23+1)/2 = 12
10 | 047          M = 160
11 | 08
12 | 00
13 | 055          Since there are 23 observations, we discard the median and
14 |              break data into two groups of 11.
15 |
16 | 00           QL is median of lower half
17 |              pos(QL) = (11+1)/2 = 6
18 | 000          QL = 118
19 | 5
20 | 0            QU is median of upper half
21 |              pos(QU) = (11+1)/2 = 6
22 |              QU = 200
23 |
24 | 05
25 | 9
26 |              five-number summary is
27 |              (80, 118, 160, 200, 399)
28 |
29 | 0

HI 399

8|0 means 80 thousand dollars
```

A simple measure of spread in a dataset is the difference between the upper and lower quartiles -- we call this the *interquartile range* (IQR):

$$IQR = Q_U - Q_L.$$

The IQR is the spread of the middle 50% of the data.

We graphically show the location of the five-number summary below. We see that the numbers divide the house prices into four parts where 25% of the data is between the low value and Q_L (80 and 118 hundred thousand dollars), 25% falls between Q_L and M (118 and 160), and so on. Here the interquartile range is equal to $IQR = Q_U - Q_L =$ 200 – 118 = 82. This means that the spread of the middle half of the house prices is 82 thousand dollars.



PRACTICE: COMPUTING A FIVE-NUMBER SUMMARY

Consider again the daily high temperatures for Atlanta for March of 2005. A stemplot of the temperatures is shown below.

```
          4 | 01
          4 | 78
          5 | 04
          5 | 555889
          6 | 01124
          6 | 5566777
          7 | 04
          7 | 5679
          8 | 0

4|0 means 40 degrees
```

1. Compute the five-number summary.
2. Compute the interquartile range IQR.
3. On the number line below, mark using X's the locations of LO, Q_L , M, Q_U , and HI.

```
+----+----+----+----+----+----+----+----+-
40   45   50   55   60   65   70   75   80  degrees F
```

4. In words, describe what the IQR is telling you about the spread in the daily temperatures.
5. Find an interval that contains the lowest 25% of the temperatures.

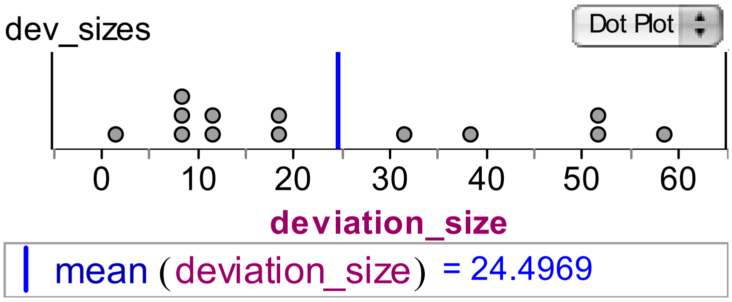
MEASURES OF SPREAD USING DEVIATIONS

Alternative measures of spread can be defined on the basis of the deviations. Recall a deviation is the difference of a data value from a particular number c. Suppose we consider the deviations of each value from the mean $\bar{x}$. In our ice cream calorie example, we show for each flavor, the calories (in a single scoop), the mean $\bar{x} = 291.54$ and the deviation = calories – $\bar{x}$. The deviation for the first value, 330, is deviation = 330 - 291.54= 38.46, the deviation for the second value, 280, is deviation = 280 - 291.54 = -11.54, and so on.

Flavor	Calories	Mean	Deviation	Absolute Deviation
Aloha Macadamia	330	291.54	38.46	38.46
Apple Crumble	280	291.54	-11.54	11.54
Bovinity Divinity	290	291.54	-1.54	1.54
Cherry Garcia™	260	291.54	-31.54	31.54
Chocolate Chip Cookie Dough	300	291.54	8.46	8.46
Chocolate Fudge Brownie	280	291.54	-11.54	11.54

Chubby Hubby™	350	291.54	58.46	58.46
Chunky Monkey™	310	291.54	18.46	18.46
Coffee Heath™ Bar Crunch	310	291.54	18.46	18.46
Concession Obsession	300	291.54	8.46	8.46
Festivus (limited edition)	300	291.54	8.46	8.46
Island Paradise	240	291.54	-51.54	51.54
Kaberry Kaboom	240	291.54	-51.54	51.54

A natural measure of spread of a dataset is the average or mean *size* of these deviations. A *size* of a deviation is simply its absolute value. We graph these deviation sizes using a dotplot:



Note that deviation sizes close to zero correspond to data values close to the mean and large deviation sizes (such as the three deviation sizes larger than 50 calories) correspond to calorie numbers that are far from the mean.

What is a typical deviation of a calorie number from its mean? A natural summary of deviation size is the Mean Absolute Deviation (or MAD for short) that is the mean of the absolute values of the deviations:

$$MAD = \frac{|330 - 291.54| + |280 - 291.54| + \cdots + |240 - 291.54|}{13} = 24.4969 .$$

In the table, we showed the absolute deviations in the rightmost column and found the mean of these values to be 24.50. So MAD = 24.50 – we can say that 24.5 calories is a typical or representative size of a deviation from the mean. In other words, on average, calories are 24.5 units from the mean. The above graph shows that the MAD is a reasonable measure of center of the deviation sizes.

A second measure of spread based on the deviations is the well-known *standard deviation*, which we abbreviate by s. This measure is based on the *squared deviations* instead of the deviation sizes. If we *square* each of the deviations, find the *sum* of the squared deviations, then the standard deviation, s, is defined to be

$$s=\sqrt{\frac{sum\ of\ \ squared\ deviations}{n-1}},$$

where *n* is the number of items in the dataset.

The table below illustrates the work in computing the standard deviation. Remembering that there are n = 13 observations, we find that

$$s=\sqrt{\frac{1479.17+133.17+\cdots+2556.37}{13-1}}=\sqrt{\frac{12369.21}{12}}=32.11.$$

Flavor	Calories	Mean	Deviation	Squared Deviation
Aloha Macadamia	330	291.54	38.46	1479.17
Apple Crumble	280	291.54	-11.54	133.17
Bovinity Divinity	290	291.54	-1.54	2.37
Cherry Garcia™	260	291.54	-31.54	994.77
Chocolate Chip Cookie Dough	300	291.54	8.46	71.57
Chocolate Fudge Brownie	280	291.54	-11.54	133.17
Chubby Hubby™	350	291.54	58.46	3417.57
Chunky Monkey™	310	291.54	18.46	340.77
Coffee Heath™ Bar Crunch	310	291.54	18.46	340.77
Concession Obsession	300	291.54	8.46	71.57

Festivus (limited edition)	300	291.54	8.46	71.57
Island Paradise	240	291.54	-51.54	2656.37
Kaberry Kaboom	240	291.54	-51.54	2656.37
				Sum of Squared Deviations = 12369.21

For these calorie numbers, we have computed two measures of spread based on the deviations MAD = 24.50 and s = 32.11. Both measures represent typical distances of the data from the mean $\bar{x}$ = 291.54 calories. Which is a better measure? The MAD is probably easier to interpret since it is a simple function of the sizes of the deviations. We will see next that the standard deviation has a nice interpretation when we have a set of data whose distribution is approximately bell shaped.

PRACTICE: MEASURES OF SPREAD USING DEVIATIONS

The following table lists the house prices (in thousands of dollars) for homes that were on sale and available for viewing in a recent weekend in the hometown of the author. The mean price of these homes is $\bar{x}$ = 171.6. The table also lists the deviations from the mean and the squared deviations.

1. Compute the absolute deviations in the table. Construct a dotplot of these absolute deviations on the following grid.

```
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
0   10  20  30  40  50  60  70  80  90 100 110 120 130 140 150 160 170 180 190 200 210 220
```

2. Find a house price that is close to the mean. Will this price have a small or large deviation size?

3. If a house price has a large deviation size, is this price close or far away from the mean?

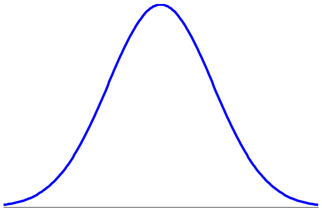
4. Compute the mean absolute deviation. Interpret what this number tells you about the spread of the house prices.

5. Complete the empty cells of the table and compute the standard deviation of the house prices s. Is the value of s close to the value of the MAD? Explain why they should be similar in size.

Price in thousands of $	Deviation	Absolute Deviation	Squared Deviation	Price	Deviation	Absolute Deviation	Squared Deviation
195	23.4		547.56	160	-11.6		134.56
135	-36.6		1339.56	245			
104	-67.6		4569.76	200			
399	227.4		51710.76	259	87.4		7638.76
107				160			
120	-51.6		2662.56	80	-91.6		8390.56
240				110	-61.6		3794.56
130	-41.6		1730.56	120	-51.6		2662.56
180	8.4		70.56	290	118.4		14018.56
180	8.4		70.56	118	-53.6		2872.96
180	8.4		70.56	135	-36.6		1339.56
100	-71.6		5126.56				

INTERPRETING S: THE 68/95/99.7 RULE FOR BELL-SHAPED DATA

Unlike the IQR, there is no general simple interpretation for the standard deviation s. However, if the data have an approximate symmetric bell or bell shape, shown here,



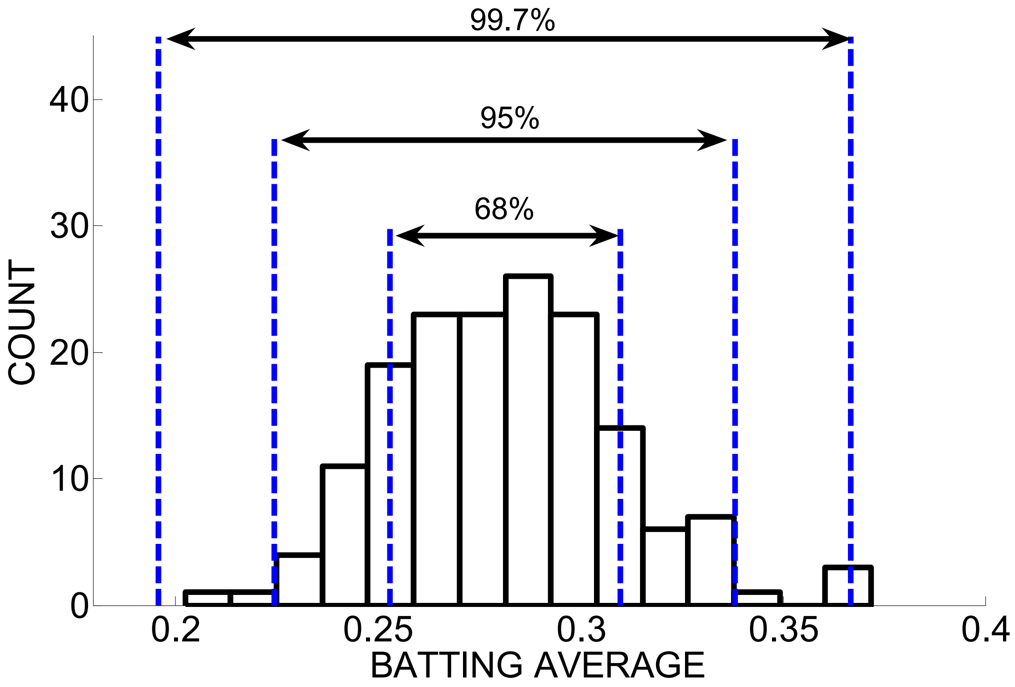
then there is a nice interpretation for s. If the data are bell-shaped, then we expect

- about 68% of the data will fall in the interval ($\bar{x}$ - s, $\bar{x}$ + s)
- about 95% of the data will fall in the interval ($\bar{x}$ - 2 s, $\bar{x}$ + 2 s)
- about 99.7% of the data will fall in the interval ($\bar{x}$ - 3 s, $\bar{x}$ + 3 s)

An example of a dataset that is bell-shaped is the collection of batting averages for all baseball players who are "regular" players during a particular baseball season. The figure below displays a histogram of the batting averages for the 162 players. We can compute the mean and standard deviation of these batting averages to be .281 and .028, respectively. Then we expect

- 68% of these batting averages to fall between .281 – .028 and .281 + .028 = .253 and .309.
- 95% of these batting averages to fall between .281 – 2 ×.028 and .281 + 2 × .028 = .225 and .337.
- 99.7% of these batting averages to fall between .281 – 3 ×.028 and .281 + 3 × .028 = .197 and .365.

These intervals are shown on the figure together with the expected percentages.



We can check the validity of the rule in this example by actually counting how many batting averages fall in the above intervals. Looking at the data, the table gives the number and percentage of values in each interval:

Interval	Number of batting intervals in interval	Percentage in interval
(.253, .309)	109	109/162 × = 67.3
(.225, .337)	156	156/162 × = 96.3
(.197, .365)	160	160/162 × = 98.8

We see that the interval percentages match up pretty well with the expected percentages 68, 95, and 99.7. This is expected since the batting averages have a distribution that is close to a bell-shape.

PRACTICE: THE 68/95/99.7 RULE

Below we apply this rule to our house prices dataset.

1. Does the dataset have a bell-shaped distribution? If not, describe the shape of its distribution.

2. The mean and standard deviation of these 23 house prices are 171.6 and 75.1, respectively. Find the interval ($\bar{x}$ - s, $\bar{x}$ + s).

3. Of the 23 house prices, how many prices fall in the interval you found in part 2?

4. What percentage of the house prices fall in the interval? Is this percentage close to the expected percentage 68?

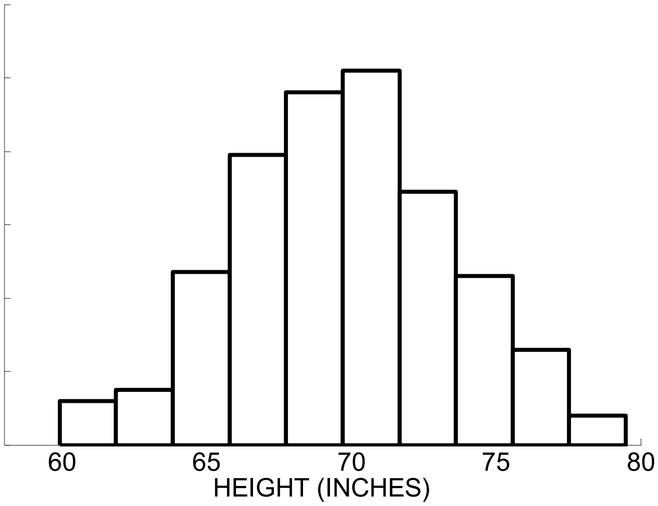
5. Find the interval ($\bar{x}$ - 2 s, $\bar{x}$ + 2 s) and find the percentage of prices that fall in this interval. Is this percentage close to 95?

6. Find the interval ($\bar{x}$ - 3 s, $\bar{x}$ + 3 s) and find the percentage of prices that fall in this interval. Is this percentage close to 99.7?

7. You may find that the percentages you found in parts 4, 5, and 6 aren't close to the values 68, 95, and 99.7 in our rule. Relating back to your answer to part 1, can you suggest a reason why the rule may not work well in this particular situation?

SPECIAL NOTE: When the data is approximately bell-shaped, then one can estimate the value of the standard deviation directly from a graph such as a histogram. To illustrate, consider the following graph of the heights (in inches) of a group of 500 men that appears to be bell-shaped. To estimate the standard deviation, we find from the graph an interval that contains approximately 95% of the heights. From the 68/95/99.7 rule, we expect 95% of the data to fall in the interval ($\bar{x}$ - 2 s, $\bar{x}$ + 2 s). By equating the width of this interval (4 s) to our estimate, we can obtain an approximate value of s.

Here the interval (61, 78) seems to contain most of the heights; this interval has a width of 78 – 61 = 17. So 4 s = 17 and solving for s, we get s = 17/4 = 4.25. To see if this is a reasonable answer, we compute the actual value of s to be 3.8 which is close to our estimate.



ACTIVITY: COLLECTING SOME DATA ON CITIES

1. Go to www.cityrating.com (click on City Guides or Weather History links).

2. For each of 20 cities of your choice (choose cities across a broad range of the U.S.), collect two quantitative variables of your choice (some possibilities might be population, percentage of women, median age, average temperature during a particular month, precipitation, etc.). Put your data in the table below.

My Variable 1 was _______________; my Variable 2 was ____________
Units of Variable 1 ____________; units of Variable 2 _______________

Number	CITY	Variable 1	Variable 2
1			
2			

3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			

3. For each variable …

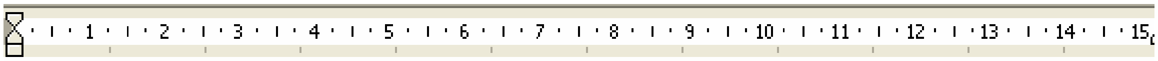
- draw a good stemplot of the data
- find a five-number summary
- find a city that represents the "average" for the variable
- find a measure of spread
- write a descriptive paragraph about the data

ACTIVITY: V IS FOR VARIATION

DESCRIPTION: In this activity, we look at the different measures of the average deviation considering the deviations of the data values from the mean. We use a "V" measurement, specifically the distance when you make a "V" with your fingers, to emphasize the focus on measures of variation. In comparing groups, we will see that comparing measures of variation tells us something very different than comparing measures of average.

This activity works best with groups of 3-5 students.

1. Measure your "V-span" as follows: With the palm of your writing hand on a flat surface, make a "V" between the index and middle fingers. Measure the distance (in centimeters) from the outside of your index finger tip to the outside of your middle finger tip when spread as far as possible.



Record the V-span (in cm) for all members of your group below. (We'll fill in the Deviation Size and Square of Deviation columns later.)

Student	V-span (cm)	Deviation Size	Square of Deviation

2. Construct a dotplot of the V-spans of your group. Label each V-span with the student's initials. Compute the mean and show the value of the mean on the graph.

3. For each V-span measurement, compute the deviation size, defined to be the absolute value of the difference of the measurement from the mean. Place these absolute deviations in the table.

4. Construct a dotplot of the deviation sizes. This graph shows you how far on average each member's V-span falls from the mean. By looking at your graph, what is a typical deviation size?

5. There are several ways of computing an "average" deviation size. One way is to compute the mean absolute deviation, called MAD, equal to the mean of the deviation sizes. Another way is to compute the standard deviation that is found using the formula

$$s = \sqrt{\frac{sum\ of\ squared\ deviations}{(number\ of\ measurements) - 1}}.$$

Compute the MAD and standard deviation for your group's measurements. (The squared deviation column in the above table can be helpful in computing the standard deviation.)

6. Suppose one group has a MAD value that is very close to zero. What does this say about the V-spans in this group?

7. Suppose Group A has four women and Group B consists of two men and two women. Which group would have a larger value of MAD -- Group A or Group B? Explain.

8. Compare the values of MAD across groups in your class. Explain why some groups have large deviations and other groups have small deviations.

9. Suppose you went to a large family gathering and measured the V-spans for all people. Now that you have the V-spans of people at this family gathering, suppose you want to compare the V-spans for the children with the V-spans for the adults. Which group (adults or children) will have the larger median V-span? Why?

10. For the same data of V-spans introduced in question 9, which group (adults or children) will have the larger MAD of the V-spans? Explain.

TECHNOLOGY LAB – DEVIATIONS, THE MEAN, AND MEASURES OF SPREAD

PART A: Deviations and the Mean

Open up a new Fathom document.

(a) The dataset baseball_ages.txt contains the ages of 20 randomly selected professional baseball players. Import this into Fathom.

(b) Define a variable m. Drag down a Slider Slider . Double-click on this Slider to change its properties.

- Change the name of the Slider from V1 to m.
- Make the Lower value 20 and the Upper value 40.

(c) Construct a dotplot of the age variable. With the graph selected, choose the menu item Graph -> Plot Value. In the Expression for Value, type in "m." You should see the value of m displayed on the graph.

(d) Now we will define a new Attribute called "deviation."

- Click in the first empty box next to "age" and type in "deviation."
- Select the deviation Attribute and choose Edit Formula from the Edit Menu.

In the Formula box, type

age – m

(For each age, the deviation will be the difference between the age and m.)

(e) We will compute the sum of the deviations by use of the Summary Table.

Drag a Summary Table from the Fathom shelf
Drag the variable "deviation" to the Summary Table – you will see the mean displayed.
Double-click on mean() – change the formula from mean() to sum().

The Summary Table shows the sum of the deviations of the ages from the value m.

(f) By moving the value of m on the Slider, find the value of m such that the sum of the deviations is equal to 0. (This is the value of m that balances the positive and negative deviations.)

Questions:

1.: If m is equal to 25, find the deviation for Andy Phillips.

2. If m is equal to 25, find the deviation for Fred McGriff.

3. What was the value of m such that the sum of deviations is equal to 0?

4. The value that you found in Q3 – is this a popular measure of "average", such as the mean or median?

5. Suppose that you wish to add three players to our group of players from the list below

Jim Thome, (born 1970)
David Bell, (born 1972)
Eric Milton, (born 1975)
Randy Wolf, (born 1976)
Doug Glanville, (born 1970)
Marlon Byrd (born 1977)

so that the sum of deviations about m remains 0. Which three players can you add?

PART B: Using Deviations to Construct Two Measures of Spread

We can define different measures of spread by using the deviations about the mean.

Make sure that m is equal to the mean of the player ages.

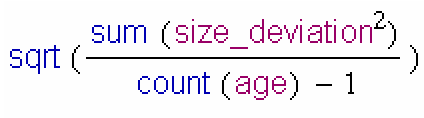
(a) Define a new Attribute called size_deviation . Select this new Attribute and choose Edit Formula from the Edit Menu. In the Formula box, type abs(deviation). This will give you the size of each deviation from the mean.

(b) Graph the sizes of the deviations using a dotplot. Write a short paragraph about these sizes, including comments about the smallest and largest values and a "typical" value.

Different measures of spread can be defined based on these deviation sizes.

(c) One possibility is to compute the mean of these sizes – this is called the MAD (for mean absolute deviation). Find this by dragging the Attribute size_deviation to a Summary Table. The MAD is equal to __________ .

(d) Another way of summarizing these sizes is by means of the "standard" deviation. To compute this, we square each deviation size, find the sum of these squared deviations, divide the result by the sample size minus one, and take the square root of the answer. In the Summary Table you just used, select Summary -> Add Formula and type the following formula.



$$\text{sqrt}\left(\frac{\text{sum}\,(\text{size\_deviation}^2)}{\text{count}\,(\text{age})\; - 1}\right)$$

.

The standard deviation here is equal to _____________.

(e) Compare the values of MAD and the standard deviation – which is larger? Looking at your dotplot of the deviation sizes, which seems to be a better "average" of the sizes? Why is one a better average of deviation size?

ACTIVITY: MEASUREMENT BIAS

DESCRIPTION: In this activity, we are introduced to the notion of measurement bias. We get experience in making measurements where a bias is present. That means that there is a known tendency to take measurements that are too small, or too large, on average. By exploring the distribution of measurements and knowing the true value, we can measure the size of the bias.

MATERIALS NEEDED: Two strings of different lengths, where the exact length of each string is known. A set of cardboard measuring instruments.

PART A: Bias in measuring the length of strings

1. Collecting the data

Look carefully at the string (marked A) your instructor is holding out straight. Without using any measuring instrument (except your eyes), estimate the length of the string to the nearest whole inch.

LENGTH OF STRING A = ______________

Your instructor will collect the estimated string lengths and give you the data for your class. You will use the data in the next part of the activity.

2. Describing the data graphically

(a) Make at least two different plots of the data on string length.

(b) Describe the plots of the data in terms of symmetry or skewness of the distribution; clusters and gaps that might be present and outliers that might be present, including a possible reason for the outliers.

3. Describing the data numerically

(a) Compute the following numerical summaries of the data: mean, median, standard deviation, and interquartile range.

(b) Which of these measures seems to provide the best description of center? Why?

(c) Which of these measures seems to provide the best description of variability? Why?

4. Collecting and summarizing another set of data

(a) Your instructor is now holding another string, string B. As before, estimate the length of the string to the nearest whole inch.

LENGTH OF STRING B = ______________

Your instructor will collect the estimated string lengths and give you the data for your class.

(b) Describe the data for string B using the graphical and numerical techniques you found most useful in the analysis of the data from string A.

5. Making comparisons

From your analysis of the two sets of data, decide which is the longer string. How much longer do you estimate it is?

6. Determining the bias

(a) Your instructor will provide you with the correct lengths for each string. Plot the correct values on the plots of the data made previously. What do you see?

(b) It is likely that the true value is not at the center of the data display. This discrepancy between the center of the measurements and the true value is called bias. Bias is a property of the measurement system, not of an individual person making an estimate. Does the "system" of estimating string lengths appear to be biased? What factors might be causing the bias?

(c) What was the effect of the bias on your answer to step 5? That is, does the bias affect the accuracy of your estimate of which string is longer and by how much? Why is this the case?

PART B: Optical illusions

Optical illusions are related to bias. There are many demonstrations of optical illusions that help prove this point, and some of them, such as the one described below, are well-suited to studies of bias in a measurement process or device.

1. You will be given a cardboard measuring instrument. The goal is to make the line with arrowtails at the end equal in length to the line with arrowheads at the end. To make this measurement, you slide the line with arrowtails out until you think the lengths of the two lines match.

When you are done, turn the card over and record the length of your line to the nearest tenth of a centimeter.

LENGTH OF LINE = ____________

2. Your instructor will collect the lengths of the lines from the class. Using a similar analysis as you did in the first activity, graph and summarize the batch of line lengths. Your instructor will give you the true length of the line. Determine the bias of this measurement.

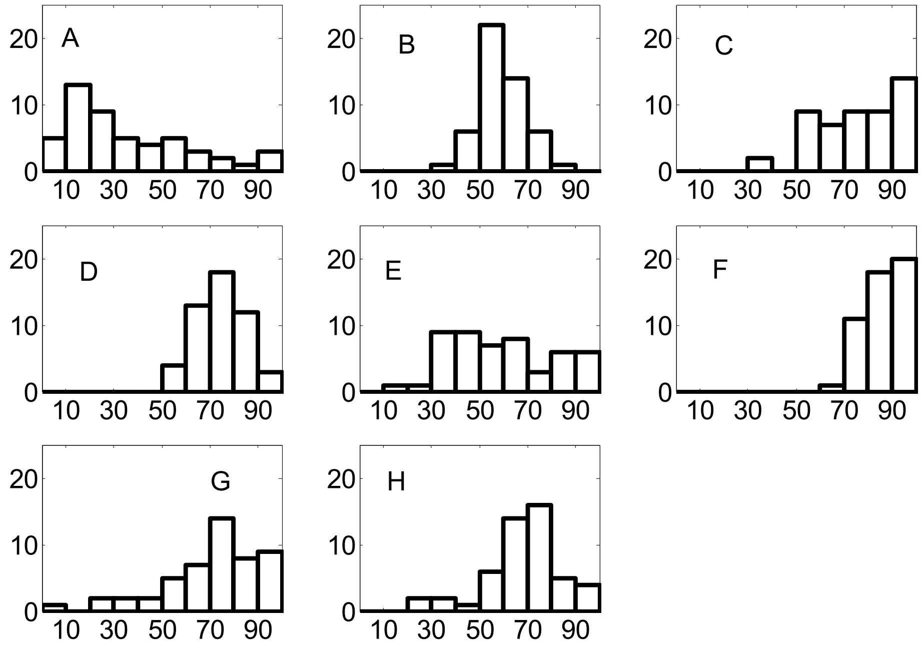
Extension

Find a printed article using data that are subject to a bias that could have a dramatic effect on the conclusions reached in the article. Summarize the conclusions in the article that are based on data, discuss possible biasing factors in the way the data were collected or analyzed, and explain how the bias in the data might affect the conclusion. (Be sure to separate bias in the data from biased reporting of the conclusions for other reasons.)

ACTIVITY: MATCHING STATISTICS WITH HISTOGRAMS

DESCRIPTION: In this activity, we will match up histograms and their corresponding summary statistics. The relative locations of the mean and median are informative about the shape of the distribution of the data. In addition, the value of the standard deviation is useful in understanding the spread of the distribution. For bell-shaped data, by the 68/95/99.7 rule, the standard deviation is helpful in understanding the proportion of data in particular intervals. Also we described how to use the 68/95/99.7 rule to estimate the standard deviation directly from a histogram.

Below are the histograms for eight datasets and following are summary statistics (the mean, median, and standard deviation) for eight datasets.



In the "Histgm" row of the below table, write down the letter of the matching histogram.

Dataset	1	2	3	4	5	6	7	8
mean	59.44	60.58	75.00	69.02	86.68	77.12	36.48	71.14
median	60.00	56.00	76.00	70.50	86.00	79.50	30.00	75.00
st dev	9.46	20.79	10.58	16.64	8.42	17.47	26.46	20.74
Histgm								

Classroom Capsule: Summarizing Risky Behavior

Overview: This activity introduces the main ways of summarizing a dataset by an "average value" and a measure of spread. This particular dataset illustrates a situation where the measures of center can be different. Two measures of spread are described; one that is based on the quartiles and a second that is based on deviations from the mean.

Objectives: The students will learn the basic principles in the computation of the mean, median, IQR, and the standard deviation. The mean and median can be different in value, and the students will understand why the two measures can be different. The students will see how one can interpret a five number summary. Also the students will gain some intuition on the standard deviation s by looking at the distribution of the sizes of the deviations.

Description:

Part 1: Measures of Center

The following table gives the percentage of young people, aged 11, 13, 15, who have been drunk two or more times for a group of 29 countries in the UNICEF study. This variable that we will call "risk" is one measure of risk of children who are living in a particular country.

Country	Pct.	Country	Pct.
Austria	15.1	Portugal	12.6
Belgium	14.5	Spain	10.2
Canada	19.8	Sweden	16.1
Czech_Republic	14.7	Switzerland	13.6
Denmark	20.1	United_Kingdom	30.8
Finland	24.7	United_States	11.6
France	8	Croatia	13.6
Germany	17.7	Estonia	23.9
Greece	10	Israel	9.3
Hungary	16.4	Latvia	16.5
Ireland	13.8	Lithuania	24.7
Italy	9.7	Malta	10.7
Netherlands	12.9	Russian_Federation	19.4
Norway	15.6	Slovenia	18.2
Poland	15.2		

1. Construct a dotplot of risk. Describe the basic shape of this data. Are there any particular countries that stand out with high or low values of risk?

2. Order the countries from smallest to largest value of risk; write below the Country and the corresponding risk percentage. (The first two countries with the smallest two values are written to start your work.)

Country	Pct.
France	8
Israel	9.3

3. Find the middle value of Pct M. This is the median that divides the risk values into a lower half and an upper half. What country has this median value?

4. Find the sum of the risk values $\sum x$ and use this sum to find the mean Pct $\bar{x}$.

5. Compare the values of the median and the mean. Which value is larger? Looking at the shape of the data distribution, what characteristics of the distribution would cause the two measures of center to be different? Explain.

6. There is one notable outlier in this distribution – what country has this outlier?

7. Suppose we remove this one outlier from the distribution. Recompute the values of the median and the mean. (To recompute the mean quickly, first compute the value of the new sum $\sum x$.)

8. Which measure of center, the median or the mean, is most changed with the removal of this outlier?

9. We say that a summary measure is resistant if it is not greatly changed with the inclusion of an outlier. Which measure of center, the median or the mean, is a resistant measure?

Part II: Measures of spread

All of the countries are listed in order from the smallest value of risk to the largest value of risk. The value of the median is marked with an M.

Country	Pct.	Deviation
France	8	
Israel	9.3	
Italy	9.7	
Greece	10	
Spain	10.2	
Malta	10.7	
United_States	11.6	
Portugal	12.6	
Netherlands	12.9	
Switzerland	13.6	

Croatia	13.6	
Ireland	13.8	
Belgium	14.5	
Czech_Republic	14.7	
Austria	15.1	M
Poland	15.2	
Norway	15.6	
Sweden	16.1	
Hungary	16.4	
Latvia	16.5	
Germany	17.7	
Slovenia	18.2	
Russian_Federation	19.4	
Canada	19.8	
Denmark	20.1	
Estonia	23.9	
Finland	24.7	
Lithuania	24.7	
United_Kingdom	30.8	

1. To start thinking about measuring the spread of this dataset, suppose you divide the risk values into a lower half (the values smaller than the median M) and an upper half (the values larger than the median). Here there are an odd number of values. If we remove the median M (corresponding to Austria), then we will have an even number of values and we can evenly divide the data into two halves. Circle these two halves of values.

2. Find the median of the lower half of values and the median of the upper half of values. Write them below. These are respectively the lower quartile and the upper quartile.

3. The five number summary consists of the smallest value, the lower quartile, the median, the upper quartile, and the largest value. Write these five numbers below.

4. On a number line, mark the locations of the five numbers.

5. These five numbers divide the dataset into quarters. Fill in the blanks in these statements.

One half of the countries have a risk value smaller than _________.

One quarter of the countries have risk values larger than _________.

Three quarters of the countries have risk values larger than _________.

Half of the countries have risk values between _______ and _________.

(There are several possible answers to the last question.)

6. The interquartile range, IQR, is the difference between the upper and lower quartiles. For these data, IQR = _________ .

7. Another way of thinking about spread in a collection is based on the notion of deviation that is defined to be a value minus the mean.

Deviation = value - $\bar{x}$,

For the risk data, $\bar{x}$ = 15.84. The deviation for Ireland would be

Deviation = Ireland's risk - $\bar{x}$ = 13.8 – 15.84 = -2.04.

This means that Ireland's risk percentage is about 2 points below the mean.

Compute the deviations for all countries and put your results in the table.

8. Suppose we are interested in a typical size of a deviation. (The size of a deviation is simply the absolute value of a deviation. For example, the size of the deviation for Ireland is +2.04.)

Construct a dotplot of the deviation sizes below.

9. Looking at the graph, find a typical deviation size.

10. One way of finding a typical deviation is to find the mean of these absolute deviations or MAD for short. Find the MAD for these data.

11. There is a second measure of computing a typical deviation size, called a standard deviation, or s for short. Use your calculator to compute s. Compare your answer with your answers in parts 9 and 10.

Share and Summarize: Here are some important points to mention when you discuss the answers to the activity.

1. It is important to focus not on the interpretation of a particular statistic, but rather the interpretation or use of this statistic.

2. The median has a simple interpretation – essentially it is the value that divides the data into halves. The interpretation of the mean isn't quite so obvious. The mean can be thought as the value that balances the distribution. (Think of data values as weights on a

number line and the mean is the value on a fulcrum that balances the weight.) Alternately, you can say the mean is the value that balances the deviations – the sum of the deviations about the mean is equal to zero.

3. Once you're talked about dividing the data into two halves to compute the median, then further division into four parts motivates the definition of the quartiles. The IQR is simply the distance between the two quartiles.

4. The idea of a deviation is important and can be described through simple examples. If one looks at a graph of the sizes of all deviations, you can talk about a typical deviation size, and that motivates the definition of the MAD and the standard deviation.

Application or Extension: Find one dataset that is symmetric and a second dataset that is strongly skewed. In each case, compute the median and the mean and compare these measures of center. When should you expect the median and mean to be of similar size, and when should you expect them to be different?

WRAP-UP

In this topic, we discussed various ways of summarizing a single dataset. For categorical data, it is helpful to find the percentages of data values in each category and the mode is the category with the highest percentage. For quantitative data, there are two primary measures of center or "average", the median and the mean. The *median* has a clear interpretation – it is the value that separates the data into halves. The *mean* is the value such that the sum of the positive deviations from that value is equal to the sum of the negative deviations. The best choice of average depends on the dataset; we saw that the mean can be influenced by a few extreme values. The *interquartile range (IQR)* is one measure of spread that has a clear interpretation – it is the width of the middle half of the data. A second type of measure is based on the deviations about the mean. The MAD is the average size of a deviation, the mean of all absolute deviations, and the *standard deviation* is based on the sum of squared deviations. The standard deviation is especially useful for bell-shaped data and we can use it to predict the proportion of data within one, two, and three standard deviations about the mean.

EXERCISES

1. **Bird Watching**

The Great Backyard Bird Count is an annual four-day event where bird watchers all over the country count birds of all species. The results of this count, published at http://www.birdsource.org/gbbc/ give a snapshot of the numbers, kinds, and distribution of birds all over the country. The following table gives the frequencies of all types of owls spotted by bird watchers in Ohio who participated in the 2006 survey.

Owl	Frequency	Percentage
Barn Owl	6	
Eastern Screech Owl	13	
Great Horned Owl	16	
Snowy Owl	9	
Barred Owl	27	
Long-eared Owl	3	
Short-eared Owl	14	

a. Find the percentage of each type of owl spotted.

b. Construct a bar graph of owl type where percentage is graphed on the vertical axis.

c. Find the mode.

d. What percentage of owls were either of the Short-eared or Long-eared varieties?

2. **Position and State Affiliation or Nationality of Professional Basketball Players**

The following table gives the position and state affiliation or nationality of all players from four teams (Cleveland, New Jersey, Detroit, and Miami) who made to the second-round of the playoffs from the Eastern Conference in 2005-2006.

Player	Position	FROM	Player	Position	FROM
Shandon Anderson	G-F	Georgia	Anderson Varejao	F	Brazil
Michael Doleac	C	Utah	Chauncey Billups	G	Colorado

Udonis Haslem	F	Florida	Kelvin Cato	C	Iowa State
Jason Kapono	F	UCLA	Dale Davis	C-F	Clemson
Alonzo Mourning	C	Georgetown	Carlos Delfino	G	Argentina
Shaquille O'Neal	C	Louisiana State	Tony Delk	G	Kentucky
Gary Payton	G	Oregon State	Maurice Evans	G	Texas
James Posey	G-F	Xavier (Ohio)	Richard Hamilton	G	Connecticut
Wayne Simien	F	Kansas	Lindsey Hunter	G	Jackson State
Dwyane Wade	G	Marquette	Jason Maxiell	F	Cincinnati
Antoine Walker	F	Kentucky	Antonio McDyess	F	Alabama
Jason Williams	G	Florida	Tayshaun Prince	F	Kentucky
Earl Barron *	C	Memphis	Ben Wallace – C	C	Virginia Union
Dorell Wright *	F	South Kent Prep HS (Lawndale, CA)	Rasheed Wallace	F	North Carolina
Drew Gooden	F	Kansas	Vince Carter	G	North Carolina
Stephen Graham	G-F	Oklahoma State	Jason Collins	F-C	Stanford
Alan Henderson	F-C	Indiana	Richard Jefferson	G-F	Arizona
Larry Hughes	G	St. Louis	Jason Kidd – C	G	California
Zydrunas Ilgauskas	C	Lithuania	Nenad Krstic	C	Serbia & Montenegro
LeBron James	F	St. Vincent-St. Mary HS (OH)	Lamond Murray	F	California
Damon Jones	G	Houston	Bostjan Nachbar	F	Slovenia
Donyell Marshall	F	Connecticut	Scott Padgett	F	Kentucky
Ronald Murray	G	Shaw	Zoran Planinic	G-F	Croatia
Ira Newble	G-F	Miami (Ohio)	Clifford Robinson	F-C	Connecticut
Aleksandar Pavlovic	G-F	Serbia & Montenegro	John Thomas	C	Minnesota
Eric Snow	G	Michigan State	Jacque Vaughn	G	Kansas
			Antoine Wright	G-F	Texas A&M

a. Construct a frequency table of the positions (G = guard, G-F = guard or forward, F = forward, F-C = forward or center, C = center) of the players. What is the mode? What proportion of players are guards?

b. Construct a frequency table of nationality categorized into America, South America, and Europe. What proportion of players are non-American?

c. For the players that are not American, are there differences in the nationality of professional baseball players and professional basketball players? Explain.

3. **Gross Sales of Julia Roberts Movies**

Here are the gross sales (in millions of dollars) for 24 movies starring Julia Roberts.

94, 16, 76, 34, 126, 10, 61, 3, 120, 31, 6, 67, 11,
64, 127, 116, 183, 101, 6, 178, 152, 101, 51, 91

a. Compute the median and mean gross sales.

b. Compare the median and mean and comment what these say about the shape of the distribution of gross sales.

c. Compute a measure of spread of these gross sales.

4. **Salaries of Basketball Players**

The table below gives the salaries (in millions of dollars) for the players on the 2003-2004 Los Angeles Lakers basketball team.

PLAYER	SALARY	PLAYER	SALARY
Shaquille O'Neal	26.5	Gary Payton	4.9
Kobe Bryant	13.5	Bryon Russell	1.1
Brian Cook	0.8	Kareem Rush	1.1
Rick Fox	4.5	Ime Udoka	0.4
Derek Fisher	3.0	Luke Walton	0.4
Horace Grant	1.1	Jamal Sampson	0.6
Devean George	4.5	Stanislav Medvedenko	1.5
Karl Malone	1.5		

a. Compute the median and mean salaries.

b. Comparing the median and mean, what do these say about the shape of the distribution of salaries?

c. If you were a basketball fan and wanted a “representative” salary of a Los Angeles Lakers player, would you be interested in the median or mean salary? Explain.

d. If you were the owner of the Lakers and concerned about the costs of running the team, would you be interested in the median or mean salary? Explain.

5. **How Long Does It Take to Score in Basketball?**

At college basketball games in the United States, it is common for the home fans for a team to remain standing until their team scores its first points. That raises the interesting question: how long does it take for a college basketball team to score its first points? The espn.com website gives game logs for games played in the 2004 NCAA men's basketball tournament. Looking at the game logs for 17 games, the author recorded the time for each team to score its first points in the game. Here are the times that were recorded (in seconds):

```
39    66    44    58   338   195    88    23    24    39    11
44    39    62     8    15   107   136   170    66    24   198
90   114    74   122    12    84    53    20    25    37    21
25
```

a. Construct a dotplot of these times.

b. Describe the basic shape of these times.

c. Compute the median and mean time.

d. If you were asked to report a typical time until the first point was scored, would you report the mean or the median? Why?

e. Suppose you are watching a game and it takes 300 seconds for a team to score its first points. Based on your work above, do you think this observation is unusual? Why?

6. **Weights of Newborns.**

The below stemplot graphs the weights in ounces for fifty babies where the number of gestation weeks exceeds 35.

```
7  | 5
8  | 3
9  | 466
10 | 4589
11 | 011123333455567789
12 | 1233336778
13 | 23444
14 | 0155588
```

```
   15 |
   16 | 0

7|5 means 75 ounces
```

a. Find the position of the median.

b. Find the median.

c. Suppose that the two largest baby weights of 148 and 160 ounces were recorded incorrectly – these weight values should have been 190 and 200 ounces, respectively. Without performing any computation, find the median of the new baby weights.

7. **Fares of Airplane Flights to Different Cities**

The round-trip air fares (in dollars) from Detroit to a number of different cities are shown in the below table.

CITY	FARE
San Francisco	310
Chicago	92
Miami	252
Denver	198
Las Vegas	242
Philadelphia	258
Boston	327
New York	170
Fargo	369
San Diego	312

a. Suppose you guess that an "average" air fare for these cities is $200. In the table below, compute the deviation of each air fare from 200, and compute the sum of the deviations. (The first two deviations have been computed for you.)

		DEVIATION
CITY	FARE	From 200
San Francisco	310	310-200 = 110
Chicago	92	92-200= -108

Miami	252
Denver	198
Las Vegas	242
Philadelphia	258
Boston	327
New York	170
Fargo	369
San Diego	312
SUM	

b. Next, suppose that you guess that the “average” air fare is $253. Compute the deviation of each air fare from 253, and compute the sum of the deviations.

		DEVIATION
CITY	FARE	FROM 253
San Francisco	310	310-253 = 57
Chicago	92	
Miami	252	
Denver	198	
Las Vegas	242	
Philadelphia	258	
Boston	327	
New York	170	
Fargo	369	
San Diego	312	
SUM		

c. Based on your work in parts a and b, which “average”, 200 or 253, is the mean of the air fares? Why?

8. **Test Scores**

Suppose a test is given to eight students and the scores are given by 70, 60, 79, 80, 75, 100, 56, 80. The mean is given by $\bar{x}$=75.

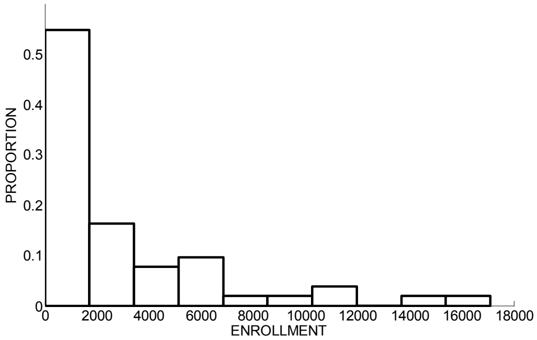
a. Find the deviation from the mean for each data value.

b. Verify that the sum of deviations from the mean is equal to 0.

c. Suppose four additional students take the test. Two of the students score 80 and 90. If the mean score for all 12 students remains at $\bar{x}$=75, find the scores of the remaining two students. (There is more than one possible answer.)

9. **College Enrollments**

The enrollments for 104 colleges are graphed in the following histogram.

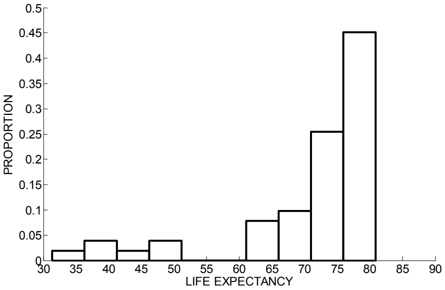


a. From the histogram, make a reasonable guess at the median enrollment. (Recall the median is the value M such that half of the enrollments are smaller than M and half of the enrollments are larger than M.)

b. Using your guess at the median M and the shape of the data, estimate the value of the mean $\bar{x}$.

10. **Life Expectancy of Selected Countries.**

The average life expectancy was recorded for a selection of countries in the Time *Almanac 2004*. A histogram of these life expectancies is displayed below.



Estimate the median and mean life expectancies for these countries.

11. **Ages of Women Participating in a Marathon**

A stemplot of the ages for 50 women participating in the 2003 Grandma's Marathon is shown below.

```
1 | 8
2 | 11111224
2 | 5566666677999
3 | 001344
```

```
    3 | 5777799
    4 | 02222234
    4 | 677
    5 | 024
    5 | 8
1|8 means 18 years old
```

a. Find a five-number summary of the ages.

b. Using the five-number summary, find an interval that contains the middle 50% of the ages.

c. The mean and standard deviation of the ages are given by $\bar{x} = 33.6$ and $s = 10.0$. Using these values, find an interval that contains approximately the middle 68% of the ages.

12. **Braking Distances of Cars**

The table below gives the braking distance (feet) of 18 cars listed in *Consumer Reports New Car Preview 2004*.

	Braking		Braking
Model	distance	Model	distance
Acura MDX	151	Nissan 350Z	116
Buick Park Avenue	137	Oldsmobile Silhouette	145
Chevrolet TrailBlazer	154	Subaru Baja	138
Dodge Neon	131	Toyota Corolla	140
Ford Taurus	151	Toyota Tundra	142
Honda Odyssey	147	Volvo S60	133
Infiniti G35	133	Cadillac DeVille	147
Lexus IS300	128	BMW 7-Series	135
Mercedes-Benz S-Class	135	Hyundai Elantra	139

a. Construct a stemplot of the braking distances.

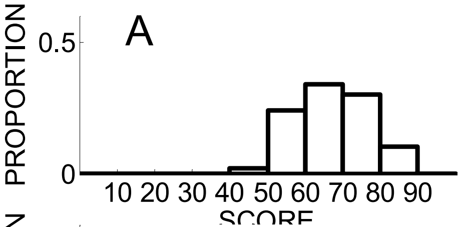
b. Find the five-number summary of the distances.

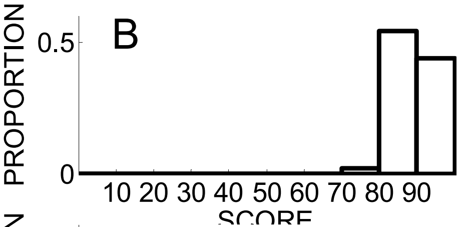
c. Find a car that has a "typical" braking distance.

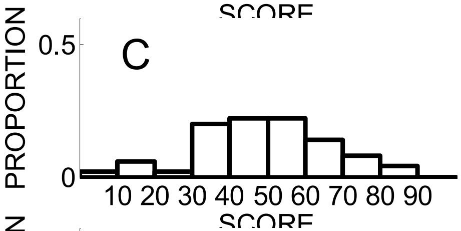
d. Suppose you are interested in a car that is tested to have a braking distance of 150 feet. Using your work in parts a and b, explain why you might not be interested in buying this car.

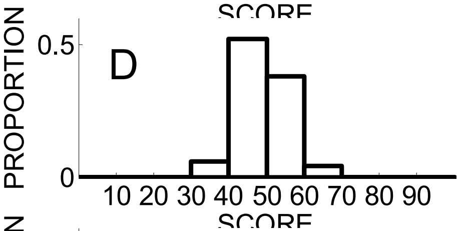
13. **Matching Graphs and Statistics**

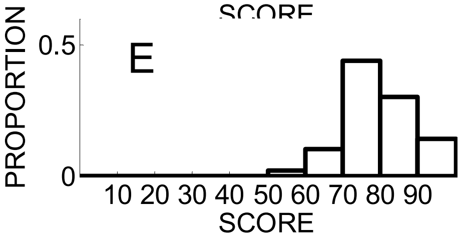
Histograms of the test scores for six classes and six sets of statistics are shown below. Write down the letter of the histogram next to the corresponding statistics.

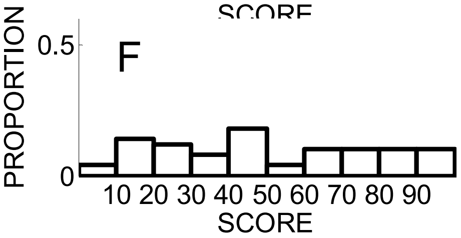












(mean, standard deviation) Histogram

stats 1 (67.7, 9.0) _____________

stats 2 (49.1, 6.3) _____________

stats 3 (51.0, 28.4) _____________

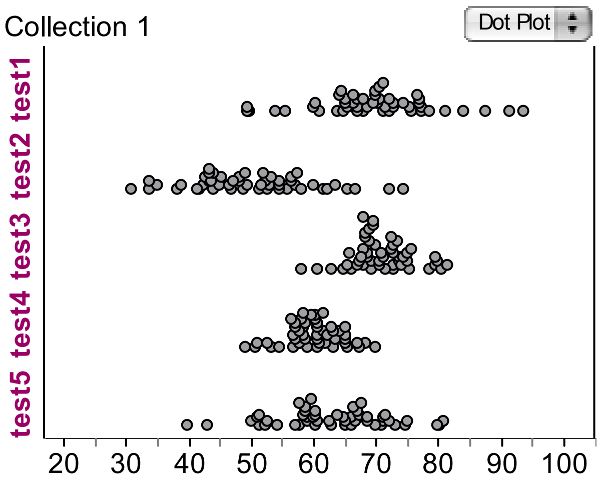
stats 4 (79.5, 9.5) _____________

stats 5 (49.5, 17.4) _____________

stats 6 (89.9, 4.9) _____________

14. **Matching Graphs and Statistics**

The below figure shows dotplots of five collections of test scores, labeled Test1, Test2, Test3, Test4, Test5 followed by five sets of statistics. Write down the name of the dataset next to the corresponding statistics.



(mean, standard deviation) Dataset

stats 1 (50.2, 9.6) ______________

stats 2 (63.1, 8.9) ______________

stats 3 (60.0, 4.5) ______________

stats 4 (69.7, 9.4) ______________

stats 5 (71.0, 4.9) ______________

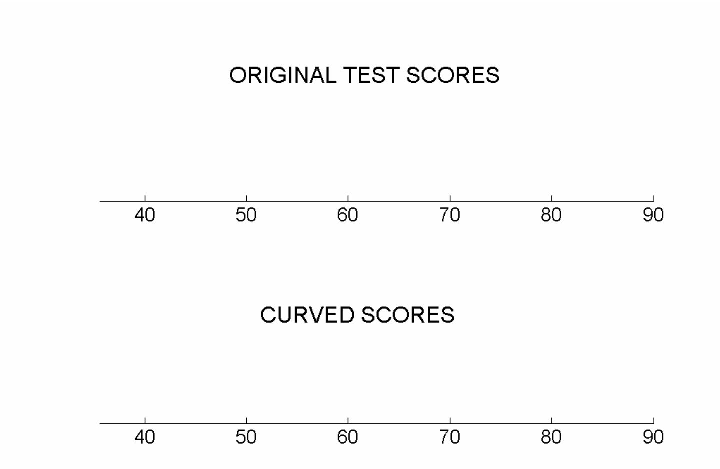
15. **Curving a Test**

Suppose that the grades on an English test for a class of ten students are given by 45, 65, 50, 44, 66, 70, 58, 40, 60, 52.

a. Construct a dotplot of the grades using the scale below.

b. Since the grades are low, the teacher decides to curve the scores by adding 15 points to each student's grade. Construct a dotplot of the "curved scores" on the display below.

c. The mean and standard deviation of the original test scores are $\bar{x} = 55$ and s =10.3. By comparing the dotplots of the original and new scores (and not by computation), find the mean and standard deviation of the new test scores.



d. Suppose instead the teacher decides to curve the grades by adding 25 points to everyone's score. The mean of the new grades would be _________ and the standard deviation of the new grades would be _________ .

16. **Rescaling a Test**

Suppose that a test has a total of 50 points and a class of ten students gets the following grades (out of 50):

30 18 36 38 24 47 47 35 38 37

a. Find the median and IQR of these new grades.

b. Suppose the teacher decides to rescale these grades by multiplying by two (so that the new grades will be out of 100 points):

60 36 72 76 48 94 94 70 76 74

Make intelligent guesses at the median and IQR of these grades and give some rationale for your guesses.

c. Find the median and IQR of these new grades.

d. Compare your answers to parts a and b; by multiplying the scores by 2, how has the median changed? How has the IQR changed?

17. **Computing a Standard Deviation**

The table below gives some of the starting calculations to compute the standard deviation of a collection of test scores. The mean score is $\bar{x} = 35$ and the "Deviation" column contains the deviations (score - $\bar{x}$). Complete the table and find the standard deviation s. Give an interpretation of s in the context of this problem.

Score	Deviation	Deviation squared
30	-5	
18	-17	
36	1	
38	3	
24	-11	
47	12	
47	12	
35	0	
38	3	
37	2	

18. **Computing a Standard Deviation**

Suppose you have five people in your family and you measure the V-span (the distance between the middle and index fingers) for all members of your family. The mean value is 6 cm. (This means that the sum of the measurements is 30 cm.) Also the smallest V-span is 3 cm and the largest V-span is 9 cm.

a. Find values of the V-spans so that the standard deviation of the measurements is as small as possible.

b. Find values of the V-spans so that the standard deviation of the measurements is as large as possible.

c. Compute the values of the standard deviations in parts b and c.

19. **Church Attendance**

The worship attendance at a church in Ohio was recorded for 209 consecutive weeks. The attendance numbers are graphed in the stemplot to the right. The mean and standard deviation of these numbers are given by $\bar{x} = 361.7$ and s = 58.5.

```
1 | 8
2 | 0
2 | 2333
2 |
2 | 6666677
2 | 8888889999999999
3 | 00000000001111111
3 | 2222222222222222223333333333
3 | 44444444444444555555555555555
3 | 666666666667777777777777
3 | 8888888888889999999999999999999
4 | 000000000000001111111111
4 | 2222222222333
4 | 4555
4 | 6777
4 | 899
5 |
5 | 233

1|8 corresponds to 180
```

a. Can you apply the 68-95-99.7 rule for this dataset? Why or why not?

b. Find an interval that you believe will contain approximately the middle 68% of the attendance numbers.

c. From the stemplot, find the proportion of values that fall in the interval you found in part b. Does this proportion agree with the 68% in the 68-95-99.7 rule?

d. Find an interval symmetric about the mean that you believe will contain approximately 95% of the numbers.

e. Find the actual number of attendance numbers that fall in the interval from part d. Comment if this proportion is close to what you would expect from part d.

20. **Snowfall in Tulsa**

The yearly snowfall in Tulsa, Oklahoma was recorded for the years 1950-2003. A stemplot of the snowfall amounts (in inches) is displayed to the right. The mean and standard deviation of these snowfall amounts are given by $\bar{x}$ = 9.5 and s = 6.0 inches respectively.

```
0 | 0011
0 | 233
0 | 4444445555
0 | 6666667777
0 | 88999
1 | 00011
1 | 223
1 | 44445555
1 | 66
```

```
   1 |
   2 | 00
   2 | 2
   2 |
   2 |
   2 | 9

0|1 corresponds to 1 inch
```

a. Is it appropriate to apply the 68-95-99.7 rule to these data? Why or why not?

b. Find the interval ($\bar{x}$ + s, $\bar{x}$ - s) and find the proportion of snowfall amounts that fall in this interval.

c. Find the interval ($\bar{x}$ + 2 s, $\bar{x}$ - 2 s) and find the proportion of snowfall amounts that fall in this interval.

d. Are the proportions you computed in parts b and c close to what you expected if you use the 68-95-99.7 rule?

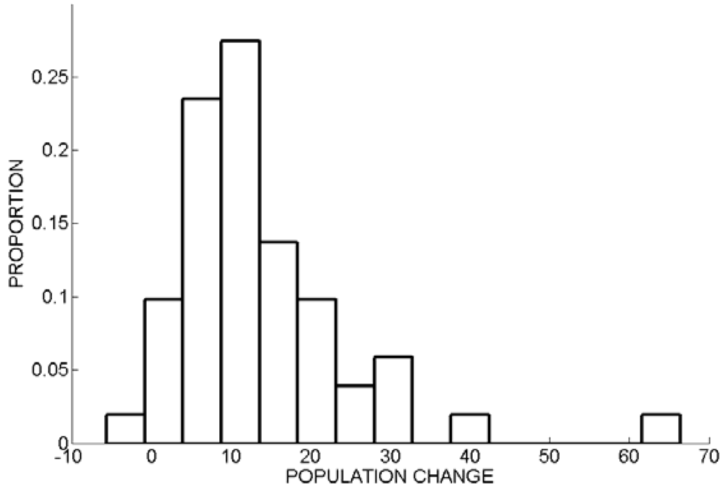
e. If the answer to part d is no, explain why the 68-95-99.7 rule may not be applicable in this situation.

21. **State Population Changes**

The table below gives the percentage change in population from 1990 to 2000 for all states in the United States.

State	%change	State	%change	State	%change	State	%change
AL	10	IL	8.7	MT	12.9	RI	4.5
AK	14	IN	9.7	NE	8.4	SC	15.1
AZ	40	IA	5.4	NV	66.4	SD	8.3
AR	13.7	KS	8.5	NH	11.4	TN	16.6
CA	13.8	KY	9.7	NJ	8.8	TX	22.8
CO	30.6	LA	5.9	NM	20.1	UT	29.7
CT	3.6	ME	3.8	NY	5.5	VT	8.2
DE	17.6	MD	10.8	NC	21.4	VA	14.4
DC	-5.6	MA	5.5	ND	0.6	WA	21.1
FL	23.5	MI	6.9	OH	4.7	WV	0.8
GA	26.4	MN	12.4	OK	9.7	WI	9.7
HI	9.3	MS	10.5	OR	20.4	WY	8.8
ID	28.5	MO	9.3	PA	3.4		

A histogram of these percentage changes in population is shown below.



a. Write a short paragraph about this dataset, discussing shape, a typical value, spread, and any unusual characteristics.

b. Looking back at the data table, find the location (region of the U.S.) of the states with the large percentage increases.

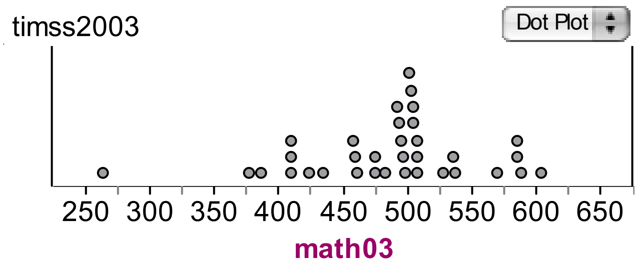
c. The mean and median of the population changes are given by $\bar{x}$ = 13.45 and M = 9.7, respectively. Explain why these two measures of center are different for this dataset.

d. Suppose that the largest population change 66.4, corresponding to Nevada, is removed from the dataset. Recalculate the mean and median for this reduced dataset.

e. We say that a measure is *resistant* if its value is not influenced by one extremely large or small observation. Are the median or mean resistant measures? Explain why or why not.

22. **International Study of Mathematics Achievement**

The Trends in International Mathematics and Science Study (TIMSS) collects data on mathematics and science achievement from countries all around the world. The display below shows a dotplot of the mean mathematics score of 8th graders from 34 countries from the 2003 study. The median and mean scores are given by 497 and 485.02, respectively.



a. What proportion of countries have scores smaller than 497?

b. Suppose the low score of 264 is removed from the dataset. Recompute the median and the mean? Are they more or less similar in value? Is this to be expected? Why?

c. Suppose South America's low score of 264 is replaced by Oz's (fictitious country) score of 150. Compute the median and mean of the new dataset.

d. Which measure (median or mean) is most affected by the unusually small score?

23. **City Temperatures**

The below table displays the average temperature (in degrees Fahrenheit) for eight American cities for each month of the year.

	San	Vero			San			
Month	Francisco	Beach	Duluth	Albuquerque	Diego	Philadelphia	Honolulu	Indianapolis
Jan	48.7	61.6	7.0	34.2	57.4	30.4	72.9	25.5
Feb	52.2	62.7	12.3	40.0	58.6	33.0	73.0	29.6
Mar	53.3	67.2	24.4	46.9	59.6	42.4	74.4	41.4
Apr	55.6	71.3	38.6	55.2	62.0	52.4	75.8	52.4
May	58.1	75.8	50.8	64.2	64.1	62.9	77.5	62.8
Jun	61.5	79.5	59.8	74.2	66.8	71.8	79.4	71.9
July	62.7	81.1	66.1	78.5	71.0	76.9	80.5	75.4
Aug	63.7	81.3	63.7	75.9	72.6	75.5	81.4	73.2
Sep	64.5	80.1	54.2	68.6	71.4	68.2	81.0	66.6
Oct	61.0	75.5	43.7	57.0	67.7	56.4	79.6	54.7
Nov	54.8	69.3	28.4	44.3	62.0	46.4	77.2	43.0
Dec	49.4	63.7	12.8	35.3	57.4	35.8	74.1	30.9

a. For each city, find the five-number summary of temperatures for the 12 months. Place your calculations in the table below. Also, for each city, find the interquartile range (IQR) of temperatures.

	Five-number summary					
CITY	LO	Q_L	M	Q_U	HI	IQR
San Francisco						
Vero Beach						
Duluth						
Albuquerque						
San Diego						
Philadelphia						
Honolulu						
Indianapolis						

b. By use of the medians, order the cities from coldest to warmest.

c. Suppose you classify a city's temperature as volatile (changing or varying across months) or stable (little change across months). What quantity would you use to measure the volatility of a city's temperature? Using this measure, order the cities from most volatile to most stable.

24. **Points Scored for Basketball Games**

The number of points scored in the first nine basketball playoff games during the 2004-5 season is recorded below for four Phoenix Suns players (Amare Stoudemire, Shawn Marion, Steve Nash, and Quentin Richardson).

Game	Stoudemire	Marion	Nash	Richardson
1	9	26	11	22
2	34	22	12	15
3	30	14	13	9
4	18	23	24	14
5	40	23	11	12
6	30	23	23	12
7	37	21	27	12
8	15	19	48	13
9	Did not play	16	34	7

a. On the average, which player scored the most points? Explain what measure you are using to measure the average.

b. Which player was the most consistent scorer for these basketball games? Explain what measure you are using to measure consistency of scoring.

TOPIC D4: COMPARING BATCHES AND RELATIVE STANDING

SPOTLIGHT: WHERE'S THE BEST PLACE TO LIVE?

Where is an ideal place to live in America? Perhaps you wish to live in a dream house in suburbs of a large city. But that might mean that you have a high cost of living and commute one hour to work. Maybe instead you would prefer to live in a small town where you are close to work. But this small town might have limited opportunities for nightlife, restaurants, theater, or activities for children. Many people prefer to live in warm climates such as Florida, but other people like colder climates such as Colorado where there are opportunities to engage in winter sports such as ice skating and skiing.

The book *Cities Ranked & Rated* recognizes that people have different goals, needs, aspirations and interests, and these qualities impact the choice of desirable locations to live. This book considers four broad categories that people may use when evaluating a possible place to relocate: economy, cost of living, climate, and character. Economy refers to the economic health and commercial aspects of a place. Cost of living refers to the costs of housing and necessities and tax burden. The character of a place refers to the area's "look and feel," its activities and services, and any negative aspects such as crime and health problems.

This book gives a rich set of facts and figures for 403 North American cities. Using these measurements, the book gives a numerical rating and ranking of the cities with respect to economy and jobs, cost of living, climate, education, health and healthcare, crime, transportation, leisure, arts and culture, and quality of life. It may not surprise you that New York City has the top rating with respect to leisure and Boston has the top-rated education. But did you know that the metro area with the strongest economy and jobs is Billings, Montana, and the least-expensive metro area is Casper, Wyoming? The top-rated area with respect to quality of life (including physical attractiveness, heritage, friendliness of residents and overall ease of living) is Madison, Wisconsin. In this book, we will use some of the data collected in this book to compare groups of cities with respect to different characteristics.

PREVIEW

Although methods of graphing and summarizing a single batch of data are useful, much of data analysis is involved with comparison of batches. We know how to compare two individuals by simply taking a difference of their individual values, such as "Susie scored 10 points higher than Joe on the math test." We would like to make similar comparative statements when we compare two or more batches of data, and this topic will describe how that can be done. Also, we'll discuss a method for describing one's relative standing in a collection of quantitative data and introduce a rule of thumb for detecting outliers.

In this topic your learning objectives are to:

- Understand how one can compare two batches of categorical data by the use of graphs and the computation of percentages.
- Understand what it means for one batch of quantitative data to be larger than another batch of data.
- Understand how one can summarize and display a batch of quantitative data by the use of five numbers.
- Understand when it is appropriate and inappropriate to compare two batches of quantitative data.
- Understand how a standardized score measures the relative standing of an observation, and understand how one can identify unusually small or large data values.

NCTM Standards

✓ In Grades 6-8, all students should collect data about a characteristic shared by two populations.

✓In Grades 6-8, all students should understand parallel box plots, and use them to display data.

COMPARING BATCHES OF CATEGORICAL DATA

In topics D2 and D3, we looked at the country of origin of Major League baseball players who were born in the year 1975. We found that approximately two-thirds of the players were born in the United States, but a sizeable proportion of players came from Latin America.

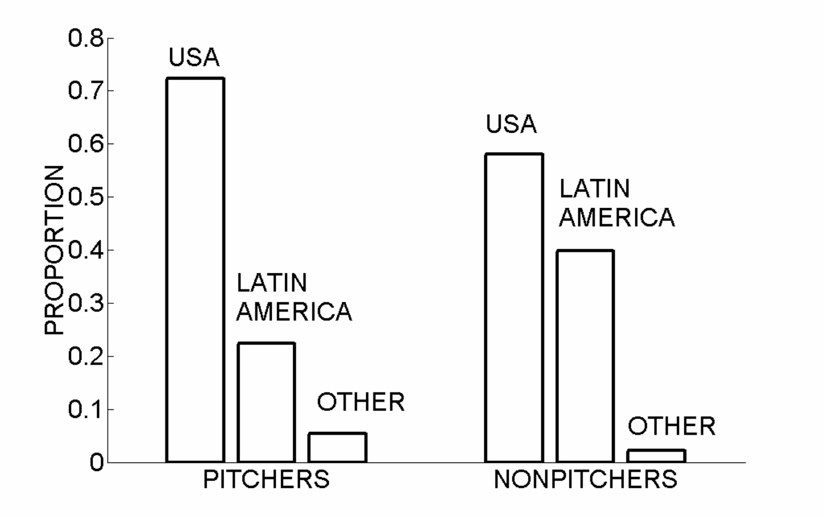
Looking at these data more carefully, it is natural to ask: do players from other countries excel in particular positions in baseball? Specifically, are there differences in the country of birth distribution between pitchers and nonpitchers?

To answer the question, we revisit our data and divide the 205 players into two groups – the 112 players who are pitchers and the 93 players who are nonpitchers. For each group, we categorize the players by the country of origin (USA, Latin American, or Other). We present these two frequency tables below. In addition, for each group, we find the proportion and percentage for each category.

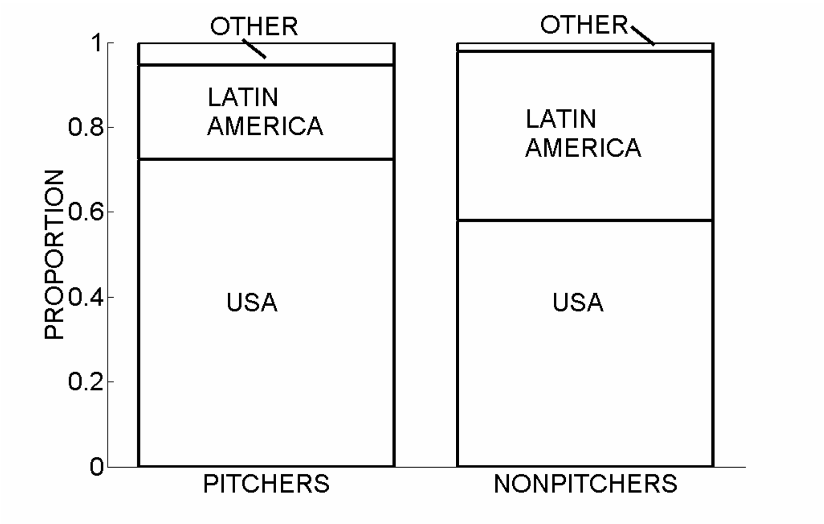
Pitchers				Nonpitchers		
Frequency	Proportion	Percentage	COUNTRY	Frequency	Proportion	Percentage
81	0.723	72	USA	54	0.581	58
25	0.223	22	Latin America	37	0.398	40
6	0.054	5	Other	2	0.022	2
112	1	100	TOTAL	93	1	100

To compare the country of origins of the pitchers and nonpitchers, we'd like first to construct a graph. In topic D2, we used a bar chart and a segmented bar chart to display a single frequency table of categorical data. These same graphs can be used to compare several frequency tables.

We first show side-by-side bar charts of the pitcher and nonpitcher tables. Since the numbers of pitchers and nonpitchers are not equal, it is best to plot the proportions (or percentages) of the two tables – in that way, the total proportion is equal to one for both groups and it is easier to make comparisons. We see differences – there are over twice as many USA pitchers than Latin American pitchers, while the proportions of USA and Latin American nonpitchers are similar in size. It appears that pitchers are more likely to be American than nonpitchers.



A similar graphical comparison can be made by use of segmented bar charts. In the figure below, we show side-by-side segmented bar charts of the category proportions. As in the earlier display, it is best to graph the proportions (instead of the frequencies) since the sum of proportions is equal to one for both tables. From this display, we clearly see the differences in the country of birth between the pitchers and nonpitchers.



Now that we have noticed that there is a difference in the country of birth in the two groups of players, how can we summarize this difference? If we focus on the percentages in the table, we read that 22% of the pitchers are Latin American in origin, and 40% of the nonpitchers are Latin American. If we look at the difference, we can say that the proportion of Latin Americans among the nonpitchers exceeds the proportion of Latin Americans among the pitchers by 40 – 22 = 18%.

PRACTICE: COMPARING BATCHES OF CATEGORICAL DATA

In the book *Cities Rated & Ranked*, 331 metropolitan areas are ranked from best to worst on the basis of quality of life. The top 30 and the bottom 30 metropolitan areas are listed in the table below. In addition, the table gives the population density (LOW or HIGH compared to the U.S. average) and the population growth in the 1990-2002 (LOW or HIGH compared to the U.S. average).

TOP 30 U.S. Metropolitan Areas

			GROWT
	STATE	DENSITY	H
Charlottesville	VA	LOW	HIGH
Sante Fe	NM	LOW	HIGH
San Luis Obispo-Alascadero-Paso Robles	CA	LOW	HIGH
Santa Barbara – Santa Maria – Lompoc	CA	LOW	LOW
Honolulu	HI	HIGH	LOW

BOTTOM 30 U.S. Metropolitan Areas

			GROWT
	STATE	DENSITY	H
Macon	GA	LOW	LOW
Owensboro	KY	LOW	LOW
Jackson	TN	LOW	HIGH
Wheeling	WV-OH	LOW	LOW
Joplin	MO	LOW	HIGH

Ann Arbor	MI	LOW	HIGH	Racine	WI	HIGH	LOW
Atlanta	GA	HIGH	HIGH	Sharon	PA	LOW	LOW
Asheville	NC	LOW	HIGH	Erie	PA	LOW	LOW
Reno	NV	LOW	HIGH	Dutchess Country	NY	LOW	LOW
Corvallis	OR	LOW	LOW	Dubuque	IA	LOW	LOW
Roanoke	VA	LOW	LOW	Waterbury	CT	HIGH	LOW
Portland – Vancouver	OR-WA	LOW	HIGH	Lewiston-Auburn	ME	LOW	LOW
Raleigh-Durham-Chapel Hill	NC	LOW	HIGH	Brownsville-Harlingen-San Benito	TX	LOW	HIGH
Bryan-College Station	TX	LOW	HIGH	Yuba City-Marysville	CA	LOW	HIGH
Lynchburg	VA	LOW	LOW	Modesto	CA	LOW	HIGH
Olympia	WA	LOW	HIGH	McAllen-Edinburgh-Mission	TX	LOW	HIGH
Norfolk-Virginia Beach-Newport News	VA-NC	HIGH	LOW	Jacksonville	NC	LOW	LOW
Colorado Springs	CO	LOW	HIGH	New Bedford	MA	HIGH	LOW
Nassau-Suffolk	NY	HIGH	LOW	Houma	LA	LOW	LOW
Pueblo	CO	LOW	HIGH	Alexandria	LA	LOW	LOW
Eugene-Springfield	OR	LOW	LOW	Fort Smith	AR-OK	LOW	HIGH
Austin-San Marcos	TX	LOW	HIGH	Anniston	AL	LOW	LOW
Lafayette	IN	LOW	LOW	Gadsden	AL	LOW	LOW
Minneapolis-St. Paul	MN-WI	HIGH	HIGH	Pine Bluff	AR-OK	LOW	LOW
Dover	DE	LOW	HIGH	Lawrence	MA-NH	HIGH	LOW
Washington	DC-MD-VA-WV	HIGH	HIGH	Kankakee	IL	LOW	LOW
Fayetteville-Springdale-Rogers	AR	LOW	HIGH	Merced	CA	LOW	HIGH
Pittsburgh	PA	HIGH	LOW	Newburgh	NY-PA	LOW	HIGH
Bloomington	IN	LOW	LOW	Stockton-Lodi	CA	LOW	HIGH
Stamford-Norwalk	CT	HIGH	LOW	Laredo	TX	LOW	HIGH

1. For each group of metropolitan areas, find the number of areas that have a low and high population density and place your frequencies in the table below.

	TOP AREAS			BOTTOM AREAS	
Population density	Frequency	Proportion		Frequency	Proportion
LOW					
HIGH					

2. Construct side-by-side bar charts of the proportions of low and high density areas in the “top” and “bottom” groups.

3. By computing proportions, investigate if the population density differs between the “top” areas and the “bottom” areas.

4. For each group of areas, find the proportion that have a low and high population growth rate. Construct two segmented bar charts to compare the proportions of low and high growth in the two groups.

5. Are the population growth rates different for the “top” and “bottom” metropolitan areas? (Compare proportions to answer the question.)

COMPARING BATCHES OF QUANTITATIVE DATA

Next we discuss how we compare two or more batches of quantitative data. First, let's talk about how we compare things. Suppose someone is interested in comparing your height with your mom's height. Now this person can say "you are taller than your mom,” but usually he or she is interested in a more informative comparison like

"you are three inches taller than your mom."

This is a typical kind of comparison -- we say that one measurement is so much greater (or smaller) than another measurement.

Other types of comparisons are in terms of ratios. For example, suppose you wish to compare your income this year with your income last year. You may say that this year's income is, say, $2000 more than last year's income. But it is more common to compare incomes in terms of ratios. For example, pay raises are usually expressed in terms of percent, so you may say that this year's income is 4% higher than last year's income.

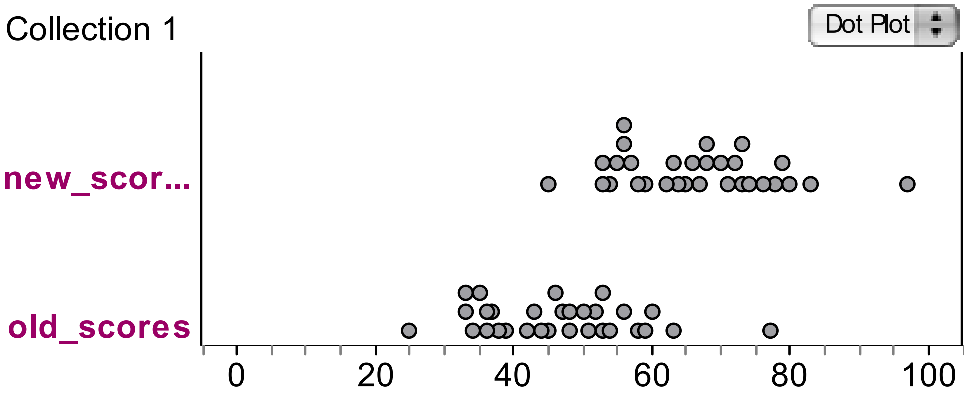
When we compare two batches of data, the easiest type of comparison is to say, for example, that

"one batch tends to be 10 more than another batch."

What does it mean to say that one batch of quantitative data is larger than a second batch?

Suppose you are given a relatively difficult exam in a particular class. The class does poorly – the median grade (out of 100 possible points) is only 46.5 and the quartiles are 37 and 53. The instructor decides to curve the exam grades by adding 20 points to each student's grade. What is the effect of this adjustment on the batch of test scores?

The figure below shows dotplots of the old test scores and the new test scores on the same scale.



How do the two batches of scores differ? Note that both sets of test scores have the same shape and same spread, but different locations. Note that one can get the distribution of new scores by moving the old distribution of scores 20 points to the right.

This example illustrates the situation when we are able to compare two batches. If two batches have approximately the same shape and same spread, then saying

"batch 1 is 10 points larger than batch 2"

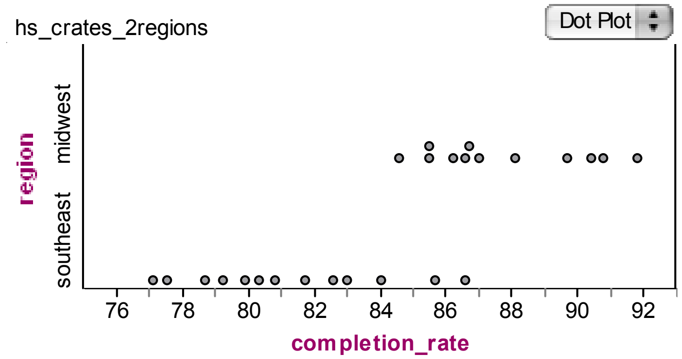
means that we can obtain batch 1 by adding 10 points to each value in batch 2. But this statement assumes that the batches have similar spreads. In practice, we have to check this assumption before we can make any comparison.

In Section D1, we looked at the high school completion rates for all states. We saw much variability in these completion rates. Some of the smallest completion rates corresponded to the Southeast states and the largest rates for states in the Midwest. That raises the question: Do Southeast states generally have lower high school completion rates than Midwest states?

To start to answer this question, we collect the high school completion rates for the two groups of states.

Midwest States		Southeast States	
Illinois	85.5	Alabama	77.5
Indiana	84.6	Arkansas	81.7
Iowa	89.7	Florida	84
Kansas	88.1	Georgia	82.6
Michigan	86.2	Kentucky	78.7
Minnesota	90.8	Louisiana	80.8
Missouri	86.6	Maryland	85.7
Nebraska	90.4	Mississippi	80.3
North Dakota	85.5	North Carolina	79.2
Ohio	87	South Carolina	83
South Dakota	91.8	Tennessee	79.9
Wisconsin	86.7	Virginia	86.6
		West Virginia	77.1

A good way of graphically comparing the two groups of completion rates is by *parallel dotplots*. On the below figure, we construct a dotplot of the Midwest states values on the Midwest line and a dotplot of the Southeast states values on the Southeast line.



A different graph of the rates for the two groups is ***side-by-side stemplots***, where one uses a common list of stems (as we have done below) and place the leaves for the Midwest states on the right and the leaves for the Southeast states on the left.

```
SOUTHEAST STATES                               MIDWEST STATES
```

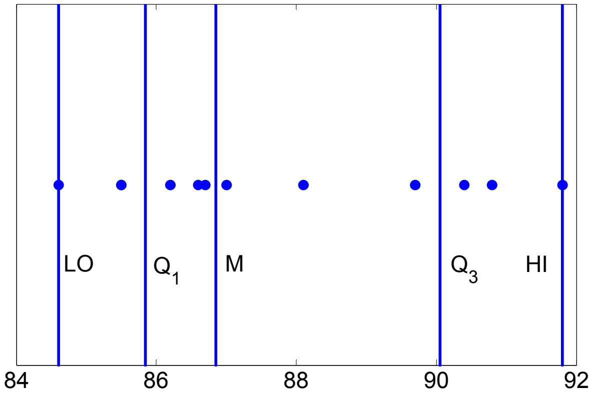
```
51 | 77 |
 7 | 78 |
92 | 79 |
83 | 80 |
 7 | 81 |
 6 | 82 |
 0 | 83 |
 0 | 84 | 6
 7 | 85 | 55
 6 | 86 | 267
   | 87 | 0
   | 88 | 1
   | 89 | 7
   | 90 | 48
   | 91 | 8

77|1 means 77.1      84|6 means 84.6
```

To compare two batches, it is useful to first summarize each batch with a five-number summary and then compare the summaries of the two batches. You can confirm that the five-number summary of the high school completion rates for the Midwest states, and the five-number summary of the rates for the Southeast states are given by

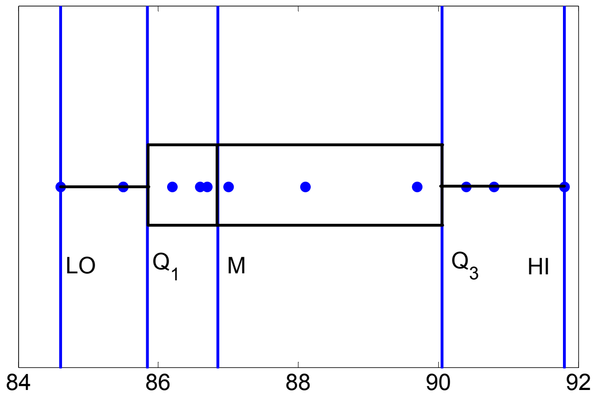
MIDWEST: (LO, Q_L, M, Q_U, HI) = (84.6, 85.85, 86.85, 90.05, 91.8)

SOUTHEAST: (LO, Q_L, M, Q_U, HI) = (77.1, 79.2, 80.8, 83, 86.6).

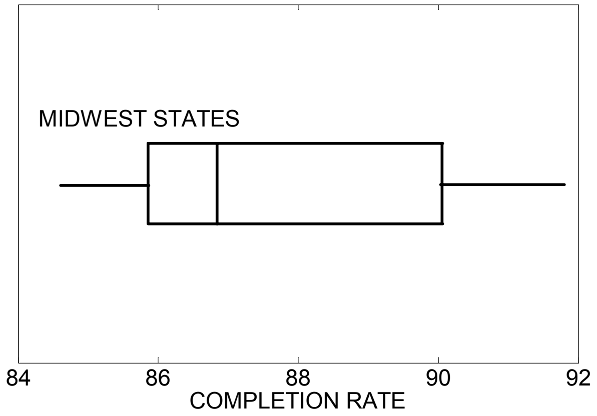
A ***boxplot*** is a graph of a five-number summary. To draw this for the Midwest states rates, we first locate the five numbers (LO, Q_L, M, Q_U, HI), on a number line below.



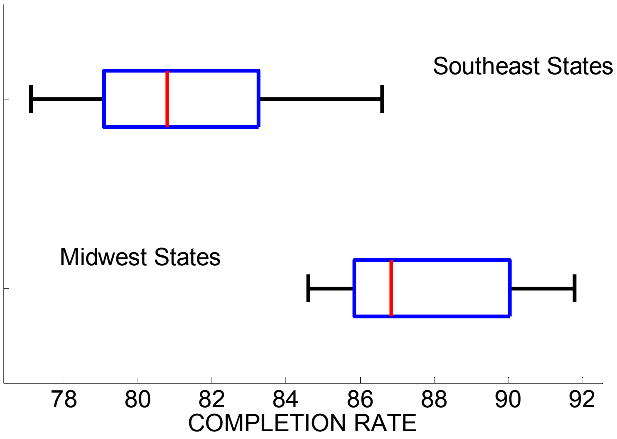
Next we draw a box extending from the lower to upper quartiles with the location of the median represented by a line inside the box. We complete the boxplot by drawing lines (sometimes called whiskers) from the outsides of the box to the locations of the LO and HI observations, respectively.



Removing the labels and the guidelines, we get the final boxplot display:



In similar fashion, we can construct a boxplot of the Southeast states completion rates. If we place both boxplots on the same graph on the same scale, we get the following display:



We can compare the two groups if the spread of the southeast completion rates is about the same as the spread of the Midwest rates. To check this, we place the quartiles and the interquartile spread of each group in the below table. We note that the two IQR's are 4.2 and 3.8, so the two batches have approximately the same spread.

Group	Q_L	Q_U	IQR
Midwest States	85.85	90.05	4.2
Southeast States	79.2	83	3.8

When two batches have equal spreads (as they do here), one can compare the two groups by finding the difference in medians. We note that the median completion rate of the Midwest states is 86.85 compared to a median rate of 80.8 for the Southeast states. This means that the high school completion rates for the Midwest states tend to be 86.85 – 80.8 = 6.05 higher than the completion rates for the Southeast states.

To emphasize what "one batch tends to be 6 units larger than a second batch" means, look at the boxplot that represents the high school completion rates for the Southeast states. Suppose we add 6 points to each of the rates for the Southeast states.

When we do this each of LO, Q_L, M, Q_U, HI will increase by 6 points and the five-number summary will change from

(77.1, 79.2, 80.8, 83, 86.6) to (83.1, 85.2, 86.8, 89, 92.6).

This new five-number summary and the corresponding boxplot represent the completion rates for the Midwest states.

SPECIAL NOTE: How can we say that two batches have approximate equal spreads? This is a difficult question to answer in general, but here are some guidelines to use in practice. (These guidelines assume that each batch is approximately mound-shaped.) We use the IQR to measure the spread of a batch and we compare the spreads of the two batches by means of the ratio IQR(batch2)/IQR(batch1). If each batch has about 50 values, then if the ratio of IQRs is between .75 and 1.35, then we can say the batches have approximately equal spreads. If we have smaller batches, each of size 25, then we can conclude "equal spreads" if the ratio of IQRs is between .6 and 1.6.

PRACTICE: COMPARING BATCHES OF QUANTITATIVE DATA

The book *Cities Rated & Ranked* produces a ranking of the top 30 and bottom 30 metropolitan areas with respect to quality of life. One key variable that may distinguish the top and bottom areas is the unemployment rate. These rates (expressed as a percentage) are shown in the below table.

TOP METROPOLITAN AREAS					BOTTOM METROPOLITAN AREAS				
4.0	3.6	4.8	5.6	2.7	6.9	5.4	10.6	3.4	3.3
7.4	2.6	4.0	4.3	4.7	5.2	5.7	7.4	6.8	8.0
8.1	6.6	4.6	3.9	3.3	7.4	4.6	5.6	8.2	8.4
2.6	3.5	5.8	4.0	4.7	6.3	4.5	5.9	4.1	5.0
4.1	4.3	4.3	5.6	4.9	4.0	4.1	13.6	11.5	9.3
3.2	3.6	3.4	3.2	3.3	4.7	9.2	6.3	4.7	11.0

1. Construct back-to-back stemplots of the unemployment rates for the top and bottom areas using the stems shown below.

```
 Bottom areas    | 2 |         Top areas
                 | 3 |
                 | 4 |
                 | 5 |
                 | 6 |
                 | 7 |
                 | 8 |
                 | 9 |
                 |10 |
                 |11 |
                 |12 |
                 |13 |

    3|2 means an unemployment rate of 3.2%
```

2. Find five-number summaries of the rates for each group of areas.

3. Is it reasonable to say that the two groups of unemployment rates have similar spreads? Explain.

4. Is it appropriate to compare the two groups of rates by computing the difference in the medians? Explain.

RELATIVE STANDING

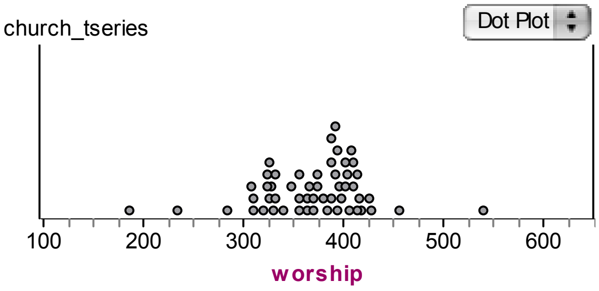
Suppose you are interested in learning about the Sunday worship attendance at a local church. You are told that the attendance for one week (a Sunday in July) was 327. You might wonder if this number represents a typical attendance for this church. You might suspect that this number is lower than average since you know that church attendance is usually smaller in the summer months due to people going on vacation. But how much smaller? Is there a way of talking about this attendance number (327) in the context of the distribution of attendance numbers for this church over many weeks?

To answer these questions, we collect the attendance numbers for this church for all of the weeks this year. This table displays the attendance numbers in chronological order. So, for example, the attendance numbers in the first five weeks in January were 234, 394, 417, 186 and 406.

MONTH	ATT	MONTH	ATT	MONTH	ATT	MONTH	ATT

January	234	April	339	July	309	October	428
January	394	April	369	July	284	October	365
January	417	April	539	July	307	October	414
January	186	April	414	July	327	October	394
January	406	April	388	July	348	October	402
February	397	May	395	August	373	November	410
February	310	May	374	August	332	November	404
February	425	May	384	August	355	November	387
February	319	May	325	August	324	November	392
March	364	June	329	September	324	December	401
March	370	June	355	September	388	December	455
March	364	June	392	September	415	December	408
March	356	June	331	September	409	December	379
						December	326

We graph the attendance numbers using a dotplot. We see a lot of variability in the numbers -- the attendances range from about 186 to 539 and a typical number in the middle of the batch is approximately 370. If the pastor of this church wonders about the summer worship attendance, the July attendance number of 327 looks a little lower than average.



One way of describing the location of 327 relative to the distribution of values is based on a *standardized score*. Using a computer, we find the mean $\bar{x}$ and standard deviation s of the weekly attendance numbers to be $\bar{x} = 368.6$, $s = 54.5$. The standardized score or z-score of a data value *x* is found by subtracting the mean and dividing by the standard deviation:

$$z = \frac{x - \overline{x}}{s}.$$

Here the standardized score of July's attendance 327 is

$$z = \frac{327 - 368.8}{54.5} = -0.77 .$$

The standardized score tells you the number of standard deviations the data value is from the mean. A *positive* z-score indicates the value is *above* the mean, and a *negative* z-score means the value is *below* the mean. Here z = -0.77 which means that the particular July attendance of 327 is approximately .8 or 4/5 of a standard deviation below the mean Sunday attendance. For a second example, note that the attendance for the first week of October was 428 has a z-score of

$$z = \frac{428 - 368.8}{54.5} = 1.09 .$$

Since this standardized score is a positive value close to one, we can say that this attendance of 428 is approximately one standard deviation above the mean.

FLAGGING POSSIBLE OUTLIERS

When we graphed the weekly attendance numbers, we saw one large cluster of values in the 300-450 range and we also saw a few unusually small and large values. Do these extreme attendance numbers deserve extra attention? Is there a method of identifying possible outliers in a dataset?

There is a useful "rule of thumb" for identifying data values that are unusually small or large. This method is based on the quartiles and the IQR that measures the spread of the middle half of the measurements. We first compute the five-number summary of the attendance numbers:

$$(\text{LO}, Q_L, \text{M}, Q_U, \text{HI}) = (186, 330, 374, 403, 539).$$

We say that an extreme observation is worthy of special attention if it falls further than one step from the lower and upper quartiles, where a step is defined to be

$$\text{STEP} = 1.5 \text{ IQR}.$$

We call such extreme observations *outliers*.

Let's illustrate this rule of thumb for our attendance data:

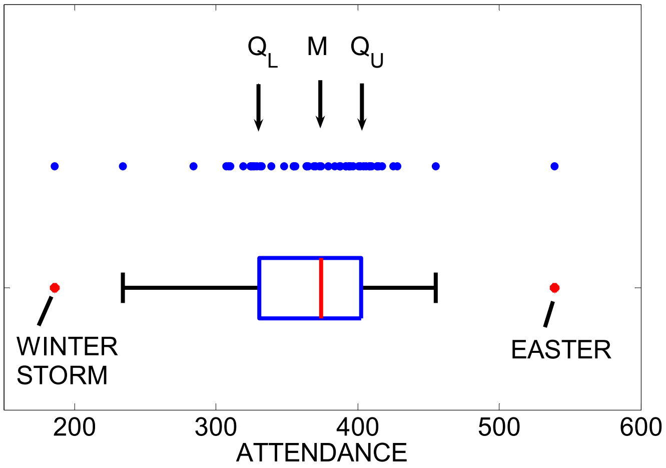
1. We compute the interquartile range IQR = Q_U - Q_L = 403 – 330 = 73.
2. A step is then equal to STEP = 1.5 IQR = 1.5 (73) = 109.5.
3. We say that an extreme observation deserves extra attention if either it

   -- falls below Q_L - STEP = 330 – 109.5 = 220.5

or

   -- it falls above Q_U + STEP = 403 + 109.5 = 512.5 .

Does this rule identify any unusually small or large attendance numbers? Looking back at our data, we see that the April attendance of 539 and the January attendance of 186 are outliers using our definition. It is common to draw a boxplot showing these outliers. The process of constructing this *modified boxplot* is displayed below. On the top of the figure, we show the observations with the locations of the median and quartiles indicated. We plot the "box" part of the boxplot as before. We then indicate the outliers by separate plotting points and draw lines (whiskers) from the box to the most extreme points at each end that are not outliers.



This rule of thumb identifies extreme data values that deserve extra attention. What are the possible explanations for unusually high or low attendance numbers? For the Christian faith, there are two major religious holidays, Christmas and Easter that will draw a large number of people for worship. Also, inclement weather may make it harder for people to attend church on particular days, causing small worship attendances. Indeed, in this example, it is not difficult to infer that the single large attendance number corresponds to an Easter Sunday in April, and the small January attendance number was likely due to a winter storm that made it difficult to travel to church.

PRACTICE: RELATIVE STANDING AND FLAGGING OUTLIERS

Again consider the unemployment rates for the top 30 metropolitan areas as reported by *Cities Rated & Ranked* .

```
          TOP METROPOLITAN AREAS
----------------------------------
    4.0     3.6     4.8     5.6     2.7
    7.4     2.6     4.0     4.3     4.7
    8.1     6.6     4.6     3.9     3.3
    2.6     3.5     5.8     4.0     4.7
    4.1     4.3     4.3     5.6     4.9
    3.2     3.6     3.4     3.2     3.3
```

1. The mean and standard deviation of the unemployment rates are given by $\bar{x} = 4.36$ % and s = 1.33 %. The unemployment rate for Bloomington, Indiana is 2.7 %. Find Bloomington's standardized score.

2. Interpret Bloomington's standardized score in terms of the number of standard deviations above or below the mean.

3. Find and interpret the standardized score for Portland, Oregon that has an unemployment rate of 8.1 %.

4. Construct a modified boxplot of these rates. Using the rule of thumb, identify any outliers among the unemployment rates among the top areas.

ACTIVITY: COMPARING MEN AND WOMEN IN THE CLASS DATASET

In Topic D1, we collected a number of different variables from all the students in the class. Choose two variables that you believe will be different between men and women in the class. (One obvious variable that would distinguish genders is height.)

For each variable

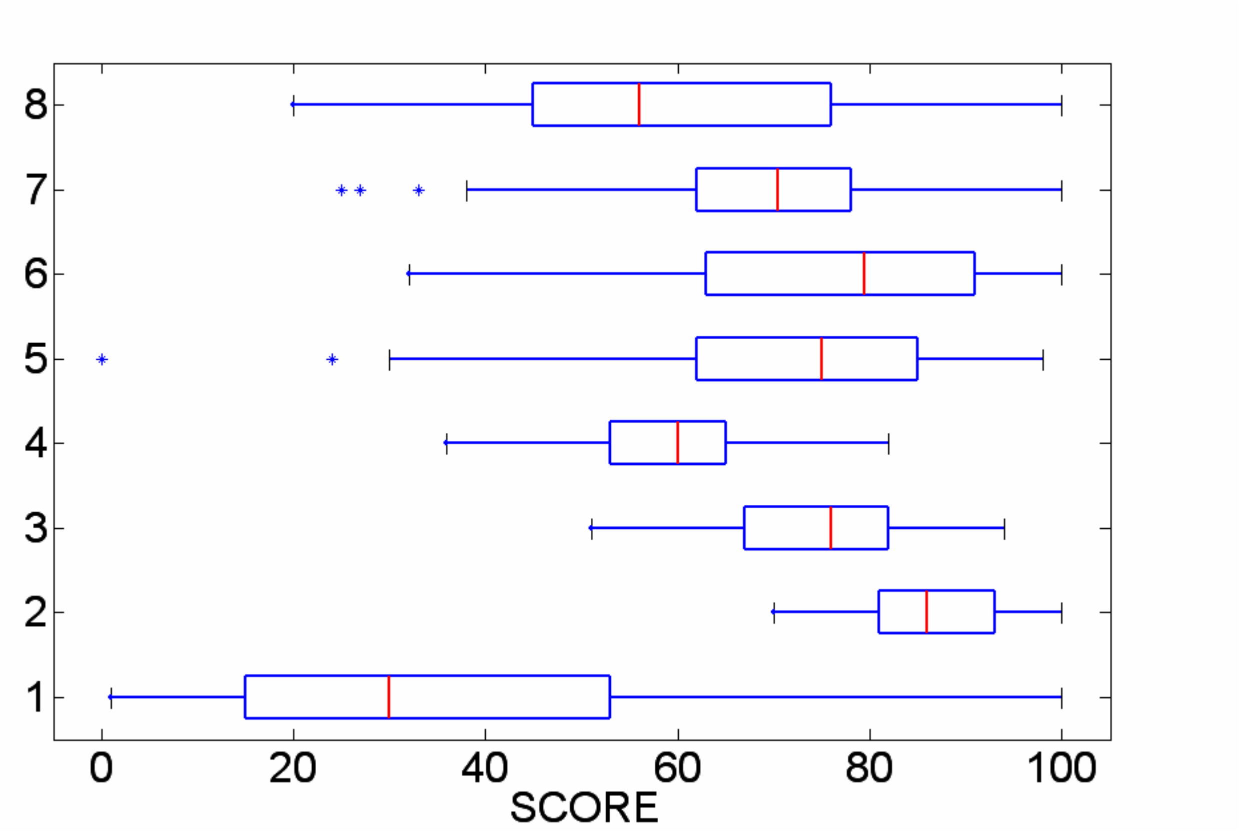
1. Construct parallel dotplots or parallel stemplots comparing men and women.
2. Find five-number summaries of each group.
3. Construct parallel boxplots.
4. Find the interquartile spread for each group. Is it reasonable to say that the men and women data have approximately the same spreads?

5. If the answer is "yes" to the previous question, make a comparison between the men and women. (It is not sufficient to say that one group tends to be larger than the second group. You should indicate how much larger.)

ACTIVITY: MATCHING STATISTICS WITH BOXPLOTS

DESCRIPTION: In this activity, we will match up boxplots and their corresponding summary statistics. The relative locations of the mean and median are informative about the shape of the data that affects the relative lengths of the "whiskers" and the two components of the "box" part of the boxplot. In addition, the value of the standard deviation is useful in understanding the length of the boxplot.

Below are the boxplots for eight datasets and following are summary statistics (the mean, median, and standard deviation) for eight datasets.



In the “Boxplot” row of the below table, write down the letter of the matching boxplot.

Dataset	1	2	3	4	5	6	7	8
mean	59.44	60.58	75.00	69.02	86.68	77.12	36.48	71.14
median	60.00	56.00	76.00	70.50	86.00	79.50	30.00	75.00
st dev	9.46	20.79	10.58	16.64	8.42	17.47	26.46	20.74
Boxplot								

ACTIVITY: COUNTING PASTA

DESCRIPTION: Suppose you manage an Italian restaurant and pasta is one of the main items you serve. Employees use different techniques of measuring half-cup servings of pasta. Some workers use a cup measure and fill pasta to the half-cup line, and other workers use half-cup spoons. You notice that these half-cup servings vary in size. As a manager, you need to find the way of measuring pasta (using a cup or spoon measure) that varies as little as possible so that you can plan and run your business better.

MATERIALS NEEDED: Several boxes of pasta shells. A set of clear one-cup measures and a set of plastic spoon measures.

1. Collecting the data

The class will be divided into an even number of teams consisting of at least two students each. The teams will then be separated into two groups.

Within the first group, each team pours shells into a cup measure and counts the number of shells.

(a) One person pours shells into the cup measure until he or she thinks it is full at the half-cup level.

(b) A second person then counts the number of shells and records it, without informing the first person the number.

(c) Each team repeats the experiment five times and computes the mean ($\bar{x}$) and standard deviation (s) for its measurements.

In the second group, students use a plastic spoon to measure a half cup of shells.

(a) One person pours shells into the spoon until he or she thinks that the spoon is full.
(b) A second person then counts the number of shells and records it, without informing the first person the number.
(c) Each team repeats the experiment five times and computes the mean and the standard deviation (s) for its measurements.

2. Making plots
(a) Construct stemplots of the mean counts of the teams for each group and combine these two plots in a back-to-back stem and leaf diagram.
(b) Construct stemplots for the standard deviations of the teams for each group and combine these in a back-to-back stem and leaf diagram.

3. Analyzing the results
(a) Based on looking at the stem and leaf diagrams for the mean counts for the two groups in Making plots (a), do you see any difference? Does one method of measurement tend to give larger number of counts of shells on average?
(b) Based on looking at the stem and leaf diagrams for the standard deviation of the counts for the two groups in Making plots (b), do you see any difference? Do both measurement processes vary by about the same amount?

Note: We often divide variability in processes into two types:

- Common cause variation is part of the system or process and affects everyone in the system.
- Special cause variation either is not part of the system or process all of the time or does not affect everyone in the system.

(c) Based on your comments in (a) and (b), which of the two measurement methods (cup or spoon) would you recommend and why?

WRAP-UP

In this topic, we were introduced to some basic methods for comparing two or more batches. When the variables are categorical, *side-by-side barcharts* or *segmented bar charts* are useful in comparing frequency distributions and groups can be compared by computing a difference in percentages. When comparing batches of quantitative data, a comparison like "one batch is 10 larger than a second batch" is possible when the two batches have approximately equal spreads. To summarize each batch we compute a *five-number summary*, and *parallel boxplots* are helpful in graphically comparing batches. To understand one's *relative standing* in a batch, it is useful to compute a *standardized score* that indicates the number of standard deviations one falls from the mean. A rule of thumb was described for flagging possible outliers for special attention and a *modified boxplot* displays these possible outliers with special plotting points.

EXERCISES

1. Bird Watching

The below table gives the frequencies of all types of owls spotted by bird watchers in the states of Ohio and Pennsylvania who participated in the 2006 survey of the Great Backyard Bird Count.

	Ohio	Pennsylvania
Type	Frequency	Frequency
Barn Owl	6	6
Eastern Screech Owl	13	30
Great Horned Owl	16	41
Snowy Owl	9	6
Barred Owl	27	13
Long-eared Owl	3	2
Short-eared Owl	14	4

a. For each state, find the percentage of each type of owl spotted.

b. Construct parallel bar charts for the percentages in the two states.

c. Find the mode for each state.

d. Find the types of owls where there is a significant difference in the percentages in the two states.

2. **Teacher's Salary and Proficiency.**

The average teacher's salary and the percentage of 8th graders "above proficiency" on a specific mathematics exam were collected for all states in the *Report Card on American Education* by the American Legislature Exchange Council. The average teacher's salary for a state was classified as "LOW," "MEDIUM" and "HIGH" and the percentage above proficiency (PCT) was broken down into three groups. The table below gives the number of states with a given salary level and range of PCT; for example, we see that 7 states have low teacher salaries and PCT below 24%.

		Percentage (PCT) above proficiency		
		PCT < 24	24 <= PCT < 32	PCT >=32
Teacher's Salary	LOW	7	5	6
	MEDIUM	5	3	9
	HIGH	2	9	4

a. Do you think there is a relationship between teachers' salaries and performance of students on a standardized test? Explain.

b. For each group of teacher's salary, find the proportion of states with a percentage above proficiency in each of the three groups.

c. Construct segmented bar charts of the proportions computed in part b, where each bar corresponds to one group of teachers' salaries.

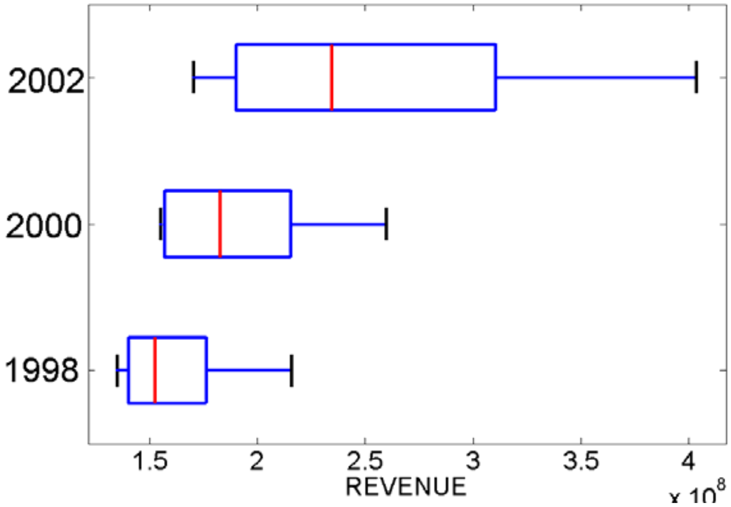
d. Based on your computations in parts b and c, do you think states with higher salaries tend to have better performances on the mathematics exam? Explain.

3. **Gross Sales for Top Movies**

The table below lists the top ten movies (in terms of gross revenue) for the years 1998, 2000, and 2002.

Movie	Revenue ($)	Movie	Revenue ($)
Spider-Man (2002)	403706375	Saving Private Ryan (1998)	216119491
The Lord of the Rings: The Two Towers (2002)	340478898	Armageddon (1998)	201573391
Star Wars: Episode II - Attack of the Clones (2002)	310675583	There's Something About Mary (1998)	176483808
Harry Potter and the Chamber of Secrets (2002)	261970615	A Bug's Life (1998)	162792677
My Big Fat Greek Wedding (2002)	241437427	The Waterboy (1998)	161487252
Signs (2002)	227965690	Doctor Dolittle (1998)	144156464
Austin Powers in Goldmember (2002)	213079163	Rush Hour (1998)	141153686
Men in Black II (2002)	190418803	Deep Impact (1998)	140459099
Ice Age (2002)	176387405	Godzilla (1998)	136023813
Chicago (2002)	170684505	Patch Adams (1998)	135014968
How the Grinch Stole Christmas (2000)	260031035		
Cast Away (2000)	233630478		
Mission: Impossible II (2000)	215397307		
Gladiator (2000)	187670866		
What Women Want (2000)	182805123		
The Perfect Storm (2000)	182618434		
Meet the Parents (2000)	166225040		
X-Men (2000)	157299717		
Scary Movie (2000)	156997084		
What Lies Beneath (2000)	155370362		

Parallel boxplots of the revenues for the three years are displayed below. (The unit for REVENUE on the graph is in hundreds of millions of dollars.)



a. From the graph, describe the distribution of each batch. This discussion should include statements about shape, typical values, spread, and any unusual characteristics.

b. Using the graph, complete the below table.

Group	Median	Q_L	Q_U	IQR = $Q_U - Q_L$
1998				
2000				
2002				

c. How does the spread of revenues change from 1998 to 2000 to 2002? (Look at the IQR column.)

d. How does the median revenue change from 1998 to 2000 to 2002 movies?

e. Can one say that revenues for one year, say 2002, are a particular dollar amount greater than the revenues for a second year such as 2000? Why or why not?

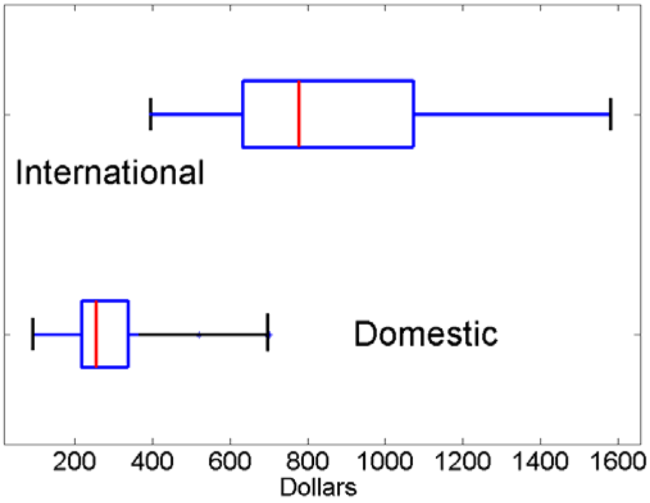
4. **Fares of National and International Flights**

The following table gives the airfares (in dollars) from Detroit to a number of U.S. and International cities. (The data were collected from orbitz.com in May 2004.)

DOMESTIC INTERNATIONAL

CITY	FARE	CITY	FARE
San Francisco	310	London	624
Chicago	92	Cape Town	1582
Miami	252	Beijing	1022
Denver	198	Paris	619
Las Vegas	242	Sydney	1200
Philadelphia	258	Amsterdam	793
Boston	327	Lima, Peru	644
New York	170	Mexico City	397
Fargo	369	Jerusalem	968
San Diego	312	Bangkok	1124
Portland	352	Tokyo	763
Raleigh	224	Milan	657
Kansas City	220		
Honolulu	701		
Fairbanks	696		
Orlando	236		
Sante Fe	521		
New Orleans	280		
Phoenix	204		
Houston	219		

Parallel boxplots of the domestic and international fares are shown in the figure below.



a. From the graph, find the median fare of the domestic flights and the median fare of the international fares.

b. By using a suitable measure from the graph, compare the spreads of the two batches of fares.

c. Would it be appropriate to compare the domestic and international fares by just comparing the medians? Why or why not?

5. **Salaries of Basketball and Baseball Players**

The table below gives the salaries (in millions of dollars) for the players on the 2003-2004 Los Angeles Lakers basketball team and the Los Angeles Dodgers baseball team.

Los Angeles Lakers		**Los Angeles Dodgers**	
PLAYER	SALARY	PLAYER	SALARY
Shaquille O'Neal	26.5	Shawn Green	16.7
Kobe Bryant	13.5	Darren Dreifort	11.4
Brian Cook	0.8	Hideo Nomo	9
Rick Fox	4.5	Todd Hundley	7
Derek Fisher	3.0	Jeff Weaver	6.2
Horace Grant	1.1	Adrian Beltre	5
Devean George	4.5	Eric Gagne	5
Karl Malone	1.5	Odalis Perez	5
Gary Payton	4.9	Paul Lo Duca	4.1
Bryon Russell	1.1	Paul Shuey	3.9
Kareem Rush	1.1	Juan Encarnacion	3.6
Ime Udoka	0.4	Kazuhisa Ishii	2.5
Luke Walton	0.4	Wilson Alvarez	1.5
Jamal Sampson	0.6	Guillermo Mota	1.5
Stanislav Medvedenko	1.5	Tom Martin	1.4
		Alex Cora	1.3
		Robin Ventura	1.2
		Dave Roberts	1.0
		Cesar Izturis	.4
		David Ross	.3
		Duaner Sanchez	.3
		Jayson Werth	.3
		Brian Falkenborg	.3
		Jason Grabowski	.3
		Wilkin Ruan	.3
		Joe Thurston	.3

a. Using the stems below, construct back-to-back stemplots of the Lakers and Dodgers salaries. (The break between the stem and leaf occurs at the decimal point. Since Shaquille O'Neal's salary is so large relative to the remaining salaries, you can place his salary on the "HI" line.)

Lakers Salaries **Dodgers Salaries**

```
| 0  |
| 1  |
| 2  |
| 3  |
| 4  |
| 5  |
| 6  |
| 7  |
| 8  |
| 9  |
| 10 |
| 11 |
| 12 |
| 13 |
| 14 |
| 15 |
| 16 |
| HI |
```

b. Find five-number summaries of each dataset. Describe the distribution of each batch of salaries using the summary numbers.

c. Compare the two groups of salaries. Can you say that one group tends to get higher salaries than the other group?

6. **Ice Cream Calories of Two Manufacturers**

The following table gives the calories of a half-cup serving of different flavors made by Ben and Jerry's and Breyers.

Ben and Jerry's		Breyers	
Flavor	calories	Flavor	calories
Brownie Batter	310	carmel fudge	160
Butter Pecan	290	vanilla	140
Cherry Garcia	260	french vanilla	150
Chocolate	260	van/choc/straw	140
Chocolate Chip Cookie Dough	270	butter pecan	170
Chocolate Fudge Brownie	270	chocolate	150
Chubby Hubby	330	mint chocolate chi	160
Chunky Monkey	300	strawberry	120
Coffee	240	rocky road	150
Coffee HEATH Bar Crunch	290	cookies & cream	160
Dublin Mudslide	270	vanilla fudge twirl	140
Everything But The...	320	peach	130

Fudge Central	300	coffee	140
Half Baked	280	cherry vanilla	140
Karamel Sutra	280	chocolate chip	160
Mint Chocolate Cookie	270	chocolate chip cookie dough	170
New York Super Fudge Chunk	310	vanilla & choc fudge checks	170
Oatmeal Cookie Chunk	280	banana fudge chunk	170
One Sweet Whirled	280	vanilla fudge brownie	160
Peanut Butter Cup	380	cherry chocolate chip	150
Phish Food	280	peanut butter & fudge	170
Pistachio Pistachio	280	dulce de Leche	150
Primary Berry raham	270	lactose free vanilla	160
Strawberry	240	mocha almond fudge	170
Uncanny Cashew	290	butter almond	160
Vanilla	240	calcium rich vanilla	130
Vanilla HEATH Bar Crunch	300	carmel praline crunch	180
Vanilla Swiss Almond	280	fresa banana	140
		homemade vanilla	140
		extra creamy vanilla	150
		extra creamy chocolate	140
		take two	150
		take two (sherbet)	130

a. Construct a stemplot of the calorie numbers for the Ben and Jerry’s flavors and construct a separate stemplot of the calorie numbers of the Breyers flavors.

b. Find a five-number summary of each batch of calorie numbers.

c. Assuming that the spreads of the two batches of calorie numbers are approximately equal, compare the two batches. Which brand of ice cream tends to have more calories and by how much (on average)?

7. **Basketball Field Goal Percentages**

If a basketball team wins many games, then one would expect that the team would have a high proportion of field goal attempts that are successful. In contrast, teams that lose games tend to have a low proportion of successful field goal attempts. But that raises the interesting question: how much better are good teams than bad teams in making field goals? To answer this question, the table below gives, for the 2003-2004 basketball season, the team field goal proportions (FG) for the top 20 ranked college basketball teams and the team field goal proportions for the bottom 20 ranked teams.

BOTTOM TEAMS	FG	TOP TEAMS	FG
Cleveland State	.408	Duke	.472
Harvard	.394	Kentucky	.470
Norfolk State	.392	St. Joseph's	.478
Albany NY	.401	Mississippi State	.468
Eastern Illinois	.444	Connecticut	.481
VMI	.403	Pittsburgh	.482
Western Illinois	.429	Oklahoma State	.517
Florida International	.404	Stanford	.485
Campbell	.394	Texas	.445
Navy	.394	Cincinnati	.457
Charleston Southern	.423	Syracuse	.465
Md. Eastern Shore	.382	Florida	.483
Bethune-Cookman	.391	Georgia Tech	.470
Army	.362	Gonzaga	.521
Howard	.399	Wisconsin	.446
Loyola-Maryland	.378	North Carolina State	.452
North Carolina A&T	.361	Kansas	.462
Nicholls State	.394	North Carolina	.461
Arkansas-Pine Bluff	.344	Maryland	.443
Dartmouth	.407	Providence	.453

a. For the Top Teams, the mean and standard deviation of the shooting proportions are $\bar{x}$ =.471 and s = .021, respectively. Find the standardized scores of North Carolina and Gonzaga and explain what these scores tell you about the relative standing of these two schools with respect to field goal shooting.

b. Using the rule of thumb, determine if there are any outliers in field goal proportion among the Top Teams, and also determine if there are any outliers in shooting among the Bottom Teams.

c. For each batch of shooting proportions, construct a suitable graph and summarize the batch by computing a five-number summary. Write a short paragraph about the distribution of proportions for each batch. Construct parallel boxplots of the two batches. From your work, compare the shooting proportions of the Bottom and Top teams.

8. **Gas Prices**

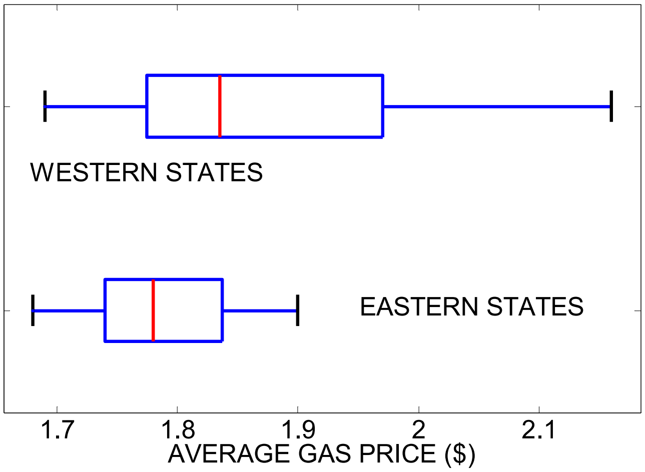
During the spring of 2004, gasoline prices rose sharply. Consumers were interested in the variation of gas prices across states and the cause of this variation. The table below gives the average gas price for Eastern States and Western States, where the dividing line for west/east was the Mississippi River.

EASTERN STATES		WESTERN STATES	
State	gas_price	state	gas_price
Alabama	1.73	Alaska	1.92
Connecticut	1.85	Arkansas	1.74
District of Columbia	1.85	Arizona	1.98
Delaware	1.76	California	2.16
Florida	1.82	Colorado	1.85
Georgia	1.7	Hawaii	2.16
Illinois	1.89	Iowa	1.77
Indiana	1.83	Idaho	1.96
Kentucky	1.77	Kansas	1.8
Massachusetts	1.78	Louisiana	1.73
Maryland	1.78	Minnesota	1.8
Maine	1.79	Missouri	1.72
Michigan	1.85	Montana	1.89
Mississippi	1.74	North Dakota	1.85
North Carolina	1.74	Nebraska	1.82
New Hampshire	1.73	New Mexico	1.78
New Jersey	1.69	Nevada	2.11
New York	1.9	Oklahoma	1.69
Ohio	1.83	Oregon	2.06
Pennsylvania	1.79	South Dakota	1.81
Rhode Island	1.82	Texas	1.7
South Carolina	1.68	Utah	1.94
Tennessee	1.74	Washington	2.03
Virginia	1.71	Wyoming	1.79
Vermont	1.76		
Wisconsin	1.89		
West Virginia	1.84		

a. The mean and standard deviation of the gas prices for the eastern states are $\bar{x}$ = \$1.79 and s = \$0.063, respectively. Find the standardized scores for the gas prices of South Carolina and New York and give an interpretation of these scores.

b. Using the rule of thumb, determine if there are any outliers among the gas prices for the western states.

Parallel boxplots of the gas prices from the western and eastern states are displayed in the figure below.



c. Compare the two batches of gas prices with respect to "average" and spread.

d. On the average, how much more expensive is gas from a western state than from an eastern state?

e. Would it be accurate to say that *all* western states have more expensive gas than *all* eastern states? If this is not true, find an eastern state that has more expensive gas than a western state.

9. **Cost of Grocery Shopping**

How much money does a consumer spend on a single trip at the grocery store? How has the single-trip cost at a grocery store changed from 2001 to 2003?

Below is a stemplot of the grocery costs (in dollars) of 50 visits in 2001 and 34 visits in 2003 made by the author. (The smallest value in the first dataset corresponds to a purchase of $10.)

Make a comparison of the two batches by (a) finding 5-number summaries of each batch, and (b) constructing parallel boxplots. Compare the two datasets with respect to averages (medians) and spreads (quartile spreads). Make a comparison by comparing medians.

PURCHASE AMOUNTS IN 2001 **PURCHASE AMOUNTS IN 2003**

```
        0 | 1 | 3
        7 | 1 | 6778
     3322 | 2 | 12
       98 | 2 | 79
    43332 | 3 | 13
998777665 | 3 | 559
    43100 | 4 | 1
      998 | 4 | 5799
  4442100 | 5 | 12
       76 | 5 | 5557
      300 | 6 |
    77655 | 6 | 56788
        0 | 7 | 1
        7 | 7 | 8
          | 8 | 12
          | 8 |
        1 | 9 |


1|0 means $10       1|3 means $13
```

10. **Car Mileages**

The table below gives the mileage (miles per gallon) for a selection of 2004 model cars.

CAR	TYPE	MPG	CAR	TYPE	MPG
Chrysler PT Cruiser	Sedan	18	Acura MDX	SUV	17
Ford Focus	Sedan	24	Buick Rendezvous	SUV	16
Honda Civic	Sedan	36	Ford Escape	SUV	17
Hyundai Accent	Sedan	26	Honda Element	SUV	20
Mitsubishi Lancer	Sedan	20	Hyundai Santa Fe	SUV	18
Pontiac Vibe	Sedan	26	Mazda Tribute	SUV	18
Saturn Ion	Sedan	24	Nissan Murano	SUV	19
Subaru Impreza	Sedan	20	Saturn Vue	SUV	18

Toyota Corolla	Sedan	29	Subaru Forester	SUV	21
Toyota Matrix	Sedan	24	Toyota Rav4	SUV	22
Volkswagen Golf	Sedan	41	Volkswagen Toureg	SUV	15
Volkswagen New Beetle	Sedan	25	Volvo XC90	SUV	18

a. Looking at the entire collection of cars, use the rule of thumb to find any possible outliers in the mileage values.

b. Draw a modified boxplot of the mileages.

c. By the use of graphs and appropriate summary statistics, compare the mileages of the Sedans with the mileages of the SUVs. Can you explain why there are substantial differences in the mileages between the two groups of cars?

11. **State Population Changes**

The table below gives the percentage change in population from 1990 to 2000 for all states in the United States. The first column indicates if the state is East or West of the Mississippi River.

	State	%change		State	%change		State	%change		State	%change
East	AL	10	East	IL	8.7	West	MT	12.9	East	RI	4.5
West	AK	14	East	IN	9.7	West	NE	8.4	East	SC	15.1
West	AZ	40	West	IA	5.4	West	NV	66.4	West	SD	8.3
West	AR	13.7	West	KS	8.5	East	NH	11.4	East	TN	16.6
West	CA	13.8	East	KY	9.7	East	NJ	8.8	West	TX	22.8
West	CO	30.6	East	LA	5.9	West	NM	20.1	West	UT	29.7
East	CT	3.6	East	ME	3.8	East	NY	5.5	East	VT	8.2
East	DE	17.6	East	MD	10.8	East	NC	21.4	East	VA	14.4
East	DC	-5.6	East	MA	5.5	West	ND	0.6	West	WA	21.1
East	FL	23.5	East	MI	6.9	East	OH	4.7	East	WV	0.8
East	GA	26.4	West	MN	12.4	West	OK	9.7	East	WI	9.7
West	HI	9.3	East	MS	10.5	West	OR	20.4	West	WY	8.8
West	ID	28.5	West	MO	9.3	East	PA	3.4			

a. Explain (before looking at the data) how you think the United States has grown in recent years. Are there particular areas of the country that have experienced high growth? Is your state one of the high growth states?

b. Suppose we define a population change as being HIGH if the percentage change exceeds 10 % and LOW if the percentage change is 10 % or lower. Compare the proportion of LOW and HIGH population changes for the Eastern and Western states using a graph and appropriate summary statistic.

c. Using parallel boxplots, compare the percentage changes for the Eastern and Western states.

12. **City Temperatures**

The table below gives the average temperature for eight cities for each month of the year.

Month	San Francisco	Vero Beach	Duluth	Albuquerque	San Diego	Philadelphia	Honolulu	Indianapolis
Jan	48.7	61.6	7.0	34.2	57.4	30.4	72.9	25.5
Feb	52.2	62.7	12.3	40.0	58.6	33.0	73.0	29.6
Mar	53.3	67.2	24.4	46.9	59.6	42.4	74.4	41.4
Apr	55.6	71.3	38.6	55.2	62.0	52.4	75.8	52.4
May	58.1	75.8	50.8	64.2	64.1	62.9	77.5	62.8
Jun	61.5	79.5	59.8	74.2	66.8	71.8	79.4	71.9
July	62.7	81.1	66.1	78.5	71.0	76.9	80.5	75.4
Aug	63.7	81.3	63.7	75.9	72.6	75.5	81.4	73.2
Sep	64.5	80.1	54.2	68.6	71.4	68.2	81.0	66.6
Oct	61.0	75.5	43.7	57.0	67.7	56.4	79.6	54.7
Nov	54.8	69.3	28.4	44.3	62.0	46.4	77.2	43.0
Dec	49.4	63.7	12.8	35.3	57.4	35.8	74.1	30.9

a. Suppose we say that a month is “HOT” if the average temperature exceeds 70; otherwise the month is “COLD.” For each city, find the proportion of HOT and COLD months for each city. Graph these proportions and discuss any differences you see between cities.

b. For four cities of your choice, compare the monthly temperatures. Construct a graph that is helpful in comparing cities and find appropriate summary statistics for comparing cities with respect to monthly temperature.

13. **Church Attendance**

This table gives the worship attendance at a church for all weeks during the four seasons of the year. For example, the attendances during the first two weeks of the Winter season were 439 and 349.

WINTER	SPRING	SUMMER	FALL
439	426	375	429
349	418	342	470
388	535	416	438
421	352	342	407
363	522	332	395
362	398	372	355
406	427	293	402
343	384	343	455
479	409	322	422
399	472	357	425
381	377	299	426
289	344	348	399

a. Suppose we say that attendance is "HIGH" for a particular week if it is 350 or higher, otherwise it is "LOW." Find the proportion of HIGH and LOW attendance numbers for each season of the year.

b. Construct a suitable graph comparing the proportions you computed in part a. What differences do you find between seasons?

c. Compare the four seasons of attendance numbers by use of five-number summaries and parallel boxplots. Do you reach the same conclusions as you found in part b?

14. **Men and Women Professional Golfers**

The datasets pgastats.txt and lpgastats.txt contain statistics for the top 30 men and top 30 women professional golfers for the 2002 season. Some of the variables included on these datasets include.

DRIVING_AVG – the average (mean) length of a drive

DRIVING_ACC – the percentage of drives that land in the fairway

GREEN_PCT – the percentage of greens that are hit in regulation

PUTTS – the average number of putts for a 18-hour round

For two of these variables, compare the men and women professional golfers using the methods described in this topic.

TOPIC D5: RELATIONSHIPS BETWEEN CATEGORICAL VARIABLES

SPOTLIGHT: THE TITANIC: WOMEN AND CHILDREN FIRST?

On April 15, 1912, the ocean liner Titanic collided with an iceberg, sank, and many of its passengers lost their lives. Many books have been written about this catastrophic event. One issue that has been discussed is the role of economic status in determining the mortality status of the passengers on this ship. In the recent movie *Titanic*, it was suggested that many of the lower-class passengers were less likely than the higher-class passengers to survive this accident. Dawson (1995) in the *Journal of Statistics Education* (vol. 3, n. 3) makes available an interesting dataset on the passengers of the Titanic. For each of the 2201 passengers, the dataset records

- the economic status (crew, first-class, second-class, third-class)
- the age (adult or child)
- the gender (male or female)
- the survival status (either yes or no)

By use of these data, we will see if there is a relationship between the economic status and survival of the Titanic passengers.

PREVIEW

In this topic, we discuss ways of exploring relationships when the variables are categorical. We begin with a two-way table of counts and by computing relevant *conditional* proportions from the table, we can discuss how knowledge of one variable is informative about the second variable.

In this topic your learning objectives are to:

- Understand how to construct a two-way table of counts from a data table.

- Understand how to compute conditional row percentages or conditional column percentages from the two-way table.
- Understand how to graph sets of conditional percentages.
- Understand how to use conditional percentages to describe the relationship between two categorical variables.

NCTM Standards

✓ In Grades 6-8, all students should collect data about different characteristics within one population.

✓In Grades 9-12, all students should display and discuss bivariate data where at least one variable is categorical.

A TWO-WAY TABLE OF COUNTS

The following table displays a partial listing of the Titanic data. Each row corresponds to the values of the four categorical variables for a passenger.

CLASS	AGE	SEX	SURVIVED
First	Adult	male	Yes
First	Adult	male	Yes
First	Adult	male	Yes
First	Child	female	No
Crew	Adult	female	Yes
Crew	Adult	female	No
Crew	Adult	female	No

Crew	Adult	female	No

The Titanic accident always will be remembered for the large number of fatalities. Let's first focus on the "Survived" variable in the dataset and construct a frequency table. The table below shows the counts of people who survived and who did not survive; this table also shows the corresponding percentages. From this table, we see that about 2/3 of the passengers did not survive this trip.

.

	Survived		Total
	no	yes	
Count	1490	711	2201
%	67.7	32.3	100

Let's consider the question posed earlier -- were passengers of certain economic classes more or less likely to survive the accident?

To answer this question, we classify the passengers in the below table by the survival status and the economic status (crew, or first, second, or third class). We see, for example, there were 122 people in first class who did not survive, and 203 people in the first class who survived.

		Survived		Total
		no	yes	
Class	Crew	673	212	885
	First	122	203	325
	Second	167	118	285
	Third	528	178	706
	Total	1490	711	2201

Although this table is useful in understanding the numbers of people who died and survived in different classes, it is difficult to compare the distributions since there were many more non-survivors in the dataset. For comparison purposes, it is better to compute the percentage of survivors and non-survivors for each class.

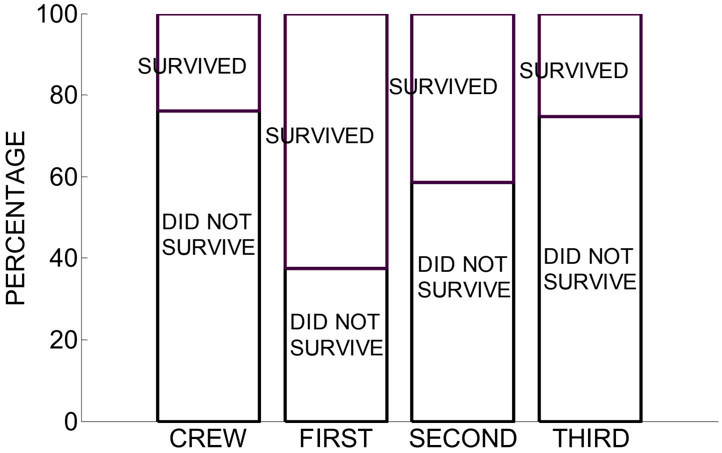
We can get these percentages from the table by dividing each basic count by the corresponding row total. For example, the percentage of first-class people who survived is 203 / 325 = 62%. If we compute these row percentages for all of the table entries,

		Survived		Total
		No	yes	
Class	crew	673/885	212/885	885
	first	122/325	203/325	325
	second	167/285	118/285	285
	third	528/706	178/706	706
	TOTAL	1490/2201	711/2201	2201

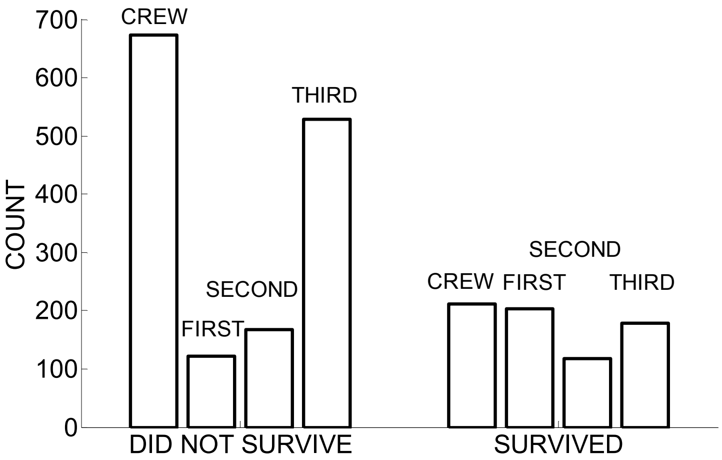
we obtain the row percentages.

		Survived		TOTAL
		no	yes	
Class	Crew	76%	24%	100%
	First	38%	62%	100%
	Second	59%	41%	100%
	Third	75%	25%	100%
	Total	68%	32%	100%

We see from the row percentages that there is a strong association between class and survival. Overall, 32% of the passengers survived. However, a majority (62%) of first-class passengers survived compared to a survival rate of 41% for second-class passengers and a survival rate of 25% for third-class passengers and crew. To display this association graphically we can use a set of stacked bar charts. Each bar corresponds to a passenger class, and the bar is divided by the survival status. This graph clearly shows the high survival rate of the first-class passengers.



In the above analysis, we divided each table count by the row total and found the row percentages. This told us the percentage of each class group that survived and did not survive. But we can look at the association in the table a different way. To compare the class of the non-survivors with the class of the survivors, we can construct side-by-side bar charts of the two sets of column frequencies, as shown below.



This graph is informative, but it is difficult to compare the two sets of bars, since there are different numbers of passengers in the two classes. It is easier to compare the two sets of frequencies if they are converted to percentages. Again we start with our basic count table

		Survived		Total
		no	yes	
Class	Crew	673	212	885
	First	122	203	325
	second	167	118	285
	Third	528	178	706
	Total	1490	711	2201

and divide the counts by the column totals

		Survived		TOTAL
		No	yes	
CLASS	crew	673/1490	212/711	885/2201
	first	122/1490	203/711	325/2201
	second	167/1490	118/711	285/2201
	third	528/1490	178/711	706/2201
	TOTAL	1490	711	2201

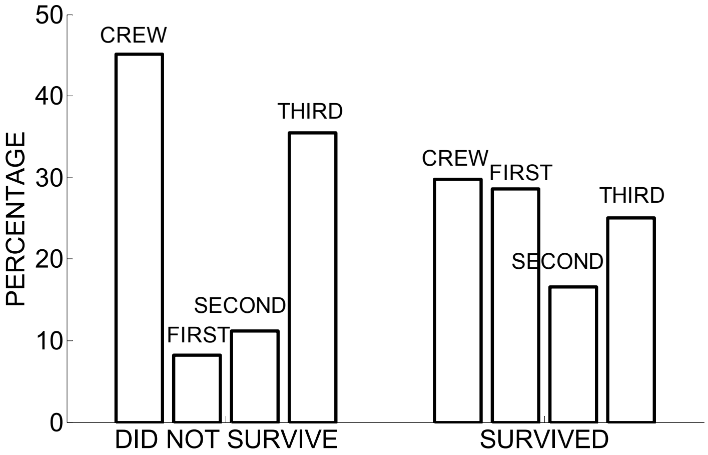
to obtain the column percentages

		Survived		TOTAL
		no	yes	

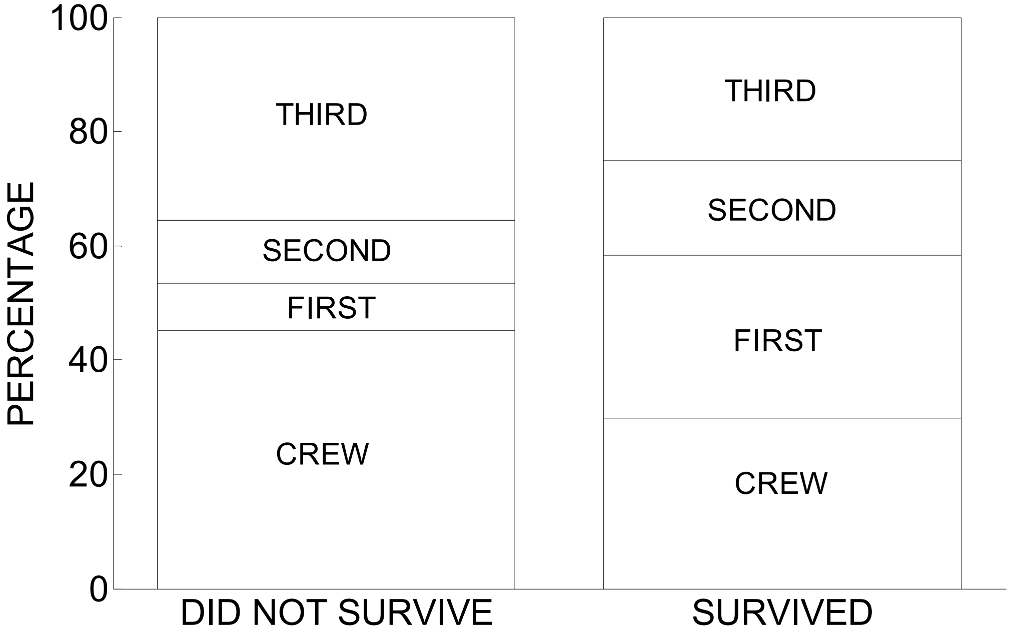
CLASS	Crew	45%	30%	40%
	First	8%	29%	15%
	Second	11%	17%	13%
	Third	35%	25%	32%
	TOTAL	100%	100%	100%

Let's interpret these column percentages. Of the 2201 passengers in the Titanic, 40%, 15%, 13%, and 32% were respectively crew, first-class, second-class, and third-class passengers. However, there was a different composition if you consider only the 711 people who survived the accident. Of this group, there were about equal percentages of crew and first-class passengers, 25% third class, and 17% that were second class. But more importantly, if the passengers who died, almost half of them were crew and only 8% were first-class.

We can display these two sets of column percentages in several ways. We can use side-by-side bar charts, where the percentage (instead of a frequency) is plotted.



Alternately, we can use two stacked bar charts, where each bar chart contains the percentages of class memberships for a survival group.



One can study the association in a two-way table by computing row percentages or column percentages. Which way to go depends on the particular application and ease of interpretation.

In this example, is it easier to talk about

- the survival rate of different classes of passengers

or is it easier to discuss

- the class status of the group of passengers who survived and the status of the passengers who did not survive?

Here the author thinks it makes more sense to talk about how survival depends on the economic status of the passenger, so I would prefer computing row percentages and using a corresponding graph (such as a set of segmented bar charts) to communicate the difference in row percentages across class.

PRACTICE: LOOKING AT THE AGE AND GENDER OF THE TITANIC PASSENGERS

It is also interesting to relate the survival of the Titanic passengers with respect to their age and their gender.

1. The below table classifies the passengers with respect to their survival (no or yes) against their age (child or adult). For each age, compute the percentage of survivors and percentages of nonsurvivors and place your results in the second table. Were children more or less likely to survive than adults? Explain.

		Survived?	
		No	Yes
Age	Child	52	57
	Adult	1438	654

		Survived?	
		No	Yes
Age	Child		
	Adult		

2. Using two stacked bar charts, show how the survival of the passenger depends on age.

3. In a similar fashion, compute relevant percentages to see if the survival status of the passengers depended on their gender. Place your percentages in the second table and describe your conclusions.

		Survived?	
		No	Yes
Gender	Female	126	344
	Male	1364	367

		Survived?	
		No	Yes
Gender	Female		
	Male		

4. Use stacked bar charts to show how the survival depends on gender.

5. We have seen that the survival status of the Titanic passengers was dependent on their economic status and also dependent on their gender. That raises the question: was there any relationship between economic status and gender? By using the below table and computing relevant row or column percentages, answer this question.

		Class			
		Crew	First	Second	Third
Gender	Female	23	145	106	196
	Male	862	180	179	540

6. Does it make sense that children and women were more likely to survive this disaster? Explain.

HANDS-ON ACTIVITY: PREDICTABLE PAIRS

DESCRIPTION: This is a class participation activity, where students will indicate, through a show of hands, if they have seen or not seen different movies. If two movies are similar in some way, then it is possible that one can use information about students seeing one movie to predict whether they have seen (or not seen) a second movie.

1. Find a movie that roughly half of the class has seen. (You can ask the class for possible suggestions of movie titles.) Write down the name of the movie.

2. Now find a second movie that you think is similar in some way to the first movie. Write down the name of this second movie. ______________________

- For the students that ***have seen*** the first movie, ask if they have or have not seen the second movie.
- For the students who ***have not seen*** the first movie, ask if they have or have not seen the second movie.

Put your data in the table below.

	Seen 1st movie?	
Send 2nd movie?	Yes	No
Yes		
No		

3. By the computation of relevant row or column percentages, answer the question: Is there a relationship between seeing the first movie and seeing the second movie? Explain.

4. Repeat the above analysis for a movie that is a romantic comedy.

5. Repeat the above analysis for an action movie with some violence.

SIMPSON'S PARADOX

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

Which hospital had the better survival rate? We can answer this question by computing the proportion of hospital A's patients who survived and the proportion of hospital B's patients who survived.

	Survival Rate
Hospital A	800/1000 = 0.8
Hospital B	900/1000=0.9

We see that Hospital B saved the higher percentage of its patients.

Before you decide that Hospital B is the better hospital, 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

Suppose that we compute the recovery rates for the two hospitals for each condition. Among those who were in good condition, we compute the recovery rates for the two hospitals by dividing the number who survived (590 and 871) by the number of good patients (600 and 900).

Good condition	Survival rate

Hospital A	590/600=0.983
Hospital B	870/890=0.967

Hospital A saved the greater percentage of its patients who had been in good condition.

Let's look next at the patients who were in poor condition. Using the counts in the second table, we compute the recovery rates for poor condition patients.

Poor patients	Survival Rate
Hospital A	210/400=0.525
Hospital B	30/100=0.30

Hospital A also saved a greater proportion of patients in good condition.

We have discovered a surprising result known as Simpson's paradox. Hospital A has the higher recovery rate for each type of patient. But when we aggregate the table, the association goes the other direction – Hospital B has the higher recovery rate for all patients. In generally, Simpson's paradox refers to the fact that aggregate proportions can reverse the direction of the relationship seen in the individual pieces.

Why are we observing Simpson's paradox in this illustration? First, note that good patients tend to survive more often than poor patients. If we look at the original set of tables, we see that 900/1000 = 90% of the patients going to Hospital B are good patients; in contrast only 400/1000 = 40% of the Hospital A patients are good. So the reason why Hospital B has an overall higher survival rate is due primarily to the fact that it is treating better patients.

PRACTICE: SIMPSON'S PARADOX

In the two-year period 1995-1996, the baseball player Derek Jeter had 195 hits in 630 at-bats (opportunities) and David Justice had 149 hits in 551 at-bats.

1. Which player had the higher batting average during this period (batting average is defined as hits divided by at-bats).

2. Suppose we break the above data by year – we get the following table.

	1995		1996	
	Derek Jeter	David Justice	Derek Jeter	David Justice
Hits	12	104	183	45
At-bats	48	411	582	140

Who was the better hitter in 1995? Who was the better in 1996?

3. Explain how this example demonstrates Simpson's paradox.

WRAP-UP

In this topic, we discussed how to summarize relationships between two categorical variables when the data are presented in a two-way table. One first computes row percentages (or column percentages) and then looks for association by comparing the sets of row percentages. If there are differences in the row percentages across rows, this indicates that the outcome of the column variable is dependent on the outcome of the row variable. Side-by-side bar charts or segmented bar charts are useful for displaying the relationship that one finds in a two-way table.

EXERCISES

1. **Movie Ratings**

On the Internet Movie Database (www.imbd.com), people are given the opportunity to rate movies on a scale from 1 (hate the movie) to 10 (love the movie). For 27337 people who rated the movie *Die Hard*, the first table classifies these people by their rating and their gender. The second table classifies the people by their rating and their age.

Rating

				5 or
		9, 10	6, 7, 8	lower
Gender	Males	9228	13154	1634
	Female	891	1862	468

			Rating	
				5 or
		9, 10	6, 7, 8	lower
	under 18	423	333	43
Age	18-29	6490	9360	1186
	30-44	2308	3878	620
	45+	518	919	193

a. Explore the relationship between gender and rating in the first table by computing relevant conditional percentages. Use a graph to show this relationship. Is this movie more popular among men or women?

b. Explore the relationship between age and rating in the second table. Is Die Hard more popular among viewers of a particular age?

c. Do you think that the sample of people who rated this movie is representative of all people who watched this movie? Explain.

2. **Movie Ratings**

The following table classifies people who rated the movie *Sleepless in Seattle* by rating and gender. By computing relevant row or column percentages, investigate if the ratings of males is different from the ratings of females. Also use a graph to display the relationship. Are you surprised by the findings?

			Rating	
		8,9,10	5,6,7	1,2,3,4
Gender	Males	2217	3649	754
	Females	1059	835	178

3. Available Vehicles to Households

The following table from U.S. Census Bureau classifies the American households for five years with respect to the number of vehicles (unit = 1000 households). We see from the table that, for example, there were 11,417 (thousand) households in 1960 that did not have a vehicle.

			Year		
Vehicles					
Available	1960	1970	1980	1990	2000
None	11,417	11,081	10,390	10,602	10,861
1	30,189	30,268	28,565	31,039	36,124
2	10,074	18,600	27,347	34,361	40,462
3 or more	1,342	3,495	14,088	15,945	18,034

a. For each year, compute the percentage of households having different number of available vehicles.
b. By comparing the column percentages, describe how the number of vehicles available to American households has changed over time.

4. Participation in Selected Sports Activities

The following table gives the number (in thousands) of Americans in 2001 that participated in selected sports activities. (These data were taken from *Statistical Abstract of the United States* 2003.)

	Age of Participant							
	7-11	12-17	18-	25-34	35-44	45-54	55-64	65 yrs
Sport	yrs	yrs	24yrs	yrs	yrs	yrs	yrs	and over
Basketball	6356	7818	3955	4397	3616	1278	422	361
Bowling	5330	5893	6806	8597	7205	3649	1265	1558
Exercise	2417	3550	6936	12332	14692	13616	8237	9438

walking								
Golf	1011	2264	3022	5197	5906	4754	2033	2450
Soccer	5867	1811	1312	1115	972	447	146	196
Tennis	728	1963	1487	2256	2157	1217	535	569

By computing relevant conditional percentages, explore the relationship between sport and age of the participant. Classify the sports into those that are enjoyed by "young" people and those that are enjoyed by "old" people.

5. **Cigarette Smoking among Americans**

The following table (from *Statistical Abstract of the United States* 2003) gives the percent of people in each category who smoke at least "some days" for the years 1990, 1995, 2000. By constructing suitable graphs, show how the tendency to smoke depends on gender and age. Also, use graphs to show how the percent of people who smoke has changed over time.

	AGE	1990	1995	2000
Male, total		28.4%	27%	25.7%
	18 to 24 years	26.6%	27.8%	25.7%
	25 to 34 years	31.6%	29.5%	29.0%
	35 to 44 years	34.5%	31.5%	30.2%
	45 to 64 years	29.3%	27.1%	26.4%
	65 years and over	14.6%	14.9%	10.2%
Female, total		22.8%	22.6%	21.0%
	18 to 24 years	22.5%	21.8%	25.1%
	25 to 34 years	28.2%	26.4%	22.5%
	35 to 44 years	24.8%	27.1%	26.2%
	45 to 64 years	24.8%	24.0%	21.6%
	65 years and over	11.5%	11.5 %	9.3%

6. **Spending Money**

How do Americans spend their money? The following table gives the total expenditures of U.S. consumers (in $1000) in several categories for three years. from *Statistical Abstract of the United States* 2003.

	1990	1995	2000
Food	4296	4505	5158
Alcoholic beverages	293	277	372
Housing	8703	10458	12137
Apparel and services	1618	1704	1856
Transportation	5120	6014	7404
Health care	1480	1732	2066
Entertainment	1422	1612	1863
Personal insurance and pensions	2592	2954	3365
TOTAL	25524	29256	34221

a. For each year, compute the percentage of the total expenditure that was spent on the different categories.

b. Construct a graph to compare the percentages you computed in (a).

c. Has the pattern of spending in the different categories changed over time? Explain.

7. **Births in the United States**

When are babies born? All of the births in the United States in the year 1978 were classified by month and day of the week. The unit is thousands of births – so, for example, there were a total of 38 thousand babies born on Sundays in January.

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	TOTAL
January	38	45	46	36	37	37	32	271
February	31	36	38	37	37	37	33	249
March	31	37	38	46	46	47	32	277
April	37	36	37	35	35	36	39	255

May	30	44	47	46	36	37	32	272
June	31	37	38	38	47	48	32	271
July	42	49	39	39	41	41	44	295
August	34	40	52	50	51	41	35	303
September	34	39	42	42	41	51	44	293
October	41	48	50	39	39	39	34	290
November	32	39	40	47	46	38	33	275
December	40	37	39	39	39	48	42	284
TOTAL	421	487	506	494	495	500	432	3335

a. Compute the proportion of babies that were born on different days of the week. Are there particular days of the week when babies are less likely to be born? Can you explain the reason why these days are not popular?

b. Compute the proportion of babies born in different months of the year. Are there particular seasons where babies are more or less likely to be born? Can you explain why?

8. **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 lived 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 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.

9. **Throwing and Batting Side of Baseball Players**

For all of the major league baseball players who where born in 1970 or later, the following table classifies the players by their throwing hand (left or right) and their batting side (left only, right only, or both sides).

		Throwing hand	
		Left	Right
Batting	Both	58	587
side	Left only	1216	610
	Right only	195	4287

a. If a baseball player is a southpaw (that is, throws with his left hand), what is the chance that he will also bat from the left side only?

b. If a player throws right, what is the chance that he will bat from the left side only?

c. A switch-hitter is a player that bats from both sides. Is it more likely for a left-hand thrower or a right-hand thrower to be a switch-hitter? Can you explain why this might be the case? (HINT: In baseball, most pitchers are right-handed and it is easier to get a hit from a right-handed pitcher if you bat from the left side.)

10**. Purchases by Gender**

In a survey, students in a statistics class specified the number of pairs of shoes they owned and the number of movie DVDs they owned. The following tables classify the responses by gender and number of pairs of shoes, and by gender and number of DVDs owned.

Number of pairs of shoes owned	Female	Male
Eight or less	53	181
Between 9 and 20	225	34
Betweeen 21 and 50	124	6
Over 50	11	0

Number of DVDs owned	Female	Male
None	16	10
Between 1 and 10	140	61
Between 11 and 20	115	57
Between 21 and 50	118	56
Over 50	32	36

a. For each gender, compute the percentage of students who own Eight or less, Between 9 and 20, Between 21 and 50, and Over 50 pairs of shoes. Graph the two sets of percentages, and explain if there is a relationship between gender and number of pairs of shoes owned.

b. For each gender, compute the percentage of students who own no DVDs, between 1 and 10 DVDs, and so on. By comparing the two sets of percentages, is there a relationship between gender and the number of DVDs owned?

11. **The Pop and Soda Controversy**

On the Pop vs. Soda web page at www.**pop**vs**soda**.com, American readers were asked to give their state of their home town and answer the question "What generic word do you use to describe carbonated soft drinks?" The following table gives the answer to the question for respondents from six different states.

State	pop	soda	coke	other
Alabama	30	260	2161	198
California	669	12843	2301	6006
Kansas	1938	483	266	212
Maine	25	993	15	57
New Jersey	96	5830	212	113
Ohio	12317	1730	329	343

a. For each state, compute the percentages of people giving the answers "pop", "soda", "coke" and "other".

b. Graph the six sets of state percentages using parallel bar plots.

c. Some states in the U.S. are regarded as "pop" states, other states are regarded as "soda" states, and other states are "coke" states. Based on your computations, describe what regions of the United States correspond to each type.

d. What is your answer to this question? Do you believe that your answer is consistent with the majority of people from your home town state? (Check the website if your state is not included in the above list.)

12. **The Ohio Graduation Test**

All students in Ohio must currently pass the Ohio Graduation Test (OGT) that measures the level of reading, writing, mathematics, science, and social studies skills of students at the end of the sophomore year. Each school is rated on the basis of their students' performance in the OGT and a school earns a state indicator if 75% or more students at that school obtain a proficient level on a particular section of the test.

All of the public high schools in Ohio were classified as either earning a state indicator or not earning a state indicator for the math and science sections of the OGT and the following table gives the corresponding two-way table of counts.

		State indicator in science?	
		No	yes
State indicator in math?	no	73	0
	yes	163	373

a. What percentage of high schools earn a state indicator in math?

b. What percentage of high schools earn a state indicator in science?

c. What percentage of high schools earned state indicators in both math and science?

d. Of the schools earning a state indicator in math, what percentage also earned a state indicator in science?

e. Of the schools not earning a state indicator in math, what percentage also earned a state indicator in science?

f. Is there an association between a school's performance on the OGT math section and the OGT science section? Explain why this association exists.

g. Suppose the high schools are divided into "small" and "large" schools depending on their enrollment. The following table classifies the schools by their size and their performance on the OGT science test. By computing relevant percentages, check if there is an association between a school's size and its performance on the science OGT test.

		State indicator in science?	
		No	Yes
Size of school	small	108	195
	large	128	178

13. **Salaries for Associate and Bachelor Degrees**

An administrator is interested in comparing the salaries of students who receive Associate Degrees with the salaries of students who receive Bachelor Degrees. For 5000 recent Associate Degree graduates, suppose we classify the graduate by Subject Area (health, business, other) and the Salary (low or high), obtaining the following table.

Associate Degree graduates:

	Salary		
Subject Area	Low	High	TOTAL
Health	250	2250	2500
Business	600	900	1500
Other	700	300	1000
TOTAL	1550	3450	5000

Likewise, suppose 5000 Bachelor Degree graduates are classified with respect to Subject Area and Salary.

Bachelor Degree graduates:

	Salary		
Subject Area	Low	High	TOTAL
Health	25	475	2500

Business	700	1300	1500
Other	1625	875	1000
TOTAL	2350	2650	5000

a. What proportion of Associate Degree graduates earn high salaries? What proportion of Bachelor Degree graduates earn high salaries? Is it accurate to say that a greater proportion of Associate Degree holders get high salaries?

b. Suppose that you are a Business major. What proportion of these majors earn high salaries if you have an Associate Degree. What proportion of these majors earn high salaries if you have a Bachelor Degree? Is it advantageous to get a Bachelor Degree?

c. Suppose that you are a Health major. Are you more likely to get a high salary with an Associate Degree or a Bachelor Degree?

d. Suppose that you are an “Other” major. Are you more likely to get a high salary with an Associate Degree or a Bachelor Degree?

e. Based on your answers to parts b, c, d, which type of degree gives you a higher salary? Why does the comparison using the totals in part a go the other direction?

f. How does this example illustrate Simpson’s paradox?

14. **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 was 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.

TOPIC D6: RELATIONSHIPS BETWEEN QUANTITATIVE VARIABLES

SPOTLIGHT: MEASURING CLIMATE

One reason why people relocate to a different city is climate. The book *Cities Ranked & Rated* presents data on the key components of climate – temperature, precipitation, cloud cover, humidity, and hazards – that people may use in deciding on a place to live. We will be using some of the data from this book to explore the relationships between different weather measurements for a group of cities.

Climate actually refers to the physical characteristics of a location that will affect the weather that we observe. What factors determine climate of a metropolitan area? One obvious factor is a place's latitude or north-south location. Generally places farther south tend to be warmer, and those further north are colder and have greater seasonal changes. The altitude of an area can have a large effect on climate. A location's higher altitude means less dense air and generally less humidity – this means less oxygen which places greater strain on the human circulatory system. The availability of nearby water (a lake or an ocean) can have a significant impact on climate. Water helps to moderate a place's temperature, but the water's moisture can affect local precipitation such as "lake effect" snows from Lake Erie or Lake Michigan. Wind direction can also affect climate – much of the climate in the United States is governed by the west to east movement of air across the middle latitudes of North America. Landforms, such as mountains and valleys, can have a large impact on climate. For example, mountain ranges can block winds, creating a drier, less humid climate. Also storm tracks affect the climate of cities along their path. Cities located near common storm tracks will experience greater swings in weather and strong storms.

The climate or physical characteristics of cities actually are fixed. But the weather we observe depends on the interaction of climate factors such as latitude, winds, and storms, and can exhibit considerable variation. *Cities Ranked & Rated* measures weather for a location in different ways. The average minimum temperature in January and the average maximum temperature in July are recorded – this gives a person an idea about the temperature range for a city. In addition, the book presents the average number of days where the high temperature exceeds 90 or the low temperature is below freezing. Precipitation is measured both by the annual inches of rain and snow combined and the number of inches of snowfall. The average number of days in the year with at least some measurable rain or snow is recorded. The July relative humidity is measured for a city – this is the moisture content of air relative to temperature and greater humidity usually refers to less comfort. Other measures of comfort are made, including the average number of mostly sunny days, and a score measuring the risk of tornados and hurricanes.

PREVIEW

Often we collect more than one variable from each individual in a dataset. We do so because we are interested in studying the relationship between the variables. When we do this, we typically have

- *a response variable* – a variable that we are mainly interested in
- *an explanatory variable* – a variable that we think might be helpful in explaining some of the variation in the response variable

In this topic we describe some general ways of looking at the relationship between two variables.

In this topic your learning objectives are to:

- Understand how a pattern in a scatterplot tells us about the direction and the strength of a relationship.
- Understand what a QCR and a correlation coefficient tell us about the relationship between two quantitative variables.

NCTM Standards

✓In Grades 6-8, all students should select, create, and use appropriate graphical representations of data, including scatterplots.
✓In Grades 6-8, all students should make conjectures about possible relationships between two characteristics of a sample on the basis of scatterplots of the data.
✓In Grades 9-12, all students should, for bivariate measurement data, be able to determine correlation coefficients using technological tools.

RELATIONSHIPS - SCATTERPLOTS

In many situations, we will collect two or more measurements from individuals. For example, we might collect the pulse rate and weight from a number of students, or we might collect different weather measurements from different cities. In this situation, we are often interested in the relationships between the measurements. We will study relationships by

- graphing the data to get a picture of the association between two measurements
- computing a summary value, called a correlation coefficient, to describe the strength of the relationship between the variables
- drawing a "best line" of fit to describe how one variable changes as a function of the other variable

Weather data

From the usatoday.com web site, there is a weather section that describes the climate of different cities in the United States. For 10 cities, the below table displays the average high daily temperatures (in degrees Fahrenheit) in January and July, the average amount of precipitation (inches of rain or melted snow) in January and July, the average dew point (a humidity measure in Fahrenheit degrees) in January and July, and the latitude in degrees. Here the data units are the cities, and we are recording seven measurements for each data unit. Note that there are several blank values in the table – the dew point was not given for two cities.

	January	July	Jan	July	Jan	July	
City	temp	temp	precip	precip	dew	Dew	Latitude
Aberdeen	20	85	0.5	3	3	59	45
Akron	33	83	2.7	4	19	61	40
Albuquerque	47	92	1.4	1.3	18	49	35
Amarillo	49	91	0.5	2.8	19	58	35
Aspen	34	80	2	1.5			39
Atlanta	52	89	4.7	5.3	32	68	33
Bakersfield	57	98	1	0	39	51	35
Bar Harbor	33	77	4.7	3.3			44
Chicago	29	84	1.7	3.6	14	62	41
Miami	76	89	2	6	58	73	25

Let's focus on two variables

- the average high daily temperature in January (called Jan temp)
- the average high daily temperature in July (called July temp)

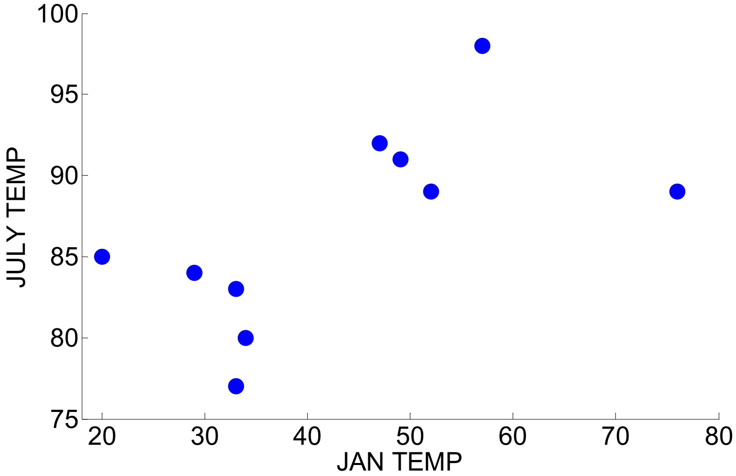
Do you think these two variables are associated? If I told you that the average daily high temperature in January of an unknown city was 50 degrees, would this give any information about the average daily high temperature of the city in July? I think the answer should be "yes." If the average high temp in January is 50 degrees, I think that it is a city in a warm climate (at least, warmer than Ohio), and so I would also expect the average high temperature in July also to be high.

Scatterplot

How can we study the relationship between Jan temp and July temp?

We begin by constructing a graph called a scatterplot. This is a generalization of the dotplot where you are plotting points along two axes.

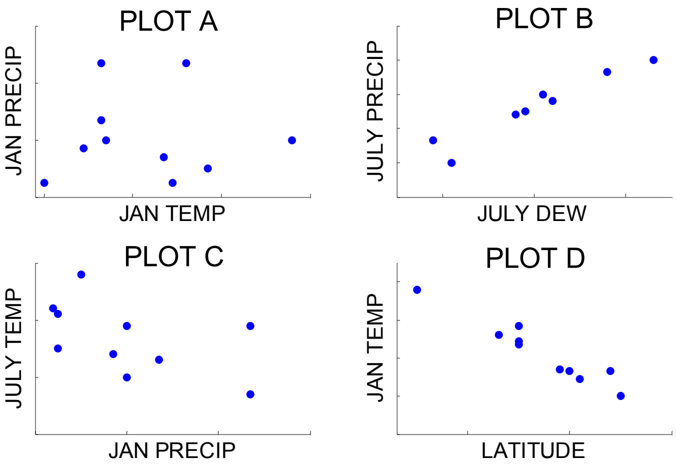
First we construct a Cartesian grid, where the values of one variable (here Jan temp) are along one axis and the values of the second variable (July temp) are along the second axis. Then for each city we place a dot corresponding to the values of the two variables. If we do this plotting for all 10 cities, we get the following scatterplot.



Scatterplot Patterns

We detect association between a pair of variables by finding a pattern in this scatterplot. What type of patterns are we looking for?

In the figure below, we show four scatterplots labeled PLOT A, PLOT B, PLOT C, PLOT D.



1. First, look at PLOT A that graphs the January temperature against the January precipitation. As we look at the points, we don't detect any general drift in the point (either upward or downward) as you scan the points from left to right. The conclusion is

that there is at most a weak relationship between a city's January temperature and its January precipitation. If we are told a city's temperature in January, this gives us little information about that city's precipitation that month.

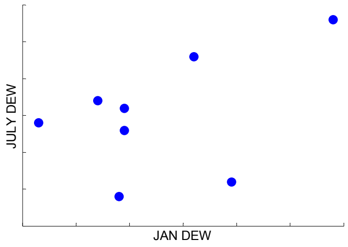
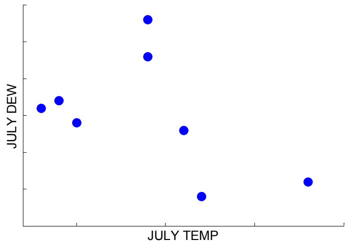
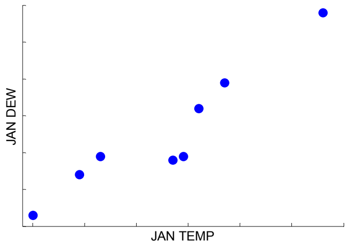
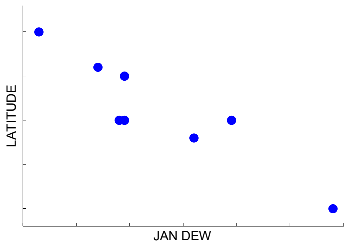
2. In PLOT B we see the points drift from the lower-left to the upper-right sections of the plot. This indicates that the two variables July dew and July precipitation are *positively associated*. Cities with small dew points in July tend to be associated with small July precipitation values, and large values of July dew and large values of the July precipitation go together. If you know anything about weather, this makes sense – cities with high humidity tend to have more rain.

3. PLOT C illustrates *negative association* between the two variables. The points tend to drift from the upper-left to the lower-right sections of the plot. This means that large values of January precipitation tend to be associated with small values of July temperature. Likewise small values of January precipitation go together with large values of the July temperature.

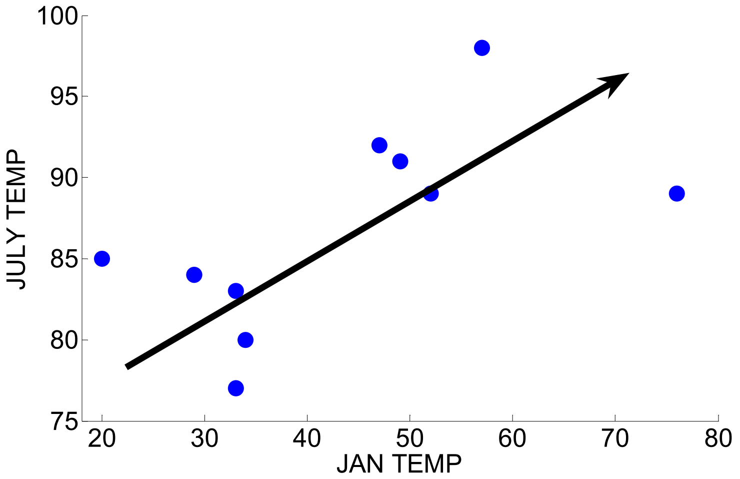
4. In PLOT D, we see that the variables latitude and January temperature are also negatively associated since the graph has the same downward drift pattern as PLOT C. But note that the points in PLOT D are more clustered about a line than the points in PLOT C. This tells us that latitude and January temperature of cities have a stronger association than January precipitation and July temperature.

Generally, when we look at a scatterplot, we identify both the *direction* and the *strength* of the association. Using our weather data, we construct four scatterplots in the figure below that illustrate different types of direction and strength of relationships. All four graphs illustrate some association between the weather variables. The left two graphs illustrate positive relationships where small (large) values of one variable are associated with small (large) values of the second variable. The right two graphs demonstrate negative relationships where small (large) values of the first variable are associated with large (small) values of the second variable. The bottom two graphs demonstrate stronger relationships where the points tend to follow a line.

	DIRECTION	
STRENGTH	Positive Association	Negative Association

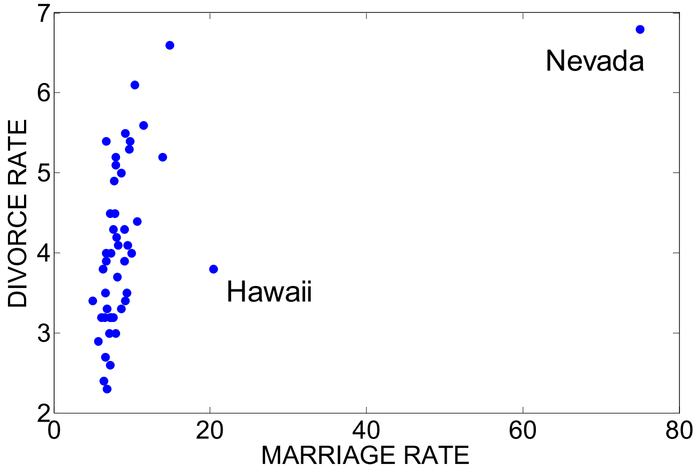
Weak		
Strong		

Let's return to our example where we constructed a scatterplot of the average January high daily temperature and the average July high daily temperature for a group of cities. What type of association do we see in this plot?



I hope you agree that the points have a positive drift, indicating that the January high temperature and the July high temperature are positively associated. This makes sense. Cities that are unusually cold in January (like Aberdeen and Chicago) tend also to be cooler in July; likewise, warm January temperatures (like those in Bakersfield and Miami) tend to be associated with warm July temperatures.

When we were describing a distribution of a single batch in topic D2, we were interested in the general shape of the data and also in observations that were far away from the general pattern. Similarly, when we look at scatterplots, we may observe outlying points that don't agree with the general pattern. For example, suppose we are interested in the relationship between a state's marriage rate (the number of marriage per 1000 people) and its divorce rate (the number of divorces for each 1000 people). We collect the marriage and divorce rates for all 50 states from 2001 data and construct a scatterplot shown below.



Generally there is a strong positive association in this graph – states with high marriage rates also tend to have high divorce rates. But there are two points corresponding to the states Hawaii and Nevada that don't follow the general pattern. Since these points seem special, it is helpful to label them in the scatterplot. Nevada is a very popular place for couples to visit to acquire a quick wedding certificate and Hawaii is a popular place to get married due to its nice climate. For both states, we observe a much higher marriage rate than one would expect based on the data from the remaining states. In practice, it is important to identify both the general relationship between two quantitative variables and the particular observations like Hawaii and Nevada that don't follow the general relationship pattern.

PRACTICE: INTERPRETING SCATTERPLOTS

A number of measurements were made on 38 1978-79 model automobiles. For each car, the variables measured were

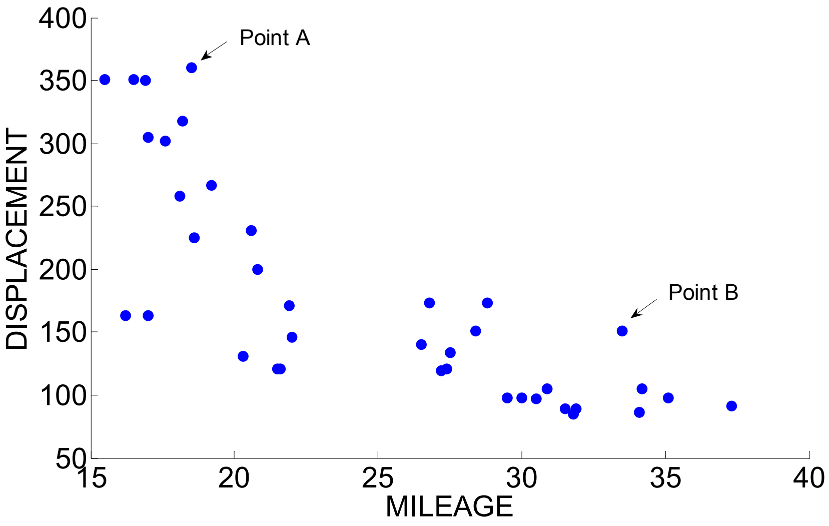
- the mileage in miles per gallon (MPG)
- the weight in thousands of pounds
- the drive ratio
- the horsepower
- the displacement of the car in cubic inches

A table of this data is shown below.

Car	MPG	Weight	Drive_Ratio	Horsepower	Displacement
Buick_Estate_Wagon	16.9	4.36	2.73	155	350
Ford_Country_Squire_Wagon	15.5	4.054	2.26	142	351
Chevy_Malibu_Wagon	19.2	3.605	2.56	125	267
Chrysler_LeBaron_Wagon	18.5	3.94	2.45	150	360
Chevette	30	2.155	3.7	68	98
Toyota_Corolla	27.5	2.56	3.05	95	134
Datsun_510	27.2	2.3	3.54	97	119
Dodge_Omni	30.9	2.23	3.37	75	105
Audi_5000	20.3	2.83	3.9	103	131
Volvo_240_GL	17	3.14	3.5	125	163
Saab_99_GLE	21.6	2.795	3.77	115	121
Peugeot_694_SL	16.2	3.41	3.58	133	163
Buick_Century_Special	20.6	3.38	2.73	105	231
Mercury_Zephyr	20.8	3.07	3.08	85	200
Dodge_Aspen	18.6	3.62	2.71	110	225
AMC_Concord_D/L	18.1	3.41	2.73	120	258
Chevy_Caprice_Classic	17	3.84	2.41	130	305
Ford_LTD	17.6	3.725	2.26	129	302
Mercury_Grand_Marquis	16.5	3.955	2.26	138	351
Dodge_St_Regis	18.2	3.83	2.45	135	318
Ford_Mustang_4	26.5	2.585	3.08	88	140

Ford_Mustang_Ghia	21.9	2.91	3.08	109	171
Mazda_GLC	34.1	1.975	3.73	65	86
Dodge_Colt	35.1	1.915	2.97	80	98
AMC_Spirit	27.4	2.67	3.08	80	121
VW_Scirocco	31.5	1.99	3.78	71	89
Honda_Accord_LX	29.5	2.135	3.05	68	98
Buick_Skylark	28.4	2.67	2.53	90	151
Chevy_Citation	28.8	2.595	2.69	115	173
Olds_Omega	26.8	2.7	2.84	115	173
Pontiac_Phoenix	33.5	2.556	2.69	90	151
Plymouth_Horizon	34.2	2.2	3.37	70	105
Datsun_210	31.8	2.02	3.7	65	85
Fiat_Strada	37.3	2.13	3.1	69	91
VW_Dasher	30.5	2.19	3.7	78	97
Datsun_810	22	2.815	3.7	97	146
BMW_320i	21.5	2.6	3.64	110	121
VW_Rabbit	31.9	1.925	3.78	71	89

A scatterplot of DISPLACEMENT and MILEAGE is shown below.

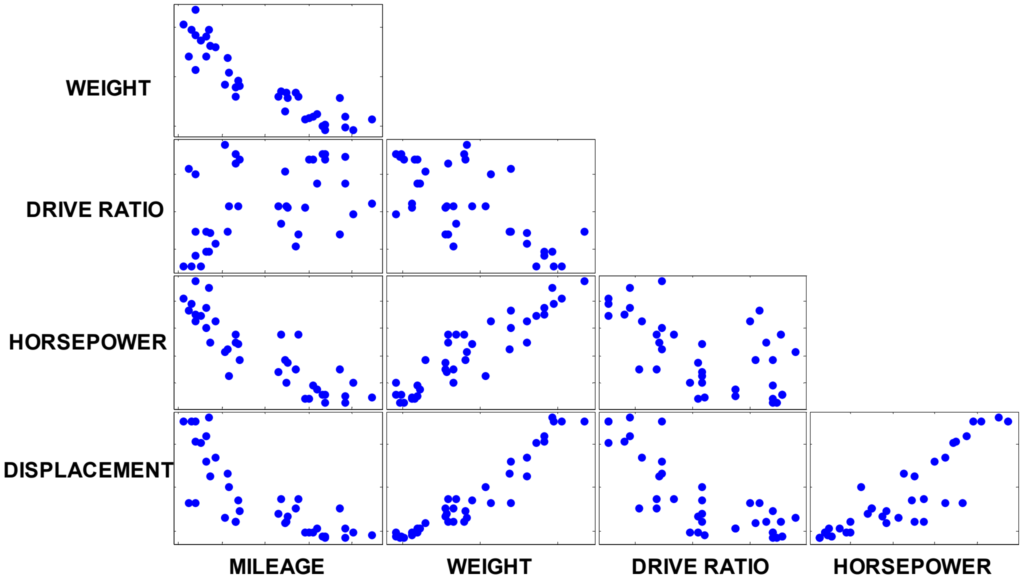


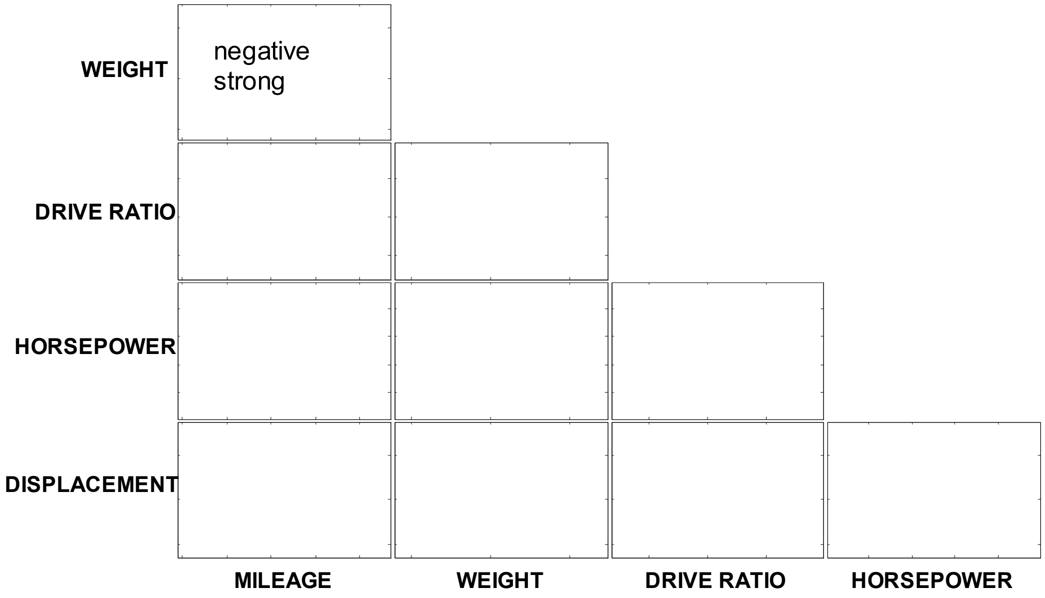
1. Identify the cars corresponding to Point A and Point B in the scatterplot.

2. What is the general pattern of the scatterplot? Is there a relationship between the displacement of the car and its mileage? Describe to a layman what this means?

3. Circle one point in the scatterplot corresponding to a car that has a small displacement and a small mileage. Is this car unusual with respect to the general pattern in the scatterplot that you described in question 2?

4. The following figure displays scatterplots of all 10 possible pairs of variables among MILEAGE, WEIGHT, DRIVE RATIO, HORSEPOWER and DISPLACEMENT. For each scatterplot, describe the general pattern (negative, positive, or little) and if the pattern indicates a positive or negative association, state if the association is strong or weak. Place your answers in the empty boxes below. (The first box has been completed for you.)



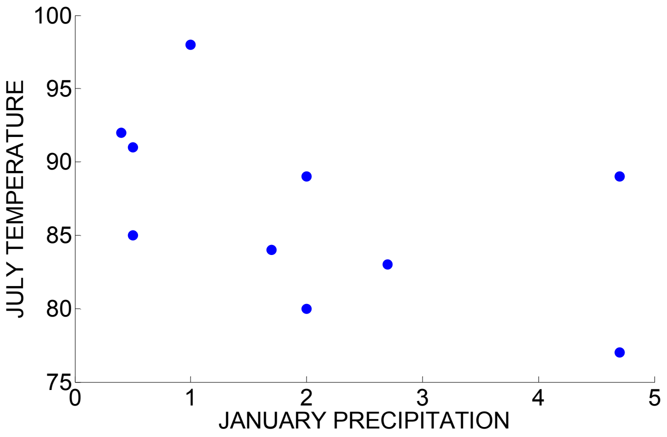


INTRODUCTION – LOOKING AT WEATHER DATA

Let's return to our weather dataset that describes the climate of different cities in the United States. We focus on the average precipitation in January (labeled "Jan precip") and the average high temperature in July (labeled "July temp") for 10 cities.

City	Jan precip	July temp
Aberdeen	0.5	85
Akron	2.7	83
Albuquerque	0.4	92
Amarillo	0.5	91
Aspen	2	80
Atlanta	4.7	89
Bakersfield	1	98
Bar Harbor	4.7	77
Chicago	1.7	84
Miami	2	89

We learned in Topic D5 that a good first step in exploring the relationship between a city's January precipitation and its July temperature is to construct a scatterplot.



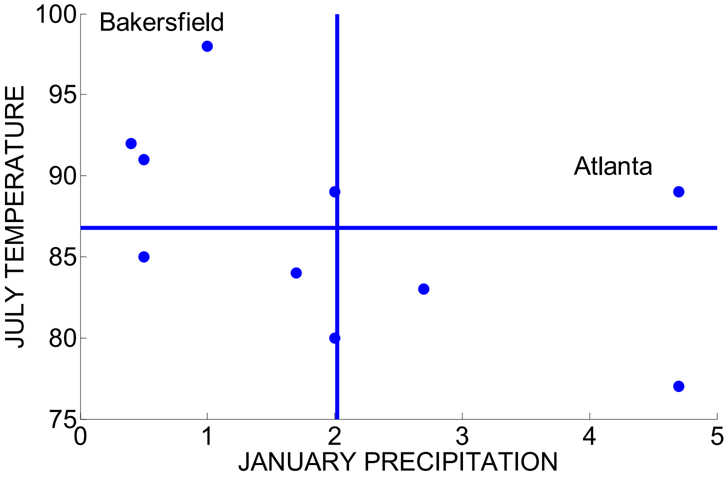
We see a negative trend in this display. This means that cities that have low precipitation in January tend to have high temperatures in July. Also, high precipitation in January and low July temperatures tend to go together.

We can summarize the relationship that we see in this scatterplot in a couple of ways.

1. First, we want to describe the *strength* of the association that we see in the plot by using a number called a *correlation coefficient*. We will describe how to compute and interpret this measure of association.
2. Second, a useful way of describing the positive association is by fitting a line to the plot. We will describe two ways of fitting a "best line" to the points called a *least-squares line* and a *median-median line*, and discuss how we can use these lines to predict the value of one variable knowing the value of the second variable. We'll talk more about best fitting lines in the second half of this topic.

A SIMPLE CORRELATION FORMULA – THE QCR

To motivate a simple formula for measuring a relationship, suppose that we divide the scatterplot into four regions by drawing horizontal and vertical lines at the respective means of the two variables. The mean of the January precipitation values is 2.02 inches and we draw a vertical line at this value. Also, we draw a horizontal line at the value 86.8 degrees (the mean July temperature value).



We have labeled two points that correspond to the January precipitation and July temperature for the cities Atlanta and Bakersfield. Atlanta has an above-average precipitation in January and an above-average temperature in July and therefore this point is to the right of the vertical line and above the horizontal line. Bakersfield, in contrast, is left of the vertical line and above the horizontal line, which means, respectively, that this city has a below-average January precipitation and an above-average July temperature.

Note that most of the points fall in the upper left and lower right sections of the plot. In these regions, the cities either have above-average precipitations and below-average temperatures, or below-average precipitations and above-average temperatures. A simple of measure of association finds the number of points in the upper right and lower left sections (quadrants I and III), subtracts the number of points in the upper left and lower right sections (quadrants II and IV), and divides the result by the number of points (n). This simple measure, called the Quadrant Count Ratio or QCR for short, is defined as

$$QCR = \frac{(\# \; of \;\; points \; in \; Quadrants \; I \;\; and \;\; III) - (\# \; of \;\; points \; in \; Quadrants \; II \;\; and \;\; IV)}{n}.$$

In our example, we count 1 point in Quadrant I, 4 points in Quadrant II, 3 points in Quadrant III, and 2 points in Quadrant IV. Also there are n = 10 points in our graph. So the Quadrant Count Ratio is given by

$$QCR = \frac{(1+3)-(4+2)}{10} = -0.2.$$

The QCR has some attractive properties as a measure of association. If all of the points are in quadrants I and III, then QCR = +1, and if all of the points are in quadrants II and IV, then QCR = -1. If most of the points fall in quadrants I and III, the measure of association will be positive. Similarly, if the points generally fall in quadrants II and IV, then the association will be negative.

In this example, the QCR is negative, reflecting a negative association between a city's January precipitation and its July temperature.

THE CORRELATION COEFFICIENT

A second, more traditional way of measuring the relationship between two quantitative variables is by means of a correlation coefficient. As in the QCR, we motivate this measure of relationship by dividing the scatterplot into four regions by drawing horizontal and vertical lines at the means of the two variables. For a correlation, we use more information that simply the number of points in the four regions.

We compute a correlation coefficient in two steps:

STEP 1: The correlation can be shown to only depend on the variables though their standardized scores introduced in topic D4. So we first find two standardized scores or z-scores corresponding to the two variables for each city. For each January precipitation value, we standardize it by subtracting its mean and dividing by its standard deviation. Likewise, for each July temperature value, we standardize by subtracting its mean and dividing by its standard deviation.

The formulas for the two standardized scores, z_x and z_y are respectively

$$z_x = \frac{January\ precip - mean(January\ precip)}{std\ dev(January\ precip)}$$

$$z_y = \frac{July\ temp - mean(July\ temp)}{std\ dev(July\ temp)}$$

We found the standardized scores for all cities and placed them in the below table. Let's check the calculations for Atlanta. The mean January precipitation (across cities) was 2.02 inches and the standard deviation was 1.61 inches; for the July temperatures the mean and standard deviation are 86.8 and 6.21 degrees, respectively. Atlanta's January precipitation was 4.7 inches, so its standardized score was

$$z_x = \frac{4.7 - 2.02}{1.61} = 1.66\,.$$

-- this means that Atlanta's precipitation in January was approximately 1.66 standard deviations above the mean. Atlanta's July temperature was 89 degrees and the corresponding standardized score is

$$z_y = \frac{89 - 86.8}{6.21} = 0.35.$$

Likewise, Atlanta's July temperature was about a third of a standard deviation higher than the mean.

STEP 2. After we find the two standardized scores for each city, we take the products of the standardized scores – these products are placed in the "Product" column of the table.

City	Jan precip	July temp	z_x	z_y	Product
Aberdeen	0.5	85	-0.94	-0.29	0.27
Akron	2.7	83	0.42	-0.61	-0.26
Albuquerque	0.4	92	-1.01	0.84	-0.85
Amarillo	0.5	91	-0.94	0.68	-0.64
Aspen	2	80	-0.01	-1.1	0.01
Atlanta	4.7	89	1.66	0.35	0.58
Bakersfield	1	98	-0.63	1.8	-1.13
Bar Harbor	4.7	77	1.66	-1.58	-2.62
Chicago	1.7	84	-0.2	-0.45	0.09
Miami	2	89	-0.01	0.35	0.00
SUM					-4.55

The correlation coefficient, denoted by r, is computed by dividing the sum of the products by the number of cases minus one. Or if you like formulas, r is given by

$$r = \frac{sum(z_x \times z_y)}{(number\ of\ cases) - 1}.$$

Here the correlation coefficient is given by

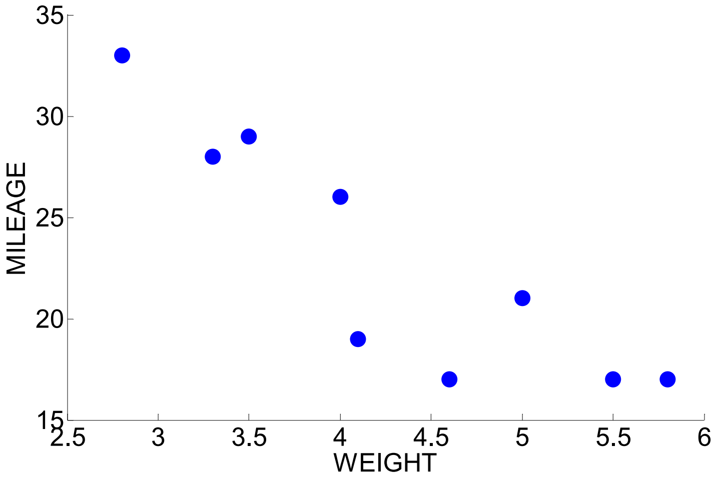
$$r = \frac{-4.55}{10-1} = -0.51$$

We'll discuss how to interpret this value shortly – we'll see that the negative value of r indicates that there is a negative association between the January precipitation and the July temperature.

PRACTICE: COMPUTING THE CORRELATION COEFFICIENT

The author collected the weight (in thousands of pounds) and the highway mileage for nine 2004 cars manufactured by Chevrolet. A table of the data is shown below together with a scatterplot of the two variables.

		Highway			
Car	Weight	MPG	z_x	z_y	Product
Avalanche	5.8	17	1.48	-0.98	-1.45
Cavalier	2.8	33			
Corvette	3.3	28	-0.97	0.82	-0.8
Impala	3.5	29	-0.77	0.98	-0.75
S-10	4.1	19	-0.19	-0.66	0.13
Suburban	4.6	17			
Tahoe	5.5	17	1.19	-0.98	-1.17
TrailBlazer	5	21	0.7	-0.33	-0.23
Venture	4	26			



1. The mean and standard deviation of the car weights are given by 4.29 and 1.02 thousand pounds, respectively. The mean and standard deviation of the mileages are 23 and 6.1 mpg. Using these values, fill in the missing standardized scores in the table.

2. Draw vertical and horizontal lines on the scatterplot corresponding to the mean weight and the mean mileage of the cars. From looking at the scatterplot, compute the value of the QCR. What does this say about the relationship between car weight and car mileage?

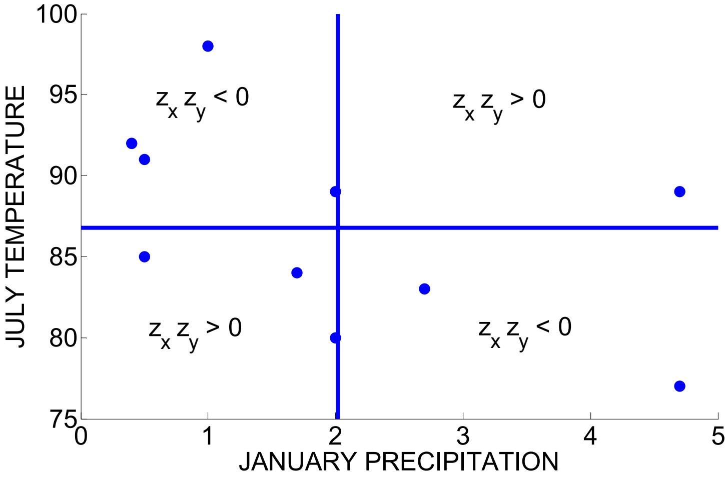
3. Compute the value of the correlation.

4. In this example, the pattern in the scatterplot was _____________ and the sign of the correlation was ________________ .

INTERPRETING THE CORRELATION COEFFICIENT

Okay, we compute the correlation (or more precisely, we have a computer compute a correlation for you). What does it mean? Here are some basic facts about the correlation coefficient r.

THE SIGN OF R: Recall that the numerator of r is the sum of the products of the standardized scores $z_x \times z_y$. This product will be positive in the upper right and lower left quadrants of the scatterplot, and the product will be negative in the two other quadrants. (See the below figure.) When the data primarily fall in the upper left and lower right sections, we will be adding up a lot of negative products, and r will be negative. This happened in our example. Similarly, when the data fall in the lower left and upper right sections, there will be many positive products of standardized scores, and r will be positive. So the sign of the correlation is informative about the general positive or negative pattern in the scatterplot.

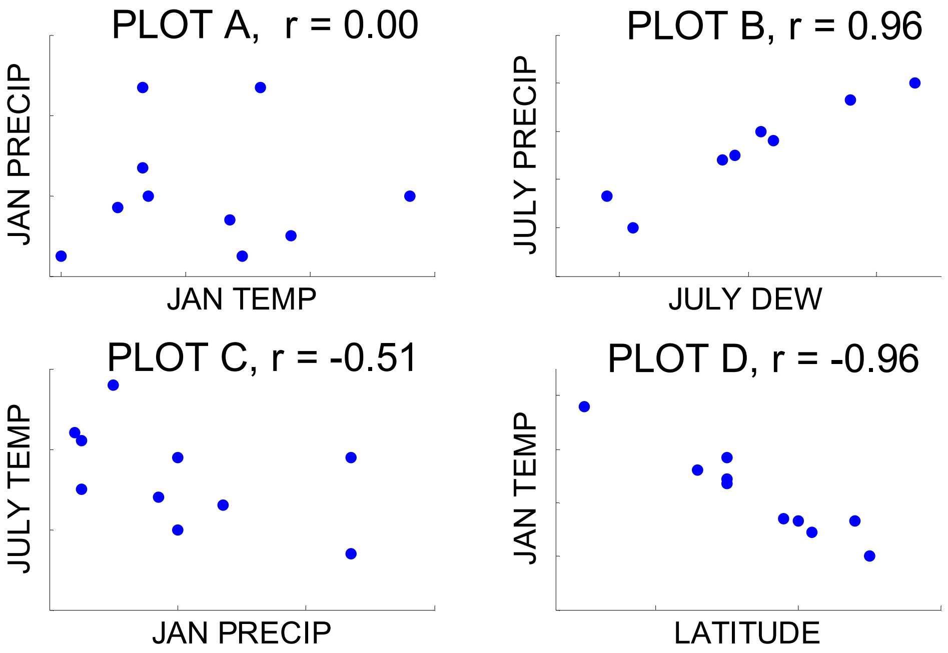


THE SIZE OF R: The value of the correlation coefficient will range from -1 to +1. If r is close to +1, then the points will fall closely to a line with positive slope. Likewise, a correlation value r close to -1 corresponds to a scatterplot where the points fall close to a line with negative slope. A correlation value close to zero means that there is little straight-line association in the graph.

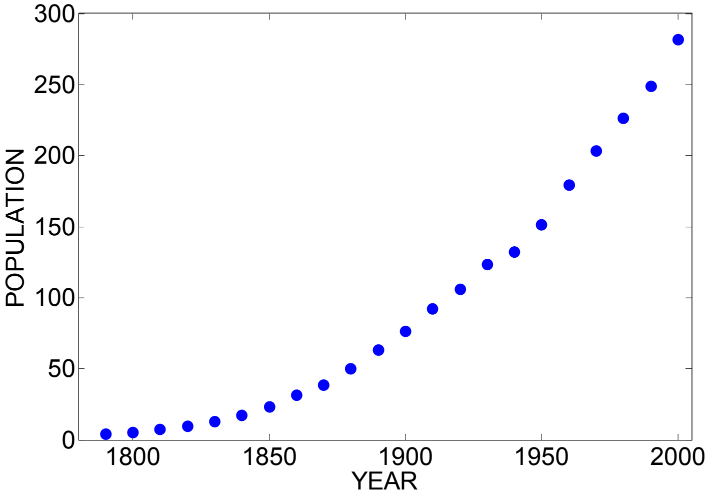
To illustrate the interpretation of r, let us revisit the weather example where we considered scatterplots of four pairs of the variables.

- In Plot A where there was little association between the January precipitation and the January temperature, the value of r is close to zero.

- In Plot B, there is a strong positive association, and r is close to one.
- Plots C and D both show negative associations between the corresponding variables and the values of r for both datasets are negative. But note that Plot D shows a stronger relationship than Plot C which is reflected in the corresponding values of the correlation. The variables latitude and January temperature have an association close to a straight line; the correlation value is r = -0.96.



To emphasize the second point, the correlation coefficient r measures the strength of a *straight-line relationship* between the two variables. It is possible that there is a strong relationship between two variables, but the relationship is not linear and so r is not the best measure of this association. For example, suppose we record the population of the United States for the years 1790, 1800, ..., 2000. Here is a scatterplot of the population against year:

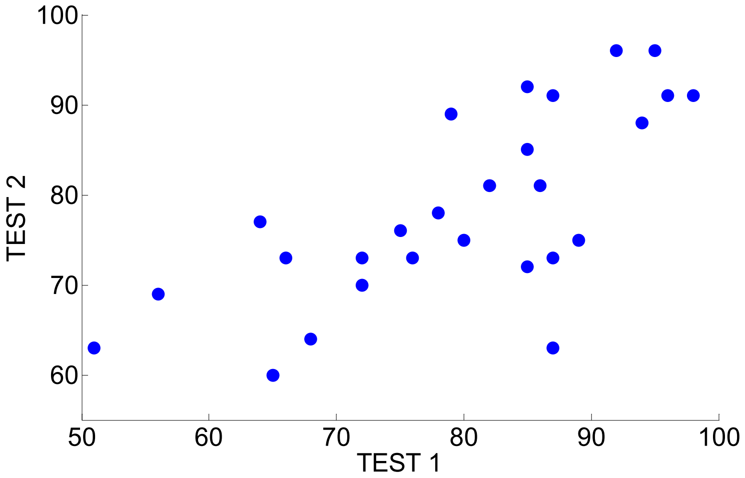


Is there a strong relationship between year and population? Yes, there is, but it isn't a straight-line relationship. In the early years, the population of the U.S. increased exponentially and this accounts for the significant curvature in the graph. The correlation coefficient for these early years is 0.959. This example illustrates several points:

- Although r measures the strength of a linear relationship, a high value of r need not mean that the variables are linearly associated.
- It is not sufficient to compute a correlation coefficient – one should graph the data *and* compute a correlation coefficient to understand the relationship between two variables.

PRACTICE – INTERPRETING CORRELATION

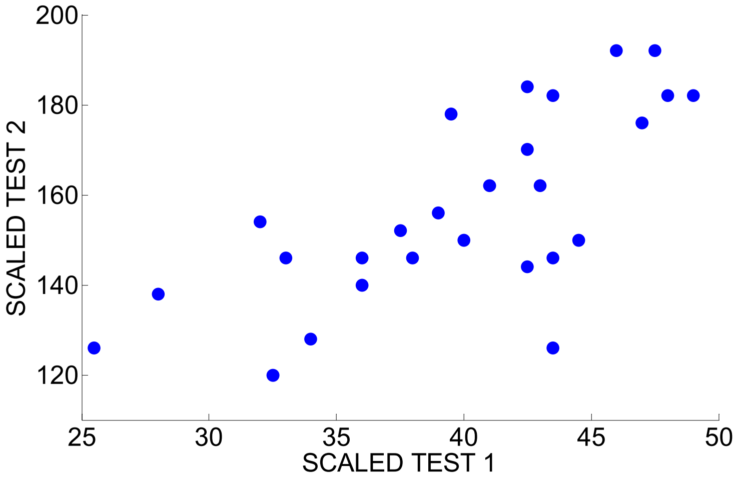
For a recent class, an instructor collected the student grades for two tests, each test scored on a scale from 0 to 100. A scatterplot of the TEST 1 score and the TEST 2 score is displayed below.



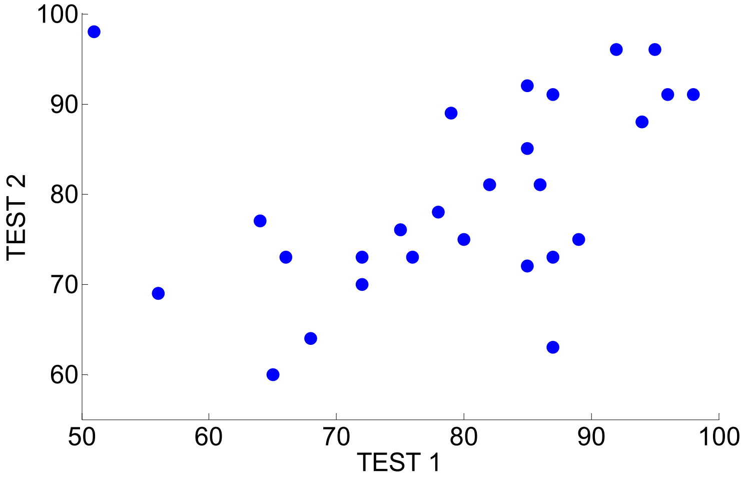
1. From looking at the pattern in the scatterplot, select the value of the correlation coefficient from the list of values below.

-.91 -.71 -.21 0 +.21 +.71 +.91

2. Suppose that the instructor decides on changing the scale of the two tests to give the tests different weights in the determination of a final grade. She decides on dividing the TEST 1 score by 2, so that the new TEST 1 range is 0 to 50, and multiplying the TEST 2 score by 2, so that the new TEST 2 range is 0 to 200. A scatterplot of the scaled TEST 1 and scaled TEST 2 scores is shown below. How would this rescaling affect the value of the correlation coefficient? Explain why there would be a change or no change in the value in r. (Hint: How does this rescaling affect the standardized scores of TEST1 and TEST2?)



3. Return to the original test scores TEST 1 and TEST 2. Suppose that there was a mistake in scoring TEST 2 for one student – his grade of 63 should have been 95. A scatterplot of the adjusted TEST 1 and TEST 2 scores is shown below. How will this change affect the value of the correlation r?

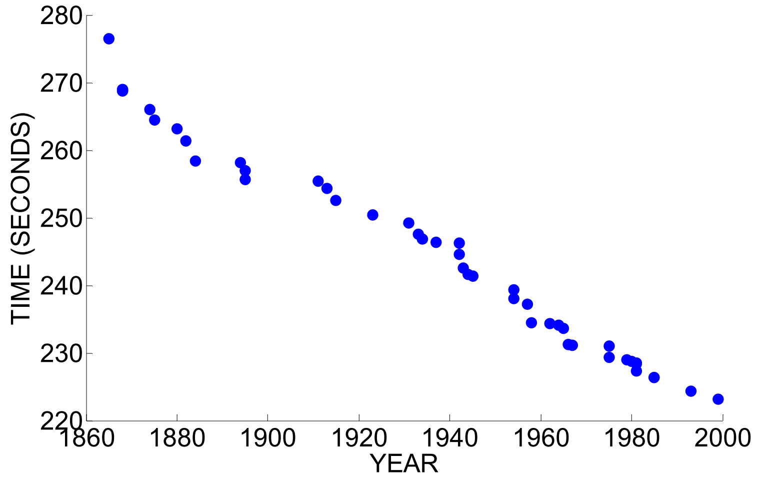


4. It turns out that the correlation value for the adjusted test scores is r = .41. (The correlation for the original test scores was r = .71.) What does this say about the sensitivity of the correlation coefficient to unusual values?

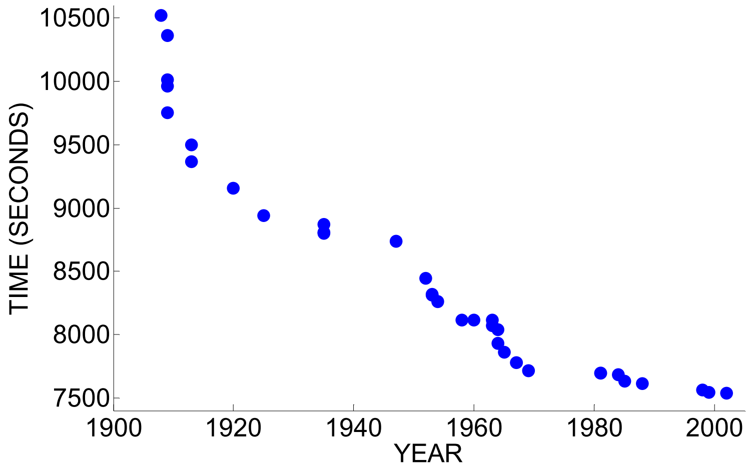
5. It is interesting to look at the pattern of record times for different athletic events. The first scatterplot below graphs the world record time in the men's mile race against the

year in which the record was set. The second scatterplot graphs the world record time in the men's marathon race and the year.

WORLD RECORD TIME FOR THE MILE



WORLD RECORD TIME FOR THE MARATHON



Which graph displays a stronger straight-line relationship? The correlation values for the two scatterplots are –0.95 and –0.99. Which value is associated with which scatterplot?

TECHNOLOGY ACTIVITY – GUESSING CORRELATIONS

DESCRIPTION: Here we will get some experience looking at scatterplots and guessing at the value of the correlation. Even if you know little initially about scatterplot patterns and the corresponding correlation values, you'll start understanding this connection when you see enough scatterplots.

PART 1: Guess the Correlation Game

For this part, you only need to know that a correlation is simply a measure of the association pattern in the scatterplot.

1. Open up the Fathom document "guess_correlation.ftm."
2. Select the scatterplot. Hit the Apple-Y key to simulate a new scatterplot.
3. Guess at the value of the correlation – put your guess in the "Guess at Correlation" column of the table.
4. Now scroll right to see the actual value of the correlation. Put this in the "Actual value of Correlation" column. Also compute your error – the distance between your guess and the actual value.
5. Scroll back (left) so you can see the scatterplot. Repeat steps 2, 3, 4 – keep going until you have done 20 guesses at the correlations for 20 scatterplots. When you are done, compute the TOTAL ERROR in your 20 guesses.

Trial	Guess at Correlation	Actual value of Correlation	Error
1			
2			
3			
4			
5			
6			
7			

8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
TOTAL			

PART 2: Using Fathom to analyze your correlation data.

To see how you did in this guessing activity,

- Open up a new Fathom document.
- Open up a new Collection and new Case Table.
- Create two variables, called GUESS and ACTUAL.
- Put the data from the table above into the Case Table.
- Construct a scatterplot of the GUESS, ACTUAL data.
- Compute the correlation between GUESS AND ACTUAL.
- Interpret what the value of this correlation is telling you.

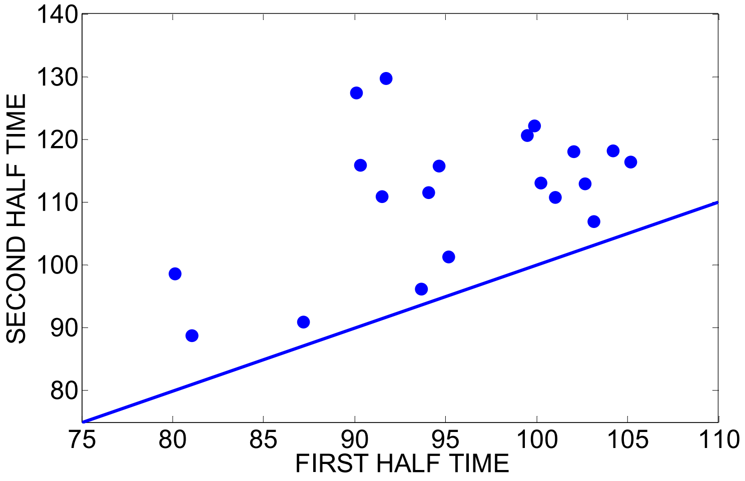
WRAP-UP

In this topic, we were introduced to different strategies for understanding the relationships between two quantitative variables. In studying this relationship, typically there is a *response variable* that is of main interest and an *explanatory variable* that we believe is helpful in understanding the variation in the response variable. In this topic, we were introduced to the *direction* and *strength* of association in a scatterplot. The QCR and the correlation coefficient are statistics that measures the strength of the relationship between the variables. The correlation only measures the straight-line relationship, so one should also construct a scatterplot to check this straight-line assumption.

EXERCISES

1. **Boston Marathon Running Times**

A random selection of runners from the 2003 Boston Marathon was selected. For each runner, the time (in minutes) to run the first half of the race and the time to run the second half were recorded. The below figure displays a scatterplot of the first-half time and the second-half time.



a. Describe the pattern of association in the scatterplot. Explain why you see a relationship between these two variables.

b. The line FIRST HALF TIME = SECOND HALF TIME is drawn on the figure. Note that all of the points fall above this line. Can you explain why?

c. Circle one point (label it "S") where the runner got tired and took much longer to complete the second half than the first half.

d. Circle a second point (label it "F") where the runner took about the same time to complete both the first and second halves of the race.

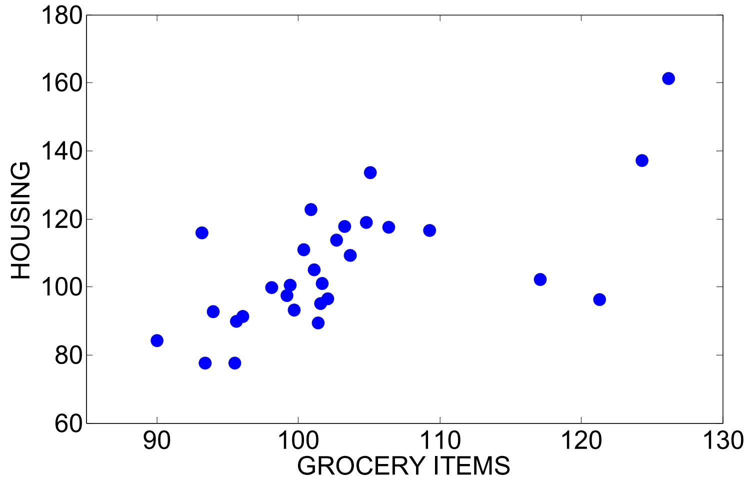
e. Circle the point corresponding to the runner that had the fastest total time in this race and circle the point corresponding to the runner with the slowest total time.

2. **Cost of Living Indices**

The ACCRA Cost of Living Index (www.costofliving.org) is a quarterly report that compares basic living expenses in various U.S. cities. The index measures the price level for consumer goods and services relative to the average, and a value of 100 represents an average price level. In a 2000 report, the index values for a number of American cities are listed. The table below lists the index values for 29 cities for the categories grocery items and housing.

City	**Grocery items**	**Housing**	**City**	**Grocery items**	**Housing**
Anchorage, Alaska	124.3	137.1	Manchester, N.H.	104.8	119.0
Phoenix, Ariz.	101.7	100.9	Albuquerque, N.M.	102.7	113.8
Sacramento, Calif.	121.3	96.2	Charlotte, N.C.	98.1	99.8
San Diego, Calif.	126.2	161.3	Cincinnati, Ohio	99.2	97.5
Colorado Springs, Colo.	103.3	117.7	Cleveland, Ohio	109.3	116.5
Jacksonville, Fla.	101.4	89.5	Oklahoma City, Okla.	95.5	77.7
Atlanta, Ga.	103.7	109.2	Salem, Ore.	100.4	111.0
Springfield, Ill.	99.7	93.1	Philadelphia, Pa.	105.1	133.6
New Orleans, La.	102.1	96.5	Memphis, Tenn.	95.6	89.9
Baltimore, Md.	94.0	92.6	Austin, Tex.	93.2	115.9
Lansing, Mich.	100.9	122.8	El Paso, Tex.	93.4	77.5
Minneapolis, Minn.	101.1	105.0	San Antonio, Tex.	90.0	84.2
Billings, Mont.	99.4	100.4	Salt Lake City, Utah	106.4	117.6
Omaha, Neb.	96.1	91.2	Cheyenne, Wyo.	101.6	95.1
Las Vegas, Nev.	117.1	102.2			

The figure below displays a scatterplot of these two index values.



a. Describe the general pattern of this scatterplot.

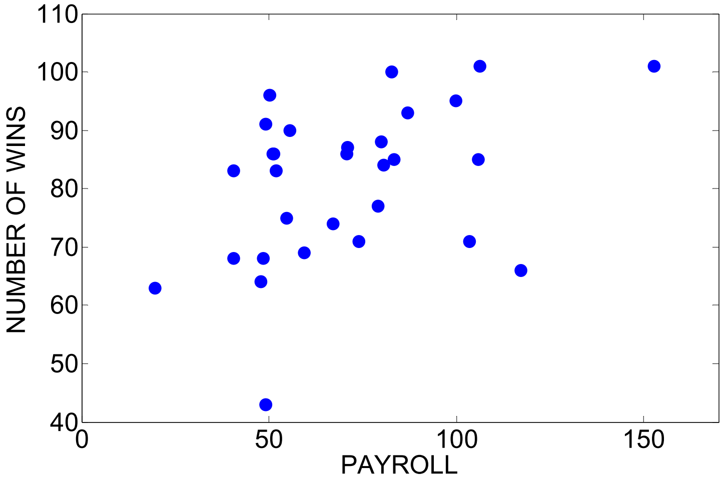
b. There are four points that seem to stand out from the main group. Identify the cities that correspond to these four points and explain how they are different from the remaining cities.

c. Circle the point corresponding to a city that has the lowest index value for grocery items and circle a second point corresponding to a city that has the lowest index value for housing.

d. Choose a city that you believe is similar in costs to the costs of your hometown. Is this city above or below average with respect to costs in this group of cities?

3. **Baseball Team's Payroll and Winning**

In professional baseball, there is currently a great variation in the amount of money that teams pay for ballplayers. That raises the question: are teams with high payrolls generally more successful in winning games than teams with low payrolls? To answer this question, the total team payroll (in millions of dollars) and the number of season wins was recorded for the 2003 baseball season. A scatterplot of payroll against number of wins is shown below.



a. Describe the general pattern in this scatterplot. Are teams with high payrolls generally more successful in winning games?

b. Circle a point (label it "A") corresponding to a team that had a high payroll but had a relatively small number of wins.

c. Circle a second point (label it "B") corresponding to a team who had a small payroll but was very successful in winning games.

d. Circle any points that seem to deviate from the general pattern in the plot. Provide how these points are outliers.

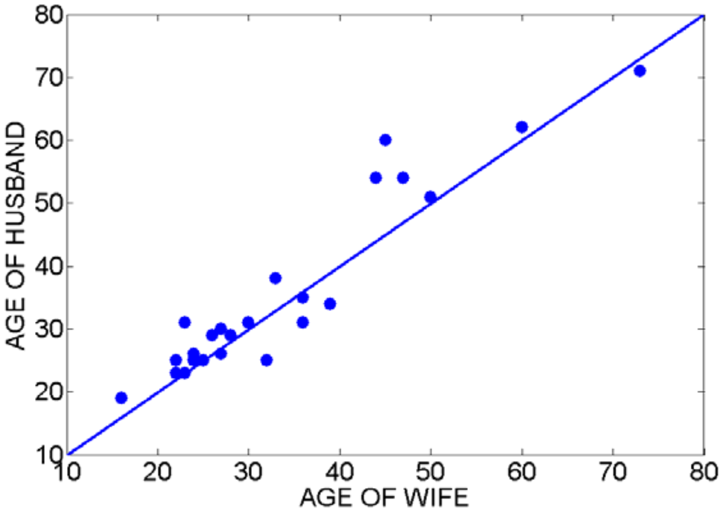
4. **Marriage Ages**

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

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

12	34	39	24	35	36

The following scatterplot displays the relationship between husband's age and wife's age. The line drawn on the scatterplot is a 45 degree-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.
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 degree line.

5. **Direction and Strength of Association**

Describe the relationship between the following pairs of variables as NEGATIVE, POSITIVE, or LITTLE. If there is a positive or negative relationship, describe the strength of the relationship (STRONG, MODERATE, or WEAK).

Pair of variables	Direction of association	Strength of association
Height and armspan		
Height and shoe size		
Height and GPA		
SAT score and college GPA		
Latitude and average January temperature of American cities		
Lifetime and weekly cigarette consumption		
Serving size and calories of fast food sandwiches		
Airfare and distance to destination		
Cost and quality rating of peanut butter brands		
Course enrollment and average student evaluation		
Number of absences and grade in a statistics class		

6. **Direction and Strength of Association**

Suppose you are a math teacher who is interested in understanding which variables are related to the student's final grade that is measured as a percentage from 0 to 100. Below is a list of variables collected on the students that may help to explain the student's final grade. For each variable, give the direction of the relationship of the variable with the student's final grade. Also rank the variables in order of most associated with final grade to least associated with final grade.

a. Student's IQ.

b. Number of absences from class.

c. Student's score on the first test

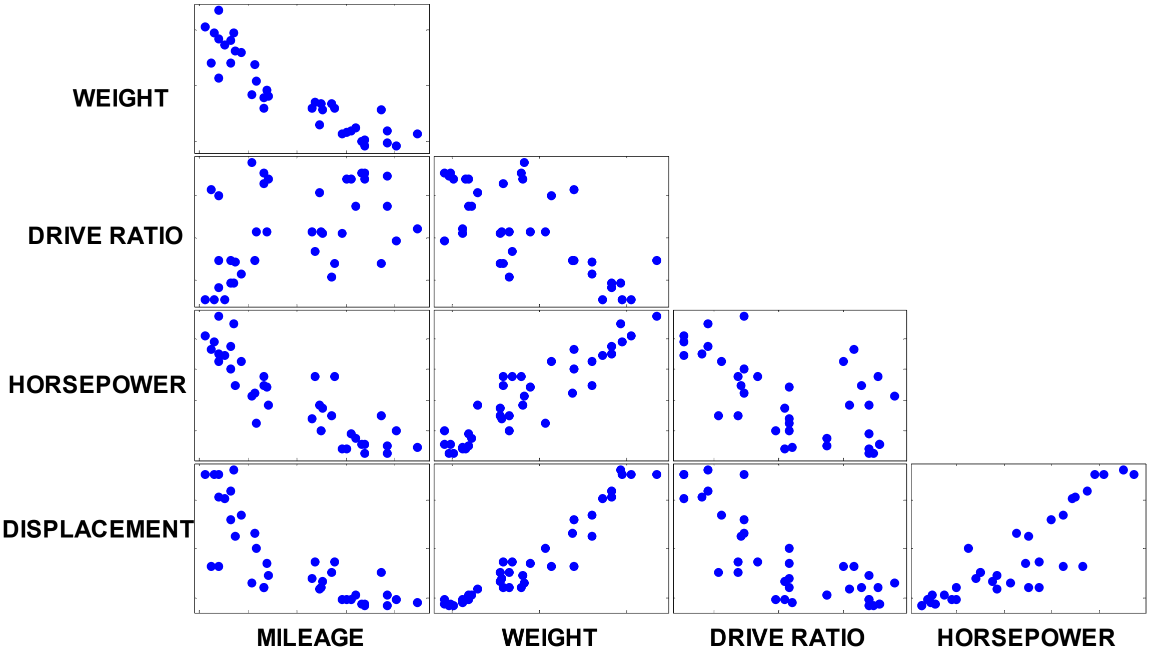
d. Student's height.

e. Number of hours that the student studies on average each week.

f. Student's grade point average.

7. **Car Measurements**

Below we have displayed a scatterplot matrix of the measurements WEIGHT, DRIVE RATIO, HORSEPOWER, DISPLACEMENT, and MILEAGE that were made on 38 1978-79 model automobiles.



Using the list of correlation values

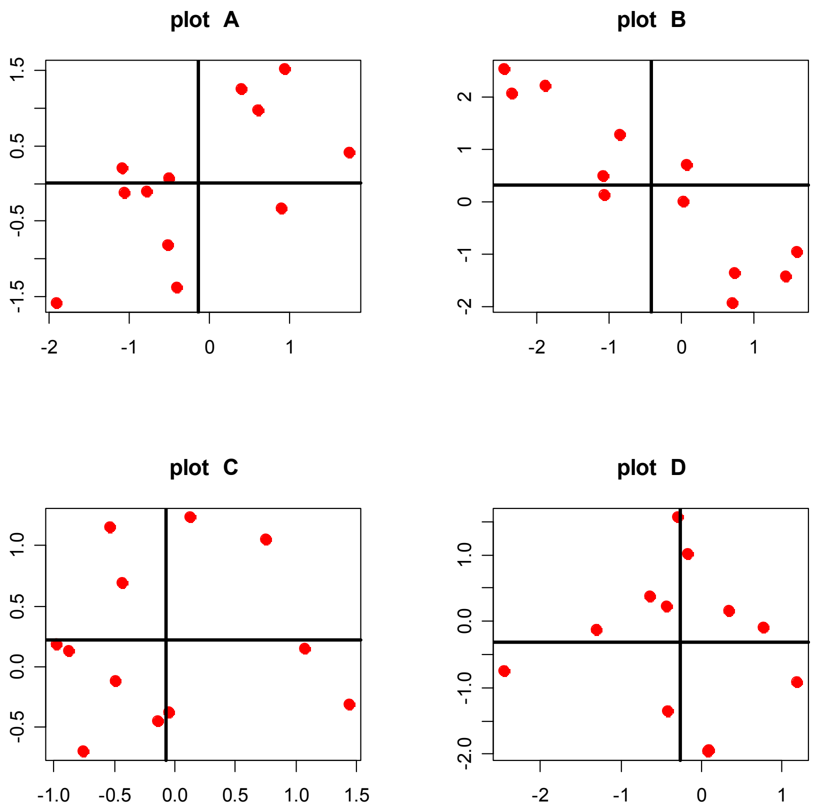
0.42, -0.69, -0.79, -0.90, 0.95 -0.80, 0.87, -0.87, 0.92, -0.59

write down the correlation above each scatterplot in the matrix.

8. **Computing Values of the QCR**

For each of the following scatterplots, (1) describe the direction and strength of the relationship between the two variables and (2) compute the value of the Quadrant

Count Ratio (QCR). The horizontal and vertical lines are drawn at the means of the two variables.



Scatterplot A:

Scatterplot B:

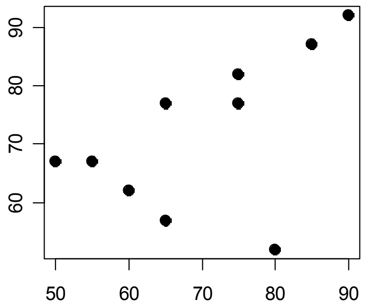
Scatterplot C:

Scatterplot D:

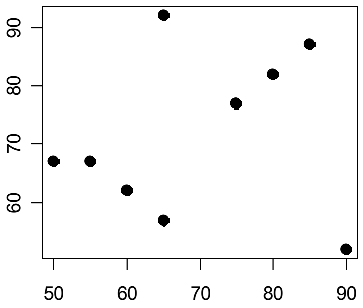
9. **Two Measures of Association**

The following display shows four scatterplots of hypothetical test scores.

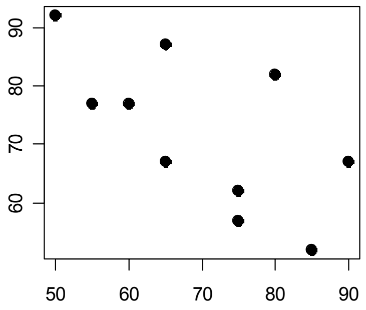
PLOT A



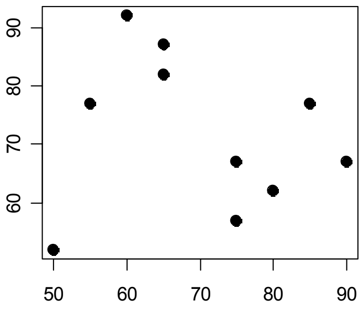
PLOT B



PLOT C



PLOT D



(a) List the four scatterplots in the order of most negatively associated to most positively associated. (Which plot is most negatively associated? Next, which plot is next most negatively associated, and so on.)

(b) For each plot, compute the value of the QCR. (The mean value of the horizontal variable is 70 and the mean value of the vertical variable is 72.)

(c) Do the values of the QCR follow the same ordering as your ordering in part a?

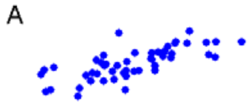
(d) The correlation values of Plots A, B, C, and D are respectively 0.53, 0.14, -0.64, and -0.14. Do these correlation values following the same ordering as your ordering in part b?

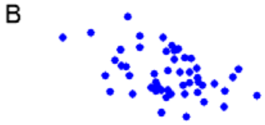
(e) Explain why the QCR and the correlation lead to different orderings of the association in the four scatterplots.

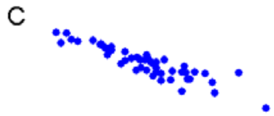
10. **Estimating Correlations**

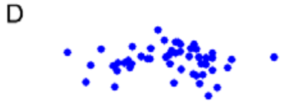
Six scatterplots are shown in the figure below. Using the list of possible correlation values, find the correlation for each scatterplot.

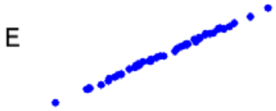
Scatterplot	Correlation	Scatterplot	Correlation
A		D	
B		E	
C		F	

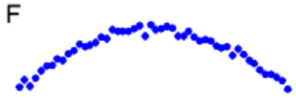












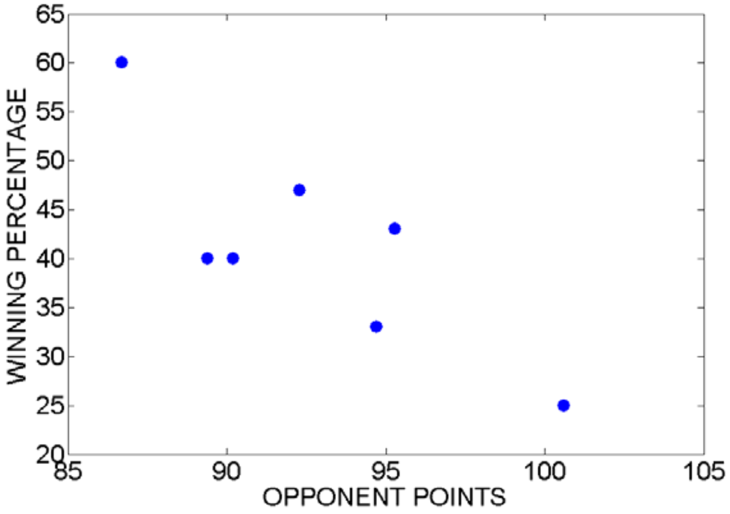
Possible correlation values: -0.99, -0.91, -0.73, -0.43, -0.14, 0.00, 0.43, 0.73, 0.91, 0.99

11. **Points Scored and Winning for Basketball Teams**

The table below gives the average number of points scored by an opponent and the number of wins and losses for seven professional basketball teams. Suppose you are interested in computing the correlation between a team's winning percentage (Win_Pct) and the points scored by the opponent (Opp). In basketball, for a team to be successful in winning, it is important to play good defense to keep the opponent from scoring many points. A scatterplot of these two variables is shown below.

Team	Opp	Wins	Losses	Win_Pct	ZX	ZY	product
Miami	89.4	21	32	40	-0.72	-0.10	0.07

Washington	94.7	16	33	33	0.43	-0.74	-0.32
Orlando	100.6	13	40	25			
New Jersey	86.7	30	20	60			
Philadelphia	90.2	21	31	40	-0.54	-0.10	0.05
New York	92.3	25	28	47	-0.09	0.54	-0.05
Boston	95.3	23	30	43	0.57	0.17	0.10



a. From the scatterplot, estimate the value of the correlation.

b. In the table, some of the standardized scores (for both opponent points and winning percentage) have been computed. Complete this table and use the results to compute the correlation. (The mean and standard deviation of opponent points are 92.7 and 4.6; the mean and standard deviation of winning percentage are 41.1 and 11.0.)

12. **Computing Correlations**

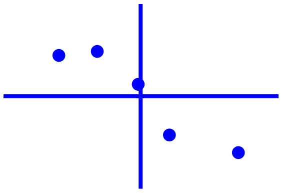
The table shows values of the standardized scores for the x and y variables for four small hypothetical datasets. Compute the correlation for each dataset. When you are completed, write the values of the correlations on the top of the corresponding scatterplots in the figure.

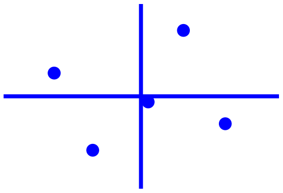
Dataset A	
ZX	ZY
-.63	-1.56
-1.20	.18
-.14	-.26
.67	1.08
1.30	.56

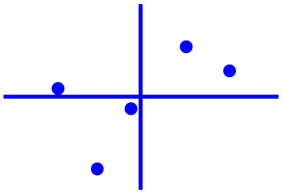
Dataset B	
ZX	ZY
-.63	.96
-1.19	.88
-.03	.25
.42	-.85
1.42	-1.23

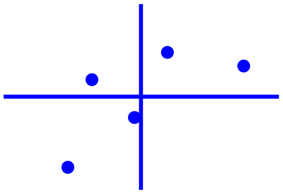
Dataset C	
ZX	ZY
-.70	-1.18
-1.26	.49
.11	-.13
.62	1.41
1.23	-.60

Dataset D	
ZX	ZY
-.71	.37
-1.06	-1.53
-.09	-.45
.38	.95
1.49	.66







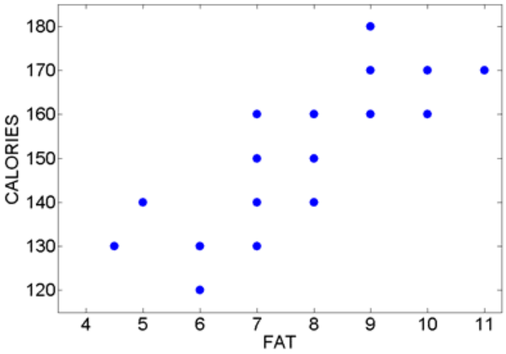


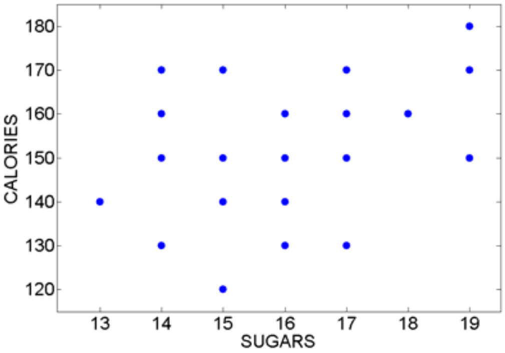
13. **Calories in Ice Cream**

The following table gives the calories, grams of fat, and grams of sugars in a ½ cup serving of each of the flavors of Breyer's ice cream.

Flavor	calories	fat	sugars	Flavor	calories	fat	sugars
carmel fudge	160	7	18	banana fudge chunk	170	9	19
Vanilla	140	8	15	vanilla fudge brownie	160	9	16
french vanilla	150	8	15	cherry chocolate chip	150	8	17
van/choc/straw	140	8	15	peanut butter & fudge	170	10	15
butter pecan	170	11	14	dulce de Leche	150	7	19
Chocolate	150	8	16	lactose free vanilla	160	9	17
mint chocolate chip	160	9	17	mocha almond fudge	170	9	15
strawberry	120	6	15	butter almond	160	10	14
rocky road	150	8	19	calcium rich vanilla	130	7	14
cookies & cream	160	8	16	carmel praline crunch	180	9	19
vanilla fudge twirl	140	7	15	fresh banana	140	5	16
Peach	130	6	16	homemade vanilla	140	7	13
Coffee	140	8	15	extra creamy vanilla	150	8	14
cherry vanilla	140	8	16	extra creamy chocolate	140	7	15
Chocolate chip	160	9	17	take two	150	8	15
Chocolate chip cookie dough	170	9	17	take two (sherbet)	130	4.5	17
vanilla & choc fudge checks	170	9	17				

The left scatterplot below graphs the fat content against the calories and the right scatterplot graphs sugar content and calories.





a. Describe the pattern of association in each graph.

b. Which graph has a stronger association?

c. Estimate the correlation in each graph.

d. Suppose someone tells you that it would be easy to manufacture a “low calorie” ice cream if they could eliminate the fat content. Based on the graphs above, can you agree with this statement? Why or why not?

14. **Body Measurements**

There was a recent study (Mohanty, Babu, and Nair, *Journal of Orthopedic Surgery*, 2001, 19-23) that investigated the relationship between different body parameters. The objective was to find particular body parameter measurements that correlate best with height. Sitting height, standing height, arm span, and leg lengths for 505 healthy women from South India between the ages of 20 and 29 were measured. The table presents correlations that were computed from these data.

Physical measurements	Correlation
Sitting height and arm span	0.561
Standing height and arm span	0.816
Sitting height and leg length	0.294
Standing height and leg length	0.842

a. Based on the table, describe the relationship between a person’s sitting height and one’s arm span.

b. Based on the table, what is the best predictor of a woman’s standing height? Explain.

c. Note that there is a relatively weak association between a woman’s sitting height and her leg length. Can you explain why there is a relatively weak association?

d. Think of another physical measurement that you would believe would have a weak relationship with standing height. Explain why the relationship would be weak.

TOPIC D7: RELATIONSHIPS - SUMMARIZING BY A LINE

SPOTLIGHT: MEASURING A CAR

Every fall, Consumers Reports publishes a magazine *New Car Preview* that reviews the new cars that are available for a particular model year. The intent of this publication is to present information about the new cars so you can make an informed decision when you are in the market for buying a car. Specifically, Consumer Reports presents a profile of a particular car model that contains much information and test data about the car. But actually how does this consumer organization measure a car?

One section of measurements about a car is labeled "Specifications." This section contains physical measurements, such as its length (inches), width (inches) and curb weight (pounds). This section also contains important fuel measurements such as its estimated mileage (in miles per gallon) in city and highway driving, and the number of gallons in its fuel tank. Another section of the profile is labeled "From the Test Track." This section presents measurements about the car that were made when Consumer Reports took the car on a test drive. On this test drive, measurements were made on the seating dimensions of the car such as the inches of rear leg room and the inches of front head room. Also in the test drive, Consumer Reports made several measurements of the acceleration and the distance required to brake on a dry surface. Measurements were made on the size of the cargo area in the trunk (in cubic feet) and the maximum load (in pounds) that the car can carry including passengers and cargo. To check the fuel economy, the mileage of the car was measured both in city and highway driving, and the annual fuel cost was estimated.

Above we have focused on quantitative car measurements. But the profile also presents many categorical measurements about the car. These measurements include ratings on the car's reliability with respect to the engine, cooling, fuel, transmission, etc. Also Consumer Reports shows the results on the car's crash tests (from good to bad) and rates the convenience and comfort of the car such as the ease in using the controls, the front-seat comfort, and the noise level.

Are all of these car measurements helpful when you make your decision on what car to purchase? Probably not. But there will likely be certain measurements that will be important to you when you compare car models. In the case of our family, we wanted to make sure that the car would fit into our garage, so the length of the car was an important measurement. Also the fuel economy measurements were important due to the current high price of gasoline.

Later in this topic, we will explore the relationships between several different car measurements as reported in this magazine.

PREVIEW

In the last topic, we were introduced to the scatterplot that graphically shows the association between two quantitative variables. In this topic we describe ways of summarizing this association. We first introduce a correlation coefficient that measures the direction and strength of the association that we see in the scatterplot. When there is an association, it is helpful to fit a line to the points and we'll describe two methods for fitting this line.

In this topic your learning objectives are to:

- Understand how one can use a line to describe a relationship and predict values of one variable given values of a second variable.
- Understand two methods of fitting a line, and when one method may be preferable to the second method when there are outliers in the data.
- Understand the computation and interpretation of residuals.
- Understand the limitations of a best fitting line in understanding a relationship.
- Understand three ways of studying the relationship between two quantitative variables.
- Understand that the best way of studying the relationship can depend on the particular dataset or way we communicate the relationship.

NCTM Standards

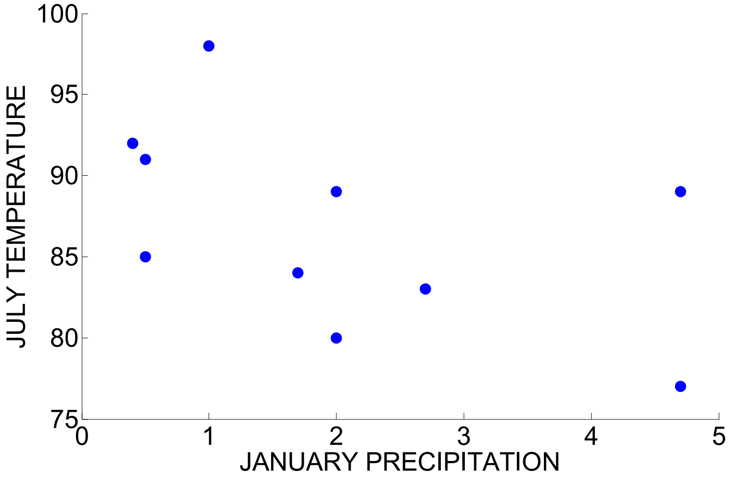
✓ In Grades 6-8, all students should make conjectures about possible relationships between two characteristics of a sample on the basis of scatterplots of the data and approximate lines of fit.

✓In Grades 9-12, all students should, for bivariate measurement data, be able to determine regression coefficients and regression equations using technological tools.

✓In Grades 9-12, all students should identify trends in bivariate data and find functions that model the data.

RELATIONSHIPS - SUMMARIZING BY A LEAST-SQUARES LINE

Again we look at the relationship average amount of precipitation (inches of rain or melted snow) in January and the average high temperature July for ten cities.



In the first part of this topic, we talked about measuring the strength of the association between two measurement variables by means of a correlation coefficient. Suppose we are primarily interested in one variable, called the *response variable* and we wish to use a second variable, called the *explanatory variable*, to predict values of the response variable. When two variables have a strong association, it is not enough to just state a correlation value – we like to know *how* the response variable changes as we change the explanatory variable. A simple way to describe this relationship is to fit a "best line" to the scatterplot, and use this line to make our predictions.

Predicting a city's average July temperature

Let's suppose that you are given the average July temperatures for the ten cities and you are interested in predicting the average temperature in July for another city, say Metropolis. Can you predict Metropolis' average July temperature? If you were not given any information about this city, then it would be reasonable to assume that Metropolis is similar to the other ten cities with respect to temperature, and so you can predict Metropolis' July temperature by using the mean July temperature for the ten cities. This mean temperature turns out to be 86.8 degrees.

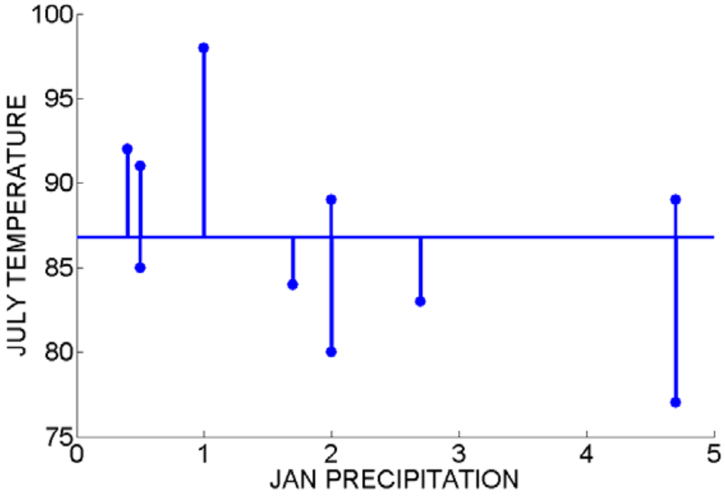
Is this a good prediction? Let's use this prediction, 86.8, to estimate the July temperature for the 10 cities. In the below scatterplot, a horizontal line is drawn at the value 86.8 degrees. This line

$$predicted\ \ July\ temperature = 86.2$$

represents our predicted July temperature. Of course, the cities' actual July temperatures are either smaller or larger than 86.8 degrees. We have drawn vertical lines from the actual July temperatures (the dots) to the predicted values (the horizontal line). We define a residual to be the difference between the actual and predicted temperature:

$$RESIDUAL = Observed\ temperature - Predicted\ temperature\,.$$

The lengths of these lines represent our errors in using 86.8 to predict the July temperatures for the 10 cities. One way of measuring the accuracy of our prediction is by the sum of the squared residuals.



We show how to compute this measure in the below table. We show the July temperature for each city, the predicted value, the residual, and the square of the residual.

	January	July		Squared
City	temp	temp	Residual	Residual
Aberdeen	20	85	-1.8	3.24
Akron	33	83	-3.8	14.44
Albuquerque	47	92	5.2	27.04
Amarillo	49	91	4.2	17.64
Aspen	34	80	-6.8	46.24
Atlanta	52	89	2.2	4.84
Bakersfield	57	98	11.2	125.44
Bar Harbor	33	77	-9.8	96.04
Chicago	29	84	-2.8	7.84
Miami	76	89	2.2	4.84
			SUM	347.6

We see from the table that the sum of the squared residuals is 347.6. This sum represents the total size of the error in using the single value, 86.8, to predict the July temperature for the ten cities.

Using January precipitation to predict July temperature

How can we get a better prediction of Metropolis' July temperature? We know from the previous discussion that a city's average January precipitation is negatively associated with a city's July temperature. So maybe if we knew Metropolis' January precipitation, then we could use this information to obtain a better prediction of the city's July temperature.

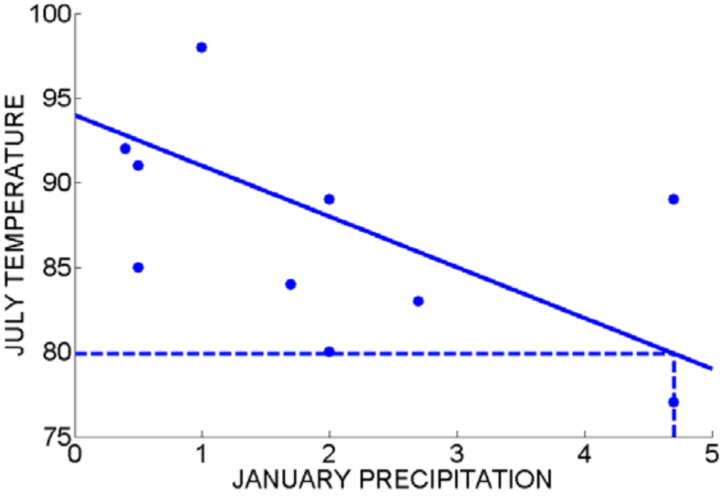
We can find a "good" predictor by fitting a line to our scatterplot. Below we have redrawn the scatterplot of the (Jan precipitation, July temperature) data. Suppose we draw the line

$$predicted\ July\ temperature = -3 \times (January\ precipitation) + 94$$

on top of the graph. (I found this line by trial and error. I wanted to find a line with a simple formula that went through the middle of the cluster of points in the scatterplot.) This formula gives us one possible prediction for a city's July temperature if we know its January precipitation. For example, Atlanta's average January precipitation is 4.7 inches. Using this line formula, we would predict Atlanta's average July temperature to be

$$-3 \times 4.7 + 94 = 79.9.$$

So we would predict Atlanta's July temperature to be 79.9 degrees. This prediction is shown graphically on the figure.



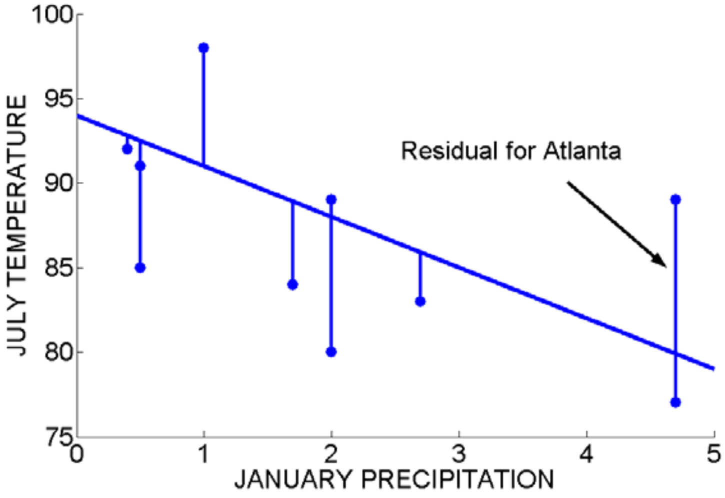
But how good is our prediction? Suppose we use this line formula

$$predicted\ July\ temperature = -3 \times (January\ precipitation) + 94$$

to predict the July temperature for all ten cities. For each city, we can compute a corresponding residual. For example, for Atlanta, our prediction was 79.9 degrees and Atlanta's actual July temperature is 89 degrees. The corresponding residual is

RESIDUAL = Actual July temperature - Predicted July temperature = 89 – 79.9 = 9.1.

On the scatterplot, the residuals are represented by the vertical lines from the observed points to the line. We have pointed out the residual for Atlanta in the figure. For this city, the large positive residual indicates that Atlanta’s actual July temperature is far above the line that represents its predicted temperature.



In the table below, we compute the predicted July temperatures for all 10 cities. As in the earlier table, we compute the residuals and the squared residuals.

	Jan	July			Squared
City	precip	temp	predicted	Residual	residual
Aberdeen	0.5	85	92.5	-7.5	56.25
Akron	2.7	83	85.9	-2.9	8.41
Albuquerque	0.4	92	92.8	-0.8	0.64
Amarillo	0.5	91	92.5	-1.5	2.25
Aspen	2	80	88	-8	64
Atlanta	4.7	89	79.9	9.1	82.81
Bakersfield	1	98	91	7	49
Bar Harbor	4.7	77	79.9	-2.9	8.41
Chicago	1.7	84	88.9	-4.9	24.01
Miami	2	89	88	1	1
				SUM	296.78

As before, we can measure the accuracy of this line prediction by the sum of squared residuals. Here this sum is 296.78. Recall that the sum of squared residuals for the first line

$$predicted\ July\ temperature = 86.2$$

was found to be 347.6. So this line prediction is better (has a smaller total error) than the use of the single value, 86.8, to predict the July temperatures.

We just drew one line (found by trial and error) to predict the July temperature from the January precipitation. Is this the best line prediction we can find in the sense of smallest sum of squared residuals?

Actually no. It turns out that one can mathematically find the line that makes the sum of squared residuals as small as possible. This line is called the *least-squares line* since it minimizes (makes "least") the squares of the residuals.

There is a relatively simple formula for this line. The least-squares line has the equation

$$\hat{y} = a + b\,x$$

($\hat{y}$ denotes predicted value of y) where the slope *b* and the intercept *a* are given by

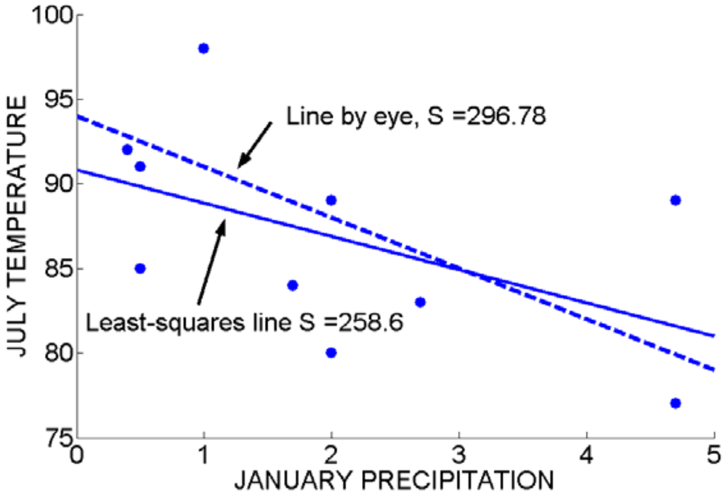
$$b = r\frac{s_y}{s_x},\ \ a = \overline{y} - b\,\overline{x}\ ,$$

where r is the correlation coefficient, $\overline{x}, s_x$ are the mean and standard deviation of the x (horizontal) variable, and $\overline{y}, s_y$ are the mean and standard deviation of the y (vertical) variable. The formula for the intercept *a* reflects the fact that the least-squares line passes through the point ($\overline{x}, \overline{y}$).

In practice, we don't find the least-squares line by hand – instead, we use a computer program. In the below figure, the least-squares line

$$July\ temperature = -1.96 \times (January\ \ precipitation) + 90.8$$

has been drawn together with the line that we fit by eye. The sum of the squared residuals of this least-squares line is equal to 258.6; recall that the sum of squared residuals about our best line by eye was 296.78. As we would expect, this sum of squared residuals is smaller than the value we got from the line that we found by trial and error.



PRACTICE: WHAT IS LEAST-SQUARES?

The below table displays the average price of unleaded gasoline (in cents) in the United States for every four years from 1976 through 2004.

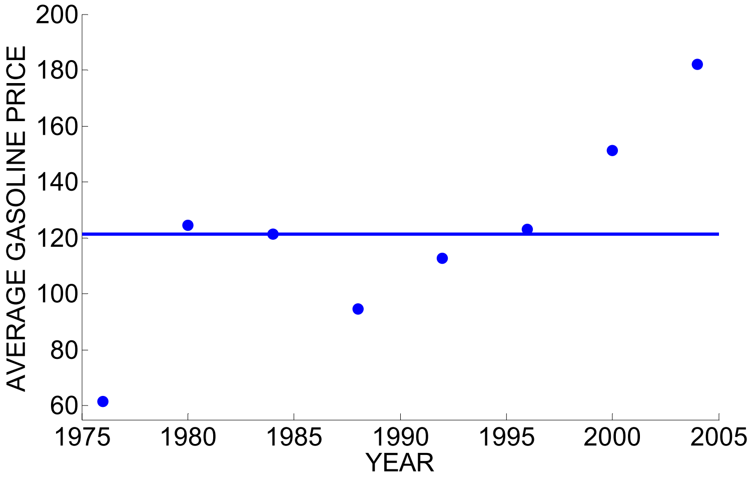
				Residual
Year	price	predicted	residual	squared
1976	61.4	121.3		
1980	124.5	121.3		
1984	121.2	121.3		
1988	94.6	121.3		
1992	112.7	121.3		
1996	123.1	121.3		

2000	151.1	121.3
2004	181.9	121.3
TOTAL		

1. Suppose we wish to predict the gasoline price using the mean value which can be computed to be 121.31 cents. The figure below shows a scatterplot of price against year with the basic prediction model

$$\textit{Average gasoline price} = 121.31$$

drawn as a horizontal line on top of the scatterplot.



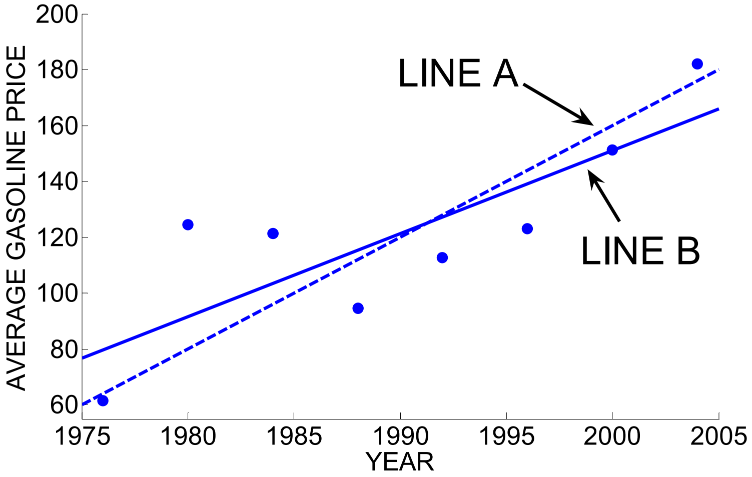
For each year in the table above, find the residual and the residual squared. Find the sum of squared residuals for this prediction of gasoline price.

2. Suppose instead that the price of gasoline is predicted using the "trial and error" formula

$$\textit{Average gasoline price} = 120 + 4(\textit{Year} - 1990).$$

(This is line A in the below figure.) In the below table, find the predicted prices for each year. Find the residuals, the squared residuals, and the sum of squared residuals.

				residual
Year	price	predicted	residual	squared
1976	61.4			
1980	124.5			
1984	121.2			
1988	94.6			
1992	112.7			
1996	123.1			
2000	151.1			
2004	181.9			
TOTAL				



3. The “least-squares” line was fit to these data – the equation of this line is

$$Average\ gasoline\ price = 121.31 + 2.98(Year - 1990).$$

which is Line B in the figure. As in part 2, find the predicted gas prices, the residuals, and the squared residuals using this line prediction. Find the sum of squared residuals.

4. To summarize your work, place the sum of squared residuals for each of the three fits in the below table. Which is the best fit, the next-best fit, and the worst fit in the sense of minimizing the sum of squared residuals? Explain why?

Fit	Sum of squared residuals
Constant value 121.31	
Line A: PRICE = 120 + 4(YEAR – 1990)	
Line B: PRICE = 121.31 + 2.98 (YEAR – 1990)	

Making appropriate and inappropriate predictions

The least-squares line allows us to make predictions about the response variable given a value of the explanatory variable. Let's return to the problem of predicting the July temperature of our hypothetical city Metropolis. Suppose we are told that this city's January precipitation is 4.0 inches. Then, using the least-squares line, we would predict the July temperature of Metropolis to average

$$\textit{July temperature} = -1.96 \times (4.0) + 90.8 = 82.96 \textit{ degrees}.$$

Suppose another city, say Emerald City, has a January precipitation of 10 inches. Using this line, we would predict the July temperature of Emerald City to average

$$\textit{July temperature} = -1.96 \times (10) + 90.8 = 71.2 \textit{ degrees}.$$

Thinking about this prediction, this seems that Emerald City is unusually cold in July. Looking back at the scatterplot, note that a January precipitation of 10 inches is off the graph – the highest precipitation value in our dataset was only 4.7 inches.

This illustrates one caution about the use of a best-line to make predictions. Based on the data, we are confident of a straight-line relationship *only* for the range of the values of the explanatory variables in the data. In this example, it only makes sense to predict July temperature for cities with January precipitation values between 0 and 4.7

inches. It would be inappropriate to use this line for values of January precipitations outside of this range.

PRACTICE: APPROPRIATE AND INAPPROPRIATE PREDICTIONS

Return to the earlier problem where we were examining the average gasoline prices of gasoline over a year. The least-squares fit was

$$\textit{Average gasoline price} = 121.31 + 2.98(\textit{Year} - 1990).$$

1. Would it be reasonable to use this line to predict the average gasoline price in 1980? Why or why not? If it is reasonable, make a prediction.

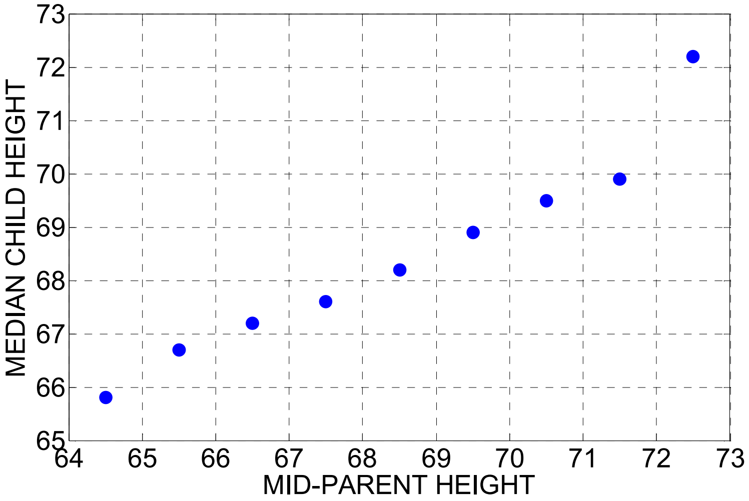
2. Would it be reasonable to use this line to predict the average gasoline price in 2050? Why or why not? If it is reasonable, make a prediction.

ACTIVITY: FITTING A LINE BY EYE TO GALTON'S DATA

DESCRIPTION: Francis Galton, in the famous 1886 paper "Regression Towards Mediocrity in Hereditary Stature," studied the degree to which children resembled their parents. For a large number of families, Galton measured the "mid-parent height" (the average of the heights of the mother and father), and the height of the child when fully grown (the "adult child height"). For each of the mid-parent heights 64.5 inches, 65.5 inches, and so on, the table below gives the median adult child height (in inches). (Galton multiplied all of the female children heights by 1.08, so that all of the children heights could be measured on the same scale.) A scatterplot of the two variables is shown to the right of the table.

MATERIALS NEEDED: A number of short pieces of spaghetti.

Mid-parent height	Median adult child height
64.5	65.8
65.5	66.7
66.5	67.2
67.5	67.6
68.5	68.2
69.5	68.9
70.5	69.5
71.5	69.9
72.5	72.2



1. Using the piece of spaghetti given to you by your instructor, fit a "good line" to the points on the scatterplot.

2. Find two points on your "good line." Write below your two points as ordered pairs.

	Mid-parent height	Median adult child height
Point 1		
Point 2		

3. Using your two points, find the slope and y-intercept of your line. (Show your work.) Write your equation of the line below.

MEDIAN CHILD HEIGHT = □ MID-PARENTS' HEIGHT + □

4. If the average of the mid-parents' heights is 66.5 inches, use your line to predict the median adult child's height.

5. From the table, the median adult child's height (when the mid-parent height is 66.5 inches) was 67.2 inches. Compute the residual.

6. To understand the difficulty of fitting lines by eye, collect for all of the students in the class the

- Computed slopes
- Computed y-intercepts

Graph the batch of slopes. Summarize this batch, including a description of the shape, typical value, and any outliers. Likewise, graph and summarize the batch of y-intercepts.

7. Can you explain why there is so much variation in the slopes and y-intercepts?

THE MEDIAN-MEDIAN LINE – A ROBUST ALTERNATIVE METHOD OF FITTING A LINE

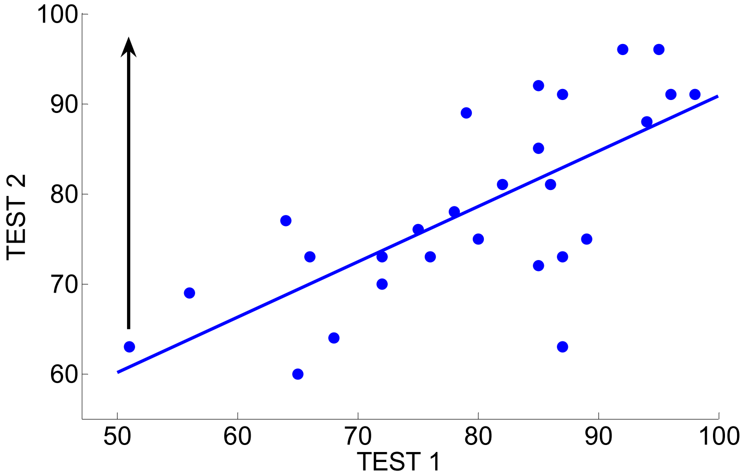
The least-squares line is the most common way of fitting a line to data. However, there is an undesirable characteristic of the least-squares method when there are outliers that are far away from the general pattern in the scatterplot. Here we demonstrate this problem and consider an alternative method of fitting a line that is less sensitive to outliers in the data.

Let's return to the dataset that contained two test scores for 27 students in a college class. (The dataset is shown below.) When we plot the Test 1 score against the Test 2 score, we see a positive association and we're interested in using a line to predict a student's Test 2 score from his Test 1 score. We fit a least-squares line that has the equation

Test 2 score = 29.44 + 0.614 Test 1 score.

Test1	Test2	Test1	Test2	Test1	Test2
82	81	95	96	78	78
79	89	98	91	89	75

92	96	87	73	85	72
96	91	80	75	75	76
72	70	76	73	56	69
86	81	72	73	87	91
66	73	94	88	51	63
87	63	64	77	68	64
85	92	65	60	85	85

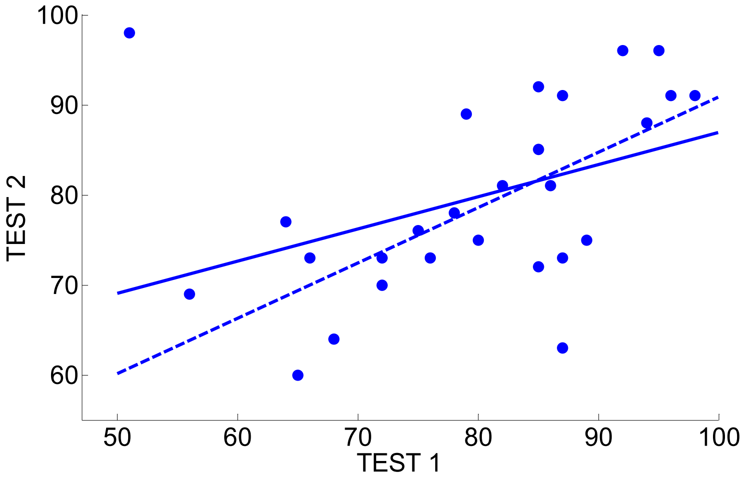


Let’s introduce an outlier in the data. There is one student who received 51 on Test 1 and 63 on Test 2 – his scores are represented by the plotting point on the left side of the figure. Suppose that this student really got a 96 on Test 2 – with this change, the point on the left side is moved in the direction of the arrow. This student is really an outlier since his point is far away from the general group of points in the scatterplot.

Does the introduction of this outlier have any effect on our least-squares line? We recompute the line to be

$$\text{Test 2 score} = 51.22 + 0.357 \text{ Test 1 score}$$

and graph this new line on the below scatterplot. Note that the least-squares fit has substantially changed – the single outlier seems to have flattened the least-squares line and the new line no longer seems to be a good fit to the general pattern of points. The problem with the least-squares line is similar to the problem of using a mean as an average of a batch of data in the presence of outliers. A least-squares line, like a mean, can be very sensitive to a few unusual data values that deviate from the main body of data.



A median-median line

There is another way of fitting a good line, the median-median line that is less sensitive to outliers. As the name suggests, this line is based on finding medians, rather than means, in regions of the scatterplot.

To compute this line, we first sort the data by the explanatory variable (here, Test 1) and divide the data into three equal-size groups (or nearly equal-size groups) by the ordered values of this variable. In this example, we have 27 points and we divide the data into the

- *Left group* – the 9 leftmost points
- *Middle group* – the 9 points in the center of the scatterplot
- *Right group* – the 9 rightmost points

The three groups of points are shown in the below table. Also in the below figure, we divide the points in the scatterplot by vertical lines, showing the left, middle, and right groups.

Left Group		Middle Group		Right Group	
Test1	Test2	Test1	Test2	Test1	Test2
51	63	76	73	87	63
56	69	78	78	87	73
64	77	79	89	87	91
65	60	80	75	89	75
66	73	82	81	92	96

	68	64	85	92	94	88
	72	70	85	72	95	96
	72	73	85	85	96	91
	75	76	86	81	98	91
Summary Point	66	70	82	81	92	91

For each group, we find a *summary point* that is the

(median of explanatory variable, median of response variable).

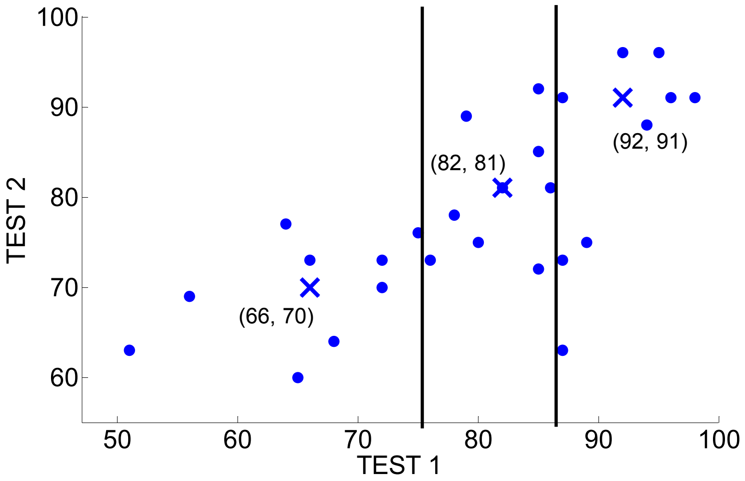
To illustrate, for the Left Group, the median of the explanatory variable Test 1, is

$$\text{median}\{51, 56, 64, 65, 66, 68, 72, 72, 75\} = 66$$

and the median of the response variable Test 2 is

$$\text{median}\{63, 69, 77, 60, 73, 64, 70, 73, 76\} = 70.$$

So the Left summary point is (66, 70). In a similar fashion, we find the summary points for the Middle and Right groups. We denote these points by (x_L, y_L), (x_M, y_M), and (x_R, y_R). They are shown in the table and plotted in the figure.



Now we can compute the median-median line:

1. The slope of the line is found by finding the slope between the Left and Right summary points.

$$b = \frac{y_R - y_L}{x_R - x_L}$$

Here the left and right summary points are (66, 70) and (92, 91) and so the slope is

$$b = \frac{91 - 70}{92 - 66} = 0.81.$$

2. Using each point, we can solve for the intercept by the usual formula $a = y - bx$. The intercept of the median-median line is the average of these three intercepts:

$$a = \frac{(y_L - bx_L) + (y_M - bx_M) + (y_R - bx_R)}{3}$$

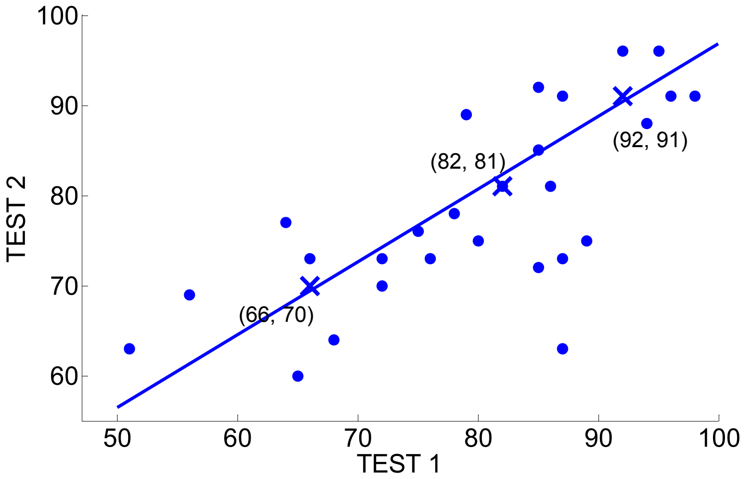
In this example, the intercept is computed to be

$$a = \frac{(70 - 0.81 \times 66) + (81 - 0.81 \times 82) + (91 - 0.81 \times 92)}{3} = 16.05.$$

Summing up, our median-median line is

Test 2 score = 16.05 + 0.81 × (Test 1 score).

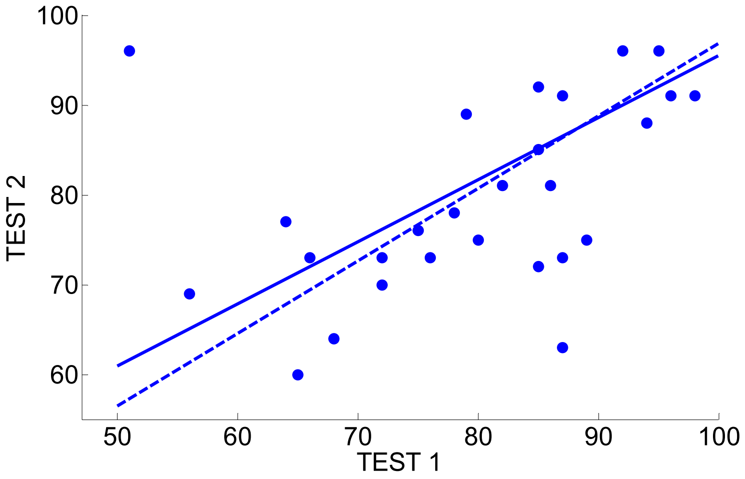
This line is displayed in the below scatterplot.



What was the point of introducing this new way of fitting a line? Remember that the least-squares line could be distorted by the introduction of a single outlying point. Suppose, as above, that we change the Test 2 grade of that one student from 63 to 96. The median-median line for this new dataset is

Test 2 score = 26.3 + 0.69 × (Test 1 score).

We've plotted the original median-median line and the new one in the below figure. Note that the median-median line has changed, but this line is much less affected by the outlier than the least-squares line. To say it a little differently, the median-median is more robust or insensitive to outliers than is the least-squares line. In practice, both lines will give similar fits with no outliers present, but the lines can give very different results when there are outliers in the data.



PRACTICE: THE MEDIAN-MEDIAN LINE

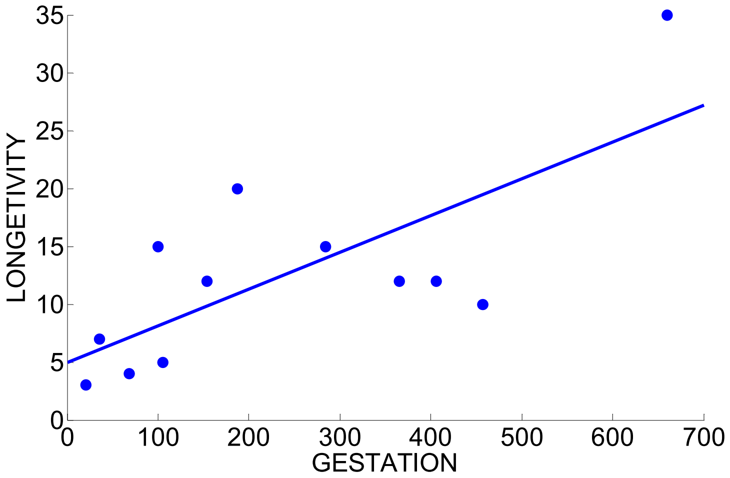
The table below gives the average gestation period (in days) and the average longevity (in years) for twelve animal species.

Animal	Gestation	Longevity	Summary points
Mouse	21	3	
Kangaroo	36	7	
Guinea pig	68	4	
Lion	100	15	
Beaver	105	5	
Sheep	154	12	
Baboon	187	20	
Cow	284	15	
Ass	365	12	
Camel	406	12	
Giraffe	457	10	
Elephant	660	35	

A scatterplot of the two variables is shown below. If we are interested in predicting an animal's longevity based on its gestation period, a least-squares fit is given by

LONGEVITY = .0318 GESTATION + 5.0.

This best line is drawn on the graph.



1. The dataset has been arranged by increasing values of the gestation period. Divide the dataset into three groups and find the summary point for each group.

2. Using the three summary points, find the equation of the median-median line.

3. Graph the median-median line on the scatterplot.

4. Are the least-squares and median-median lines similar in this example? If they are different, describe any special features of this dataset that might cause the two fits to be different.

TECHNOLOGY ACTIVITY – Fitting a "best line"

DESCRIPTION: In this Fathom activity, we will get some experience fitting lines to a scatterplot. Although there are several reasonably good straight-line fits, we will see the least-squares line is the one that is best in minimizing the sum of squared residuals.

1. Suppose you plan to fly from Detroit to St. Louis over Thanksgiving break. How much do you expect your round-trip plane fare to cost?
2. Suppose instead you decide to fly from Detroit to Seattle? How much do you think this will cost?
3. Why are your answers to 1. and 2. different?

4. Besides the fact that different plane trips cover different distances, why is there so much variation in plane fares? What other factors besides distance determine the size of a plane fare?

PART I: Fitting a line to remove the tilt in the scatterplot

Recently the author collected the lowest fares from Detroit to a number of U.S. cities. Also, I found the distance (in miles) of each city from Detroit. The data can be found in the Fathom document **airfares1.ftm**.

In this document, you will see

- A scatterplot of the distance (MILES) against the plane FARE.
- A line that passes through the point ($\bar{x}, \bar{y}$) and has slope equal to .7.
- A graph of the residuals of this particular line fit.

5. By playing with the slider, find a line that seems to "best fit" the points. (You find the line that removes the "tilt" in the residual plot.) Write the equation of your line below (put your slope in the box).

$$\text{FARE} = 283.1 + \boxed{}\,(\text{MILES} - 821.4)$$

6. For your line, find the residuals for the cities Chicago and Denver. (Look at the RESIDUAL column of the data table.) Verify (using your calculator) that these two residuals have been computed correctly.

7. Suppose you plan on flying from Detroit to Miami over the break. Predict what the airfare will be. (You can find a web site that gives the distance between cities.)

PART II: Understanding a "least-squares" line.

Open the Fathom document **airfares2.ftm**. You will see the same dataset.

8. Construct a scatterplot of MILES AND FARE.
9. Place a moveable line on the plot by selecting the scatterplot and then selecting the menu item Graph > Moveable Line.
10. By selecting the menu item Graph > Show Squares, you will see squares that correspond to the residuals from the line. You want to make the sum of the areas of the squares (the sum of squared residuals) as small as possible. Move the line and try to make the Sum of Squares as small as you can. Write down your final line and the sum of squares.

FINAL LINE: SUM OF SQUARES:

11. Now see how close your line is to the "least-squares line." (Select the menu item Graph > Least Squares Line.) Write down the least squares line and the sum of squares.

LEAST-SQUARES LINE: SUM OF SQUARES:

12. Compute the difference between the sum of squares for your line and the sum of squares of the least-squares line. This measures how well you did in part 10.

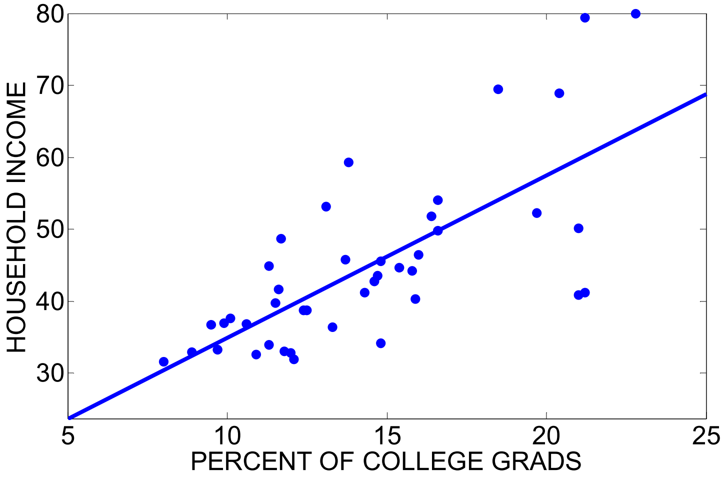
PLOTTING RESIDUALS

When we fit a line to a scatterplot relating two quantitative variables, we like to make the residuals small. The least-squares line is the one that makes the sum of squared residuals as small as possible. But the residuals, the differences between the observed and predicted responses, are important quantities in themselves. By graphing the residuals against the explanatory variable, we can see if our "best line" is a good description of the data.

In our cities dataset, the percentage of adults who completed four years of college and the average household income (in thousands of dollars) has been collected for 42 cities. Generally we think there is a relationship between these two variables: if a greater percentage of adults are college educated, one might think that this would result in a higher average household income. Our objective here is to describe this relationship using a least-squares fit and then use a plot of the residuals to look for interesting points that deviate from the straight-line pattern.

We begin with a scatterplot of the two variables shown below. As expected, we see a positive relationship between the percentage of college graduates and household income. By use of a least-squares line, we get the relationship

HOUSEHOLD INCOME = 12.47 + 2.25 (PCT OF COLLEGE GRADS)



How can we interpret this least-squares line? The slope of the best line is b = 2.25 which means that if a city has a college graduation percentage that is 1% higher, then we would expect its average household income to increase by 2.25 thousand dollars. (Recall that a slope is the increase in the y-variable for a unit increase in the x-variable.)

Although the line is a general description of the relationship between college graduation percentage and household income, note that there are a number of data points that are far from the line. We can focus on these unusual points by plotting the residuals against the graduation percentage.

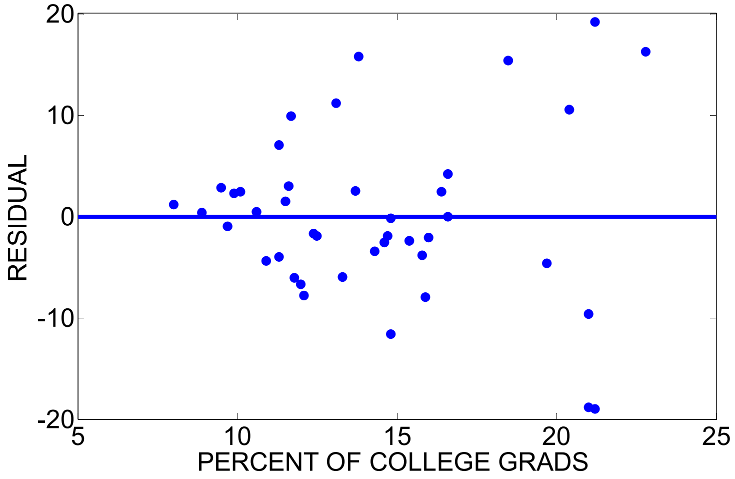
Let's review the computation of the residual. The first city Charlottesville has a college graduation percentage of 21 and an average household salary of $50.1 thousand. Using the line, we would predict this city's household income to be

PREDICTED INCOME = 12.47 + 2.25 x 21 = 59.7.

So the residual for Charlottesville would be

RESIDUAL = OBSERVED INCOME – PREDICTED INCOME = 50.1 – 59.7 = -9.6

Suppose we compute these residuals for all cities and graph the residuals against the explanatory variable (college graduation percentage). We get the following residual plot. (We typically add a horizontal line at the value RESIDUAL = 0 so it will be easy to spot the negative and positive residual values.)

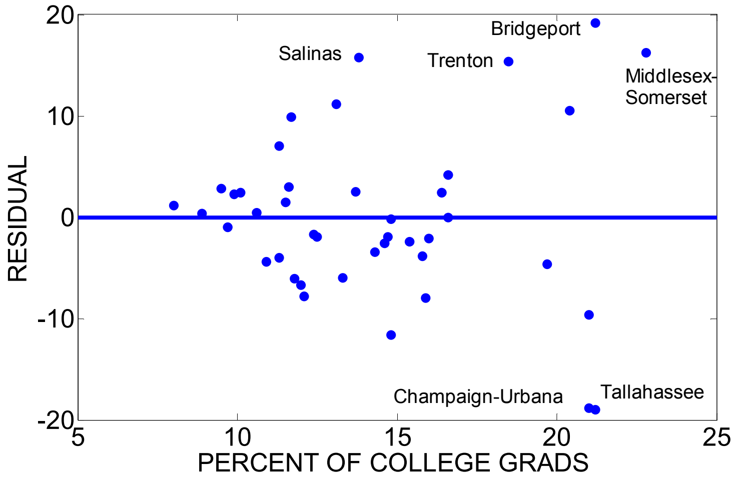


The residual plot focuses our attention on the deviations of the observed household incomes from the predicted values. When we look at a residual plot, we look for

- Systematic patterns of positive and negative residuals. For example, you might notice that there are many negative residuals on the left side of the plot and positive residuals on the right side of the plot. This indicates that the straight-line may not be a suitable fit to the data and an alternative method of describing the relationship may be necessary.
- Large positive and large negative residuals. In a residual plot, we look for large values that indicate points that are not close to the fitted line. We identify these

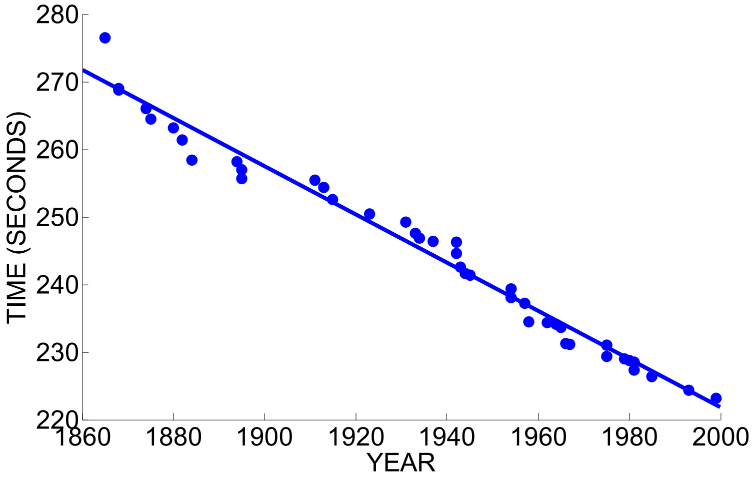
outlying points and think of any reason why these points don't follow the general pattern.

What do we see in our residual plot? There doesn't appear to be any systematic pattern of positive and negative residuals. Most of the residuals fall between –10 and 10 thousand dollars, but there are four large positive and two large negative residuals that seem to stand out. Looking back at our data, we see that the four large positive residuals correspond to Middlesex-Somerset (NJ), Bridgeport (CT), Salinas (CA), and Trenton (NJ). For these cities, their household incomes are much higher than one would predict on the basis of their college graduation rate. This is not surprising, since all four of these cities are located in expensive parts of the country (New Jersey, Connecticut, and California) where the cost of living is high. And the two large negative residuals correspond to Tallahassee (FL) and Champaign-Urbana (IL). These cities have household incomes that are smaller than one would predict on the basis of their college graduation rate. This is also not surprising since these are both college towns and the high college graduation rate reflects the college environment rather than the wealth of the general population. This discussion of unusually large residuals from the least-squares fit suggests that a city's average household income depends on more variables than just the education background of the population.



PRACTICE: PLOTTING RESIDUALS

Let's revisit the plot of the world record in the men's mile run that is graphed against the year where the record was obtained. Generally we see a straight-line pattern in the scatterplot – a least-squares fit is TIME = 933.94 - .36 YEAR .

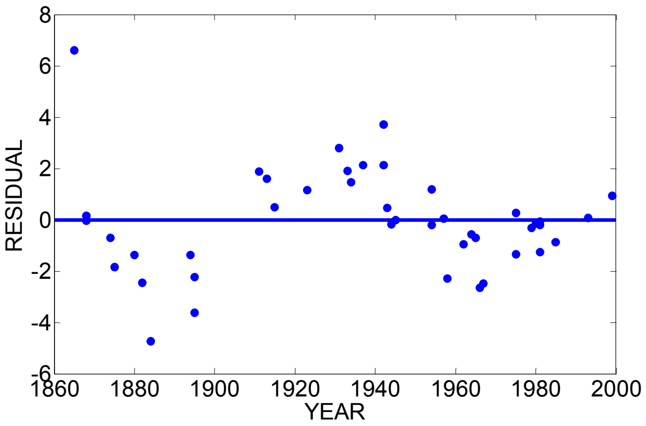


1. Based on the least-squares line, how much on average has the world record time decreased each year?

2. In the year 1895, the world record was 255.6 seconds. Use the least-squares equation to find the predicted time and compute the residual.

3. In the year 1937, the world record was 246.4 seconds. Compute the residual.

4. The residuals were computed for all points and a plot of the residuals against the year is displayed below.



Do you see any pattern in this residual plot? (Are there ranges of years where the residuals tend to be positive and when the residuals tend to be negative?) What does this say about the suitability of a straight-line relationship for the world record data?

5. Circle the point with the largest positive residual and the point with the largest negative residual. Is there any possible explanation for these large residuals?

6. As this book is written, the current world record for the mile run was set in 1999. Based on your work in this problem, does this surprise you? Explain.

TECHNOLOGY ACTIVITY: EXPLORING SOME OLYMPICS DATA

Open the Fathom file **summer_olympics.ftm**. This data gives the winning time in the men's 100, 200, 400, and 800 meter runs (in seconds) for each of the Summer Olympics from 1952 through 2004.

Focus on one race (either the 100m, 200m, 400m, or 800m) and see how the winning time for one race has changed from 1952 to 2004.

In your data exploration,

- Construct a scatterplot of the winning time (vertical) against the year (horizontal).
- Fit a good line to the points.
- Construct a residual plot.

Write a paragraph describing what you learned in this data analysis. In this paragraph, you should

- Give the equation of your best fit line and explain what it means. (How much on average is the winning time decreasing for each year?)
- Predict the winning time of the race in the next summer Olympics.
- Discuss any interesting features of the residual plot. Were there any particular years where the residual is unusually small or large? (Look for unusually small or large values in the residual plot.)

- The 1968 Olympics was unusual since it was held in Mexico City, a city at a high elevation, and many track and field records were set, such as the long jump. Do you notice anything unusual in your data in the 1968 time?

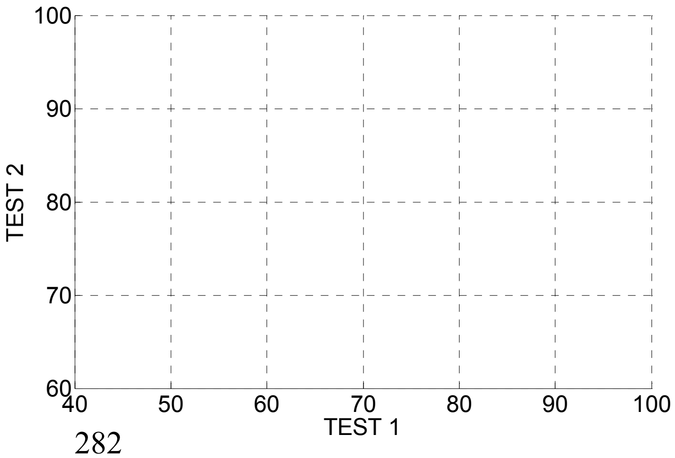
ACTIVITY: REGRESSION TO THE MEAN

DESCRIPTION: This activity demonstrates the "regression effect" that is generally unknown to many people. If you look up the word "regress" in the dictionary, it will tell you the word means to "go back." Suppose we collect two measurements from people that are positively correlated. We will see that there is a general tendency for a person's second measurement to go back, or regress to the mean.

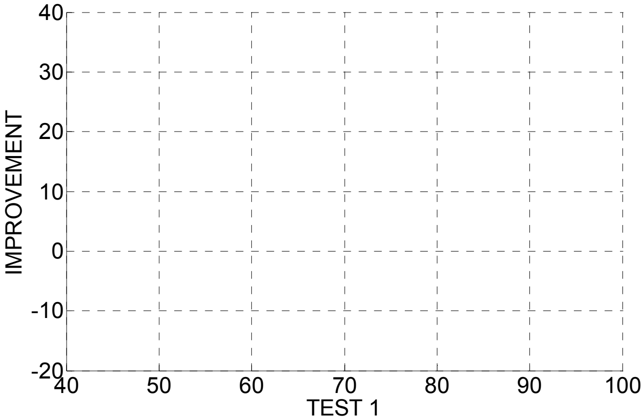
1. The following table gives the scores of 13 students on two tests in a statistics class. Construct a scatterplot of the Test 1 score (horizontal axis) against the Test 2 score on the below figure. Describe any relationship you see in the scatterplot. Is this relationship to be expected? Why?

STUDENT	TEST 1	TEST 2	IMPROVEMENT
1	96	87	
2	48	71	
3	75	80	
4	74	92	
5	88	97	
6	100	97	
7	51	77	
8	82	73	
9	80	87	
10	86	84	
11	76	64	
12	57	94	
13	56	79	

2. Next, compute the improvement TEST 2 – TEST 1 for each student and put the values in the table in the IMPROVEMENT column.



3. Construct a scatterplot of the improvement values (vertical axis) against the Test 1 scores (horizontal axis) on the axis below.



4. Do you see a pattern in this plot? Complete the following sentences. Students who had poor scores in Test 1 tended to ___________ and students who did well in Test 1 tended to ___________________.

5. Some of you may have heard about the so-called "sophomore slump" in sports. This happens when a player does well in his/her rookie year and then slumps in the sophomore year. Some baseball people believe that this player may be struggling due to the pressure of maintaining the first-year performance. Based on what you have learned in this activity, is there an alternative explanation for the sophomore slump? Explain.

OPTIONAL ACTIVITY: For twelve players in a particular professional sport (such as baseball, basketball, or football), find their statistics for two consecutive seasons. Use this data to demonstrate or refute the regression effect.

DIFFERENT WAYS OF LOOKING AT RELATIONSHIPS

A typical human sleeps about eight hours a day. Is this a common amount of sleep among mammals? Or do humans sleep longer or shorter relative to other mammals?

To answer these questions, we need to obtain some relevant data. The table below gives the following characteristics of 24 types of animals:

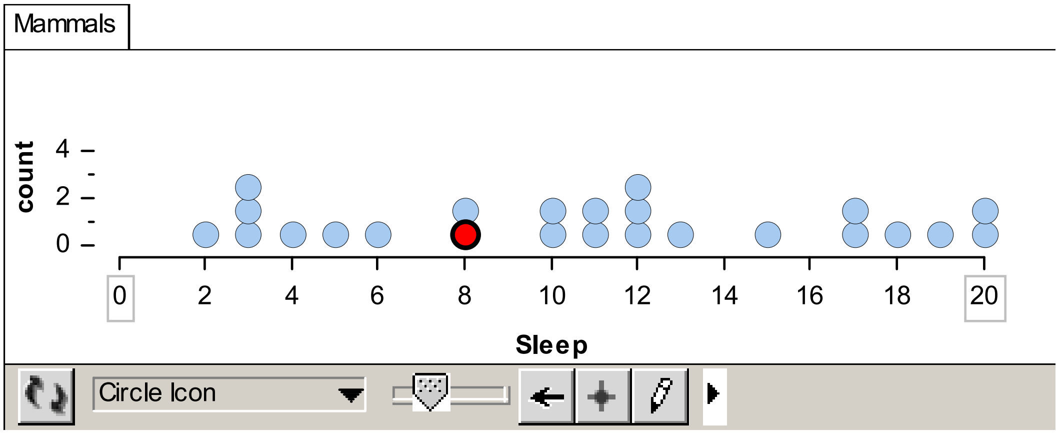
- MAMMAL -- the name of the animal
- SLEEP – the daily hours of sleep
- LIFESPAN – the life expectancy in years
- HEIGHT – the adult height in meters
- MASS – the adult mass in kilograms

Mammal	Sleep	LifeSpan	Height	Mass
African Elephant	3	70	4	6400
Asian Elephant	4	70	3	5000
Big Brown Bat	20	19	0.1	0.02
Bottlenose Dolphin	5	25	3.5	635
Cheetah	12	14	1.5	50
Chimpanzee	10	40	1.5	68
Domestic Cat	12	16	0.8	4.5
Donkey	3	40	1.2	187
Giraffe	2	25	5	1100
Gray Wolf	13	16	1.6	80
Grey Seal	6	30	2.1	275
Ground Squirrel	15	9	0.3	0.1
Horse	3	25	1.5	521
House Mouse	12	3	0.1	0.03
Human	8	80	1.9	80
Jaguar	11	20	1.8	115
Lion	20	15	2.5	250
N. American Opossum	19	5	0.5	5
Nine-Banded Armadillo	17	10	0.6	7
Owl Monkey	17	12	0.4	1

Pig	8	10	1	192
Rabbit	11	5	0.5	3
Red Fox	10	7	0.8	5
Spotted Hyena	18	25	0.9	70

Distribution of sleeping times

We focus on the relevant variable, SLEEP, that will help to answer our question. A dotplot of SLEEP for all 24 animals is shown below.



We see a lot of variation in the sleeping times -- some animals sleep, on average, 20 hours a day, and others sleep only 3 hours a day. Where does man stand in this sleeping time distribution? We label man's sleeping time with a black dot. We see that man's hours of sleep, 8, falls in the low end of this sleep distribution.

Since we observe a large spread of sleeping times it is natural next to try to explain why there is so much variation. Are there other variables in the dataset that might help in explaining the differences in sleep?

Let's consider LIFESPAN as a possible variable to help explain the variation in the response variable SLEEP. If we knew the lifetime of a particular animal, would this information be helpful in predicting its sleeping time?

Relating a categorical variable with a measurement variable

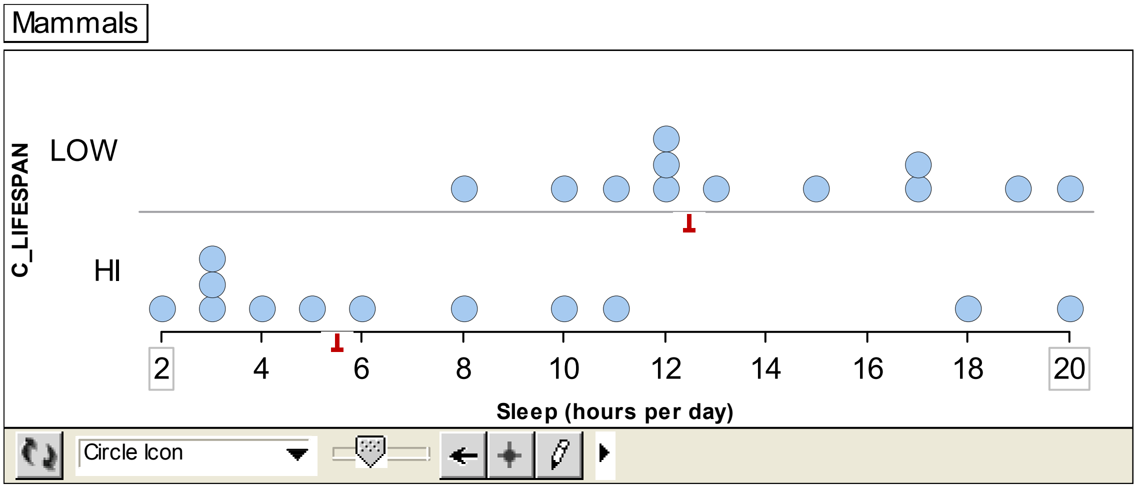
One way of studying the relationship between SLEEP and LIFESPAN is to divide the animal lifespans in two groups – the short-living animals and the long-living animals – and then compare the sleeping times of the two groups. This will be a familiar analysis for us, since we just described methods of comparing two batches in Topic D4.

We compute the median lifespan of the animals to be M = 17.5 years. We break the animals into two groups – the group living less than 17.5 years and the group living longer than 17.5 years – and then compare the sleeping times of the groups.

Here are the two groups of sleeping times:

SHORT-LIVING ANIMALS		LONG-LIVING ANIMALS	
(LIFESPAN < 17.5 YEARS)		(LIFESPAN > 17.5 YEARS)	
Mammal	Sleep	Mammal	Sleep
Cheetah	12	African Elephant	3
Domestic Cat	12	Asian Elephant	4
Gray Wolf	13	Big Brown Bat	20
Ground Squirrel	15	Bottlenose Dolphin	5
House Mouse	12	Chimpanzee	10
Lion	20	Donkey	3
N. American Opossum	19	Giraffe	2
Nine-Banded Armadillo	17	Grey Seal	6
Owl Monkey	17	Horse	3
Pig	8	Human	8
Rabbit	11	Jaguar	11
Red Fox	10	Spotted Hyena	18

In the below figure, we plot parallel dotplots of the two groups of sleeping times; we have also computed the medians of each group and marked these values on the display. Reading from the graph, we see that the short-living animals sleep, on average, about 12.5 hours, and the long-living animals sleep on average about 5.5 hours. So there is indeed a relationship between lifespan and sleep – the short-livers tend to sleep about 7 hours a day longer than the long-livers.



Relating two categorical variables

The above analysis helps us to understand that sleeping time is indeed related to the lifespan of an animal. But there are other ways to describe this relationship and these alternative ways may be helpful in explaining this phenomenon to others or in understanding the relationship in more detail.

Suppose we categorize both the response variable SLEEP and the explanatory variable LIFESPAN into two groups. Above we defined short and long livers by the animals that lived shorter and longer than the median lifespan. Similarly, we can compute the median sleeping time M = 11 hours and define

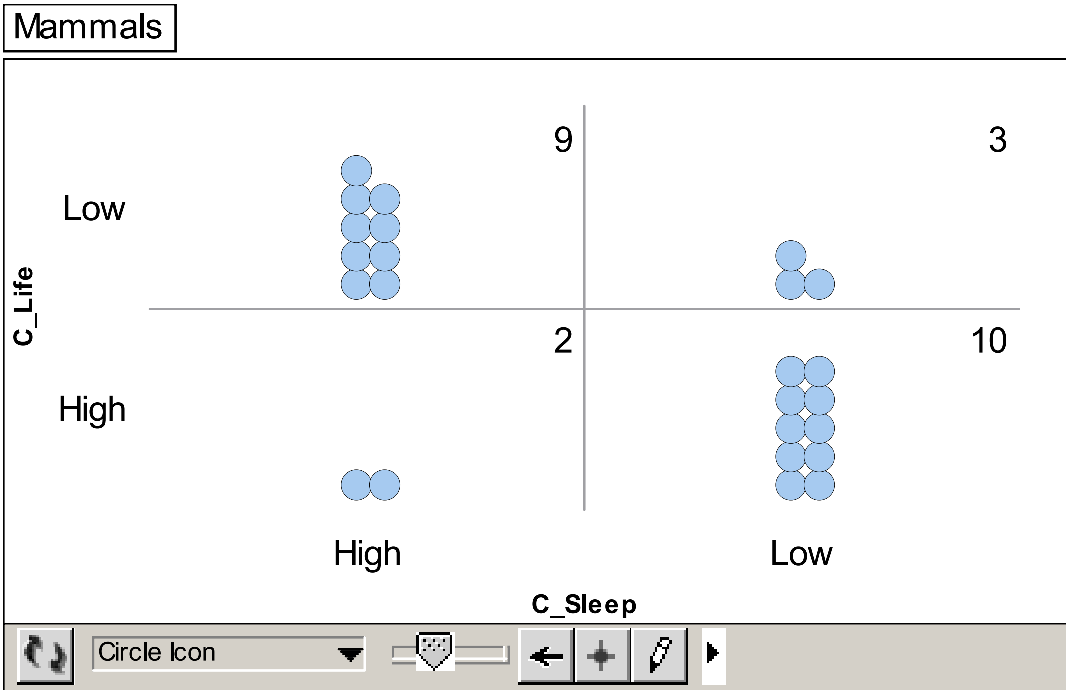
- Light-sleepers – animals that sleep 11 or fewer hours a day
- Heavy-sleepers – animals that sleep more than 11 hours a day

In the table below, we categorize all animals as HIGH or LOW on each of the two variables.

		Type of		Type of
Mammal	Sleep	Sleep	LifeSpan	LifeSpan
African Elephant	3	LOW	70	HIGH
Asian Elephant	4	LOW	70	HIGH
Big Brown Bat	20	HIGH	19	HIGH

Bottlenose Dolphin	5	LOW	25	HIGH
Cheetah	12	HIGH	14	LOW
Chimpanzee	10	LOW	40	HIGH
Domestic Cat	12	HIGH	16	LOW
Donkey	3	LOW	40	HIGH
Giraffe	2	LOW	25	HIGH
Gray Wolf	13	HIGH	16	LOW
Grey Seal	6	LOW	30	HIGH
Ground Squirrel	15	HIGH	9	LOW
Horse	3	LOW	25	HIGH
House Mouse	12	HIGH	3	LOW
Human	8	LOW	80	HIGH
Jaguar	11	LOW	20	HIGH
Lion	20	HIGH	15	LOW
N. American Opossum	19	HIGH	5	LOW
Nine-Banded Armadillo	17	HIGH	10	LOW
Owl Monkey	17	HIGH	12	LOW
Pig	8	LOW	10	LOW
Rabbit	11	LOW	5	LOW
Red Fox	10	LOW	7	LOW
Spotted Hyena	18	HIGH	25	HIGH

Once we have categorized animals with respect to the two variables, we can divide the animals into four groups – the ones that are LOW on both variables, ones LOW on lifespan and HIGH on sleep, one HIGH on lifespan and LOW on sleep, and those that are LOW on both variables. In the following Tinkerplots display, each animal is represented by a dot, and the dots are divided into the four groups.



When both variables are categorical, then we can describe the relationship by the computation of percentages.

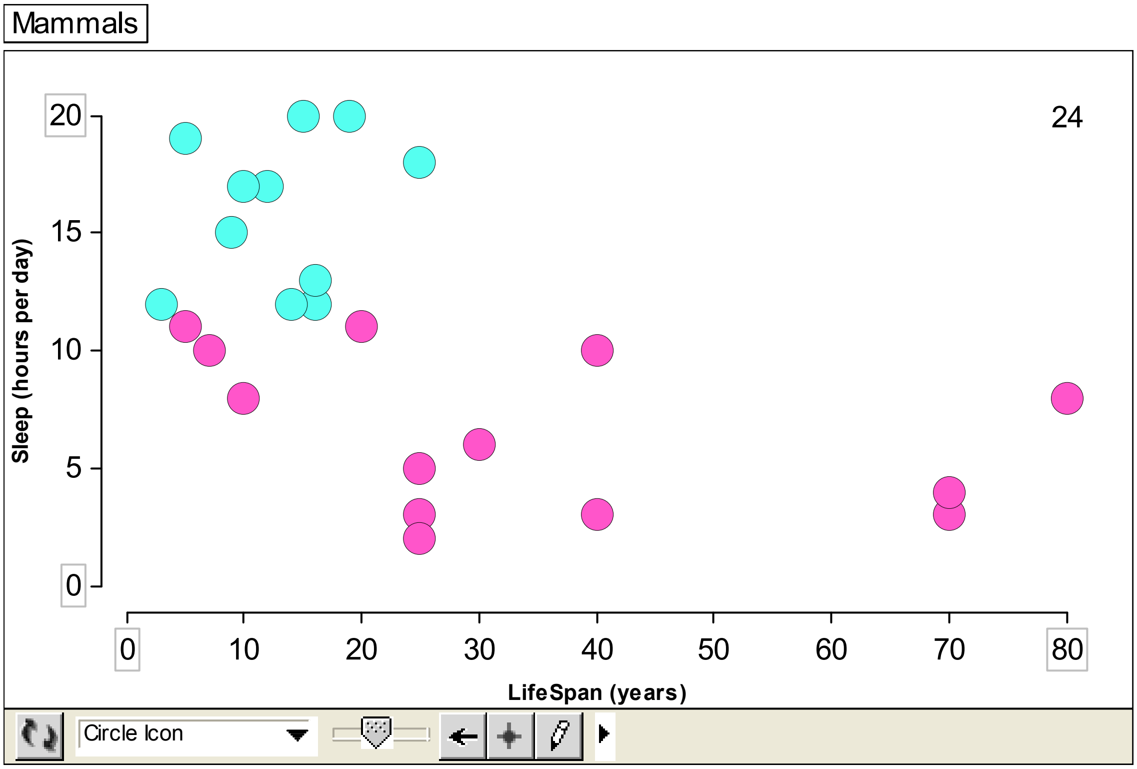
- Of the 12 short-living animals, we see from the figure that 9 or 9/12 = 75% are heavy-sleepers.
- In contrast, of the 12 long-living animals, we see that 2 or 2/12 = 16% are heavy sleepers.
- Since 75% is much higher than 16%, we can say that short-living animals are more likely to be heavy sleepers than long-living animals.

In Topic D7, we will focus on relationships between two variables that are both categorical.

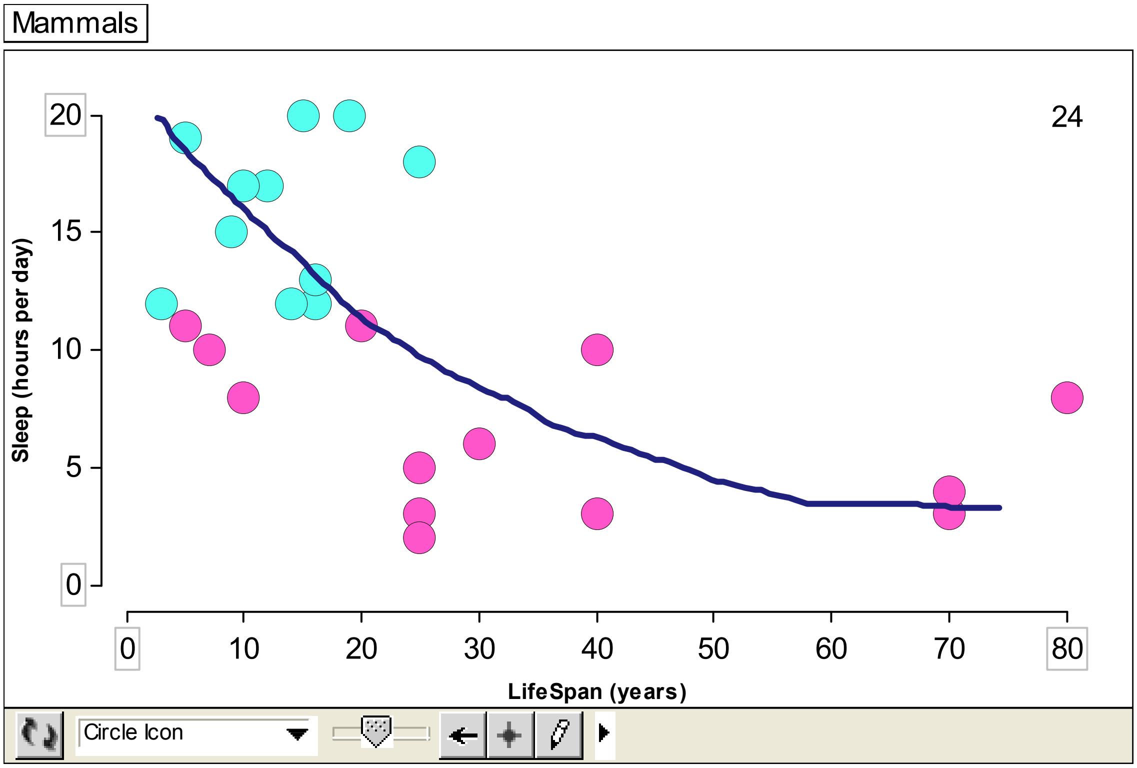
Relating two measurement variables

In both of the above analyses, we categorized animals as having short or long lifespans and compared the two groups of animals with respect to sleeping time. Is it possible to relate lifespan and sleeping time without this categorization?

When one has two measurement variables, a useful initial way of studying their relationship is by means of a scatterplot. Here one draws a grid of possible values of sleeping time (vertical) and lifespan (horizontal) and the data values are represented by dots placed at the corresponding values of the two variables. We get the scatterplot display shown below.



We show a pattern in the scatterplot by drawing a curve through the points as we look at the display from left to right.



We see that the drawn curve has a negative trend – this means that as the animals' lifespans increase, the sleeping times decrease. In particular, we see animals that live only about 10 years tend to sleep about 15 hours and animals with lifespans of 70 years tend to sleep about 5 hours.

What is the best way of studying relationships?

We have illustrated three methods for understanding how an animal's sleeping time relates to its lifespan. Which is the best way for understanding this relationship?

Actually, there is not a "best" method in general. We use different descriptions depending on the problem on how we plan on communicating the relationship. The use of the scatterplot might seem like the best method because one loses information about sleeping time lifespan when one categorized them into HI/LOW groups. But it can be difficult to summarize the pattern of the relationship in a scatterplot and easier to describe the relationship when the variables are categorized and summarized as counts in a two-way table. For any method we use, it is important to be able to summarize how the knowledge of the explanatory variables helps us predict the values of the response variable.

PRACTICE: DIFFERENT WAYS AT LOOKING AT RELATIONSHIPS

In the previous example, suppose we are interested in explaining the difference in sleeping times by the MASS variable (the weight of the animal in kilograms).

1. The median mass of the 24 animals is 70 kilograms. Divide the animals into two groups – the ones who weigh at most 70 kilograms and the ones who weigh more than 70 kilograms. In the below table, write the names and sleeping times for the animals in the two groups.

MASS AT MOST 70 KG	
Animal	Sleeping Time

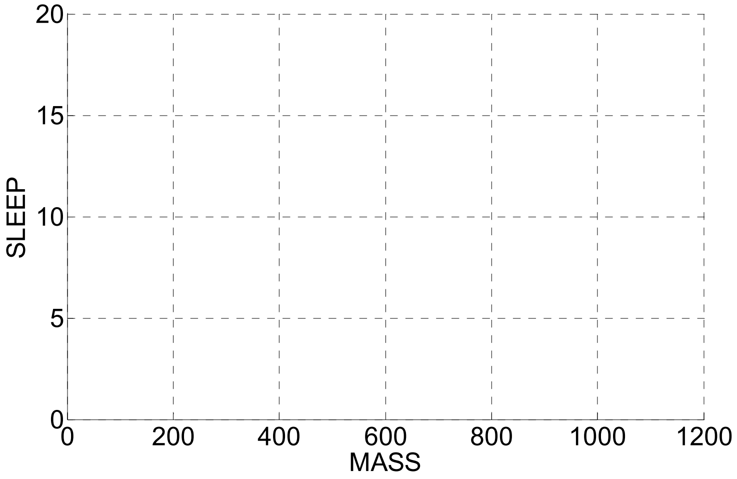
MASS MORE THAN 70 KG	
Animal	Sleep Time

2. By use of parallel boxplots, compare the sleeping times of the two groups of animals. Does one group of animals tend to sleep more on average than the other group? Explain.

3. Recall that the median sleeping time of the animals was 11 hours. As in the example, divide the animals into two groups – the light-sleepers (hours at most 11 hours) and the heavy-sleepers (hours exceeding 11 hours). Construct a two-way table of counts that categorizes the animals by mass and sleep. Using the data in this table, describe the relationship between mass and sleep.

		Sleeping Time	
		Low	High
MASS	Low		
	High		

4. As a final way of studying the relationship between mass and sleeping, construct a scatterplot of the two variables on the grid below. (To produce a reasonable looking display, we have limited the range of the horizontal axis from 0 to 1200 and two points will be off the graph.) Describe the general pattern in the scatterplot. As animals get heavier, do they tend to have lower or higher sleeping times?



TECHNOLOGY ACTIVITY – USING TINKERPLOTS TO STUDY RELATIONSHIPS

In this lab, you will use Tinkerplots to explore the relationship between a pair of variables.

1. First, you want to find two variables in a dataset where you suspect there is a relationship. A good source of interesting datasets can be found at DASL (the data and story library) at the website

`http://lib.stat.cmu.edu/DASL`

To find a interesting dataset where there is a relationship between a pair of variables, go to LIST ALL METHODS and find datasets that use the methods

Boxplot or Correlation or Regression or Contingency Table

2. BEFORE YOU GRAPH THE DATA, describe the dataset you will be using, discuss the variables you will be looking at, and what relationship you think you might find between these two variables. Why are you interested in this dataset?

[WRITE THIS DESCRIPTION ON THE TINKERPLOTS DOCUMENT.]

3. Copy the web address for the dataset that you are interested in on the Clipboard. Launch Tinkerplots. To get the data into Tinkerplots, go to

Import from URL from the File menu

and paste the web address of your dataset into the dialog box. Your data should be loaded into Tinkerplots.

4. Experiment with different graphs on Tinkerplots until you find one that is helpful for showing the relationship between the two variables.

5. Explain (AGAIN ON THE TINKERPLOTS DOCUMENT) what you have learned about the relationship between the two variables using the graph.

6. Repeat this work (steps 1 through 5) using two variables in a different dataset.

WRAP-UP

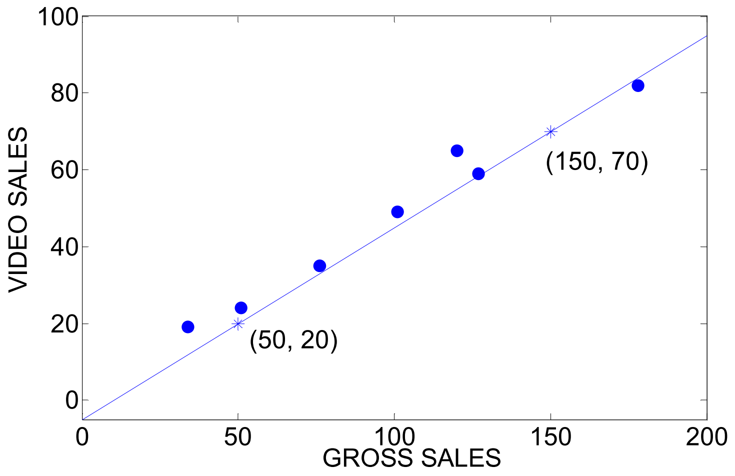
In this topic, we continued our study of relationships of two quantitative variables. The least-squares line is the standard method for fitting a straight-line to data. It is called least-squares since it is the line that makes the sum of squared residuals as small as possible. An alternative method of fitting a line, a median-median line, was described that is less sensitive to outliers than this least-squares method. When fitting a line, it is also helpful to construct a residual plot that focuses on the vertical distances of the points from the line. We look for patterns in the residual plot that indicate that a line may not be the best description of the relationship between the two variables. We demonstrated the regression effect where there is a general tendency with correlated data for the second observation to move back or regress towards the mean. We concluded by describing three approaches for studying the relationship between two quantitative variables. One approach for studying the relationship is to categorize each variable and make a comparison by comparing percentages as described in Topic D5. A second approach is to divide the explanatory variable into several categories and then compare the batches of values of the response variable using the methods described in Topic D4. A third strategy is to construct a *scatterplot* of the two variables and then use methods such as the computation of a correlation or the use of a best-line fit to understand the relationship.

EXERCISES

1. **Gross and Video Sales for Movies Starring Julia Roberts**

The table below lists the gross sales and video sales (both in millions of dollars as of March 2004) for seven movies featuring Julia Roberts. A scatterplot of these two variables follows.

	GROSS	VIDEO
MOVIE	SALES	SALES
Conspiracy Theory (1997)	76	35
Dying Young (1991)	34	19
Hook (1991)	120	65
My Best Friend's Wedding (1997)	127	59
Pelican Brief, The (1993)	101	49
Pretty Woman (1990)	178	82
Something to Talk About (1995)	51	24



a. A line is fitted through the points (50, 20) and (150, 70), say. Find the equation of this line.

b. In the table below, compute the residuals using the fitted line that you found in (a).

Fitting the line from (a).

	GROSS	VIDEO	Fitted	Residual
MOVIE	SALES	SALES	value	
Conspiracy Theory (1997)	76	35		
Dying Young (1991)	34	19		
Hook (1991)	120	65		
My Best Friend's Wedding (1997)	127	59		
Pelican Brief, The (1993)	101	49		
Pretty Woman (1990)	178	82		
Something to Talk About (1995)	51	24		

c. We can judge the goodness of the fit of this line by the sum of squared residuals. Compute the sum of squared residuals for the line you found in (a).

d. The “least-squares” line for these data is

$$\text{VIDEO SALES} = 2.39 + 0.46 \text{ GROSS SALES}$$

Compute the residuals for the least-squares line in the table below.

Fitting the least-squares line

MOVIE	GROSS	VIDEO	Fitted	Residual

	SALES	SALES	value
Conspiracy Theory (1997)	76	35	
Dying Young (1991)	34	19	
Hook (1991)	120	65	
My Best Friend's Wedding (1997)	127	59	
Pelican Brief, The (1993)	101	49	
Pretty Woman (1990)	178	82	
Something to Talk About (1995)	51	24	

e. Compute the sum of squared residuals for the least-squares line. Is this value larger or smaller than the sum of squared residuals for the line from part (a)? Are you surprised by this result?

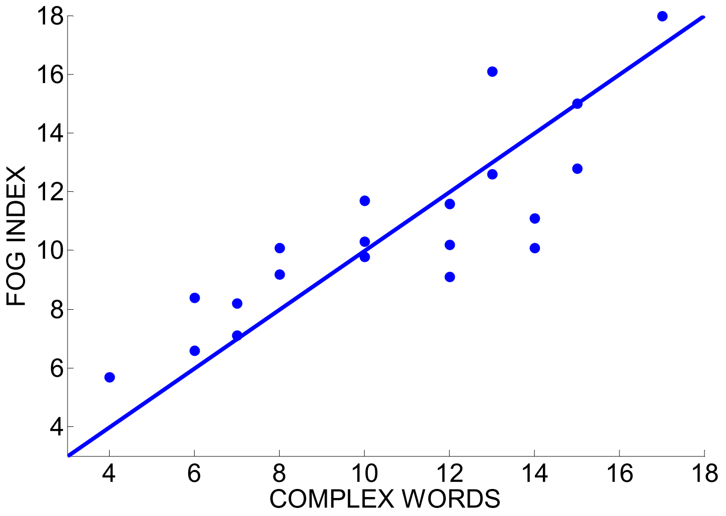
2. **Book Statistics**

The amazon.com website gives "text statistics" for many of the books it sells. For a particular book, the website displays the "fog index", the number of years of formal education required to read and understand a passage of text, and the "complex words", the percentage of words in the book with three or more syllables. The following table displays the complex words and fog index for a selection of 20 popular books.

	Complex	Fog	Residual
Book	Words	Index	
The Da Vinci Code	12	9.1	
Marley & Me	8	9.2	
The World is Flat	15	15	
Freakonomics	14	11.1	
Misquoting Jesus	13	16.1	
Power of Thinking Without Thinking	12	11.6	
The Mermaid Chair	7	8.2	
Memoirs of a Geisha	8	10.1	
The Five People You Meet in Heaven	6	6.6	
The Kite Runner	7	7.1	
In Cold Blood	10	9.8	

A Million Little Pieces	4	5.7
The Tipping Point	13	12.6
The Glass Castle	6	8.4
Collapse	17	18
Confessions of an Economics Hit Man	15	12.8
Curve Ball	14	10.1
A Mathematician at the Ballpark	12	10.2
Moneyball	10	10.3
Jim Cramer's Real Money	10	11.7

a. Suppose we are interested in predicting the Fog Index from Complex Words using the simple line formula FOG = COMPLEX. The graph below shows a scatterplot and this line. For each book, find the residual = actual Fog Index – predicted Fog Index and place the residuals in the table. Compute the sum of squared residuals.



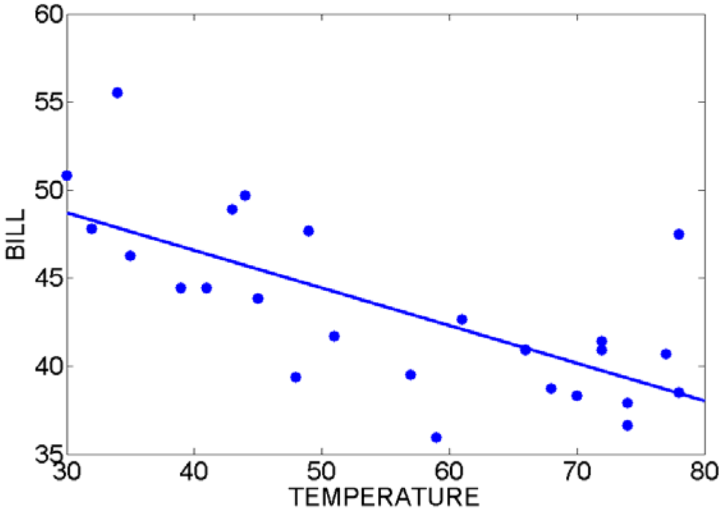
b. Find the book that has the residual of the largest size. Circle the point on the graph that has this largest residual.

c. The least-squares fit to these data is FOG = 2.91 + .73 COMPLEX. Find the residual for *The DaVinci Code* from this least-squares fit.

d. Suppose we compute the sum of squared residuals for the least-squares fit. Will this value be smaller or larger than the value of the sum of squared residuals you found in part a? Explain.

3. **Electricity Bills**

A homeowner collected the average temperature of a month (degrees Fahrenheit) and the amount of her electricity bill (in dollars) for that month. The figure below displays a scatterplot of temperature against the bill amount.



a. 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?

b. Using the summary statistics shown below, determine the equation of the least squares (regression) line for predicting the electric bill from the average temperature. Record the equation of the line.

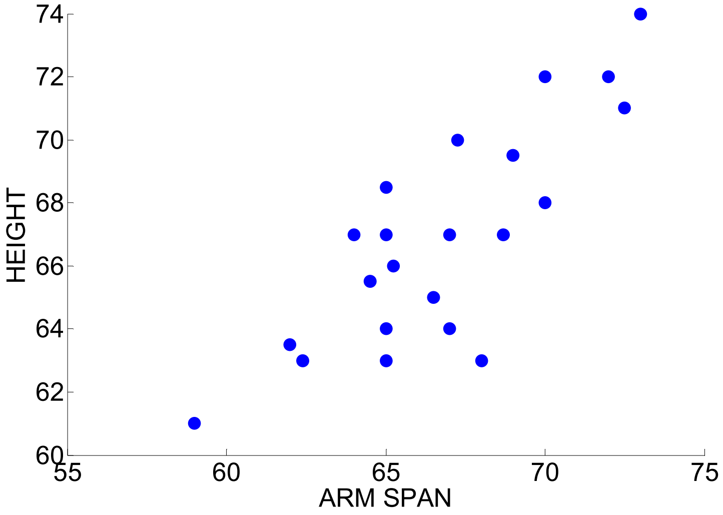
Variable	Temperature	Bill	Correlation
Mean	55.88	43.18	-0.695
Standard deviation	16.21	4.99	

c. In March of 1992, the temperature was 41 degrees and the electricity bill was $44.43. Use the equation you found in part b to determine (by hand) the fitted value and residual for March of 1992.

d. From looking at the graph, identify the points that have unusually large residual values. Were the electric bills higher or lower than expected for their average temperature?

4. **Height and Arm Span of Students**

It has been found that a person's arm span is strongly related to height. To investigate the nature of this relationship, arm spans and heights (both measured in inches) were measured for a class of 22 college students. A scatterplot of arm span and height is shown below.



a. The mean and standard deviation of arm span are given respectively by $\bar{x} = 66.73$ inches and $s_x = 3.52$ inches, the summary values for height are $\bar{y} = 68.86$ inches and s_y =3.48, and the correlation between the two variables is r = .813. Using these statistics, find the equation of the least-squares line predicting height from arm span.

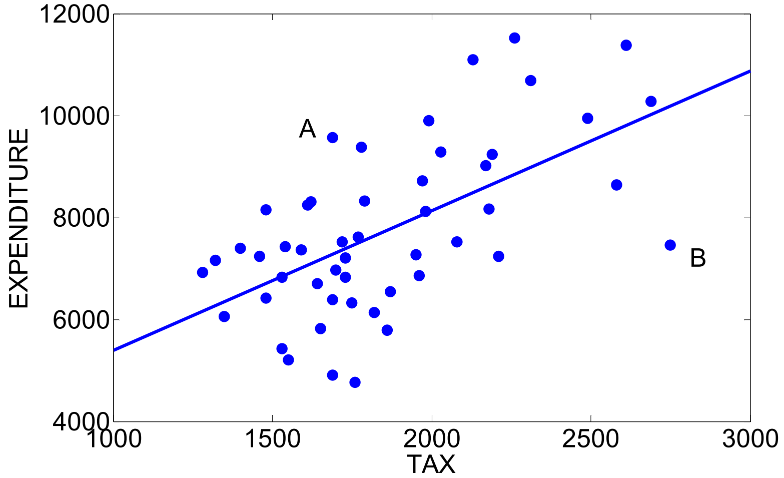
b. Graph the equation of the least-squares line on the scatterplot.

c. Suppose one student's arm span is 10 inches longer than another student's arm span. Predict how much taller the first student will be compared to the second student.

5. **Tax Revenue and Spending on Schools**

The 2004 Supplement of the *World Book Encyclopedia* gives the tax revenue per capita (the amount of tax collected per resident) and the public school expenditure per pupil for each of the 50 states. Below a scatterplot of TAX vs. EXPENDITURE is displayed – a best-line (least-squares fit) is

EXPENDITURE = 2.74 TAX + 2650.



a. The slope of the best-fit line is _______ . This means that if a state decides to tax the residents an additional $100 per resident, one would predict that the expenditure per pupil would increase by _______________.

b. The state of Ohio collected $1720 of tax per resident. Using the best-line equation, predict how much their expenditure should be per pupil.

c. Actually, it turns out that Ohio's expenditure per pupil was $7520. Compute the residual or the error in your prediction. (Remember that a RESIDUAL = ACTUAL y VALUE – PREDICTED y value.)

d. For North Dakota, the (TAX, EXPENDITURE) = (1760, 4770). Compute the residual.

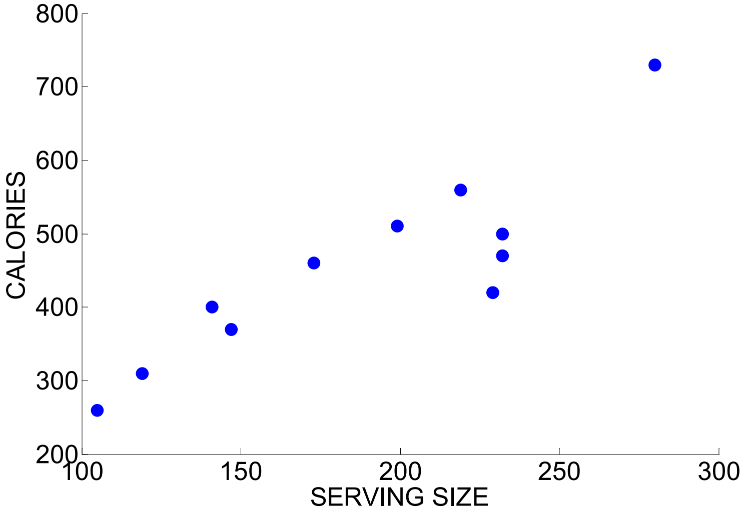
e. Two points are labeled in the graph. From the graph, compute (approximately) the value of the residual.

POINT "A" RESIDUAL: __________ POINT "B" RESIDUAL: __________

6. **Nutrition at a Fast-Food Restaurant**

McDonalds restaurant publishes nutritional information about all of the sandwiches they sell. The below table shows the serving size (in grams) and calories for a number of sandwiches. Suppose you are interested in understanding the relationship between the two variables. A scatterplot is shown below.

Sandwich	Serving size (gm)	Calories
Hamburger	105	260
Cheeseburger	119	310
Double Cheeseburger	173	460
Quarter Pounder® with Cheese+	199	510
Double Quarter Pounder® with Cheese++	280	730
Big Mac®	219	560
Big N' Tasty®	232	470
Filet-O-Fish®	141	400
McChicken ®	147	370
Premium Grilled Chicken Classic Sandwich	229	420
Premium Crispy Chicken Classic Sandwich	232	500



A least-squares line to these data is given by

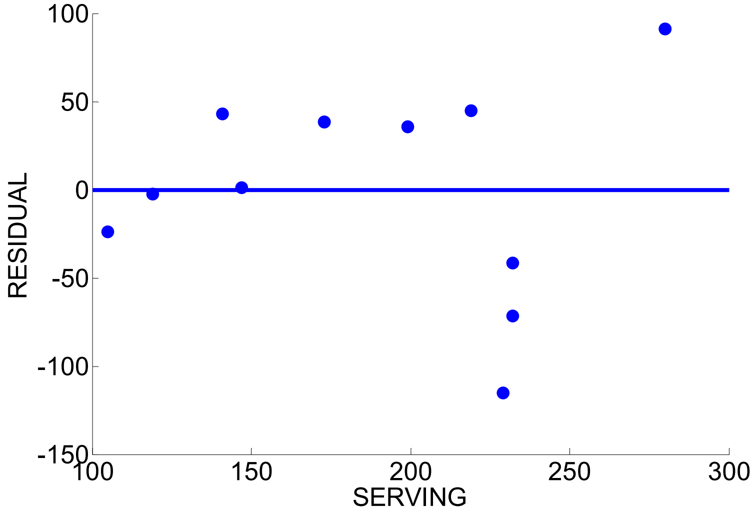
CALORIES = 71.0166 + 2.0274 SERVING_SIZE .

a. Suppose a sandwich's serving size is 200 grams. Predict the number of calories of this sandwich.

b. Suppose you are given the option to "super-size" a sandwich by making it 100 grams larger. How many extra calories will be in this super-size sandwich?

c. Compute the residual for Filet-O-Fish.

d. The residuals are computed for all sandwiches and a graph of the residuals against the serving size is displayed below. Are there any unusual points in this residual graph? Looking back at the data table, find the sandwiches that have these unusual residuals. Are these sandwiches different from the remaining sandwiches?



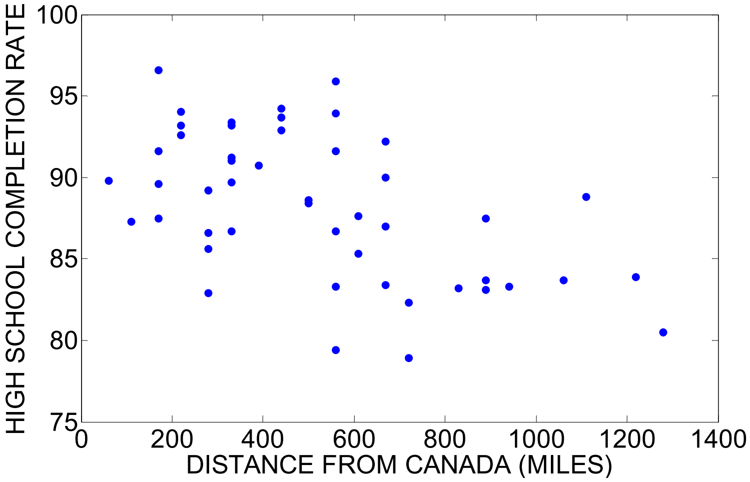
7. **High School Completion Rates**

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 below gives the adult high school completion rate (as a percentage) for each of the continental 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, the author got out his 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.

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	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	670
Kentucky	83.3	560	South Dakota	93.2	330
Louisiana	83.9	1220	Tennessee	82.3	720
Maine	94	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	670	Wisconsin	93.4	330
Montana	91.6	170	Wyoming	91.6	560

The figure below plots the distance from Canada (horizontal axis) against the completion rate (vertical axis).



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?

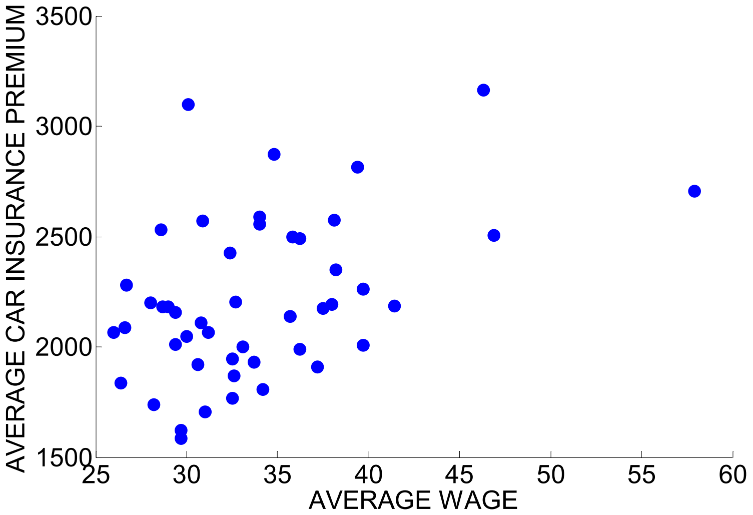
d. Circle a second point on the graph which corresponds to a state which is far from Canada and has a relatively large completion rate value. Label this point B. Which state does this point correspond to?

e. 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? (Such a variable is called a *lurking variable*.)

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?

8. **Car Insurance Premiums and Average Salary**

One major expense in owning a car is insurance. There is a large variation in the cost of car insurance across states and is natural to wonder about the cause for this variation. For each state, two variables are recorded: the average annual car insurance premium in the year 2005 and the average wage in the year 2002. A scatterplot of these variables is shown below.



a. Is there a relationship between a state's average wage and its average car insurance premium? Describe the direction and strength of this association.

b. Make an intelligent guess at the value of the correlation r.

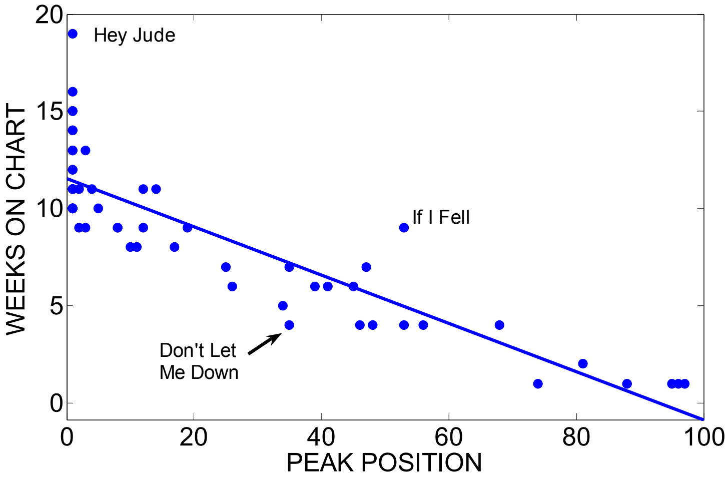
c. Would it be accurate to say that some state have high car insurance premiums since the workers in those states have higher incomes and therefore can afford these high premiums?

d. Is there a lurking variable present that might explain both the difference in wages and the difference in car insurance premiums? Explain.

9. **Beatles' Hit Songs**

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. 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 below 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 WEEKS = 11.54 - 0.124 PEAK . This line is placed on top of the scatterplot in the figure.



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.

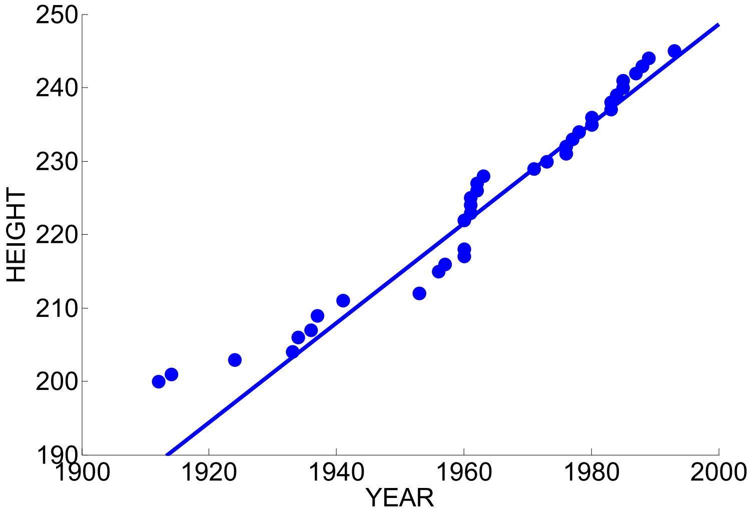
c. The point corresponding to the song ``Hey Jude" is labeled on the scatterplot. This song peaked at number 1 and stayed 19 weeks on the hit chart. Compute the residual for this song.

d. Two other songs, ``If I Fell" and ``Don't Let Me Down," are also labeled 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.

10. **High Jump Record**

In most track and field events, there has been a general increase in the world record. To illustrate, the below figure graphs the word record achievement (in meters) in the men's high jump competition as a function of year.

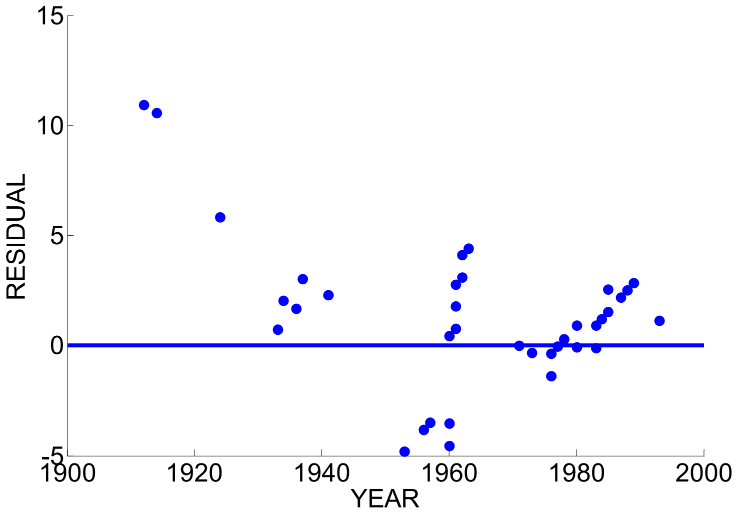


The pattern of achievement in the height of the record high jump can be described by the line HEIGHT = 0.6766 YEAR – 1104.6.

a. Generally, how much has the world record for the high jump changed for each year?

b. Can you use this line to predict the record in the high jump in the year 2050? Explain.

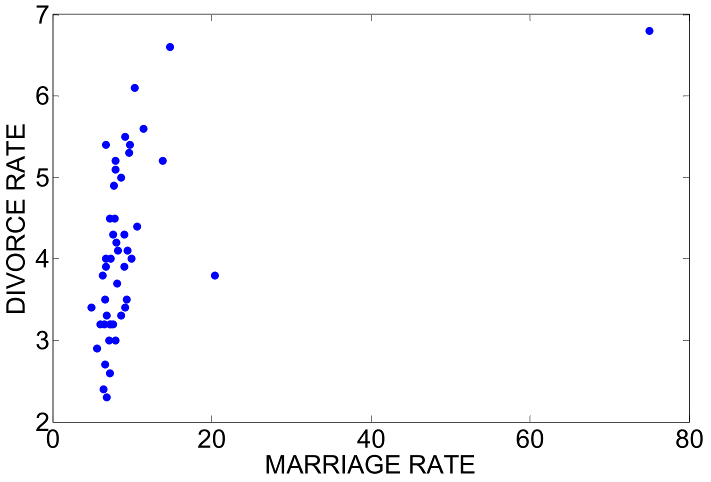
The below figure plots the residuals from this line fit as a function of year. This residual graph has a distinctive pattern.



c. There are four clusters of points in the residual graph corresponding to “before 1940”, “1955 to 1960”, “1960s”, and “1970 to 2000”. Describe the pattern of each cluster of points and explain what is says about the change in the high jump record.

11. **Marriage and Divorce Rates**

The *Statistical Abstract of the United Sates 2003* lists the marriage and divorce rates for the year 2001 of all of the states of the United States and the District of Columbia. To investigate a relationship between the two rates, we construct a scatterplot pictured below, where the divorce rate is plotted on the vertical axis and the marriage rate on the horizontal axis.



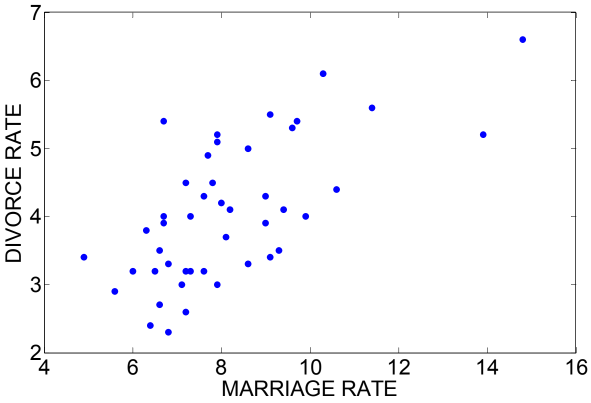
a. Describe in a few sentences any pattern or distinctive aspects of this scatterplot. Does a relationship appear to exist between the divorce rate and the marriage rate?

b. The correlation coefficient for this dataset is calculated to be r = .468. Explain in words what this value means.

c. Suppose that the marriage rates for Nevada and Hawaii (the unusual points in the scatterplot) are removed from the data. How would this change in the dataset affect the value of the correlation? Would the value of r go up, go down, or stay about the same?

d. Actually, the new value of r with Nevada and Hawaii removed is equal to r = .659. Why is this new value so different from the old value?

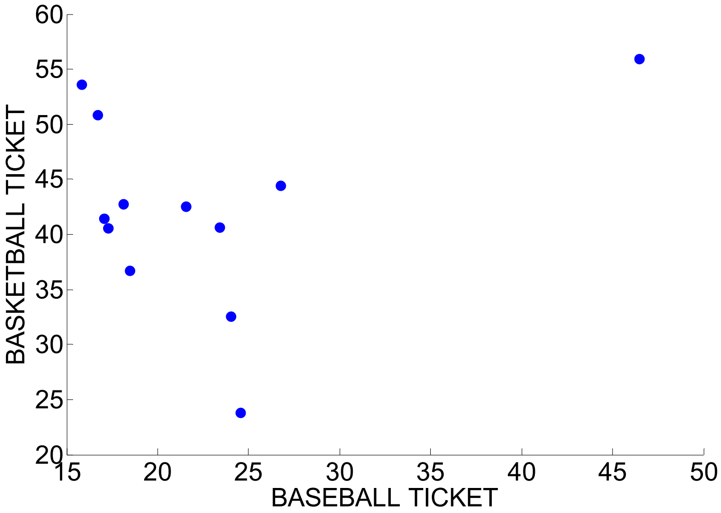
e. Suppose that we redraw the scatterplot with the Nevada and Hawaii points removed --- the new scatterplot is shown below. Does this change have an affect on the appearance of the scatterplot? Which scatterplot do you prefer --- the first one or the second one? Why?



12. **Cost of Sports Tickets**

There is variation in the cost of living among different cities in the United States. For example, it is more expensive to live in an eastern city, say Boston, than a city in the Midwest such as Minneapolis. Also one might think there is a relationship between the cost of two items for different cities. To investigate this, the below table displays the average cost of a ticket for professional baseball and professional basketball games for 12 cities. A scatterplot of these data is also drawn.

TEAM	MLB_Ticket	NBA_Ticket
BOSTON	46.46	55.93
SAN FRANCISCO	24.53	23.82
PHILADELPHIA	26.73	44.47
SEATTLE	24.01	32.54
TORONTO	23.4	40.67
DETROIT	18.48	36.75
CLEVELAND	21.54	42.52
FLORIDA	16.7	50.87
MINNESOTA	17.26	40.6
ATLANTA	17.07	41.43
TEXAS	15.81	53.6
MILWAUKEE	18.11	42.78



a. There is one outlying point in this graph. Identify the city that corresponds to this outlier.

b. Suppose the outlying point in the graph is removed. Describe the relationship in the graph and estimate the value of the correlation.

c. If one includes the outlying point, how do you think the value of the correlation will change? Why?

d. Use a statistics computer package or a calculator to compute the correlation for the full dataset and the dataset with the outlier removed. Is the value of the correlation sensitive to the inclusion of the outlying point? Explain.

13. **Gross and Video Sales for Movies Starring Julia Roberts**

In Exercise 1, the gross sales and video sales for seven Julia Roberts movies was considered.

a. Construct a median-median line predicting the video sales from the gross sales.

b. In Exercise 7, the least-squares line was given to be VIDEO SALES = 2.39 + 0.46 GROSS SALES. Compare the median-median and least-squares lines by finding the predicted values for each movie.

MOVIE	GROSS SALES	VIDEO SALES	Predicted value least-squares	Predicted value median-median
Conspiracy Theory (1997)	76	35		
Dying Young (1991)	34	19		
Hook (1991)	120	65		

My Best Friend's Wedding (1997)	127	59
Pelican Brief, The (1993)	101	49
Pretty Woman (1990)	178	82
Something to Talk About (1995)	51	24

c. Based on your work from part b, do the least-squares and median-median lines give similar predictions? Explain.

14. **Book Statistics**

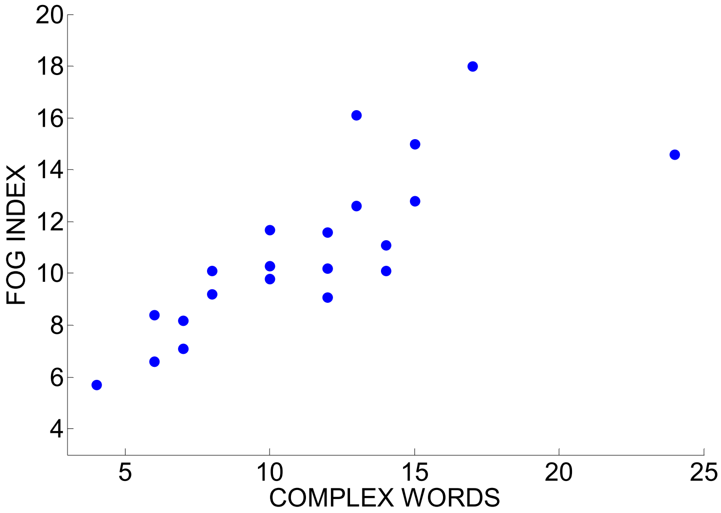
Suppose you are interested in predicting the Fog Index of a book from its Complex Words. (The data is given in Exercise 2.)

a. Find the median-median line by first finding the three summary points and then computing the slope and intercept.

b. The least squares fit to these data is FOG = 2.91 + .73 COMPLEX. Compare the median-median and least-squares fit by computing the predicted values for the following three books. Are the two lines similar with respect to their predictions for these books?

Book	Complex Words	Prediction of FOG from least-squares	Prediction of FOG from median-median
A Million Little Pieces	4		
The Glass Castle	6		
The Mermaid Chair	7		

c. Suppose that the book *Ordinal Data Modeling* with Complex Words = 24 and Fog Index = 14.6 is added to the dataset. A scatterplot of COMPLEX and FOG is drawn below – the new book is labeled on the graph. Recompute the median-median line for this new dataset. Is the median-median line sensitive to the addition of this new book? Explain.



d. The least squares fit to the original dataset was 2.91 + .73 COMPLEX and the least squares fit to the new dataset is FOG = 4.6 + .552 COMPLEX. Is the least squares line sensitive to the addition of this new book? Explain.

e. Based on your work from parts d and e, what is one advantage of the median-median line over the least-squares line?

15. **Batting Averages**

The table below gives the batting average (AVG) for 13 baseball players that played in the 2003 World Series between the New York Yankees and the Florida Marlins.

	2003 AVG	2002 AVG	Improvement
Jorge Posada	0.281	0.268	
Jason Giambi	0.250	0.314	
Alfonso Soriano	0.290	0.300	
Derek Jeter	0.324	0.297	
Bernie Williams	0.263	0.333	
Raul Mondesi	0.272	0.232	
Nick Johnson	0.284	0.243	
Ivan Rodriguez	0.297	0.314	

Derrek Lee	0.271	0.270
Luis Castillo	0.314	0.305
Mike Lowell	0.276	0.276
Juan Pierre	0.305	0.287
Juan Encarnacion	0.270	0.271

(a) Construct a scatterplot of the 2002 batting averages against the 2003 averages. Comment on any pattern in the plot.

(b) For each player, compute the improvement in batting average from the 2002 to the 2003 season (IMPROVEMENT = 2003 AVG – 2002 AVG). Place your values in the table.

(c) Construct a scatterplot of the improvement (vertical axis) against the 2002 AVG (horizontal axis). Describe the pattern of this pattern.

(d) Explain why this example illustrates the regression effect. Describe this concept in words that would be understandable by a layman.

(e) Suppose a baseball player has a great season and hits for a very high batting average (much higher than his previous seasons). Do you expect this player to have the same batting average the next season? Explain.

16. **Measuring School Achievement**

In many states, scores on standardized exams such as the ACT are used to rank high schools. The following table gives the average scores on an ACT exam for 14 Illinois high schools for the years 2001 and 2003.

High School	act2001	act2003	Improvement
J STERLING MORTON WEST HIGH SCH	19.9	18.7	
HAMPSHIRE HIGH SCHOOL	23.5	21.1	
COULTERVILLE HIGH SCHOOL	22.6	18.7	
VIRGINIA SR HIGH SCHOOL	19.7	19.1	
MT ZION HIGH SCHOOL	22.7	21.4	

BOGAN HIGH SCHOOL	17.1	16
BROWN COUNTY HIGH SCHOOL	21	19.4
ROCKFORD EAST HIGH SCHOOL	20.5	17.7
GENEVA COMMUNITY HIGH SCHOOL	23.7	21.5
CENTRAL HIGH SCHOOL	23.3	20.4
CARROLLTON HIGH SCHOOL	21.3	18.3
GLENBARD NORTH HIGH SCHOOL	22.7	20.8
VANDALIA COMMUNITY HIGH SCHOOL	21.5	19.2
ARCOLA HIGH SCHOOL	23	19.8

a. Construct a scatterplot of the 2001 ACT score and the 2003 ACT score. Describe the relationship you see in the scatterplot. Are you surprised by this relationship? Explain.
b. For each school, compute the improvement (2003 ACT score) – (2001 ACT score). Place the improvement values in the table.
c. Construct a scatterplot of the improvement values (vertical axis) against the 2001 ACT scores (horizontal axis). Describe the relationship you see in the graph.
d. Explain how this example illustrates regression to the mean.

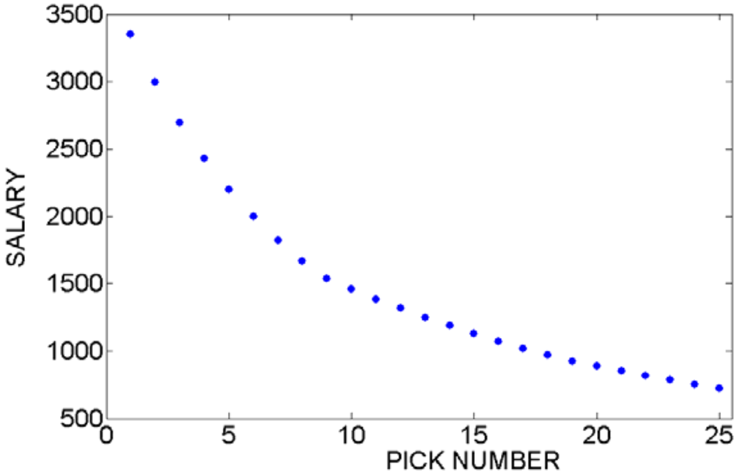
17. **Basketball Rookie Salaries**.

The table below pertains to basketball players selected in the first round of the 2003 National Basketball Association draft. It lists the draft number (the order in which the player was selected) of each player and the annual salary, in thousands of dollars, of the contract that the player signed. (Data is from the website InsideHoops.com.)

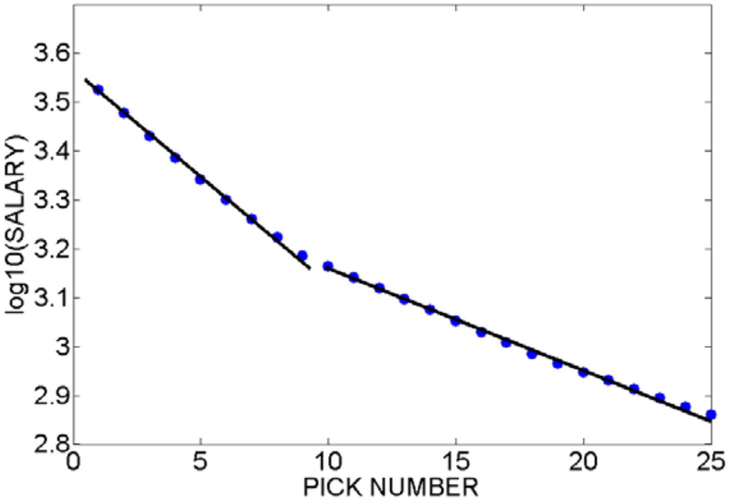
Pick		**Pick**		**Pick**	
Number	**Salary**	**Number**	**Salary**	**Number**	**Salary**
1	3349	**11**	1384	**21**	851
2	2997	**12**	1315	**22**	817
3	2691	**13**	1250	**23**	784
4	2426	**14**	1187	**24**	753
5	2197	**15**	1128	**25**	723
6	1996	**16**	1071		
7	1822	**17**	1018		
8	1669	**18**	967		

9	1534	**19**	923
10	1457	**20**	886

A scatterplot of salary against pick number is shown below.



Note that there is strong curvature in the plot and fitting a line will not provide a good description of the decreasing pattern. One can simplify the pattern by plotting the logarithm (base 10) of salary against pick number, obtaining the following scatterplot:



Note that this scatterplot can be fitted by two lines – one line for picks 1 through 9 and a second line for picks 10 through 25.

(a) Find the equation of the line fit to pick numbers 1 through 9. (The line will have the form log10(salary) = a + b (pick number).)

(b) Find the equation of the line fit to pick numbers 10 through 25.

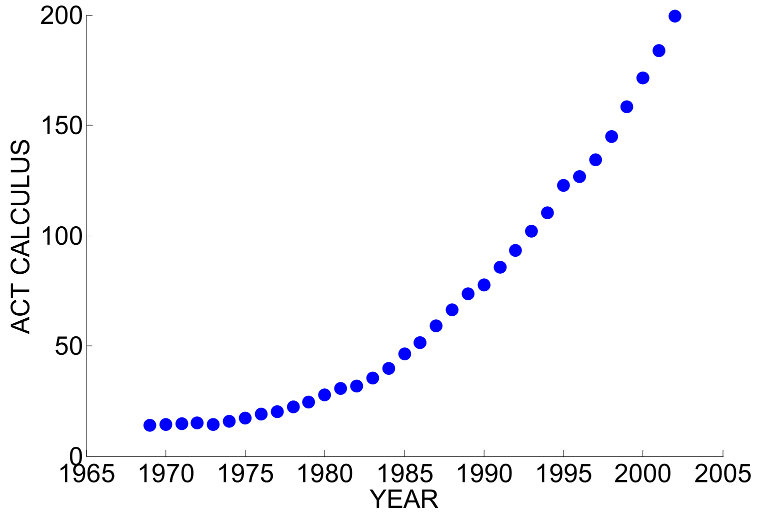
(c) By exponentiation of both sides of the equation, find the prediction equations for salary.

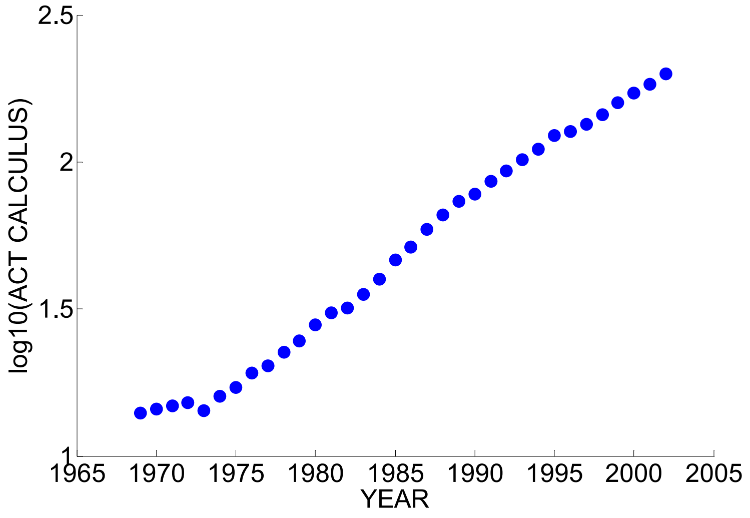
(d) What yearly salary would the regression line predict for the player drafted at number 5? How about for number 20?

(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?

18. **Advanced Placement Tests**

The number of students who take the AP Calculus exam has increased since the introduction of this high school course in ???. The below graph displays the number of students taking this exam (ACT CALCULUS) for each year from 1969 to 2002. There is a strong association between YEAR and number of exam takers, but the association is not linear. If one transforms the y variable by a log (base 10) transformation, then the display ??? shows that the relationship between year and log(students) is approximately of the straight-line type.





a. Fit a median-median line to the (year, log10(students)) data.

b. If a line to these data has the form $\log(STUDENTS) = a + b\,YEAR$,

the model for STUDENTS has the form $STUDENTS = 10^{a+b\,YEAR} = 10^{a}(10^{b})^{YEAR}$. By using this form of the model, on average, what is the percentage increase in the number of students taking the AP calculus test each year?

19. **Brain Weight and Body Weight**

Are the brain weights of animals related to their body weights? In other words, does it require a larger brain to govern a heavier body? The brain weights (in grams) and the body weights (in kilograms) of 26 animals are given in the table below. Since there is large variation in the data, it is convenient to record the log (base 10) brain weight and log body weight. Here you will explore the relationship between brain and body weight three ways.

a. Categorize each body weight as HIGH (larger than the median) or LOW (smaller than the median). Likewise, categorize each brain weight as HIGH (larger than the median) and LOW (smaller than the median). Construct a two-way count table. Using this table, describe the relationship between brain weight and body weight.

b. Consider the log brain weights of the LOW body weight animals, and the log brain weights of the HIGH body weight animals. By using 5-number summaries and parallel boxplots, compare the two groups of log brain weights.

c. Construct a scatterplot of the log body weights (horizontal axis) and the log brain weights (vertical axis). Describe the general pattern in the scatterplot. Are there any points in the scatterplot that seem to be different from the general pattern? (Which animals do these correspond to?)

d. Which way (a, b, or c) do you think is best for describing the relationship between brain and body weights? Why?

	Body wt		Brain	
Species	(kg)	Log_body_wt	wt (g)	Log_brain_wt
Mountain beaver	1.35	0.13	8.1	0.91
Cow	465	2.67	423	2.63
Gray wolf	36.33	1.56	119.5	2.08
Goat	27.66	1.44	115	2.06
Guinea pig	1.04	0.02	5.5	0.74
Diplodocus	11700	4.07	50	1.7
Asian elephant	2547	3.41	4603	3.66
Donkey	187.1	2.27	419	2.62
Horse	521	2.72	655	2.82
Polar monkey	10	1	115	2.06
Cat	3.3	0.52	25.6	1.41
Giraffe	529	2.72	680	2.83

Human	62	1.79	1320	3.12
African elephant	6654	3.82	5712	3.76
Triceratops	9400	3.97	70	1.85
Rhesus monkey	6.8	0.83	179	2.25
Kangaroo	35	1.54	56	1.75
Hamster	0.12	-0.92	1	0
Mouse	0.023	-1.64	0.4	-0.4
Rabbit	25	1.4	12.1	1.08
Sheep	55.5	1.74	175	2.24
Chimpanzee	52.16	1.72	440	2.64
Brachiosaurus	87000	4.94	154.5	2.19
Rat	0.28	-0.55	1.9	0.28
Mole	0.122	-0.91	3	0.48
Pig	192	2.28	180	2.26

20. **Lengths of Movies**

The table below gives the title, year made, and length (in minutes) for 40 movies randomly selected from the *Leonard Maltin's Movie and Video Guide* (1996). One is interested if there is a relationship between the year the movie is made and its running length.

a. Suppose the year of the movie is categorized as “old”, made before 1960, and “new” made in the year 1960 or later. By use of five-number summaries and parallel boxplots, compare the times of the old and new movies.

b. Suppose you also categorize the movies as “short”, 90 minutes or shorter, and “long”, over 90 minutes. Construct a two-way table categorizing movies by year (old or new) and length (short or long). Based on this table, describe the relationship between year and length.

c. Construct a scatterplot of year (horizontal scale) against length (vertical scale). Describe the general pattern of the scatterplot.

Title	Year	Length	Title	Year	Length

The Twinkle in God's Eye	1955	73	Sleep My Love	1948	97
Dakota	1988	97	City Lights	1985	85
Evergreen	1934	90	Hambone and Hillie	1984	89
The Raven	1963	86	The Great Waldo Pepper	1975	107
Hitler--Dead or Alive	1943	70	You Only Live Twice	1967	116
The Ravine	1969	97	The Unholy Three	1930	72
Hold Back Tomorrow	1955	75	The Boogens	1981	95
Lady Dracula	1973	80	Jason's Lyric	1994	119
Kronos	1957	78	Divided Heart	1954	89
Descr A Ticklish Affair	1963	89	The Cockeyed Miracle	1946	81
She Demons	1958	80	The Siege at Red River	1954	81
Okinawa	1952	67	The Stone Boy	1984	93
Bachelor Apartment	1931	77	The Mutineers	1949	60
The Romantic Age	1949	86	Flash and the Firecat	1975	84
Valley of the Dragons	1961	79	The Amazing Transparent Man	1960	58
The Miracle Worker	1962	107	Night of the Dark Shadows	1971	97
Shout at the Devil	1976	119	Windwalker	1980	108
Our Man in Havana	1960	107	House Party 3	1994	94
Falcon Strikes Back	1943	66	Blondie Has Servant Trouble	1940	70
Smash Up: The Story of a Woman	1947	103	Seminole Uprising	1955	74

TOPIC C1: OBTAINING DATA BY SAMPLING

PREVIEW

Everyday when you read the newspaper or your favorite news source from the Internet, you encounter articles that describe conclusions from data obtained from a sample. Here is a portion of an interesting recent article reporting a study looking into American's current use of e-mail.

"Might as Well Face It... We're Addicted to E-Mail"
(reported as a American Online Press Release on May 26, 2005)

DULLES, VA -- Are we a nation obsessed with e-mail? Do we check it first thing in the morning and all day long? Does it keep us up at night? Can we go more than three days without it? AOL recently announced the results of its e-mail addiction survey, which takes a look at the new behaviors and routines that have formed among millions of Americans for whom e-mail is an essential part everyday life.

The survey asked Americans about their e-mail habits, including everything from how often they check personal e-mail at work to whether or not they've ever checked e-mail while in church. The survey found that e-mail users today rely on e-mail as much as the phone for communication, spend about an hour a day on e-mail, and that 77 percent of e-mail users have more than one e-mail account -- all pointing to the fact that e-mail has forever changed the way we communicate.

Signs that we're hooked on e-mail:

We wake up and check it. Forty one percent check e-mail first thing in the morning, 18 percent check it right after dinner, 14 percent say they check e-mail right when they get home from work, and 14 percent do so right before they go to bed.
We can't make it through the night. Forty percent of e-mail users have checked their e-mail in the middle of the night.
We can't live without it! More than one in four (26 percent) say they haven't gone more

than two to three days without checking their e-mail.

We have multiple accounts. Most e-mail users have two or three e-mail accounts (56 percent). The average user has 2.8 accounts.

We check e-mail anytime, anywhere. E-Mail users have checked their e-mail in a variety of locations, including:

-- In bed in their pajamas (23 percent)
-- In class (12 percent)
-- In a business meeting (8 percent)
-- At a Wi-Fi hotspot, like Starbuck's or McDonald's (6 percent)
-- At the beach or pool (6 percent)
-- In the bathroom (4 percent)
-- While driving (4 percent)
-- In church (1 percent)

Survey Methodology: These results are based on online surveys conducted by Opinion Research Corporation with 200 residents per city in the top twenty cities nationwide; respondents were 18 years of age and older. America Online, in partnership with Opinion Research Corporation, conducted online surveys with 4,012 respondents 18 and older in the top 20 cities around the country to measure e-mail usage.

Let's look at this article using the critical perspective introduced in topic D1.

1. What were the main questions addressed by the statistical study?

This study was designed to learn about the e-mail habits of Americans. Specifically the researchers wanted to learn about the frequency of e-mail use, and when and where e-mail messages were received.

2. What data were collected to answer the questions?

If you read the bottom of the article, we learn that a sample of over 4000 individuals was taken, and each individual was asked questions about his or her e-mail usage.

3. How did they collect the data?

How did they obtain these data from these 4000 individuals? It says at the bottom of the article that online surveys were taken. These were surveys conducted from people who were currently on their computers and logged into the Internet.

4. What variables were measured in the data?

We don't know exactly what variables were measured from these individuals. But based on the reported results, it seems that the variables were primarily categorical. There were likely multiple choice responses to questions like "How often do you access your email?" And "Have you accessed e-mail in the middle of the night?"

5. What were the conclusions drawn from this statistical study? Do you believe that the conclusions are valid based on the information provided?

Do we think that the conclusions about Americans' e-mail use are valid based on the data that were collected? Thinking about this carefully, we see that the conclusions are about millions of adult Americans, but these conclusions are based on a small sample of 4000 people. Is it possible to draw valid conclusions based on such a small sample? Even if it is possible to draw valid conclusions from this sample, does it matter *how* the sample is taken?

In this topic, we focus on the general problem of taking a sample. We will see that the process of choosing a suitable sample is not trivial, and so we should be careful about the conclusions in the media drawn from questionable sampling methods.

NCTM Standards

✓In Grades 9-12, all students should understand the differences among various kinds of studies and which types of inferences can legitimately be drawn from each.

✓In Grades 9-12, all students should know the characteristics of well-designed studies, including the role of randomization in surveys.

POPULATION AND SAMPLE

We take a sample to make conclusions about a larger group of individuals that we call the *population*. Usually we think of a population as the entire group of individuals such as the population of citizens in the United States. Here we are using the word in a more restrictive sense. A population is basically the group that we are interested in learning about. If we are interested in the prices of new homes in our Ohio community, then the population would consist of the entire collection of house prices, not of Ohio, but in the small area that we consider our community. If a sample of voters is taken before an American presidential election to make a prediction about the election outcome, what is the population? It wouldn't be all Americans, since people under the age of 18 aren't eligible to vote. The population also wouldn't be all American voters, since typically a large fraction of voters don't actually vote in the election. Actually, the population in this example would be the large group of individuals who actually will be voting in the election.

Read the following two articles that describe conclusions based on sample surveys and think about the population in each article.

ARTICLE 1: "Survey shows 9% decline in teen drug use"

WASHINGTON (AP) — The percentage of Americans using illicit drugs declined slightly last year, though the results were more pronounced for youths, according to a survey released Thursday.

For people ages 12-17, there was a 9% drop in illicit drug use between 2002 and 2004, the federal government announced.

"Today's survey confirms the welcome trend on teen drug use," said John P. Walters, director of the Office of National Drug Control Policy.

Overall, 19.1 million Americans used illicit drugs last year, or 7.9%. The numbers were basically the same for the surveys taken in the previous two years, when about 8% of Americans reported using illicit drugs within the previous month.

The National Survey on Drug Use and Health is an annual survey of close to 70,000 people. It measures drug and alcohol use through several categories, including age, ethnicity and type of drug.

The survey showed that illicit drug use dropped from 11.6% to 10.6% among youths ages 12-17 from 2002 to 2004.

However, binge drinking, which is defined as five or more drinks in one setting, increased in the same age group — from 10.6% to 11.1% from 2003 to 2004.

Officials with the Marijuana Policy Project said federal officials "consistently pay no attention to this alarming situation." The group seeks the regulation of marijuana in the same manner alcohol is regulated.

"Unlike occasional marijuana use, binge drinking can actually kill you," said Bruce Mirken, director of communications for the organization.

Federal officials said they are not ignoring binge drinking. The government is focused on getting the word out to young people that alcohol is not safe, said Leah Young, a spokeswoman for the Substance Abuse and Mental Health Services Administration.

The survey also showed that 70.3 million Americans used tobacco products, which is a slight decline from the previous two years. The percentage of tobacco users among Americans ages 12 and older dropped from 30.4% to 29.2%.

ARTICLE 2: “Heavy workers, hefty price”

USA TODAY, September 11, 2005

Obese employees have much higher weight-related medical expenses and miss more work than their colleagues who maintain a healthy weight, a study shows. It places the annual cost at an additional $460 to $2,500 per obese person — those who are 30 or more pounds over a healthy weig

The price of obesity at a company with 1,000 people on staff is about $285,000 a year in medical costs and absenteeism. Roughly 30% of that cost comes from increased absences among heavyset employees, according to the study in the September/October issue of the *American Journal of Health Promotion.*

This adds to the growing body of research on the high cost of extra pounds. Another study, released this summer, showed that obesity has fueled a dramatic increase in the amount spent on treating medical conditions such as diabetes, heart disease and high cholesterol.

For the latest analysis, economists with RTI International, a non-profit think tank, and the Centers for Disease Control and Prevention, examined two national surveys that track absences and medical information on more than 20,000 full-time employees, ages 18 to 64. Among the findings, adjusted for 2004 dollars:

•Normal-weight men miss an average of three work days a year, compared with five days for men who are 60 or more pounds over a healthy weight.

•Normal-weight women miss about 3.4 days a year vs. 5.2 days for women who are obese, that is 30 to 60 pounds overweight, and 8.2 days for extremely obese, 100 or more pounds over a healthy weight.

•The average medical expenditure for a normal-weight man is $1,351 a year. Men who are 30 to 60 pounds overweight cost $462 more based on added medical costs and absenteeism. Extremely obese men cost $2,027 a year more.

•Average medical expenditures for normal-weight women are $1,956. Women who are 30 to 60 pounds overweight cost $1,372 more when medical costs and missed work are included. Women who weigh 60 to 100 pounds too much cost $2,485 more.

Overall, it's "going to take a concentrated effort to reduce these costs," he says.

Some companies are offering worksite wellness programs or incentives for maintaining a healthy weight or trying to lose weight, such as an extra day off work or paying a greater percentage of those employees' insurance premiums.

What are the populations in the studies described in these two articles? In the first article on teen drug use, several populations are described. The main population discussed in the article is the large group of American youths ages 12-17 and other statements are made regarding the population of all Americans. In the second article on the “price of obesity”, the population would be full-time employees at American companies.

Typically we are interested in some numerical characteristic of the population -- this number is called a *parameter*. If we are interested in the *proportion* of a population, we denote this proportion by the letter p. In the case where we are interested in a population *mean*, we refer to this mean by the Greek letter μ. In the two articles, here are some examples of parameters:

- The proportion p of American youths ages 12-17 that use illicit drugs.
- The mean cost of medical care μ for an overweight employee who works full-time.

Do we know the values of the population proportion p and the population mean μ? No! Usually we won't be able to inspect every member of the population. For example, it would be difficult to contact every American youth between the ages of 12 and 17, and certainly not possible to find the medical care for every American full-time employee.

Although values of parameters such as p and μ are unknown, we do compute numerical quantities from a sample, and we use these quantities to learn about the population parameters. A *statistic* is a quantity computed from a sample. We will let $\hat{p}$ denote a sample proportion and $\bar{x}$ denote a sample mean.

In our articles, it was reported that 10.6% of the American youth ages 12-17 used illicit drugs and the average medical expenditure for a normal-weight man is $1,351 a year. Are these numerical values (10.6% and $1,351) parameters or statistics? In both cases, these values were computed on the basis of samples taken from the population, and so both values are statistics. The statistic $\hat{p}$ = .106 is the proportion of the sampled American youth that use illicit drugs, and $\bar{x}$ = $1351 is the mean medical cost for the sample of normal-weight men.

PRACTICE: POPULATION AND SAMPLE

For the following examples, identify the population and the sample. In addition, define in word the parameter for this problem and the statistic computed from the data.

1. A CBS poll in October 2005 questioned 808 adults nationwide on the origin of life. Of this sample, 67% believed that it was possible to believe in both God and evolution.
2. The rising college debt of students is currently a big concern. In the year 2002, the average debt load for undergraduates had reached $18,900, according to a survey by lender Nellie Mae.

3. The Tempe, Arizona police department surveyed its residents in 2003 to learn about a number of issues, such as satisfaction with the police department, fear of crime, and opinions on traffic safety and quality of life. A total of 853 citizens were contacted by phone and asked to complete a survey. One of the findings was that 70% of the respondents rated the safety of their neighborhoods as high or very high.

4. The Northeastern University Institute on Race and Justice examined more than 1.3 million traffic tickets from April 2001 through June 2003. One interesting finding was that in Boston, men received 74 percent of the tickets, although they account for 47 percent of the driving age population. (Note: This example is different from the first three examples. Is "74 percent" a measurement of the sample or the population?)

EXAMPLES OF BAD SAMPLING METHODS – 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.

This is an illustration of a sampling problem. Here the population is all American adults and the goal is to learn something about this population by means of the sample who called-in to express their opinion about Elvis. But is this collected data representative of the population? In other words, do we think that 56% is an accurate reflection of beliefs of all American adults on this issue?

There are some obvious flaws with this sample survey. People who call the 1-900 number probably are very interested in Elvis and likely have a strong opinion regarding the likelihood of him being alive. In contrast, people who don't respond by calling may have a limited interest in Elvis and have a very different opinion about the alive/dead issue. For these reasons, the opinions of the people in this sample may be a poor reflection of the opinion of the population of all American adults on this issue.

President Landon?

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.

How could the *Literary Digest*'s prediction be so wrong? Let's look carefully at the manner which this magazine took their sample. Questionnaires were sent to people who were listed in phone books and vehicle registration lists. This doesn't seem to be a bad way of sampling – after all, didn't most people own phones and cars? Actually, in 1936, the answer would be no. At this time, phones and cars were relatively expensive and owned primarily by the wealthier segment of the American population. This sampling method was missing the large group of Americans who didn't own phones and cars. Also the Republican Alf Landon appealed more to the wealthy Americans and Franklin Roosevelt was more popular among the poorer segment of the population. Now it is clear how the *Literary Digest* overestimated the support for the Republican candidate.

Both the Elvis and *Literary Digest* examples 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.

PRACTICE: BAD SAMPLING METHODS

1. In spring 2006 there was a general concern that some of the great home run hitters in the previous ten years used steroids. On the espn.com website, readers were asked the question "Who do you consider baseball's single-season home run leader?" Of 41,070 voters that responded, 50% voted for Roger Maris' season of 61 home runs in 1961 and only 33% voted for Barry Bonds' record-breaking season of 73 home runs in 2001.

(a) What is the relevant population in this example?

(b) Do you believe that 41,070 is a large enough sample to learn about the proportion of all baseball fans who would vote for Maris' accomplishment?

(c) Explain how this is a voluntary response survey.

(d) Explain how this particular sampling method could be viewed as biased in this example.

2. Back in the 1970's a young couple wrote to the advice columnist Ann Landers undecided whether or not to start a family. This couple requested Ms. Landers to ask her readers the question "If you had it to do all over again, would you have children?" Over 10,000 responses were received and 70 percent wrote "No. If I had to do it all over again, I would not have children."

(a) What is the population in this example?

(b) Is this an example of a voluntary response survey? Explain.

(c) Explain how this particular sampling method could be biased. Do you believe that the actual proportion of American parents who would answer No is smaller or larger than 70 percent?

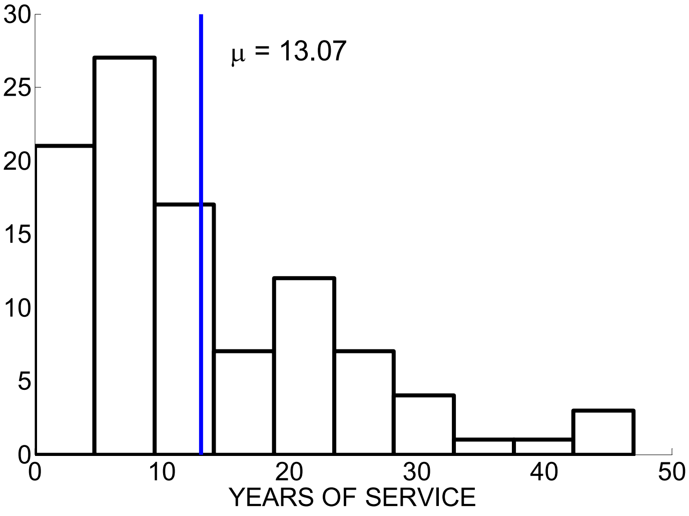
SIMPLE RANDOM SAMPLING

Suppose you are doing a report on the United States government and you are focusing your report on the people who comprise the U.S. Senate. One natural question that comes to mind is "How long are the tenures of the U.S. senators?" You are aware that a single term of a senator is six years, but it is possible for a senator to be elected for a few or many terms. What is a typical number of years of service for a current senator?

In this situation, the population of interest is the 100 people who are currently in the U.S. Senate. The table below shows the population. For each senator, we show the party (Democrat, Republican, or Independent), the gender, the state that the senator represents, and the number of years of service. A representative number of years of service can be measured by μ, the mean value of this population. In the figure below, a histogram of the population of senators is displayed and the value of the mean μ = 13.56 years is shown by a vertical line.

No	Name	Party	Gender	State	Yrs	No	Name	Party	Gender	State	Yrs
1	Akaka	Dem	male	Hawaii	16	51	Inouye	Dem	male	Hawaii	43
2	Alexander	Rep	male	Tennessee	3	52	Isakson	Rep	male	Georgia	1
3	Allard	Rep	male	Colorado	9	53	Jeffords	Indep	male	Vermont	17
4	Allen	Rep	male	Virginia	5	54	Johnson	Dem	male	South Dakota	9
5	Baucus	Dem	male	Montana	28	55	Kennedy	Dem	male	Massachusetts	44
6	Bayh	Dem	male	Indiana	7	56	Kerry	Dem	male	Massachusetts	21
7	Bennett	Rep	male	Utah	13	57	Kohl	Dem	male	Wisconsin	17
8	Biden	Dem	male	Delaware	33	58	Kyl	Rep	male	Arizona	11
9	Bingaman	Dem	male	New Mexico	23	59	Landrieu	Dem	female	Louisana	9
10	Bond	Rep	male	Missouri	19	60	Lautenberg	Dem	male	New Jersey	3
11	Boxer	Dem	female	California	13	61	Leahy	Dem	male	Vermont	31
12	Brownback	Rep	male	Kansas	10	62	Levin	Dem	male	Michigan	27
13	Bunning	Rep	male	Kentucky	7	63	Lieberman	Dem	male	Connecticut	17
14	Burns	Rep	male	Montana	17	64	Lincoln	Dem	female	Arizona	7
15	Burr	Rep	male	North Carolina	1	65	Lott	Rep	male	Mississippi	17
16	Byrd	Dem	male	West Virginia	47	66	Lugar	Rep	male	Indiana	29
17	Cantwell	Dem	female	Washington	5	67	Martinez	Rep	male	Florida	1
18	Carper	Dem	male	Delaware	5	68	McCain	Rep	male	Arizona	19
19	Chafee	Rep	male	Rhode Island	7	69	McConnell	Rep	male	Kentucky	21
20	Chambliss	Rep	male	Georgia	3	70	Menendez	Dem	male	New Jersey	0
21	Clinton	Dem	female	New York	5	71	Mikulski	Dem	female	Maryland	19
22	Coburn	Rep	male	Oklahoma	1	72	Murkowski	Rep	female	Arkansas	3
23	Cochran	Rep	male	Mississippi	28	73	Murray	Dem	female	Washington	13
24	Coleman	Rep	male	Minnesota	3	74	Nelson	Dem	male	Florida	5
25	Collins	Rep	female	Maine	9	75	Nelson	Dem	male	Nebraska	5
26	Conrad	Dem	male	North Dakota	19	76	Obama	Dem	male	Illinois	1
27	Cornyn	Rep	male	Texas	3	77	Pryor	Dem	male	Arizona	3
28	Craig	Rep	male	Idaho	15	78	Reed	Dem	male	Rhode Island	9

29	Crapo	Rep	male	Idaho	7	79	Reid	Dem	male	Nevada	19
30	Dayton	Dem	male	Minnesota	5	80	Roberts	Rep	male	Kansas	9
31	DeMint	Rep	male	South Carolina	1	81	Rockefeller	Dem	male	West Virginia	21
32	DeWine	Rep	male	Ohio	11	82	Salazar	Dem	male	Colorado	1
33	Dodd	Dem	male	Connecticut	25	83	Santorum	Rep	male	Pennsylvania	11
34	Dole	Rep	female	North Carolina	3	84	Sarbanes	Dem	male	Maryland	29
35	Domenici	Rep	male	New Mexico	23	85	Schumer	Dem	male	New York	7
36	Dorgan	Dem	male	North Dakota	14	86	Sessions	Rep	male	Alabama	9
37	Durbin	Dem	male	Illinois	9	87	Shelby	Rep	male	Alabama	19
38	Ensign	Rep	male	Nevada	5	88	Smith	Rep	male	Oregon	9
39	Enzi	Rep	male	Wyoming	9	89	Snowe	Rep	female	Maine	11
40	Feingold	Dem	male	Wisconsin	13	90	Specter	Rep	male	Pennsylvania	25
41	Feinstein	Dem	female	California	14	91	Stabenow	Dem	female	Michigan	5
42	Frist	Rep	male	Tennessee	11	92	Stevens	Rep	male	Arkansas	38
43	Graham	Rep	male	South Carolina	3	93	Sununu	Rep	male	New Hampshire	3
44	Grassley	Rep	male	Iowa	25	94	Talent	Rep	male	Missouri	3
45	Gregg	Rep	male	New Hampshire	13	95	Thomas	Rep	male	Wyoming	11
46	Hagel	Rep	male	Nebraska	9	96	Thune	Rep	male	South Dakota	1
47	Harkin	Dem	male	Iowa	21	97	Voinovich	Rep	male	Ohio	7
48	Hatch	Rep	male	Utah	29	98	Warner	Rep	male	Virginia	27
49	Hutchison	Rep	female	Texas	13	99	Wyden	Dem	male	Oregon	10
50	Inhofe	Rep	male	Oklahoma	12	0	Vitter	Rep	male	Louisana	1



Suppose we wish to learn about the population by taking a sample of five senators from the table, collecting their years of service, and then using the sample mean $\bar{x}$ as an intelligent guess at the mean μ. How should we take this sample?

Scanning the table, your author chooses the following five senators that are familiar to him.

DeWine, Voinovich, Bunning, Kennedy, Lugar

The first two senators are from the author's home state, Bunning is a former baseball pitcher, Kennedy is very famous, and Lugar is a senator from a neighboring state. Collecting the years of service from these senators (11, 7, 7, 44, 29), we obtain a sample mean of

$$\bar{x} = \frac{11+7+7+44+29}{5} = 19.6 \text{ years.}$$

Is this a reasonable guess at the mean of the years of service for all senators? No – actually your author has illustrated a bad way of taking a sample of senators. How were these senators selected? They were chosen either because they came from a local state or they were known to the author. By selecting senators in this manner, one tends to sample well-known people who have spent a large number of years in the Senate and ignore senators who have been recently elected to office. Samples chosen in this example that are "familiar" to the sampler will be biased. That means if this procedure of choosing familiar people is done repeatedly, then the sample means of years of service will tend to be larger than the population mean, and so one gets a distorted picture of the years of service of all senators.

In practice, it is desirable to choose a sample in a way that will not be biased in a particular direction. We want to use a sampling procedure that will reflect the population in the long run. That means that if we use a good sampling procedure repeatedly to generate many samples, then the collection of samples will tend to reflect the population that we are interested in.

One type of good sample is called a simple random sample or SRS for short. This sample is taken in such a way that every possible sample of the same size will be equally likely to be the sample that is chosen.

We can select a SRS by the use of any device that will simulate random numbers such as the table of random digits in the book. The method of taking this sample is described as follows.

1. [Label the population] We first assign a number label to each element of the population. In this example, our population consists of 100 senators, so we can assign a unique two-digit label to each person. If the population consisted of more than 100 elements, then more digits would have to be used.
2. [Simulate a set of labels] Here we are interested in taking a random sample of five senators without replacement from the population. We generate five groups of two-digits from a random digit table or computer package. For the sample, the author obtained the random digits 78, 28, 21, 02, and 50.
3. [Find your sample] Next one finds the population members with these labels. Looking back at the population, we see that these labels correspond to the senators Reed, Craig, Clinton, Alexander, and Inofe.
4. [Compute the estimate of interest from the sample] We are interested in learning about the years of service of the senators. So we collect the service lengths for the sample of five and compute the sample mean $\bar{x}$. Here the lengths are 9, 15, 5, 3, and 12, with a mean of $\bar{x}$ = (9+15+5+3+12)/5=8.8.

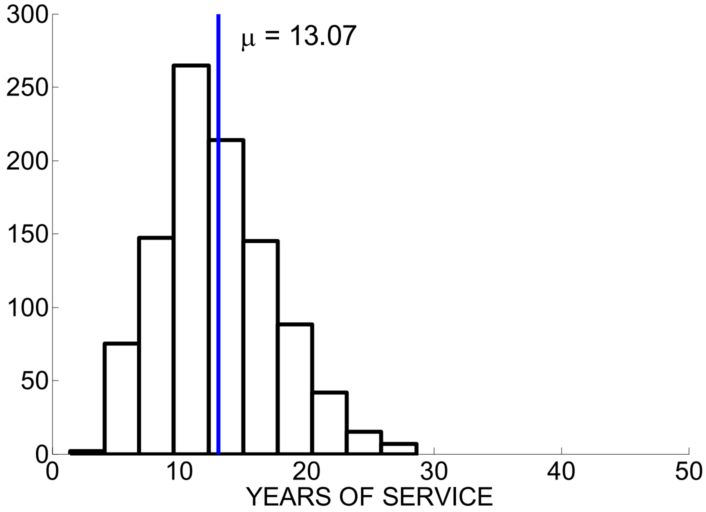
We repeat this procedure in the below table. The table shows the random digits generated in each sample, the sample collected, and the value of the sample mean.

Sample	Random digits	Sample (years of service)	Sample mean $\bar{x}$
1	78 28 21 02 50	Reed (9), Craig (15), Clinton (5), Alexander (3), Inofe (12)	(9+15+5+3+12)/5=8.8
2	16 94 82 83 64	Byrd (47), Talent (3), Salazar (1), Santorum (11), Lincoln (7)	(47+3+1+11+7)/5=13.8
3	00 67 64 24 14	Vitter (1), Martinez (1), Lincoln (7), Coleman (3), Burns (17)	(1+1+7+3+17)/5=5.8

4	34 70 98 15 2	Dole (3), Menendez (0), Warner (27), Burr (1), Alexander (3)	(3+0+27+1+3)/5=6.8
5	91 01 73 97 64	Stabenow (5), Akaka (16), Murray (13), Voinovich (7), Lincoln (7)	(5+16+13+7+7)/5=9.6

What is so special about this method of taking a sample? Does this procedure guarantee that my sample will be representative of the population? No. It is possible by chance that my sample will look very different from the population. For example, notice from the table that Sample 3 selected senators with only 1, 1, 7, 3, and 17 years of experience and the sample mean $\bar{x}$=5.8 is much smaller than the population mean μ = 13.07. So it is possible that a particular SRS will not be a good representative of the population.

But a SRS has a desirable property if this same procedure is used to select many samples. The author had the computer take many SRSs of size five from the population of senators. For each sample, the sample mean $\bar{x}$ was computed, and after all the sampling was done, we had a collection of 1000 sample means. The below figure displays a histogram of the sample means and the population mean μ is displayed with a vertical line. Note that we see much variation in the values of the sample means – it is possible that one could compute a sample mean as small as 5 years or as large as 25 years. This confirms the earlier statement that the sample need not look like the population. But note that the distribution of sample means is centered about the value of the population mean. This means that the SRS procedure will produce sample means that will look like the population mean, on average.



PRACTICE: SIMPLE RANDOM SAMPLING

Suppose you are interested in learning about the population proportion p of U.S. senators who are female.

1. The population is already labeled with the two-digit numbers 00 through 99. Using a random digit table, take a SRS of size 8 from this population. Complete the table.

Two-digit labels from the table	
Names of the sampled senators	

2. From the sample, compute the sample proportion of female senators $\hat{p}$.

3. Repeat this process nine more times – each time, take a SRS of size 8 from the population and compute the sample proportion of female senators $\hat{p}$. Write down the ten values of $\hat{p}$ from the ten samples.

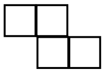
4. Construct a dotplot of the values of $\hat{p}$.

5. Here the actual value of the population proportion p is .14. Looking at the collection of sample proportions $\hat{p}$ you computed, what is the connection between this collection and the value of p?

HANDS-ON ACTIVITY: RANDOM RECTANGLES

DESCRIPTION: Suppose you live in a community with 100 homes. You are interested in finding out the mean size of these homes but it is impossible to collect the sizes for all 100 homes. So you instead decide to take a sample of five homes, compute the mean of the sizes of these homes, and use this as an estimate of the average size for all 100 homes. Does it matter how you take this sample? In this activity, we will compare taking "selected" samples with "random samples." Hopefully, we will see that it can be difficult to self-select a sample that will tend to be representative of the population.

MATERIALS NEEDED: None.

On the following page, we see the 100 homes that are laid out in a 10 by 10 grid. We measure the size of a house by the number of squares in the house rectangle. (For example, the size of the home  is 4.)

1. [Selected sampling]. First, circle five homes that you believe are representative of the sizes for the 100 homes.
2. Find the sizes of these five homes and record them in the below table. Find the mean size of these homes.

Home	1	2	3	4	5
Size					

Mean size (for selected sample) = ____________

3. [Random sampling.] A random sample is a sample chosen in such a way that all possible samples have the same chance of being chosen. We can choose a random sample in this example by the use of a 10-sided die. (Equivalently, a table of random digits can be used.) To select a home,

you choose two whole numbers between 0 and 9 – the first number corresponds to the row, and the second number, the column, of the grid

you find the home in the particular row and column in the grid

For example, suppose the rolls of the 10-sided die are 5 and 3. You find the home in row 5 and column 3 of the grid – we see that the size of this home is 1.

Using this method, take a random sample of five homes. Put the sizes of these homes in the table and compute the mean.

Home	1	2	3	4	5
Size					

Mean size (from random sample) = ____________

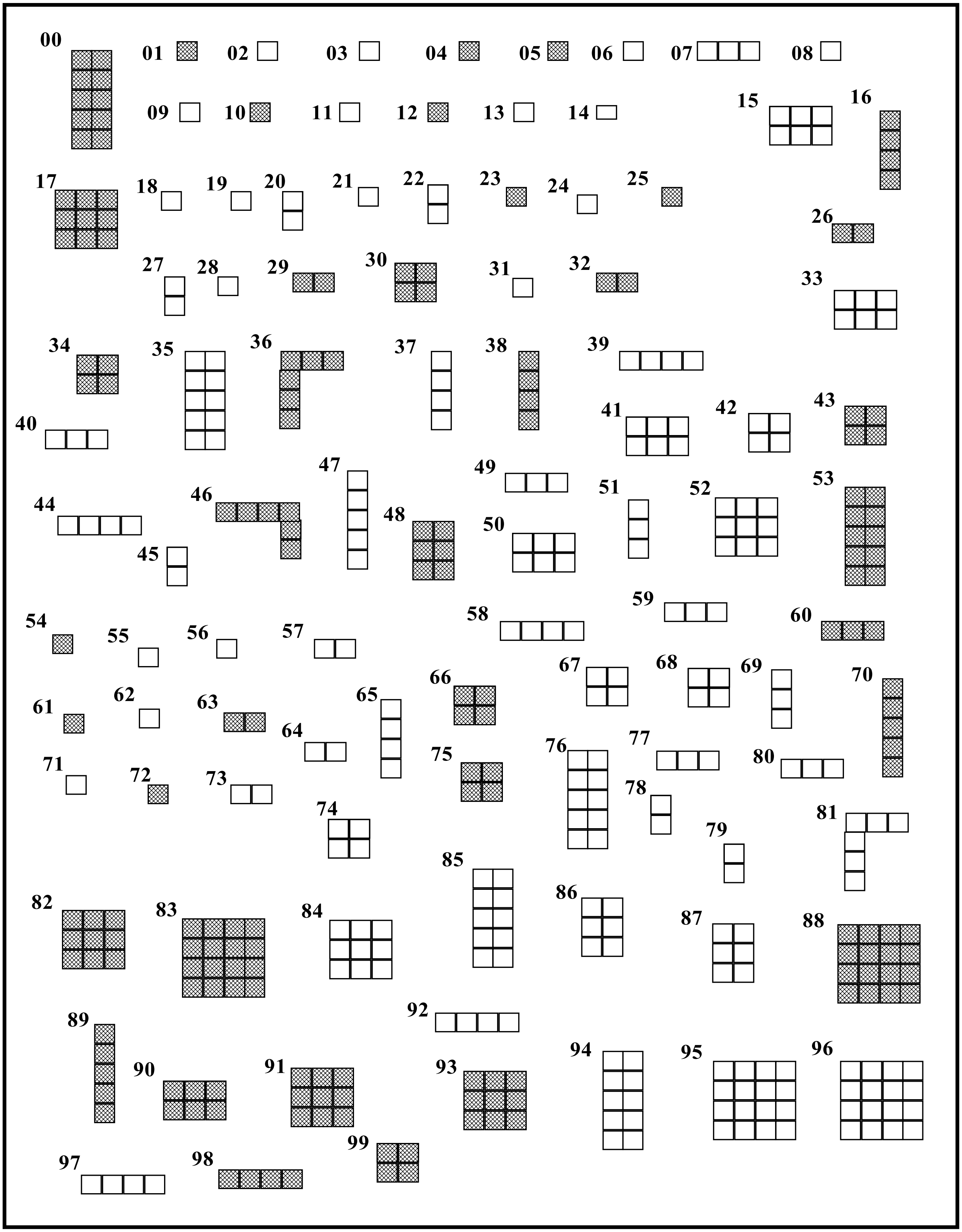
4. Collect all of the means from the "selected" samples taken by the students in the class. Construct a graph (stemplot or dotplot) of these sample means. Compute the mean and quartile spread of these sample means.

5. Likewise, collect all of the means from the random samples that are taken. Construct a graph and compute the mean and quartile spread of these sample means.

6. Compare the two batches of selected and random sample means and comment on any differences you can find.

7. The mean size of all 100 homes is actually 4.2 squares. Looking back at your work from 4. and 5., which type of sampling (selected or random) tends to give you sample means that are close to the population mean?

8. Suppose one of your friends is interested in learning about the mean number of hours per week that undergraduate students spend on homework. She plans on asking five of her friends that she believes represent the student body. Explain to your friend why there are problems with this particular sampling strategy.


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

FATHOM ACTIVITY – BIASED SAMPLING OF RECTANGLES

This is a continuation of the Random Rectangles activity where we let Fathom take random samples of rectangles from the population. By only selecting rectangles from a portion of the page, we illustrate the concept of bias. Also we see how the choice of sample size will affect the distribution of the sample means of the samples we select.

PART A: Setup of the Fathom document

1. Import the dataset `randomrectanglenew.txt` into Fathom. This collection has three attributes:

Number = the number assigned to the rectangle

Area = the area of the rectangle (number of small squares in the rectangle)

Dark = whether the rectangle is dark (1) or light (0)

2. Define two sliders

- Define the slider lo_num that goes between 0 and 50 with only integer values.
- Define the slider sample_size that goes between 1 and 10 with only integer values.

3. Restrict the collection to the rectangles with labels low_num or higher
Select the Collection and choose Add Filter from the Object menu.
In the formula box, type "number >= low_num".

4. Take a random sample of size sample_size from the collection.
Select the collection and choose Sample Cases from the Collection menu.
Inspect the Sample of Collection. Select the Until condition and type "count(number)=sample_size" in the formula box.

5. For each sample of rectangles, compute the sample mean of the areas.
Inspect the Sample of Collection and select the Measures tab.
Define a new measure called "samplemean" that is defined to be "mean(area)"

6. Take many samples and construct a histogram and summary table of the collection of sample means.
Select the Sample of Collection and choose Collect Measures from the Collection menu.
Inspect the Measures collection. Replace existing cases and collect 500 measures.

Construct a histogram of the sample means.
Find the mean and standard deviation of the sample means.

PART B: Running the Fathom simulation

In this example the mean of the areas of the population of rectangles is equal to μ = 4.2. We illustrate learning about this population mean by the use of unbiased and biased samples of different sample sizes.

1. Unbiased sampling with a small sample size.

Set the value of the slider variable low_num to 0 and the value of the slider samplesize to 4. So we taking random samples of size 5 from the entire population. Collect 500 measures.

Look at the distribution of sample means $\bar{x}$ from these 500 samples. What is the general shape of this distribution? What is the mean of this distribution? We say that a sample mean is unbiased if the average value of the sample mean over many samples is equal to the population mean. Does the sample mean appear to be unbiased? Why?

2. Unbiased sampling with a large sample size.

Keep the value of the slider variable low_num to 0 and change the value of the slider samplesize to 10. Collect 500 measures.

We are now taking samples of size 10 (instead of 5). Does this procedure still seem to be unbiased? Why? Compare the standard distribution of the $\bar{x}$'s (for n = 10) with the standard deviation of the $\bar{x}$ (for n = 5). Which distribution (n = 5 or n = 10) has the smaller standard deviation? This means that for n = 10, the sample mean $\bar{x}$ tends to be ________ to the population mean μ than for n = 5. (Fill in the blank.)

3. Biased sampling with a small sample size.

Now change the slider value low_num to 50 and the value of the slider samplesize to 4. Here we are taking random samples of size 5 only from the rectangles numbered 50 or above. Collect 500 measures.

What is the shape of the distribution of sample means? What is the mean value of $\bar{x}$? Based on this mean value, is this procedure unbiased? If it is not unbiased, what is the size of the bias?

4. Biased sampling with a large sample size.

Keep the slider value low_num to 50 and change the value of the slider samplesize to 10. We are still sampling only from the rectangles numbered 50 or higher, but we are now taking samples of size 10. Collect 500 measures.

By looking at the mean value of $\bar{x}$, decide if the procedure is biased, and if so, determine the size of the bias. Comparing the distribution of sample means $\bar{x}$'s with part 3, what is the difference in the sampling distributions?

5. Summing up.

(a) Generally, how does the sampling distribution of $\bar{x}$ change as you take a larger sample size? Is this true for biased and unbiased sampling?

(b) How can you tell if a method of sampling is biased? How do you measure the size of the bias?

PRACTICAL CONCERNS IN SAMPLING

We have illustrated the advantages of taking a simple random sample. This method will eliminate bias and we can control the variability of the sample results by adjusting the sample size. However, it can be difficult to administer SRS's in practice. Indeed, sampling in the real world is a more complex and less reliable than taking a SRS by use of a random digit table. Here we illustrate some potential problems in sampling.

Suppose you are interested in conducting a survey at your school to learn about the general satisfaction of students about the technology services provided on campus. You plan on calling a sample of students randomly selected from the campus phonebook and asking each student a series of questions related to the technology services.

One issue to be concerned with is the possibility of *undercoverage* or systematically missing a segment of the population in your sampling. Here you are sampling students who have telephones in their dormitory or apartment. Given the popularity of cell phones, perhaps a particular segment of the population doesn't use the phones present in their residences. This means that your sampling will exclude the students who exclusively use cell phones. If this is a real problem, then you should think of an alternative sampling method that can reach students with cell phones.

Another issue of concern is the likely possibility that many students will not answer your phone calls. This problem of *nonresponse* is an important issue, especially when surveys are sent by snail-mail. Nonresponse can produce a bias in your sampling procedure, especially when the opinions of responders may be different from the opinions of people who are not responding. For example, if students have having problems with technology, then they might have problems with their phone services. In this case, their dissatisfaction with technology may be positively related to their nonresponse, and so the opinions of responders may not reflect the general campus support of technology.

In situations where one wishes to ask about potentially sensitive topics, then the issue of *response error* is relevant. Suppose you were interested in the proportion of students who smoked marijuana on campus. If students were directly asked the question "Do you smoke marijuana?" some students may be uncomfortable in giving an honest response. In this case, there could be a bias in your sampling procedure and the proportion of "yes's" in your sample would likely be smaller than the actual proportion of marijuana users in the population. In the case where a response error is possible, then perhaps the question could be rephrased so it would not invade the student's privacy.

PRACTICE: PRACTICAL CONCERNS IN SAMPLING

In each situation below there is a problem with the sampling procedure. Explain if the problem is due to undercoverage, nonresponse, or response error. Also describe if there is a bias in the procedure, and how you think the sample result will be different from the population.

1. You are interested in learning if a particular community is supportive of the upcoming school levy. You plan on visiting people's homes during a morning during the week and asking each person if he or she is supportive of the levy.
2. You wonder if a person's support of the levy has some relationship with the household income. You will make phone calls, asking each person about his/her household income and his/her opinion about the school levy.

3. Since you are reluctant to ask people the school level question directly, you plan on leaving surveys in people's mailboxes with stamped envelopes and asking the people to mail back their responses.

Errors in sampling – bias and lack of precision (variability)
The population size doesn't matter.
To decide the goodness of a sample ask "What would happen if we took a large number of samples from the same population?"

EXERCISES

EXERCISE 1. Read the following article that appeared in the San Francisco Chronicle on April 30, 2002 and answer the questions that follow the article.

"Poll Finds Disparity in Schools"

A new Harris poll of 1,071 California teachers released today provides fresh ammunition for critics who complain that the state's public education system is divided between haves and have-nots.

Poor minority students are four times more likely than their wealthier white counterparts to experience high teacher turnover, and twice as likely to have old or insufficient textbooks at school, according to the poll.

Ethnic background and family circumstances also are strong predictors of whether students will encounter broken plumbing and even mice, rats and cockroaches in their classrooms, according to the poll of teachers at 1,018 schools throughout the state.

"The evidence demonstrates that there are two different environments in the state's schools," pollster Lou Harris said. "One is substantially more conducive to learning, while the other lags far behind."

The Harris Poll ranked each of the 1,018 schools in its poll according to its "risk" level. Those with the most students who were poor and spoke little English are at highest risk.

Two groups of teachers' responses were then compared: those from the 51 percent of schools with the least risk, and those from the 20 percent of schools with the greatest risk.

The poll found these key facts about the higher-risk schools:

-- Teachers were four times as likely to say turnover was a "serious problem."
-- They were four times as likely to have little parental involvement.
-- They were 2.7 times as likely to have old textbooks or too few of them.

-- They were twice as likely to have inadequate physical conditions. However, the poll found they were less likely to have overcrowded classrooms -- 15 percent of the higher-risk schools were crowded, compared with 18 percent of the larger group of schools.

(a) What was the population in this survey? What was the sample?

(b) Give examples of two variables measured in this survey.

(c) Is there any information in the article about the manner in which the sample was taken?

(d) The teachers from the "high risk" schools were 2.7 times more likely to have old textbooks than teachers from the "low risk" schools. Is 2.7 an example of a parameter or a statistic?

(e) Do you believe that the conclusions of the article are valid? What questions might you ask of the pollster that would help you judge if the conclusions from the study were valid?

EXERCISE 2. Read the following article that appeared in the San Francisco Chronicle on April 30, 2002 and answer the questions that follow the article.

"Survey Reveals Majority of College Students Breaking Even or Flat Broke While in School"

January 3, 2005 — SAN JOSE, Calif. —The plight of the college student continues. Half.com by eBay, The World's Online Marketplace®, commissioned a study, which finds students struggle financially during the school year with 55 percent saying they are either broke or just breaking even. More than 70 percent of students believe their job takes away from their study time. Shopping wisely, finding savings and making money through non-traditional channels is a growing trend demonstrated by nearly 40 percent of students who sell textbooks online.

"With only 36 percent of students describing their financial status as secure, students are always looking for ways to save and earn money on the necessities," said Mike Aufricht, vice president and general manager of Half.com. "Many students sell their previous semester's textbooks to help pay for their upcoming semester's books which helps lower the total expenditure delegated towards textbooks. Half.com provides students an opportunity to both save and make money on textbooks."

The Survey.com study was conducted in December 2004 with 500 college students between the ages of 19 and 25 as participants.

(a) What was the population in this survey? What was the sample?

(b) Give examples of two variables measured in this survey.

(c) Is there any information in the article about the manner in which the sample was taken?

(d) Is the number 55 percent in the article a parameter or a statistic?

(e) Do you believe that the conclusions of the article are valid? What questions might you ask of the pollster that would help you judge if the conclusions from the study were valid?

EXERCISE 3. Suppose you are interested in learning about the proportion of students at your school who are satisfied with the current parking on campus. You wish to take a survey of 100 students to learn about the general satisfaction about parking.

(a) Describe the population and the parameter of interest.

(b) Suppose you decide to sample by calling 100 people who have recently communicated with the parking office on campus. Do you believe this sample would represent the population of interest?

(c) Suppose instead that you decide to survey students by meeting them at the doors of the Student Union. Describe the pros and cons of this particular sampling method.

(d) Is it possible to take a SRS by labeling all students and then taking a sample by use of the random digit table? Explain.

EXERCISE 4. For each of the following situations, identify the population and the sample.

(a) The computer manufacturer Dell is interested in learning about the proportion of laptop computers manufactured in the past five years that need repair within the warranty period. From their records, they select 1000 computers and find that 228 needed some repair in the warranty period.

(b) A company is interested in test marketing a new brand of mix for baking bread in Columbus, Ohio. They wish to learn about the proportion of households in Columbus that currently use a bread machine. They take a phone survey of 120 households and find that 30% have used their bread machine in the last month.

(c) A ski resort is thinking about introducing a new learn-to-ski program for kids under 7 years old. For the planning of this program, they would like to estimate the number of children between ages 4 and 7 who would attend this ski resort in the next year. They take a survey of 1000 families who went skiing at this report and found that 15% of them has one or more child of ages between 4 and 7.

EXERCISE 5. Suppose there is an election for the president of your student body and there are two candidates Joe and Sally. You take a sample of 100 students by email two weeks before the election and you find that 70% of these students prefer Sally. The election is held and Joe wins by a large margin.

(a) What the relevant population in this example? Was a sample taken from this population?

(b) Based on your answer to part (a), can you explain why the survey results were poor in predicting the actual winner of the election?

EXERCISE 6. Suppose you need to select a sample of five students from your class to represent the class on a committee. Here are the names of the twenty students in your class:

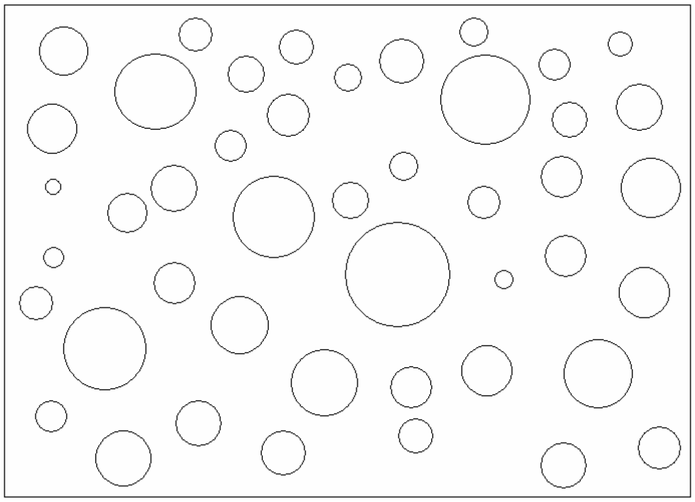
Angie, Jennifer, Tom, Carl, Natalie, Brian, Joshua, Taylor, Steven, Aaron
Bethany, Brittany, Ashley, Nicole, Ben, Craig, Laura, Joy, Gale, Jan

(a) Give a number label to each student in the class.

(b) By use of the table of random digits, take a SRS of size five. Write down the labels selected and the people in your sample.

EXERCISE 7. Suppose you are interested in estimating the mean diameter of circles shown in the diagram below.

(a) Assign to each circle a label and then use the table of random digits to select a sample of four circles. Measure the diameter of each circle you select to the nearest half of a centimeter and find the sample mean of your diameters.



(b) Repeat this process four more times, obtaining a total of five sample means of size five. Construct a dotplot and find the mean and standard deviation of the sample means.
(c) Now take five samples, each of size 10, from the collection of circles. For each sample, compute the sample mean.
(d) Construct a dotplot and find the mean and standard deviation of the sample means.
(e) Compare the distribution of sample means of size 5 from part (b) with the distribution of sample means of size 10 from part (d). Which distribution has the smaller spread? Is this what you would expect? Why?

EXERCISE 8. In each part, identify the population, the sample, and the variables measured. Each sampling situation contains a source of likely bias. State the reason for the bias and the direction of the likely bias. (That is, in what way will the sample conclusions be different from the truth about the population?)
(a) Your congressman is interested in the opinion of her constituents on a proposed bill requiring intelligent design to be taught in the public schools. Her staff reports that letters on this issue have been received by 200 constituents and 120 support the teaching of intelligent design.
(b) Suppose the Toledo police department is interested in learning about the attitudes of the black residents regarding police service. A questionnaire is prepared and a sample of 250 addresses in a predominantly black neighborhood is chosen. Police officers go to each address to ask the questions of an adult living there.

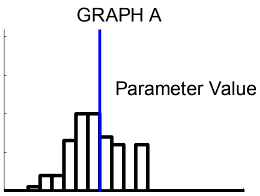
EXERCISE 9. Suppose you take repeated samples from a population with a given population mean μ. The four histograms show the sampling distribution of the sample mean in four situations. Some of the situations correspond to simple random samples and some correspond to other types of samples. The value of the population mean is shown by a vertical line.

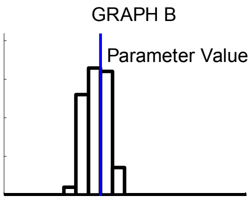
(a) Which graph illustrates sampling with small variability with no bias?

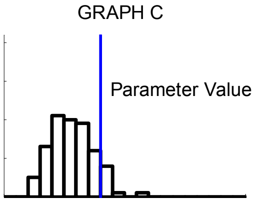
(b) Which graph illustrates sampling with small variability with a bias?

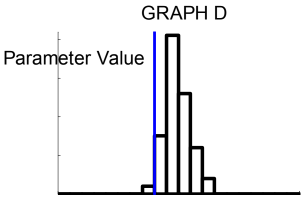
(c) Which graph illustrates sampling with large variability with no bias?

(d) Which graph illustrates sampling with large variability and a bias?









EXERCISE 10. A student is interested in learning about the average amount of TV watched per day by the people who live in her community. For each of the sampling methods, discuss any potential problems with the method and describe how the results may be biased.

(a) The student plans to have face-to-face interviews with 100 people that she finds at the local shopping mall.

(b) The student plans to have an internet poll sponsored by the local newspaper. On the front page newspaper's web site, a question will be posed regarding the quantity of TV watching.

(c) The student plans on making phone calls to 100 households randomly selected from a phonebook.

EXERCISE 11. [about movie ratings on imbd.com]

User ratings for Crash (winner of best picture oscar in 2005)

Votes	**Percentage**	**Rating**
18258	38.9%	10
11427	24.4%	9
7600	16.2%	8
3469	7.4%	7
1585	3.4%	6
884	1.9%	5
542	1.2%	4
494	1.1%	3
406	0.9%	2
2238	4.8%	1

46903

TOPIC C2: OBTAINING DATA BY EXPERIMENTS

Outline

Examples from newspapers and web sites:

USA Today headlines: March 5, 2006

Mentoring helps teens cope

Experts mixed on diet soda

Lift weights, attack belly fat

Hispanics lacking in education

Class too easy for dropouts?

NCTM Standards

✓In Grades 9-12, all students should understand the differences among various kinds of studies and which types of inferences can legitimately be drawn from each.

✓In Grades 9-12, all students should know the characteristics of well-designed studies, including the role of randomization in surveys.

AN APPLE A DAY: DIFFERENT WAYS OF COLLECTING DATA

We have all heard the expression “an apple a day keeps the doctor away”. Is it true? Are there health benefits associated with a diet of apples? You plan on collecting some data to shed some light on this issue. Here are four possible different designs of this study.

Study 1. You take a random sample of individuals and identify which do and do not eat apples regularly. You follow these individuals for six months and observe who require a visit to the doctor and who does not.

Study 2. You take a random sample of physicians and ask each if they have noticed any health benefits from eating apples.

Study 3. You take a random sample of individuals, randomly assign half to eat an apple a day for the next six months and the other half not to, and then see who require a visit to the doctor and who does not.

Study 4. You recall your Uncle Charlie who loved apples and was never sick a day in his life, while Uncle Dave despised apples and was often ill.

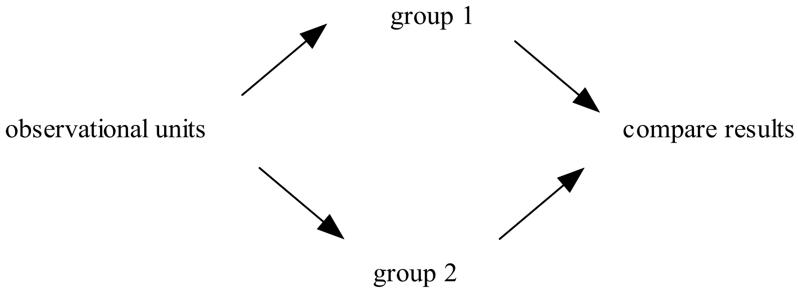
These four examples illustrate four types of studies:

- *Anecdotes* are remembrances of some incidences known to the researcher
- *Surveys* are polls or questionnaires given to people to learn about their opinions or practices.
- *Observational studies* are recordings of information on people in a passive manner.
- *Experiments* are controlled studies, where the investigator deliberately imposes some conditions on the subjects or experimental units and observes and records the results.

Your recollections about the apple habits of your two uncles in design 4 would be an illustration of anecdotes. Although this information is interesting, it has little value from a scientific perspective. It tells you little about the general health benefit of eating apples. You can likely find people who express anecdotes on both sides of a particular issue.

The questionnaire given to doctors in design 2 would be an example of a survey. Although the information from the survey results is helpful in understanding doctors' opinion about the benefit of eating apples, it does not directly measure the subjects of interest, the patients. Also the doctors may be generally unaware of the diets of their patients.

Study designs 1 and 3 have things in common. In both designs, the focus is on the eating habits of a collection of individuals. We call these people the *observational units*. In both studies, the individuals are divided into groups, those that eat and don't eat apples. After a period of time, we will compare the health of the two groups. The below diagram can represent both of the studies.



The *response variable* is the variable in the study that we are primarily interested in. Here the focus is on the health of the individuals and so health would be the response variable. In the description of the study, we measure health by the frequency of office visits. The *explanatory variable* is the variable that we think may be helpful in understanding the differences in the response variable. Since we are looking at the health benefits of apples, the explanatory variable would be the regular consumption of apples. In the diagram, we wish to compare the health of the individuals of group 1, those who eat apples, with the health of group 2, those who do not eat apples.

Although there are similarities between study designs 1 and 3, there is an important distinction that makes one an observational study and the other an experiment. In design 1, we are simply observing which individuals eat apples and which do not, and then comparing the office visits of the two groups. Since we are passively observing this data, we call this an observational study. In contrast, in design 3, we are exhibiting some control by randomly assigning the individuals to be "apple eaters" or "non-apple eaters". Since we are deliberating imposing a condition on the subjects, we call this design an experiment.

Suppose in study 1 that the individuals who eat apples tend to visit the doctor less regularly than the individuals who don't eat apples. Could you conclude that apples do have a health benefit? Actually no. We are not disputing that the one group had fewer office visits than the second group. But we really don't know if the decrease in office visits is due to eating apples since this is just an observational study. Perhaps the individuals who eat apples eat a better diet than the individuals who don't eat apples, and it is the better diet (not the apples) that caused the fewer office visits. Or maybe the apple eaters exercise more regularly than the non-apple eaters and the difference in exercise is the explanation for the difference in office visits. There are actually many reasons why the two groups would differ in office visits in this observational study.

Study 3 is different from study 1 in that subjects are randomly assigned to the two groups. By the random assignment, one hopes that the two groups of individuals are similar with respect to other variables, such as diet or health that might influence the response. In this study, one is able to see the direct effect of the explanatory variable on the response. In this experiment, if the apple eaters have fewer office visits than the non-apple eaters, then it is possible to say that the apples had a positive impact on the health. We will see that one can establish these types of cause and effect relationships between variables only through designed experiments.

PRACTICE: DIFFERENT WAYS OF COLLECTING DATA

1. Classify each of the following studies as an anecdote, an observational study, or an experiment.

(a) You currently have tennis elbow and one of your tennis friends recommends that you do push-ups before playing tennis to relieve the pain.

(b) A physical therapist has 50 patients who are currently suffering from tennis elbow. He randomly assigns half of the patients to a special exercise program, and the other half

to a no-exercise program. By asking each person if they have less pain in their elbow, she will compare the exercise and no-exercise programs.
(c) You collect data from 100 people who have had tennis elbow in the past six months. You find out that the people who regularly exercised their elbow tended to have a quicker recovery from tennis elbow compared to the people who did not exercise their elbow.

2. For the following studies, identify the observational units, the response variable, and the explanatory variables. Also give if the study is an observational study or an experiment.
(a) To investigate the relationship between a milk diet and losing weight, a researcher looked at the diets of 54 overweight people. There was no significant difference in loss of weight between the people with a high-dairy diet and the people with a low-dairy diet.
(b) People who come from big families appear to have an increased risk of heart disease. Researchers studied 4,286 women and 4,252 men aged 60 to 79 who had from zero to five or more children. Those who had more than two children had a higher risk of developing coronary hear disease and the risk becomes larger with each additional child.
(c) Does coaching help in increasing one's SAT exam? Twenty-one students were selected from public, suburban schools near New Haven who had previously taken the SAT exam. Participants received 30-35 hours of instruction over four weeks. When the SAT exam was retaken, the students scored 60 points higher on the math subtest. A control group that did not participate in the coaching had an average gain of only 13 points on a SAT retest.

MUSIC TRAINING AND MATH ACHIEVEMENT

Suppose a high school principal is concerned about her students' low grades on the mathematics component of the state's graduation test. She is aware of the studies that relate music and mathematics achievement. She has read that musicians achieve higher grade point averages than non-musicians, music majors are more likely to be admitted to medical school than non-music majors, and music study can enhance higher brain functions. Could it be true that music education at her school can contribute to higher grades on the mathematics graduation test?

To help answer this question, the principal examines the mathematics graduation test scores of 30 students, 15 who are currently in the music program in the high school and 15 students who are not currently taking music. Here is the data:

Math graduation test scores:

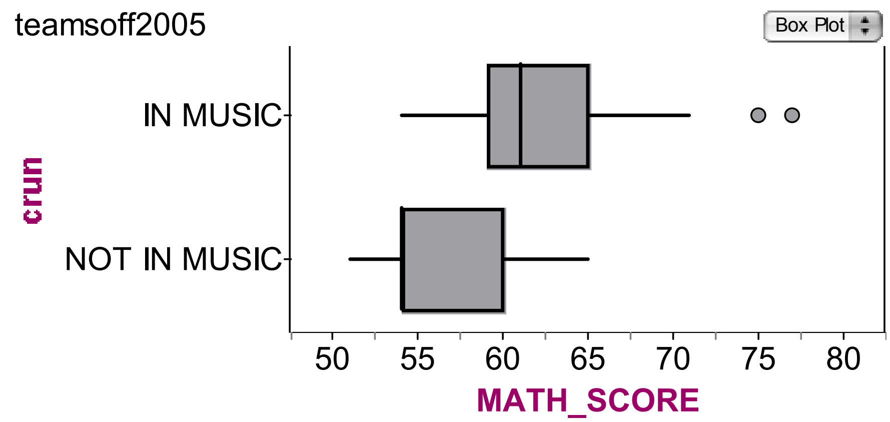
For students in music program

65, 71, 65, 61, 61, 77, 75, 59, 62, 60, 59, 56, 59, 54, 47

For students not in music program

60, 54, 65, 55, 61, 54, 57, 61, 54, 52, 54, 54, 53, 51, 55

Parallel boxplots of the math scores for the two groups are presented below.

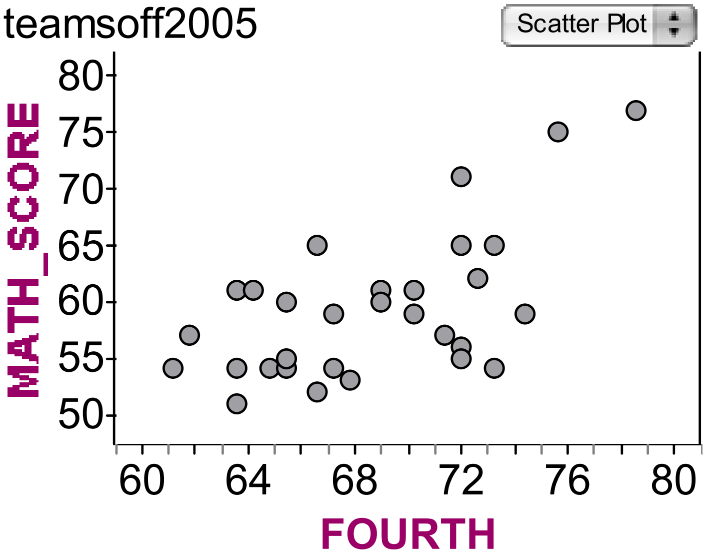


In this example, the response variable is a student's mathematics score on the graduation test and the explanatory variable is whether or not the student is in the music program. From the figure, it appears that the students in the music program tend to have higher math scores. Can the principal conclude from these data that music training does cause an increase in mathematics achievement? No, since this is an example of an observational study. The principal is not controlling whether or not a student is in the music program; instead she is just observing the math score and music background of the students and noticing a relationship.

Since this study hasn't proven that music training is helpful in boosting math achievement, what could explain the higher scores of the students in the music program? Here are some possible explanations.

1. Students in the music program may generally be more studious than students outside of the music program, and therefore do better on math.
2. The students who have strong academic achievement are more likely to participate in music programs than students with lower academic achievement.
3. Students in the music program are more committed to practicing their instrument, and students committed to practicing are also more likely to work hard on homework.

The principal decides to investigate the second explanation. Perhaps the students with high scores on the mathematics graduation test were strong in mathematics throughout their school years. To investigate this possibility, the principal collects scores of a 4th grade mathematics achievement test for the same 30 students taken before they entered the music program. She constructs a scatterplot of the 4th grade math score and the math graduation test score.



There is a positive relationship in this plot indicating that students who did well on the 4th test also did well in the graduation test.

An observational study does not control for possible effects of variation that are not considered in the study but could have an effect on the response variable. These unmonitored variables are often called *lurking variables*. In this observational study, the principal was interested in the effect of music education on mathematics achievement. But the study did not account for the possibility that the two groups scored differently on the math graduation test because the groups had different mathematics abilities. Here mathematics ability can be viewed as a lurking variable. It is difficult to say if the one group is better due to the higher mathematical ability or the music education. Generally lurking variables have effects on the response variable that are *confounded* with those of the explanatory variable. A *confounding variable* is one whose effects on the response variable are indistinguishable from that of the explanatory variable. Here the effect of mathematical ability of the students is confounded with the positive effect of music education.

PRACTICE: OBSERVATIONAL STUDIES AND LURKING VARIABLES

For the following observational studies, discuss the validity of the conclusions given. If appropriate, describe possible lurking variables that are not considered in the study but may have an effect on the response variable.

1. A school principal observes that most of the students currently involved in the music program come from higher socio-economic backgrounds. She concludes that the students from higher socio-economic backgrounds have more musical ability than the students from lower socio-economic backgrounds.

2. A convenience store wants to increase their sales and so they begin a new advertisement campaign in March. After two months, the store notices that their monthly sales have improved by 20% and so they believe the ad campaign has been successful.

3. A college wishes to learn about the success of a new on-line course for introductory statistics. In one semester, there were two sections of the class – the “traditional” class is taught in the usual classroom setup and one class is taught on-line. At the end of the semester, all of the students are given the same examination and the on-line students scored, on average, 10 points higher than the students in the traditional class.

AN EXPERIMENT TO DETECT THE MOZART EFFECT

We saw in our earlier example that it was not possible to conclude that music education causes an increase in mathematical achievement through an observational study. But there is scientific evidence that music can enhance abstract reasoning skills. Here we describe the details of a statistical experiment designed to investigate the effects of listening to classical music.

Researchers Gordon Shaw and Frances Rauscher predicted that listening to music would increase one's ability to do spatial-temporal reasoning. Spatial-temporal reasoning is the ability to create, maintain, transform, and relate complex mental images, without any help through pictures or verbal advice. Math, science, chess and music all involve this type of reasoning.

These researchers worked with 79 college students in this experiment. At the beginning, all of the students took a special test that measured the level of their spatial-temporal reasoning. This test consisted of a series of paper-folding and cutting (PF&C) questions. On a particular question, there were written directions on how to fold and cut a piece of paper in a series of steps. The subject was then asked to decide, among a set of five choices, what the paper would look like when folded.

After this initial test, the subjects were divided by random assignment into three groups called the "Silence", "Mixed", and "Mozart" groups. As one might expect, the Silence group sat in silence in a room and didn't receive any stimulation. The Mixed group of students listen to different types of media for the next day including a "minimalist" piece, a story on tape, and a dance piece. The Mozart group listed to the first movement of Mozart's Sonata (K. 448). After this experience, all of the students were given another PF&C test of their spatial-temporal reasoning. For each group of subjects, the researchers computed the number of correct answers before and after the simulation.

This example is a good illustration of a randomized comparative experiment. Here the researchers started with 79 students – these would be the experimental units. Since the focus is on spatial-temporal reasoning, the response variable would be the number of items correct on the special test designed to measure this type of reasoning.

The explanatory variable would be the type of simulation provided to the students. Here there are three categories of the explanatory variable that we call *treatments*.

- Treatment 1 would be the absence of stimulation – the subjects given this treatment were really given the silent treatment.
- Treatment 2 would be the stimulation consisting of the mixed media.
- Treatment 3 would be the Mozart simulation

How are the students assigned to the three treatments? It is important that the subjects were assigned to treatments by random allocation. This means that each subject is equally likely to receive any one of the three treatments.

How do we perform this random allocation? It can be done using a similar method to choosing a simple random sample. Suppose for sake of illustration, that there were only 15 students in the study and we wished to randomly assign them to the Silence, Mixed, and Mozart treatments. Here are the names of the 15 subjects:

Bob, Shirley, Alan, Carl, Jill, Juli, Denise, Mary, Bryan, Hannah, Emily, Alex, Anna, David, Ben

We assign each subject with a two-digit label.

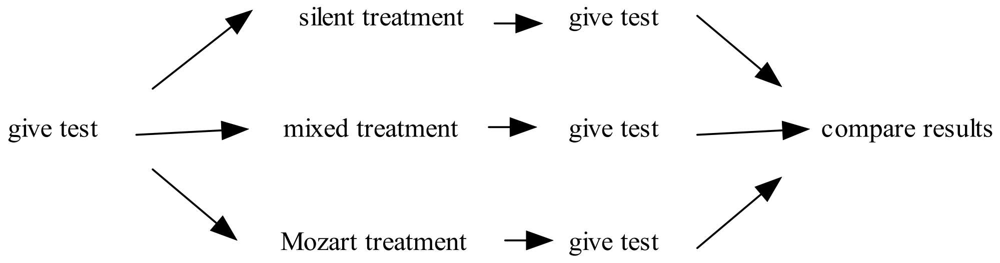
Bob,	Shirley,	Alan,	Carl,	Jill,	Juli,	Denise,	Mary,	Bryan,	Hannah,	Emily,	Alex,	Anna,	David,	Ben
01	02	03	04	05	06	07	08	09	10	11	12	13	14	15

We use the random digit table to randomly allocate the subjects to the treatments. Since we have two-digit labels, we find two-digit random numbers between 01 and 15 from a table of random digits or a computer.

1. We first find the random numbers 04, 02, 10, 12, 01, corresponding to Carl, Shirley, Hannah, Alex, and Bob. These five people would be assigned to the Silent treatment.
2. Next, we choose 09, 14, 06, 13, 05 from the digit table. The students with these labels, Bryan, David, Juli, Anna, and Jill would be assigned to the Mixed treatment.

3. There are five students left – Alan, Mary, Denise, Ben and Emily – and these students would be assigned to the Mozart treatment.

The following diagram can describe the experiment:



What were the results of this experiment? For the silent and mixed groups, the average number of correct answers on the PF&C test was approximately the same for the first and second tests. However the Mozart group showed an improvement in the number of correct answers from the first test to the second test. The conclusion from this study is that special types of music can improve the spatial-temporal reasoning of students.

PRACTICE: DESIGNING EXPERIMENTS

Suppose you are interested in seeing if a new gasoline additive is effective in increasing the mileage of cars. Suppose you have the following cars to use in an experiment (cars are listed in order according to their size).

Ford Focus, Honda Civic, Mini Cooper, Chevrolet Malibu, Nissan Maxima, Toyota Camry, Pontiac Vibe, Toyota Matrix, Ford Taurus, Subaru Outback, Mazda6, Pontiac Bonneville, Ford Explorer, Hummer H2

You plan on dividing the cars into two groups. One group will be the control group where each car will drive 200 miles of "in-town" driving. The cars in the treatment group will use the gasoline additive and also drive 200 miles "in-town". You plan on comparing the mileages of the two groups of cars to see if there is any effect of using the gasoline additive.

1. Suppose the first seven cars are assigned to the control group and the remaining cars are in the treatment group. Do you see any potential problems with this assignment of cars to groups? Would there be a lurking variable that might explain differences in mileages between the two groups?
2. By use of a random digit table, divide the cars into the two groups. Explain what labels you assign to cars and the random digits you find in the table.
3. By using this random assignment of cars to groups, are you sure that the two groups of cars will be similar with respect to any extraneous variables that might influence mileage? Explain.

BASIC PRINCIPLES OF EXPERIMENTS

Why do researchers conduct experiments instead of just conducting observational studies? We have seen that it is not possible to separate out the effects of the explanatory variables from lurking variables in an observational study. But a well-designed experiment is able to control for the effects of lurking variables. Because of this control, it is possible to say in a designed experiment that the explanatory variable caused the effect. Here we summarize the basic principles of experiments by discussing a recent article on how to reduce belly fat of women.

In the Associated Press article "Study: Lifting Weights Attacks Belly Fat," researchers are interested in reducing intra-abdominal fat in women. This is one of the more unhealthy forms of fat since it is linked with the occurrence of heart disease. The researchers want to see if a particular weight-training program is effective in preventing the increase in this type of fat.

Suppose the researchers conduct the following experiment. They recruit 100 women who are overweight and have them participate in the weight-training program. They measure the weight of the intra-abdominal fat of the women before and after the program. This experiment can be represented by the following diagram.

Experimental units -> Treatment -> Response

Suppose it turns out that, on average, that the weight of the fat of the women has decreased in the two-year period. Does this mean that the treatment is effective? No. This experiment has the same problem as an observational study. There are many possible reasons for the decrease in the fat. Possibly the women had healthier diets or they participated in some other weight loss program. In this experiment it is impossible to separate the effects of the weight-training program from the other lurking variables.

The above problem introduces the first important characteristic of a well-designed experiment – *control*. To see the effect of a treatment, one should set up two groups of subjects – a control group and a treatment group that is similar to the control group except for the presence of the treatment.

In our experiment, the article describes the use of a control group.

> "In (the study), 164 overweight and obese Minnesota women ages 24 to 44 were divided evenly into two groups. One group participated in a two-year weight-training program and the other was simply given a brochure recommending exercise of 30 minutes to an hour most days of the week. Both groups were told not to change their diets in a way that might lead to weight changes."

Here the control group would consist of the women given the brochure recommending exercise and the treatment group would consist of the women who participated in the weight-training program. Note that all of the women were asked not to change their diets, so the main difference between the two groups was the presence of the treatment.

The second principle of a well-designed experiment is *randomization*. One should randomly assign people to the control and treatment groups. We use chance to choose groups in order to eliminate any systematic bias in assigning the subjects to groups. In this example, that might be differences in the women, such as the seriousness of their obesity or their age, that might affect the differences in weight loss observed between the two groups. By randomly assigning people to groups, it is likely for a given person to be assigned to the control or treatment groups, and so the two groups will likely be balanced with respect to any lurking variables that could influence the response variable. In this article, it was not specifically stated that patients were randomly assigned to the two groups, but we would hope that randomization was used.

The last characteristic of a good experiment is *replication*. This means that we should perform the experiment on as many subjects as possible. In this article, 164 women participated in the study. This is a pretty large sample – wouldn't it be simpler to just use, say 30 women in the study?

The experiment could have been run with only 30 women, but the results may not have been helpful in showing the benefit of the weight-training treatment. When one observes the weight loss measurements for the women, one will notice chance variation. There will be natural variation in the weight losses of the women due to the individual characteristics of the women. When the weight losses for small groups of women from the control and treatment groups are collected, it is possible that the chance variation will be large and overwhelm any effect due to the treatment. Instead if measurements from larger samples of women are collected, the size of the chance variation will be reduced and it will be easier to detect the differences between the two groups.

This study of 164 presumably was large enough to show a significant treatment effect. As reported in the article

> "Women who did the weight-training for two years had only a 7 percent increase in intra-abdominal fat, compared to a 21 percent increase in the group given exercise advice".

Since this was a published result, this observed difference in the two groups (7 percent versus 21 percent) was so large that it would rarely occur by chance. In this case, we say that the result is *statistically significant*. This phrase is commonly used in reports of investigations in many fields of study. It tells you that they have found good evidence for the effect they were seeking.

PRACTICE: ASPIRIN AND HEART DISEASE

In the 1980's, there was an important medical study investigating the role of aspirin in preventing heart attacks. A large number (22,000) of male physicians was divided into two groups. One group of doctors took a buffered aspirin every other day, and the other group took a placebo. (A placebo is a pill that looks just like the aspirin pill but has no therapeutic value.) After a period of time, the number of heart attacks for both groups was recorded and the placebo group suffered almost twice as many heart attacks.

1. Is this an example of a designed experiment? Why?
2. How does this experiment illustrate the principle of control?
3. Why is it important for the one group to take a placebo instead of having no pill?
4. Although it is not mentioned in the above description, how could this experiment illustrate the principle of randomization?
5. Describe the idea of replication in the context of this example. What does it mean if it is claimed that the results of this study are statistical significant?

ACTIVITY: JUMPING FROGS

DESCRIPTION: One useful way of obtaining useful data is through a designed experiment. This activity will introduce the construction of a simple designed experiment to learn about the best design of an origami frog to maximize its jumping distance.

MATERIALS NEEDED: Each student will get a sheet of paper to construct his/her origami frog. Half of the sheets should be "small" (6 by 9 inches) and half should be "large" (8 ½ by 11 inches). Of the small sheets, half of the sheets should be standard printer paper, and half should be the thickness of construction paper. Likewise, the large sheets should be equally divided between thin and thick sheets. Also, a set of rulers with centimeter scaling are needed for the measurements.

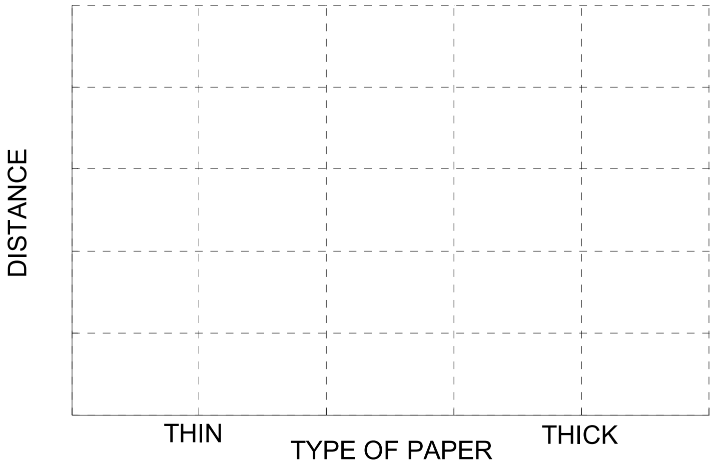
1. Your instructor will give you a single sheet of paper. Using the paper, construct an origami frog using the design on the following page.
2. Practice having your frog jump until you develop a consistent pattern of jumping. After this practice, have your frog jump one time and measure the forward movement (in cm).
3. After all of the students in the class have made measurements, record all of the measurements in the following table.

		Size of Paper	
		Small	Large
Paper thickness	Thin		
	Thick		

4. For each square of the table, find the mean jumping distance – place your means in the below table.

		Size of Paper	
		Small	Large
Paper thickness	Thin		
	Thick		

5. Plot your means in the following graph. Plot the means for the small sheets using one symbol and plot the means for the large sheets using a different symbol.



6. From your work, describe how the jumping distance compares for thin and thick paper. Is there a difference in the jumping distance for small and large sheets? How would you design a frog that jumps a long distance?

EXERCISES

EXERCISE 1. (Bananas and health) Bananas are known to be good for your health as they are a good source of potassium. In a study of 5600 people aged over 65, it was found that those with the lowest intake of potassium were 50% more likely to suffer a stroke.

(a) Is this an example of an observational study or an experiment? Explain.

(b) Identify the observational units, the response variable, and the explanatory variable.

(c) Are there any potential lurking variables present in this study? Explain.

EXERCISE 2. (Fasting and health) Results from a recent study indicate that periodic fasting can have similar health benefits as sharply cutting back on calories, even when fasting doesn't mean eating less overall. A team of researchers reported that mice that were fed only every other day - but could gorge on the days they did eat - saw similar health benefits to ones that had their diet reduced by 40 percent.

(a) Identify the observational units, the response variable, and the explanatory variable.

(b) It is likely that these results were based on a randomized comparative experiment. Describe how this experiment might be conducted.

EXERCISE 3. (Cell phones and driving) There have been many studies investigating the possible problems with people who use their cell phones during driving. Researchers for the Virginia Tech Transportation Institute and the National Highway Traffic Safety Administration tracked 100 cars and their drivers for a period of a year. On the basis of data collected, these researchers concluded that talking on cell phones caused far more crashes and near-crashes than other distractions.

(a) Is this an experiment or an observational study? Explain.

(b) Identify the observational units, the response variable, and the explanatory variable.

(c) By searching on the internet, find another study looking into the relationship of cell-phone use and traffic accidents. How was this study conducted? By comparing the results of the two studies, which study do you think is more persuasive for convincing a person not to use a cell phone while driving?

EXERCISE 4. (Prayer can speed recovery?) Researchers report that prayer may reduce the number of complications experienced by hospitalized heart patients. All of the patients received standard medical care, but unbeknownst to the patients, half of the patients received prayers for healthy recovery for a period of four weeks. Medical charts of the patients were examined, following their health histories between hospital admittance and discharge. It was found that the prayed-for individuals had significantly lower complication rates than those not prayed for.

(a) Is this an experiment or an observational study? Explain.

(b) Identify the observational units, the response variable, and the explanatory variable.

(c) Are there any possible lurking variables that might explain the different in complication rates between the two groups?

EXERCISE 5. (Reducing test stress) Since standardized tests are more important in determining a school's performance, school officials are interested in ways of reducing test anxiety. The institute HeartMath studied a group of 800 10th graders. They found that 55% of this group often had high levels of test anxiety and their test scores were lower than those of students who used the company's techniques to improve relaxation and mood.

(a) As described above, is this study an experiment or an observational study? Explain.

(b) Can you think of any lurking variables that would explain why the students using the company's techniques would have higher test scores than the group of 10th graders?

(c) Describe how one might construct a randomized comparative experiment to assess the effectiveness of the company's stress-reducing methods.

EXERCISE 6. (The placebo effect.) The placebo effect is the phenomenon that a patient's symptoms can be reduced by an ineffective treatment, called a placebo, because the individual thinks the treatment will work.

(a) There was a study in the 1950's where a group of patients with back pain were given an ineffective treatment. One quarter of these patients reported a reduction of pain. Explain this result using the placebo effect.

(b) Suppose a new treatment for back pain has been developed. It is tested on a group of 100 patients with this ailment and 30% experience some reduction of pain. Is this evidence that the new treatment is effect?

(c) Describe a better experiment for evaluating the benefit of the new treatment for back pain.

EXERCISE 7. (Randomly assign patients to treatments) Suppose you are interested in evaluating a new method for memorizing material. You wish to compare your new method with a standard method. You have the following 20 people to participate in the experiment – half will be assigned to the new method and half will be assigned to the standard method. Describe how to use the random digit table to randomly assign the people to the two groups.

Joe, Charley, Jill, Sue, Angie, Zoe, Helen, Bob, Bill, Jeff, Aaron, Kara, Beth, Lynn, Hal, Cliff, Warren, Juan, Neal, Steve

EXERCISE 8. (Evaluating a new teaching method) Suppose there are two teachers of geometry, Mr. Schmidt and Ms. Brown, at your local high school. Suppose that Ms. Brown wishes to try a new method for teaching geometry at her school where there are no lectures and the students learn by working on directed activities in small groups. At the end of the year, all students are given a standardized test on geometry and the mean score of Ms. Brown's students is ten points higher than the mean score of Mr. Schmidt's students.

(a) Is this convincing evidence that the new method of teaching geometry is better than the traditional method?

(b) Describe some lurking variables that might explain the better performance for Ms. Brown's class.

(c) Describe how one could compare the two methods of teaching by a randomized comparative experiment.

EXERCISE 9. (Vitamin C and colds) Suppose 10% of a given group of adults did not colds one winter. Next winter, each adult is asked to take 1 gram of Vitamin C each day, and 20% do not have colds. Explain why this result is not good evidence that Vitamin C is helpful in preventing colds. Give some possible lurking variables whose effects may be confounded with the effect of taking Vitamin C.

EXERCISE 10. (Does child-care attract employees?) Comparing two companies A and B. Two versions of brochure. Change to different subject, say swim clubs?

EXERCISE 11. (Designing a taste test.)

EXERCISE 12. (Testing a new drug to relieve pain from arthritis). Could administer drug and record responses. Could compare effectiveness with placebo or aspirin.

EXERCISE 13. (Designing a new cake mix) – sensitivity of taste with respect to time in oven and temperature of oven.

EXERCISE 14. Comparing three weight-loss programs and have 15 overweight females to participate in test.

TOPIC P1: PROBABILITY - A MEASUREMENT OF UNCERTAINTY

SPOTLIGHT: HOW RISKY IS …?

The magazine *Discover* once had a special issue on “Life at Risk.” In an article, Jeffrey Kluger describes the risks of making it through one day:

“Imagine my relief when I made it out of bed alive last Monday morning. It was touch and go there for a while, but I managed to scrape through. Getting up was not the only death-defying act I performed that day. There was shaving, for example; that was no walk in the park. Then there was showering, followed by leaving the house and walking to work and spending eight hours at the office. By the time I finished my day -- a day that also included eating lunch, exercising, going out to dinner, and going home -- I counted myself lucky to have survived in one piece.”

Is this writer unusually fearful? No. He has read mortality studies and concludes “there is not a single thing you can do in an ordinary day -- sleeping included -- that isn't risky enough to be the last thing you ever do.” In the *Book of Risks* by Laudan, we learn that

- 1 out of 2 million people will die from falling out of bed.
- 1 out of 400 will be injured falling out of bed.
- 1 out of 77 adults over 35 will have a heart attack this year.
- The average American faces a 1 in 13 risk of suffering some kind of injury in home that necessitates medical attention.
- 1 out of 7000 will experience a shaving injury requiring medical attention.
- The average American faces a 1 out of 14 risk of having property stolen this year.
- 1 out of 32 risk of being the victim of some violent crime.

- The annual odds of dying in any kind of motor vehicle accident is 1 in 5800.

Where do these reported odds come from? There are simply probabilities calculated from the counts of reported accidents. Since all of these accidents are possible, that means that there is a risk to the average American that they will happen to him or her. But fortunately, you need not worry – many of these reported risks are too small to really take seriously or change your style of living.

PREVIEW

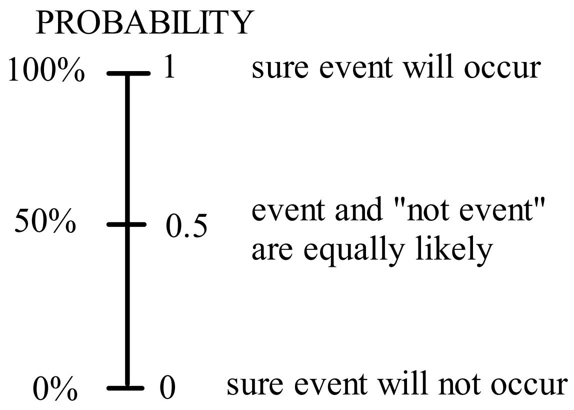
Everywhere we are surrounded by uncertainty. If you think about it, there are a number of things that you are unsure about, like

- what is the high temperature next Monday?
- how many inches of snow will our town get next January?
- what 's your final grade in this class?
- will you be living in the same state twenty years from now?
- who will win the U.S. presidential election in 2008?
- is there life on Mars?

A probability is simply a number between 0 and 1 that measures the uncertainty of a particular event.

Although many events are uncertain, we possess different degrees of belief about the truth of an uncertain event. For example, most of us are pretty certain of the statement "the sun will rise tomorrow", and pretty sure that the statement "the moon is made of green cheese" is false.

We can think of a probability scale from 0 to 1.


PROBABILITY
100%
1
sure event will occur
50%
0.5
event and "not event" are equally likely
0%
0
sure event will not occur

We would give the statement "the sun will rise tomorrow" a probability close to 1, and the statement "the moon is made of green cheese" a probability close to 0. It is harder to assign probabilities to uncertain events that have probabilities between 0 and 1. In this topic, we first get some experience in assigning probabilities. Then we will discuss three general ways of thinking about probabilities.

In this topic your learning objectives are to:

- Understand the three interpretations of probability.
- Understand what interpretations of probability are appropriate in a given situation.
- Understand how to compute approximate probabilities given data from repeated experiments.
- Understand how to obtain subjective probabilities by use of a calibration experiment.

WARM-UP ACTIVITY – SOME QUESTIONS ON PROBABILITY

For each of the following questions,

- specify the probability (as best as you can) and
- explain why you gave this particular probability value.

1. Suppose you have a bag with 4 white and 8 red balls. You choose a ball at random from the bag. What is the probability that the ball you chose is white?

2. Suppose you walk into your college bookstore blindfolded and bump into a fellow student. What is the probability the student is female?

3. ___________ [YOUR TEAM] is playing ________ [ANOTHER TEAM] soon in _________ [GIVE SPORT]. What is the chance that your team will win?

4. You drop a thumbtack 20 times on the floor and it lands with the point-side up 12 times. What is the probability that the tack will land point-side up?

5. Suppose you toss a coin 20 times and get 19 heads. What is the chance that the next toss is heads?

6. What is the chance that you will be married when you are 25 years old?

7. If you roll two dice, what is the probability that the sum of the two dice rolls is equal to 5?

8. Suppose you are going to interview a high school math teacher. What is the chance that this teacher is male?

9. What is the chance that you will complete your college education (that is, graduate) in five years or less?

10. When a meteorologist reports that there is a 50% chance of rain tomorrow, what does this mean?

11. What's the chance that two people in our class have the same birthday (month and day)?

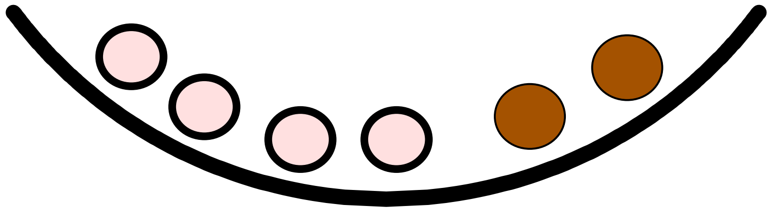
After working this activity, you should realize that probabilities are hard to measure. But there are three ways of thinking about probabilities, the classical, frequency, and subjective viewpoints, that are helpful in this measurement problem. We discuss each interpretation of probabilities in the remainder of this topic.

THE CLASSICAL VIEW OF A PROBABILITY

Suppose that we observe some phenomena (say, the rolls of two dice) where the outcome is random. Suppose we can write down the list of all possible outcomes, and we believe that each outcome in the list has the same probability. Then the probability of each outcome will be

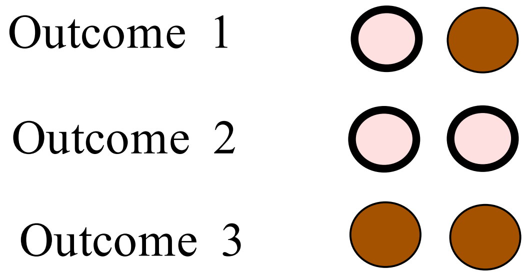
$$P(outcome\ ) = \frac{1}{number\ \ of\ outcomes}\,.$$

Let's illustrate this classical view of probability by a simple example. Suppose you have a bowl with 4 white and 2 red balls

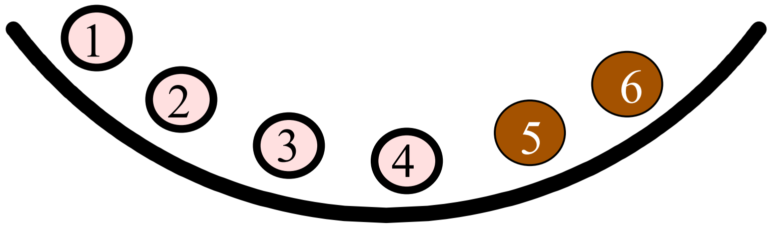


and you draw two balls from the bowl at random. We assume that the balls are drawn *without replacement* which means that you don’t place a ball back into the bowl after it has been selected. What are possible outcomes? There are different ways of writing down the possible outcomes, depending if you decide to distinguish the balls of the same color.

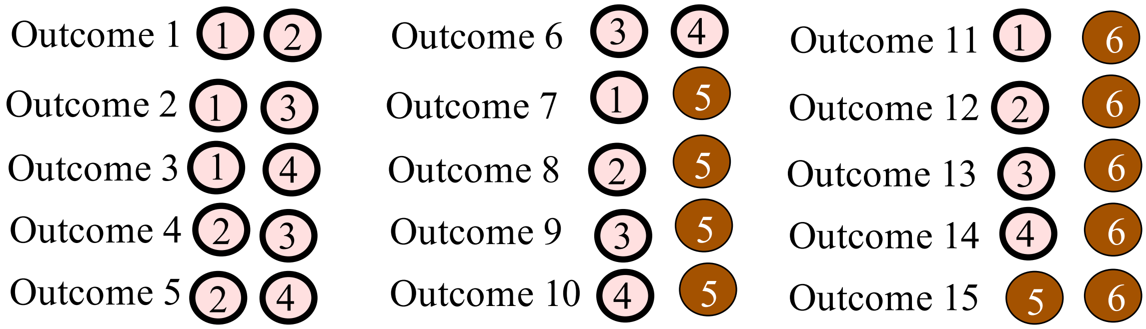
WAY 1: If we ***don't distinguish*** between balls of the same color, then there are three possible outcomes – essentially we choose 0 red, 1 red, or 2 red balls.



WAY 2: If we ***do distinguish*** between the balls of the same color, we label the balls in the bowl



and then we can write down 15 distinct outcomes of the experiment of choosing two balls.



Which is the more appropriate way of listing outcomes?

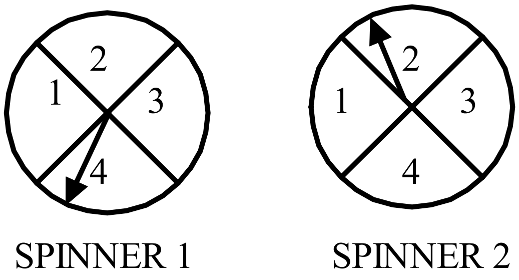
To apply the classical view of probability, we have to assume that the outcomes are all equally likely. In the first list of three outcomes, we can't assume that they are equally likely. Since there are more white than red balls in the basket, it is more likely to choose two white balls (○ ○) than to choose two red balls (● ●). So it is incorrect to say that the probability of each one of the three possible outcomes is 1/3.

That is, the probabilities of choosing 0 red, 1 red, and 2 reds are not equal to 1/3, 1/3, and 1/3.

On the other hand, since we are choosing two balls at random from the basket, it makes sense that the 15 outcomes in the second listing (where we assumed the balls distinguishable) are equally likely. So we can apply the classical notion and assign a probability of 1/15 to each of the possible outcomes. In particular, the probability of choosing two red balls (which is one of the 15 outcomes) is equal to 1/15.

PRACTICE: THE CLASSICAL VIEW OF A PROBABILITY

Suppose you have two spinners, shown below.



(a) Suppose you spin both spinners and record the ***sum*** of the two spins. For example, if SPINNER 1 lands “2” and SPINNER 2 lands “3”, the sum of the two spins is 2 + 3 = 5. Write down all of the possible sums of two spins.

(b) Assume that the possible outcomes in part (a) are equally likely. Then what would be the probability that the sum of spinners is equal to 2?

(c) Suppose instead that you record the ***Spinner 1 result and the Spinner 2 result***. (One possibility is that Spinner 1 lands 2 and Spinner 2 lands 3.) Write down all of the possible outcomes below of the two spinners. (You should have a list of 16 outcomes.)

SPINNER 1 SPINNER 2

(d) If the outcomes in (c) are assumed equally likely, what would be the probability that both spinners land 1?

(e) If you compare your answers to parts (b) and (d), note that your answers are different. That is, the probability that the SUM is equal to 2 in part (b) is not equal to the probability that both spinners land 1 in part (d). Why?

(f) In part (b), we assumed that all possible SUMS were equally likely, and in part (d), we assumed that all possible values of (SPINNER1, SPINNER2) were equally likely. Which assumption is *not* correct? Why?

THE FREQUENCY VIEW OF A PROBABILITY

The classical view of probability is helpful only when we can construct a list of outcomes of the experiment in such a way where the outcomes are equally likely.

The frequency interpretation of probability can be used in cases where outcomes are equally likely or not equally likely.

This view of probability is appropriate in the situation where we are able to ***repeat the random experiment many times under the same conditions.***

Getting out of jail in Monopoly

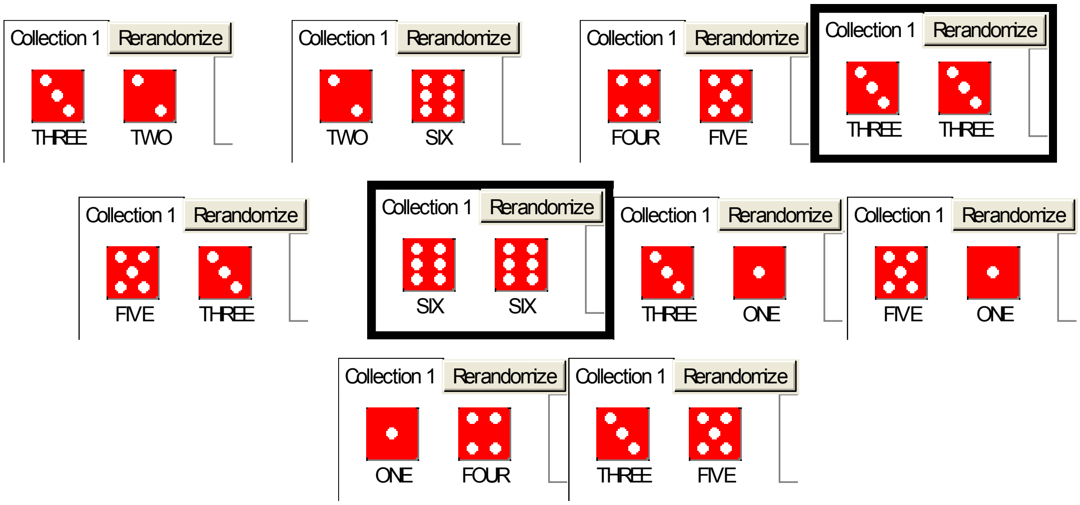
Suppose we are playing the popular game Monopoly and we land in jail. To get out of jail on the next turn, we can either pay $50 or roll "doubles" when we roll two fair dice. Doubles means that the faces on the two dice are the same. If we think that it is relatively likely to roll doubles, then we may elect to roll two dice instead of paying $50 to get out of jail.

What is the probability of rolling doubles when you roll two dice?

In this situation, we can use the frequency notion to approximate the probability of rolling doubles. We can imagine rolling two dice many times under similar conditions. Each time we roll two dice, we observe if we get doubles or not. Then the probability of doubles can be approximated by the relative frequency

Prob(doubles) = (# of doubles)/(number of experiments).

We used *Fathom* to simulate the rolling of two dice -- the results of the first 10 experiments are shown in the table below. For each experiment, we record if there is a match or no match in the two numbers that are rolled.



In these first 10 experiments, we note that we obtained a match (doubles) exactly two times. (Those are the outcomes that are boxed.) So

Prob(match) is approximately 2/10 = 0.2.

What happens if we roll the two dice more times? Then our approximate probability, the relative frequency, will get closer to the actual probability of doubles. Let us illustrate this by rolling two dice on Fathom 100 times, and then 10,000 times.

The table below tabulates the results of 100 rolls of two dice. For each roll of dice, we indicate if there was a match (YES) or not (NO).

Die 1	Die 2	Match?	Die 1	Die 2	Match?	Die 1	Die 2	Match?	Die 1	Die 2	Match?
1	6	NO	1	6	NO	6	4	NO	1	6	NO
1	3	NO	4	3	NO	5	3	NO	3	5	NO
3	3	YES	2	1	NO	6	4	NO	2	2	YES
5	4	NO	5	1	NO	6	6	YES	1	4	NO
6	5	NO	1	5	NO	6	5	NO	1	1	YES
2	2	YES	3	2	NO	2	2	YES	1	6	NO
6	2	NO	4	2	NO	4	6	NO	3	1	NO
1	2	NO	5	2	NO	1	3	NO	5	6	NO
3	2	NO	4	5	NO	3	4	NO	6	3	NO
2	6	NO	3	5	NO	4	2	NO	5	4	NO
4	6	NO	1	1	YES	2	3	NO	3	3	YES
4	4	YES	2	1	NO	1	5	NO	6	3	NO
6	4	NO	5	2	NO	2	2	YES	2	6	NO
3	5	NO	3	4	NO	4	3	NO	5	6	NO
2	1	NO	3	5	NO	5	5	YES	2	5	NO
1	1	YES	1	2	NO	1	1	YES	5	3	NO
4	3	NO	2	4	NO	6	5	NO	3	1	NO
5	1	NO	5	1	NO	2	1	NO	6	3	NO
3	4	NO	5	4	NO	3	4	NO	1	2	NO
2	1	NO	1	3	NO	2	3	NO	6	4	NO
6	5	NO	4	4	YES	6	4	NO	3	2	NO
1	1	YES	2	1	NO	1	5	NO	3	4	NO
6	4	NO	6	4	NO	2	5	NO	4	6	NO
5	1	NO	3	3	YES	6	5	NO	2	2	YES
4	3	NO	1	6	NO	2	6	NO	4	4	YES

We see from the table that we observed a match 18 times (there are 18 Yeses in the table), so

Prob(match) is approximately 18/100 = 0.18.

Let's now roll the two dice 10,000 times on computer -- this time, we observe 1662 matches, so

Prob(match) is approximately 1662/10000 = 0.1662.

Approximate and actual probabilities

Is 0.1662 the actual probability of getting doubles? No, it is still only an approximation to the actual probability.

However, as we continue to roll dice, the relative frequency

(number of doubles)/(number of experiments)

will approach the actual probability Prob(doubles).

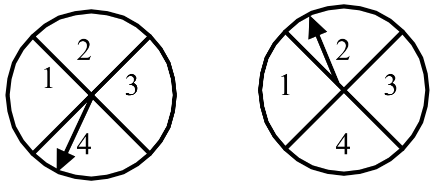
Here the actual probability of rolling doubles is

Prob(doubles) = 1/6,

which is very close to the relative frequency of doubles that we obtained by rolling the dice 10,000 times. (In this example, one can show that are 6 x 6 = 36 equally likely ways of rolling two distinguishable dice and there are exactly six ways of rolling doubles. So using the classical viewpoint, the probability of doubles is 6/36 =1/6.)

PRACTICE: THE FREQUENCY VIEW OF A PROBABILITY

Recall the spinner example considered earlier, where we spun the following two spinners:



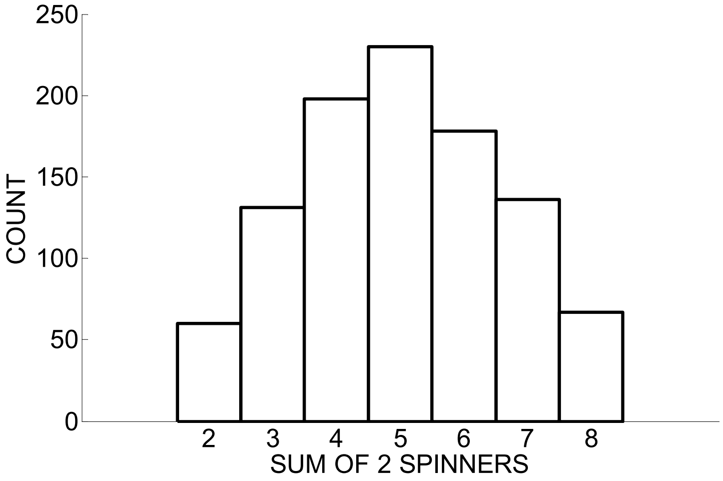
SPINNER 1 SPINNER 2

(a) Suppose we spin the two spinners 20 times with the following results. (A "3" on Spinner 1 and a "2" on Spinner 2 is represented by (3, 2).)

(4, 1)	(4, 1)	(4, 2)	(4, 2)	(4, 4)
(1, 2)	(4, 1)	(2, 1)	(3, 4)	(4, 3)
(3, 4)	(2, 1)	(3, 1)	(1, 2)	(2, 1)
(2, 1)	(1, 3)	(4, 3)	(2, 2)	(4, 3)

Using these data, compute the (approximate) probability the SUM of the two spinners is equal to 5.

(b) The two spinners were spun 1000 times – each time, the SUM of the two spins was recorded. The histogram below displays the results of the 1000 experiments.



Using the histogram, compute the approximate probability the SUM of the two spinners is equal to 5.

(c) The table below shows all of the possible outcomes of rolling the two spinners and the value of the sum S for each outcome. Assuming that the outcomes are equally likely, find the actual probability S is equal to 5. Compare your answer with the approximate probability from part (b).

Outcome	S	Outcome	S	Outcome	S	Outcome	S
(1, 1)	2	(2, 1)	3	(3, 1)	4	(4, 1)	5
(1, 2)	3	(2, 2)	4	(3, 2)	5	(4, 2)	6
(1, 3)	4	(2, 3)	5	(3, 3)	6	(4, 3)	7
(1, 4)	5	(2, 4)	6	(3, 4)	7	(4, 4)	8

ACTIVITY: TOSSING AND SPINNING A POKER CHIP

To apply the frequency notion of probability, it is important that we are able to perform our random experiment repeatedly under similar conditions. It might seem that this is obvious, but actually it is hard to repeat an experiment exactly the same way. We illustrate this through the simple experiment of flipping a poker chip.

MATERIALS NEEDED: A set of standard plastic poker chips and a container of silly putty or gum. (If poker chips are not available, you can do these experiments with all quarters or all nickels.)

1. Look at your poker chip and decide which side is “heads”. Consider the experiment of flipping a poker chip 20 times and counting the number of heads. Before you do this, the instructor and class should decide exactly what it means to “flip a chip.” Here are some guidelines for flipping (the class may wish to adjust these rules):

- you flip the chip in the air
- the chip should flip at least five times in the air before coming down
- the chip has to land on the desk
- any experiments with insufficient flips or where the chip falls on the floor are ignored

First try out your flipping method until you have a style that you can repeat many times. Then flip the chip 20 times and record the number of heads you observe.

2. Next, consider the experiment of spinning the chip 20 times and recording the number of heads. To spin the chip, you hold the chip level with one finger and flick the chip with the other finger so that the chip spins on the table with at least five spins. Practice spinning until you are comfortable with the method. Then spin the chip 20 times, recording the number of heads.

3. Suppose you add a small amount of putty to the heads side of the chip. Do you think this change will modify the probability of heads when flipping? What if you spin – do you think the probability of heads will change?

4. Put the putty on the heads side and flip the chip 20 times – record the number of heads.

5. Now spin the modified chip 20 times – record the number of heads.

6. Combine the class data for the four experiments – flip regular chip, spin regular chip, flip modified chip, and spin modified chip. For each dataset, construct a graph of the number of heads, and find a typical number of heads. (You will have four graphs and summary values corresponding to the four experiments.)

7. Summarize your results. What is the probability of heads when the chip is flipped? What is the probability when the chip is spun? Does the putty affect the probability of heads for the flipped chip? Does the putty change the probability of heads on a spin?

THE SUBJECTIVE VIEW OF A PROBABILITY

We have described two ways of thinking about probabilities:

The classical view. This is a useful way of thinking about probabilities when one can list all possible outcomes in such a way that each outcome is equally likely.

The frequency view. In the situation when you can repeat a random experiment many times under similar conditions, you can approximate a probability of an event by the relative frequency that the event occurs.

What if you can't apply these two interpretations of probability? That is, what if the outcomes of the experiment are not equally likely, and it is not feasible or possible to repeat the experiment many times under similar conditions?

In this case, we can rely on a third view of probabilities, ***the subjective view***. This interpretation is arguably the most general way of thinking about a probability, since it can be used in a wide variety of situations.

Suppose you are interested in the probability of the event

"Your team will win the conference title in basketball next season."

You can't use the classical or frequency views to compute this probability. Why?

Suppose there are eight teams in your conference. Each team is a possible winner of the conference, but these teams are not equally likely to win -- some teams are stronger than the rest. So the classical approach won't help in obtaining this probability.

The event of your team winning the conference next year is essentially a one-time event. Certainly, your team will have the opportunity to win this conference in future years, but the players on your team and their opponents will change and it won't be the same basketball competition. So you can't repeat this experiment under similar conditions, and so the frequency view is not helpful in this case.

What is a subjective probability in this case? The probability

Prob(Your team will win the conference in basketball next season)

represents your belief in the likelihood that your team will win the basketball conference next season. If you believe that your school will have a great team next year and will win most of their conference games, you would give this probability a value close to 1. On the other hand, if you think that your school will have a relatively weak team, your probability of this event would be a small number close to 0. Essentially, this probability is a numerical statement about your confidence in the truth of this event.

There are two important aspects of a subjective probability.

1. A subjective probability is ***personal***. Your belief about your team winning the basketball conference is likely different from my belief about your team winning the conference since we have different information. Perhaps you are not interested in basketball and know little about the teams and I am very knowledgeable about college basketball. That means that our beliefs about the truth of this event will be different and so our probabilities would also be different.

2. A subjective probability ***depends on your current information*** or knowledge about the event in question. Maybe you originally think that this probability is .7 since your school had a good team last year. But when you learn that many of the star players from last season have graduated, this changes your knowledge about the team, and you may now assign this probability a smaller number.

Measuring Probabilities Subjectively

Although we are used to expressing our opinions about uncertain events, using words like

likely, probably, rare, sure, maybe,

we are *not* used to assigning probabilities to quantify our beliefs about these events. To make any kind of measurement, we use a tool like a scale or ruler. Likewise, we need tools to help us assign probabilities subjectively. In the next activity, we illustrate a special tool, called a *calibration experiment*, that will help to determine our subjective probabilities.

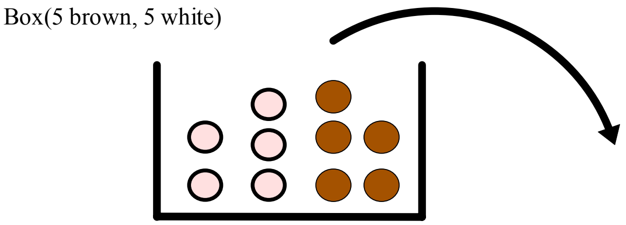
A CALIBRATION EXPERIMENT

Consider the event

W = "a woman will be President of the United States in the next 20 years".

We are interested in your subjective probability of W. This probability is hard to specify precisely since we haven't had much practice doing it. We describe a simple procedure that will help in measuring this probability.

First consider the following calibration experiment – this is an experiment where the probabilities of outcomes are clear. We have a collection of balls, 5 red and 5 white in a box and we select one ball at random.



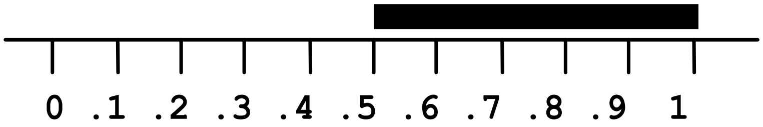
Let B denote the event that we observe a red ball. Since each of the ten balls is equally likely to be selected, I think we would agree that Prob(B) = 5/10 = .5.

Now consider the following two bets:

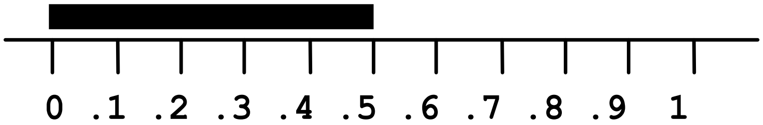
- BET 1 – If W occurs (a women is president in the next 20 years), you win $100. Otherwise, you win nothing.
- BET 2 – If B occurs (you observe a red ball in the above experiment), you win $100. Otherwise, you win nothing.

Based on the bet that you prefer, we can determine an interval that contains your Prob(W):

- If you prefer BET 1, then your Prob(W) must be larger than Prob(B) = .5 – that is, your Prob(W) must fall between .5 and 1.

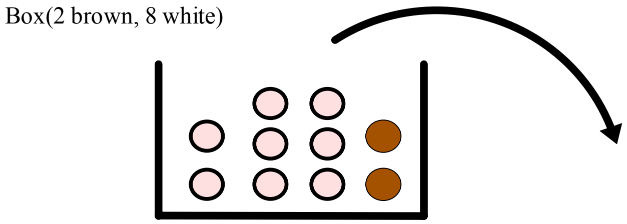


- If you prefer BET 2, then your Prob(W) must be smaller than Prob(B) = .5 – that is, your probability of W must fall between 0 and .5.

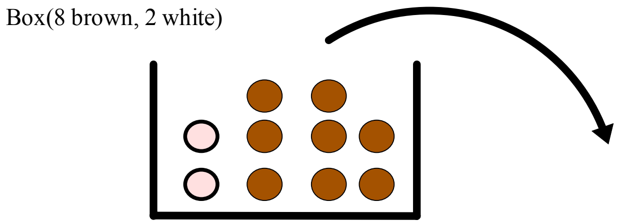


What you do next depends on your answer to part (b).

- If your Prob(W) falls in the interval (0, .5), then consider the "balls in box" experiment with 2 red and 8 white balls and you are interested in the probability of choosing a red ball.



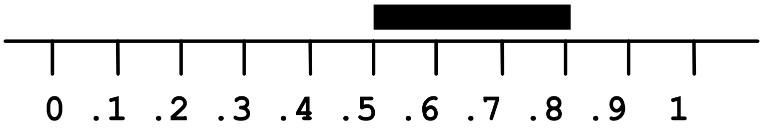
- If instead your Prob(W) falls in the interval (.5, 1), then consider the "balls in box" experiment with 8 red and 2 white balls and you are interested in the probability of choosing a red ball.



Let's suppose that you believe Prob(W) falls in the interval (.5, 1). Then you would make a judgment between the two bets

- BET 1 – If W occurs (a women is president in the next 20 years), you win $100. Otherwise, you win nothing.
- BET 2 – If B occurs (you observe a red ball with a box with 8 red and 2 white balls), you win $100. Otherwise, you win nothing.

I decide to prefer BET 2, which means that my probability Prob(W) is smaller than 0.8. Based on the information on the two comparisons, you now believe that Prob(W) falls between .5 and .8.



In practice, you will continue to compare BET 1 and BET 2, where the box has a different number of red and white balls. By a number of comparisons, you will get an accurate measurement at your probability of W.

PRACTICE: A CALIBRATION EXPERIMENT

1. Consider the following "balls in box" experiments. What is the probability of drawing a red if the box contains

(a) 5 red and 5 white?

(b) 2 red and 8 white?

(c) 7 red and 3 white?

(d) 0 red and 10 white?

2. Consider the statement

A: “I will get married in the next five years"

We want to determine your personal probability that A is true, call this PROB(A).

(If you are already married, choose a different statement where the truth of the statement is uncertain.)

Consider the following two bets:

BET 1:

- If you are married in the next five years, then you win $20.
- If you are not married in the next five years, you win nothing.

BET 2:

- If you draw a red in a balls-in-box experiment with 5 red and 5 white balls, then you win $20.
- If you draw a white, you win nothing.

(a) Which bet (BET 1 or BET 2) do you prefer?

(b) Based on your answer to (a), do you think PROB(A) > .5 or PROB(A) < .5?

3. Let's continue to make this comparison for more balls-in-box experiments. Each row of the table below gives two choices. The left choice is BET 1: you win $20 if you are married in five years 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 box with a certain number of reds and whites; otherwise you win nothing. For each pair of bets, circle the choice which you prefer

BET 1	BET 2
$20 if you are married in five years Nothing if you are not married in five years	$20 if draw red in box with 0 red, 10 white Nothing if draw white
$20 if you are married in five years Nothing if you are not married in five years	$20 if draw red in box with 2 red, 8 white Nothing if draw white
$20 if you are married in five years Nothing if you are not married in five years	$20 if draw red in box with 4 red, 6 white Nothing if draw white

$20 if you are married in five years Nothing if you are not married in five years	$20 if draw red in box with 6 red, 4 white Nothing if draw white
$20 if you are married in five years Nothing if you are not married in five years	$20 if draw red in box with 8 red, 2 white Nothing if draw white
$20 if you are married in five years Nothing if you are not married in five years	$20 if draw red in box with 10 red, 0 white Nothing if draw white

Based on your choices you made above, you should have a better idea about your probability of being married in five years. 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
MY PROBABILITY OF BEING MARRIED IN FIVE YEARS
```

4. Repeat the process in part 3 with a different statement.

PRACTICE: VIEWPOINTS OF PROBABILITY

Consider again the probability questions considered in the opening activity. For each question, ***give the viewpoint*** (classical, frequency, or subjective) that is helpful in obtaining the probability and ***explain how you use*** the viewpoint in the calculation.

1. Suppose you have a bag with 4 white and 8 red balls. You choose a ball at random from the bag. What is the probability that the ball you chose is white?
2. Suppose you walk into your college bookstore blindfolded and bump into a fellow student. What is the probability the student is female?
3. ___________ [YOUR TEAM] is playing ________ [ANOTHER TEAM] soon in _________ [GIVE SPORT]. What is the chance that your team will win?

4. You drop a thumbtack 20 times on the floor and it lands with the point-side up 12 times. What is the probability that the tack will land point-side up?
5. Suppose you toss a coin 20 times and get 19 heads. What is the chance that the next toss is heads?
6. What is the chance that you will be married when you are 25 years old?
7. If you roll two dice, what is the probability that the sum of the two dice rolls is equal to 5?
8. Suppose you are going to interview a high school math teacher. What is the chance that this teacher is male?
9. What is the chance that you will complete your college education (that is, graduate) in five years or less?

WRAP-UP

In this topic, we were introduced to three different ways of thinking about probabilities. The classical viewpoint of probability allows one to compute probabilities based on the structure of the problem. To apply this viewpoint, it is convenient to write down the collection of possible outcomes in such a way that each outcome is equally likely to occur. The frequency viewpoint is useful in the situation where one can repeat the random experiment many times under identical conditions, and the probability of an event is approximately equal to the fraction of experiments where the event occurs. The subjective viewpoint is helpful in the situation where one expresses his or her opinion about the likelihood of a one-time event. A calibration experiment provides a means for measuring subjective probabilities by comparing the situation with a different experiment with known probabilities. These three viewpoints allow us to express uncertainty in a wide variety of situations.

EXERCISES

1. **Probability Viewpoints**

In the following problems, indicate if the given probability is found using the classical viewpoint, the frequency viewpoint, or the subjective viewpoint.

a. Joe is doing well in school this semester – he is 90 percent sure that he will receive an A in all of his classes.

b. Two hundred raffle tickets are sold and one ticket is a winner. I purchased one ticket and the probability that my ticket is the winner is 1/200.

c. Suppose that 30% of all college women are playing an intercollegiate sport. If I contact one college woman at random, the chance that she plays a sport is .3.

d. Two Polish statisticians in 2002 were questioning if the new Belgium Euro coin was indeed fair. They had their students flip the Belgium Euro 250 times, and 140 came up heads.

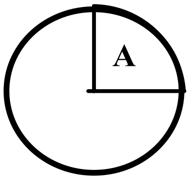
e. Many people are afraid of flying. But over the decade 1987-96, the death risk per flight on a US domestic jet has been 1 in 7 million.

f. In a roulette wheel, there are 38 slots numbered 0, 00, 1, …, 36. There are 18 ways of spinning an odd number, so the probability of spinning an odd is 18/38.

2. **Probability Viewpoints**

In the following problems, indicate if the given probability is found using the classical viewpoint, the frequency viewpoint, or the subjective viewpoint.

a. The probability that the spinner lands in the region A is ¼.



b. The meteorologist states that the probability of rain tomorrow is .5. You think it is more likely to rain and you think the chance of rain is ¾.

c. A football fan is 100% certain that his high school football team will win their game on Friday.

d. Jennifer attends a party, where a prize is given to the person holding a raffle ticket with a specific number. If there are eight people at the party, the chance that Jennifer wins the prize is 1/8.

e. What is the chance that you will pass this English class? You learn that the professor passes 70% of the students and you think you are typical in ability among those attending the class.

f. If you toss a plastic cup in the air, what is the probability that it lands with the open side up? You toss the cup 50 times and it lands open side up 32 times, so you approximate the probability by 32/50

3. **Equally Likely Outcomes**

For the following experiments, a list of possible outcomes is given. Decide if one can assume that the outcomes are equally likely. If the equally likely assumption is not appropriate, explain which outcomes are more likely than others.

a. A bowl contains six marbles of which two are red, three are white, and one is black. One marble is selected at random from the bowl and the color is observed.

Outcomes: {red, white, black}

b. You observe the gender of a baby born today at your local hospital.

Outcomes: {male, female}

c. Your school's football team is playing the top rated school in the country.

Outcomes: {your team wins, your team loses}

d. A bag contains 50 slips of paper, 10 that are labeled "1", 10 labeled "2", 10 labeled "3", 10 labeled "4", and 10 labeled "5". You choose a slip at random from the bag and notice the number on the slip.

Outcomes: {1, 2, 3, 4, 5}

4. **Equally Likely Outcomes**

For the following experiments, a list of possible outcomes is given. Decide if one can assume that the outcomes are equally likely. If the equally likely assumption is not appropriate, explain which outcomes are more likely than others.

a. You wait at a bus stop for a bus. From experience, you know that you wait, on average, 8 minutes for this bus to arrive.

Outcomes: {wait less than 10 minutes, wait more than 10 minutes}

b. You roll two dice and observe the sum of the numbers.

Outcomes: {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}

c. You get a grade for a English course in college.

Outcomes: {A, B, C, D, F}

d. You interview a person at random at your college and ask for his or her age.

Outcomes: {17 to 20 years, 21 to 25 years, over 25 years}

5. **Flipping a Coin**

Suppose you flip a fair coin until you observe heads. You repeat this experiment many times, keeping track of the number of flips it takes to observe heads. Here are the numbers of flips for 30 experiments.

1	3	1	2	1	1	2	6	1	2
1	1	1	1	3	2	1	1	2	1
5	2	1	7	3	3	3	1	2	3

a. Approximate the probability that it takes you exactly two flips to observe heads.

b. Approximate the probability that it takes more than two flips to observe heads.

c. What is the most likely number of flips?

6. **Driving to Work**

You drive to work 20 days, keeping track of the commuting time (in minutes) for each trip. Here are the twenty measurements.

25.4, 27.8, 26.8, 24.1, 24.5, 23.0, 27.5, 24.3, 28.4, 29.0

29.4, 24.9, 26.3, 23.5, 28.3, 27.8, 29.4, 25.7, 24.3, 24.2

a. Approximate the probability that it takes you under 25 minutes to drive to work.

b. Approximate the probability it takes between 25 and 28 minutes to drive to work.

c. Suppose one day it takes you 23 minutes to get to work. Would you consider this unusual? Why?

7. **A Man Sent to the Moon**

Consider your subjective probability P(M) where M is the event that the United States will send a man to the moon in the next twenty years.

a. Let B denote the event that you select a red ball from a box of five red and five white balls. Consider the two bets

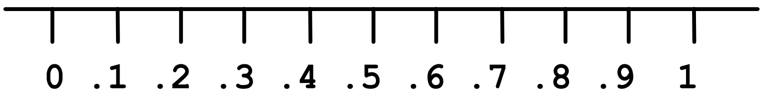
- BET 1 – If M occurs (United States will send a man to the moon in the next 20 years), you win $100. Otherwise, you win nothing.
- BET 2 – If B occurs (you observe a red ball in the above experiment), you win $100. Otherwise, you win nothing.

Circle the bet that you prefer.

b. Let B represent choosing red from a box of 7 red and 3 white balls. Again compare BET 1 with BET 2 – which bet do you prefer?

c. Let B represent choosing red from a box of 3 red and 7 white balls. Again compare BET 1 with BET 2 – which bet do you prefer?

d. Based on your answers to (a), (b), (c), circle the interval of values that contain your subjective probability P(M).



8. **What State Will You Be Living in the Future?**

Consider your subjective probability P(S) where S is the event that at age 30 you will be living in the same state as you currently live. Let B denote the event that you select a red ball from a box of five red and five white balls. Consider the two bets

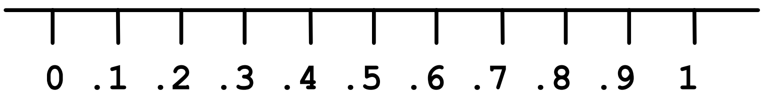
- BET 1 – If M occurs (you live in the same state at age 30), you win $100. Otherwise, you win nothing.
- BET 2 – If B occurs (you observe a red ball in the above experiment), you win $100. Otherwise, you win nothing.

Circle the bet that you prefer.

b. Let B represent choosing red from a box of 7 red and 3 white balls. Again compare BET 1 with BET 2 – which bet do you prefer?

c. Let B represent choosing red from a box of 3 red and 7 white balls. Again compare BET 1 with BET 2 – which bet do you prefer?

d. Based on your answers to (a), (b), (c), circle the interval of values that contain your subjective probability P(M).



9. **Frequency of Vowels in *Huckleberry Finn***

Suppose you choose a page at random from the book *Huckleberry Finn* by Mark Twain and find the first vowel on the page.

a. If you believe it is equally likely to find any one of the five possible vowels, fill in the probabilities of the vowels below.

Vowel	a	e	i	o	u
Probability					

b. Based on your knowledge about the relative use of the different vowels, assign probabilities to the vowels.

Vowel	a	e	i	o	u
Probability					

c. Do you think it is appropriate to apply the classical viewpoint to probability in this example? (Compare your answers to parts a and b.)

d. On each of the first fifty pages of *Huckleberry Finn*, your author found the first five vowels. Here is a table of frequencies of the five vowels:

Vowel	a	E	i	o	u
Frequency	61	63	34	70	22
Probability					

Use this data to find approximate probabilities for the vowels.

10. **Purchasing Boxes of Cereal**

Suppose a cereal box contains one of four different posters denoted A, B, C, and D. You purchase four boxes of cereal and you count the number of posters (among A, B, C, D) that you *do not* have. The possible number of "missing posters" is 0, 1, 2, and 3.

a. Assign probabilities if you believe the outcomes are equally likely.

Number of missing posters	0	1	2	3
Probability				

b. Assign probabilities if you believe that the outcomes 0 and 1 are most likely to happen.

Number of missing posters	0	1	2	3
Probability				

c. Suppose you purchase many groups of four cereals, and for each purchase, you record the number of missing posters. The number of missing posters for 20 purchases is displayed below. For example, in the first purchase, you had 1 missing poster, in the second purchase, you also had 1 missing poster, and so on.

1, 1, 1, 2, 1, 1, 0, 0, 2, 1, 2, 1, 3, 1, 2, 1, 0, 1, 1, 1

Using these data, assign probabilities.

Number of missing posters	0	1	2	3
Probability				

d. Based on your work in part c, is it reasonable to assume that the four outcomes are equally likely? Why?

TOPIC P2: SAMPLE SPACE AND ASSIGNING PROBABILITIES

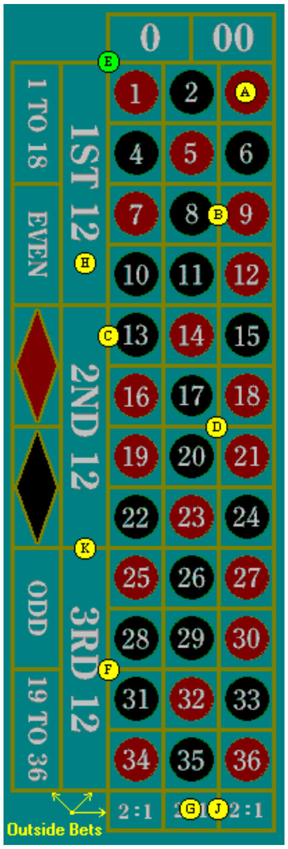
SPOTLIGHT: THE CASINO GAME OF ROULETTE

Roulette is one of the most popular casino games. The name roulette is derived from the French word meaning small wheel. Although the origin of the game is not clear, it became very popular during the 18th century when Prince Charles introduced gambling to Monaco to alleviate the country's current financial problems. The game was brought to America in the early part of the 19th century and is currently featured in all casinos. In addition, roulette is a popular game among people who like to game online.

The American version of the game that we discuss in this book varies slightly from the European version. The American roulette wheel contains 38 pockets, numbers 1 through 36 plus zero plus double zero. The wheel is spun and a small metal ball comes to rest in one of the 38 pockets.

A typical roulette table is pictured below. Players will place chips on particular locations on the table, predicting where the ball will land when after the wheel is spun and the ball comes to a stop. The dealer places a mark on the winning number. The players who have bet on the winning number are rewarded while the players who bet on losing numbers lose their chips to the casino.

In a later topic, we will introduce different types of bets that a player can make in roulette and learn about the average winnings per pay if you place a particular bet. It is important to remark that American roulette is a game that favors the casino, and a player will on average lose money by placing any type of bet. It may be a nice form of entertainment, but it certainly is not a way to earn money.



PREVIEW

A ***random experiment*** is the name for some process where the outcome is random or unknown. Examples of some random experiments are

- tossing four coins and observing the number of heads that land up
- watching cars that pass a given intersection and counting the number of cars until you see a white one
- recording the time (in minutes) it takes you to get to work
- observing the amount of money you will win on a lottery ticket bought today

Before we can talk about probabilities, we first need to get a good handle on possible outcomes of the random experiment. A list of all possible outcomes is called the ***sample space*** and denoted by S. In this topic, we describe different ways of writing down sample spaces. Probabilities are numbers assigned to outcomes of the sample space that satisfy basic rules. We will get some experience in specifying probabilities and using the rules to derive other properties about probabilities.

In this topic your learning objectives are to:

- Understand how to write down sample spaces for a variety of random experiments.
- Understand how to specify reasonable sets of probabilities satisfying certain assumptions.
- Understand basic properties of probabilities.

NCTM Standards

✓ In Grades 6-8, all students should understand and use appropriate terminology to describe complementary and mutually exclusive events.

✓In Grades 6-8, all students should compute probabilities for simple compound events, using such methods as organized lists, and tree diagrams.

✓In Grades 9-12, all students should understand the concepts of sample space and probability distribution and construct sample spaces in simple cases.

WARMUP ACTIVITY: WRITING DOWN SOME SAMPLE SPACES

For each of the following situations, write down a list of all possible outcomes.

1. TOSSING FOUR COINS. Suppose you toss four coins and you record the number of coins that show heads. What are the possible outcomes?

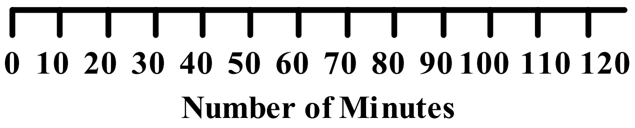
2. WATCHING CARS. Suppose you are standing at a particular intersection and you record the number of cars you observe before you see a white one. What are different outcomes for the number of cars you observe?

Note: This second example illustrates an infinite sample space, since the number of possible outcomes is infinite.

3. HOW LONG IS YOUR COMMUTE? Suppose you are a commuter and your home is located 25 miles from campus. Generally it takes you 40 minutes to get to school, but there is some variation in this commuting time. On a quiet day with little traffic, it is possible that it will only take 35 minutes to get to work. On other days, there is construction on the road and you can get stuck in traffic, and it will take close to an hour to work.

(a) Is it possible to write down all possible commuting times? Why or why not?

(b) When the outcome is continuous-valued, one convenient way of representing the sample space is by use of a number line. On the number line below, shade the region of possible commuting times.



4. WINNING IN A LOTTERY. What happens when you buy a lottery ticket? Well, your ticket could be a loser (and you win nothing), or you may win different dollar amounts depending on how closely your ticket number agrees with the winning number. Suppose that the possible winning dollar amounts are $10, $100, or $50,000 (the last prize corresponds to the big jackpot). You record the amount of money you win from one ticket. Write down the sample space.

DIFFERENT REPRESENTATIONS OF A SAMPLE SPACE

A sample space lists all possible outcomes of a random experiment. There are different ways to write down the sample space, depending on how we think about outcomes.

Let's illustrate the variety of sample spaces by the simple experiment

"roll two fair dice"

Each die is the usual six-sided object that we are familiar with, with the numbers 1, 2, 3, 4, 5, 6 on each side. When we say "fair dice", we are imagining that each die is constructed such that the six possible numbers are equally likely to come up when rolled.

What can happen when you roll two dice? The collection of all outcomes that are possible is the sample space. But there are different ways of representing the sample space depending on what "outcome" we are considering.

First, suppose we are interested in the *sum of the numbers* on the two dice. (This would be of interest to a gambler playing the casino game craps.) What are the possible sums? After some thought, it should be clear that the smallest possible sum is 2 (if you roll two ones) and the largest possible sum is 12 (with two sixes). Also every whole number between 2 and 12 is a possible sum. So the sample space, denoted by S, would be

S = {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}.

Suppose instead that you wish to record *the rolls on each of the two dice*. One possible outcome would be

(4 on one die, 3 on the other die)

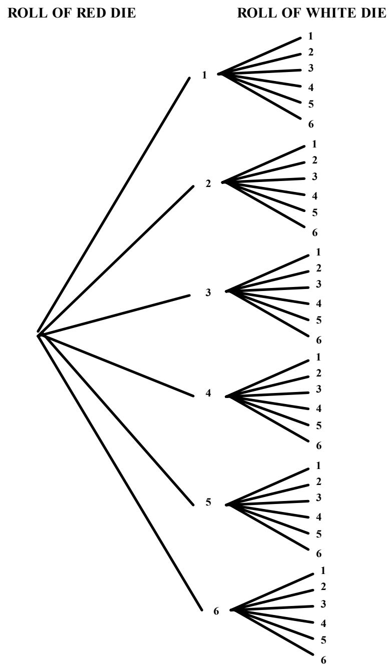
or more simply (4, 3). What are the possible outcomes? Here are the twenty-one possibilities:

(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6)
(2, 2), (2, 3), (2, 4), (2, 5), (2, 6)
(3, 3), (3, 4), (3, 5), (3, 6)
(4, 4), (4, 5), (4, 6)
(5, 5), (5, 6)
(6, 6)

Notice that we aren't distinguishing between the two dice in this list. For example, we just wrote down (2, 3) once, although there are two ways for this to happen - either the first die is 2 and the second die is 3 or the other way around.

Last, suppose you can distinguish the two dice -- perhaps one die is red and one die is white -- and you are considering all of the possible rolls of both dice. We illustrate three ways of showing the sample space in this case.

One way of representing possible rolls of two distinct dice is by a *tree diagram* shown on the next page. On the left side of the diagram, we represent the six possible rolls of the red die by six branches of a tree. Then, on the right side, we represent the six possible rolls of the white die by six smaller branches coming out of each roll of the red die. A single branch on the left and a single branch on the right represent one possible outcome of this experiment.



An alternative way of representing the possible outcomes of rolling two distinct dice uses a rectangular grid or table. The six possible rolls of the white die are the rows of the table, the six possible rolls of the red die correspond to the columns of the table, and the possible outcomes are represented by the x's in the table.

		Roll on red die					
		One	Two	Three	four	five	six
Roll on white die	One	x	X	X	x	x	x
	Two	x	X	X	x	x	x
	Three	x	X	X	x	x	x
	Four	x	X	X	x	x	x
	Five	x	X	X	x	x	x
	Six	x	X	X	x	x	x

There are still other ways to represent the outcomes of this experiment of rolling two distinct dice. Suppose we write down an outcome by the ordered pair

(roll on white die, roll on red die).

Then the possible outcomes are listed below.

(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6)

(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6)

(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6)

(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6)

(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6)

(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)

Since these are ordered pairs, the *order* of the numbers does matter. The outcome (5, 1) (5 on the red, 1 on the white) is different from the outcome (1, 5) (1 on the red die and 5 on the white die).

We have illustrated three representations of the sample space of possible rolls of two dice. These representations differ by how we record the outcome of rolling two dice. We can either (1) record the sum of the two dice, (2) record the individual rolls, not distinguishing the two dice, or (3) record the individual rolls, distinguishing the two dice. Which one is the best sample space to use? Actually, all of the sample spaces shown above are correct. Each sample space represents all possible outcomes of the experiment of rolling two dice and you can't say that one sample space is better than another sample space. But we will see that in particular situations some sample spaces are more convenient than other sample spaces when we wish to assign probabilities. (In this case the sample space with distinguishable dice is desirable from the viewpoint of computing probabilities since the outcomes are equally likely.)

When we write down sample spaces, we use whatever method we like. We can use a tree diagram or a rectangular grid, or we might like to list the outcomes. The important thing is that we have shown all of the possible outcomes in S.

PRACTICE: DIFFERENT WAYS OF REPRESENTING SAMPLE SPACES

Suppose you take a survey and you ask the question: "Do you think the world is safer today than it was ten years ago?" and each person will either respond "yes" or "no". You ask this survey question to three people.

1. If you are only interested in the number of people who say yes, write down the sample space.

2. Suppose you record the answer of person 1, person 2, and person 3 with a Y for yes and an N for no. To illustrate, a "yes" by person 1, a "no" by person 2, and a "yes" by person 3 is recorded by YNY. Write down the sample space by listing all of the possible outcomes.

3. For the same situation as part 2, use a tree diagram to represent the sample space.

4. Is it reasonable to assume that the outcomes in the sample space in part 2 are equally likely? Why or why not?

ASSIGNING PROBABILITIES

When we have a random experiment, the first step is to list all of the possible outcomes in the sample space. The next step is to assign numbers, called probabilities, to the different outcomes that reflect the likelihoods that these outcomes can occur.

To illustrate different assignments of probabilities, suppose my daughter goes to an ice cream parlor and plans to order a single-dip ice cream cone. This particular parlor has four different ice cream flavors. Which flavor will my daughter order?

First, we write down the sample space -- the possible flavors that my daughter can order. Underneath we will place probabilities to these four possible outcomes that reflect my beliefs about her likes and dislikes.

	Flavor			
	Vanilla	Chocolate	Butter Pecan	Maple Walnut
Probability				

Can our probabilities be any numbers? Not exactly. Here are some basic facts (or laws) about probabilities:

- Any probability we assign must fall between 0 and 1
- The sum of the probabilities across all outcomes must be equal to 1.
- We can give an outcome a probability of 0 if we are sure that that outcome will never occur.
- Likewise, if we assign a probability of 1 to an event, then that event must occur all the time.

With these facts in mind, we consider some possible probability assignments for the flavor of ice cream that my daughter will order.

SCENARIO 1: Suppose that my daughter likes to be surprised. She has brought a hat in which she has placed many slips of paper -- 10 slips are labeled "vanilla", 10 slips are labeled "chocolate", and 10 slips are "butter pecan", and 10 are "maple walnut". She makes her ice cream choice by choosing a slip at random. In this case, each flavor would have a probability of 10/40 = ¼ .

	Flavor			
	Vanilla	Chocolate	Butter Pecan	Maple Walnut
Probability	¼	¼	¼	¼

SCENARIO 2: Let's consider a different set of probabilities based on different assumptions about my daughter's taste preferences. I know that she really doesn't like "plain" flavors like vanilla or chocolate, and she really likes ice creams with nut flavors. In this case, I would assign “Vanilla” and “Chocolate” each a probability of 0, and assign the two other flavors probabilities that sum to one. Here is one possible assignment.

	Flavor

	Vanilla	Chocolate	Butter Pecan	Maple Walnut
Probability	0	0	.7	.3

Another possible assignment of probabilities that is consistent with these assumptions is

	Flavor			
	Vanilla	Chocolate	Butter Pecan	Maple Walnut
Probability	0	0	.2	.8

SCENARIO 3: Let's consider an alternative probability assignment from a different person's viewpoint. The worker at the ice cream shop has no idea what flavor my daughter will order. But she's been working at the shop all day and she has kept a record of how many cones of each type have been ordered – of 50 cones ordered, 10 are vanilla, 14 are chocolate, 20 are butter pecan, and 6 are maple walnut. If she believes that my daughter has similar tastes to the previous customers, then it would be reasonable to apply the frequency viewpoint to assign the probabilities.

	Flavor			
	Vanilla	Chocolate	Butter Pecan	Maple Walnut
Probability	10/50	14/50	20/50	6/50

Each of the above probability assignments used a different viewpoint of probability as described in Topic P1. The first assignment used the classical viewpoint where each of the forty slips of paper had the same probability of being selected. The second assignment was an illustration of the subjective view where my assignment was based on my opinion about the favorite flavors of my daughter. The last assignment was based on the frequency viewpoint where the probabilities were estimated from the observed flavor preferences of 50 previous customers.

PRACTICE: ASSIGNING PROBABILITIES

Topic P2: Sample Space and Assigning Probabilities

Suppose you are interested in assigning probabilities to each of the final possible grades (A, B, C, D, and F) in one class that you are currently taking. Assign probabilities to all five grades based on the given information. (It is possible that there are multiple probability assignments that are possible. In this case, you need to give only one of the possible assignments.)

1. Assign probabilities to all 5 grades if the only possible grades are A or B.

Grade	A	B	C	D	F
Probability					

2. Assign probabilities to all 5 grades if each possible grade is equally likely.

Grade	A	B	C	D	F
Probability					

3. Suppose 200 students previously took this class – 30 got A's, 80 got B's, 50 got C's, 30 got D's, and 10 failed. If you believe you are similar in ability to these 200 students, assign probabilities.

Grade	A	B	C	D	F
Probability					

4. Assign probabilities if the only possible grades are B and C, and B is twice as likely as C.

Grade	A	B	C	D	F
Probability					

5. (Continuation of part 3.) Assign probabilities if you believe you are better than a typical student taking this class.

Grade	A	B	C	D	F
Probability					

6. In each of the parts 1 – 5, explain if you used the classical, subjective, or frequency viewpoint in assigning the probabilities.

A MORE FORMAL LOOK AT PROBABILITY

In this topic, we have discussed probability in an informal way. We assign numbers called probabilities to outcomes in the sample space such that the sum of the numbers over all outcomes is equal to one. Here we look at probability from a more formal viewpoint. We define probability as a function on events that satisfies three basic laws or axioms. Then it turns out that all of the important facts about probabilities, including some facts that we used above, can be derived once we are given these three basic axioms.

Suppose that we write the sample space for our random experiment as S. An event, represented by a capital letter such as A, is a subset of S. Events, like sets, can be combined in various ways. We write

- $A \cap B$ as the event that both A and B occur (the *intersection* of the two events)
- $A \cup B$ as the event that either A or B occur (the *union* of the two events)
- $\bar{A}$ as the event that A does not occur (the *complement* of the event A)

To illustrate these set operations, suppose you choose a student at random from your class and record the month when she or he was born. The student could be born during 12 possible months and the sample space S is the list of these months:

S = {January, February, March, April, May, June, July, August, September,
October, November, December}.

Suppose event L is the event that the student is born during the last half of the year and F is the event that the student is born during a month that is four letters long. So

L = { July, August, September, October, November, December }, and
F = {June, July}.

We can illustrate various set operations:

- $L \cap F$ is the event that the student is born during the last half of the year AND is born in a four-letter month = {July}.
- $L \cup F$, in contrast, is the event that the student is EITHER born during the last half of the year OR born in a four-letter month = {June, July, August, September, October, November, December}.
- $\bar{L}$ is the event that the student is NOT born during the last half of the year = { January, February, March, April, May, June}

PRACTICE: SET OPERATIONS

Suppose in a simple lottery game, you choose two different numbers from the set {1, 2, 3, 4, 5}. The order that you choose the numbers is not important.

1. Write down the sample space *S* (you should have 10 different outcomes).

2. Let *O* denote the event that you choose the number 1, and let *S* denote the event that the sum of the two numbers is equal to six. Write down the set of outcomes for *O* and the set of outcomes for *S*.

3. Write down the outcomes in

(a) $O \cap S$

(b) $O \cup S$

(c) $\bar{O}$

4. We say that two events are mutually exclusive if they have no outcomes in common. Are *O* and *S* mutually exclusive? Why or why not?

5. Find an event *A* such that *A* and *O* are mutually exclusive.

The Three Probability Axioms

Now that we have set up a sample space S and events, we can define probabilities that are numbers assigned to events. There are three basic laws or axioms that define probabilities:

- **Axiom 1**: For any event A, $P(A) \geq 0$. (That is, all probabilities are nonnegative values.)
- **Axiom 2**: $P(S) = 1$ (That is, the probability that you observe something in the sample space is one.)
- **Axiom 3**: Suppose you have a sequence of events $A_1, A_2, A_3, \ldots$ that are mutually exclusive. (This means that for any two events in the sequence, say A_2 and A_3, the intersection of the two events is the empty set – that is, $A_2 \cap A_3 = \varnothing$.) Then you find the probability of the union of the events by adding the individual event probabilities:

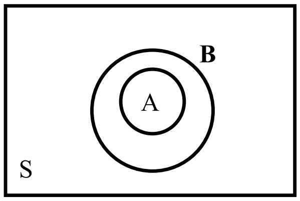
$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$$

Given the three basic axioms, we can prove some additional facts about probabilities. We call these additional facts *properties* – these are not axioms, but rather additional facts that can be derived knowing the axioms. Below we state several familiar properties about probabilities and prove how each property follows logically from the axioms.

Property 1: A subset of B $A \subset B$, $P(A) \leq P(B)$

This property says that if you have two sets, such that one set is a subset of another set, then the probability of the first set can't exceed the probability of the second set. This may seem pretty obvious, but how can we prove this from the axioms?

Proof: We start with a Venn diagram where a set A is a subset of set B.



Note that we can write the larger set B as the union of A and $\overline{A} \cap B$ -- that is,

$$B = A \cup (\overline{A} \cap B)$$

Note that $A \cap B$ and $\overline{A} \cap B$ are mutually exclusive (they have no overlap). So we can apply Axiom 3 and write

$$P(B) = P(A) + P(\overline{A} \cap B)$$

Also, by Axiom 1, the probability of any event is nonnegative. So we have shown that the probability of B is equal to the probability of A plus a nonnegative number. So this implies

$$P(B) \geq P(A),$$

which is what we wish to prove.

Property 2: $P(A) \leq 1$

This is pretty obvious – we know probabilities can't be larger than 1. But how can we prove this given our known facts that include the axioms and Property 1 that we just proved?

Proof: Actually this can be shown as a consequence of Property 1. Consider the two sets A and the sample space S. Obviously A is a subset of the sample space – that is,

$$A \subset S$$

So applying Property 1,

$$P(A) \leq P(S) = 1$$

(We know P(S) from the second axiom.) So we have proved our result.

PRACTICE: PROVING SEVERAL PROPERTIES OF PROBABILITIES.

In this practice, we will outline the proofs of several properties of probabilities. Each step in the proof is written down and you are asked to justify why this step is true.

Property 3: $P(\varnothing) = 0$

This third property says that the probability of the empty set (the event consisting of no members) is equal to zero.

Proof: In the following, we outline the steps of the proof and you are asked to give the rationale for each step.

Step 1: $\varnothing = \varnothing \cup \varnothing$

Why is this true?

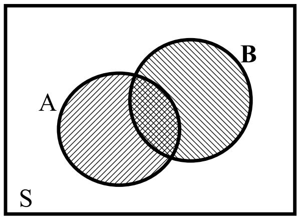
Step 2: $P(\varnothing) = P(\varnothing) + P(\varnothing)$

Why is this true?

Step 3: $P(\varnothing) = 0$

Why is this true?

Property 4: $P(A) = P(A \cap B) + P(A \cap \overline{B})$



A Venn diagram showing the two sets A and B is shown above.

Step 1: Write set A as the union of two sets that are mutually exclusive.

Step 2: Apply Axiom 3 to the union statement in Step 1.

THE COMPLEMENT AND ADDITION RULES

There are two additional properties of probabilities that we will find useful in computation. Both of these properties will be stated without proof, but an outline of the proofs will be given in the exercises.

The first property, called the *complement rule*, says that the property of the complement of an event is simply one minus the probability of the event.

Complement rule: For an event A, $P(\overline{A}) = 1 - P(A)$.

The second property, called the *addition rule*, gives a formula for the probability of the union of two events.

Addition rule: For two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Both of these rules are best illustrated by an example. Let us revisit our example where we are interested in the birth month of a child selected from our class. As before, we let L represent the event that the student is born during the last half of the year and F denote the event that the student is born during a month that is four letters long.

There are 12 possible outcomes for the birth month. One could assume that each month is equally likely to occur, but actually in the U.S. population, the numbers of births during the different months do vary. Using data from the births in the U.S. in 1978, we obtain the following probabilities for the months. We see that August is the most likely birth month with a probability of .091 and February (the shortest month) has the smallest probability of .075.

MONTH	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
PROB	.081	.075	.083	.076	.082	.081	.088	.091	.088	.087	.082	.085

Using this probability table, we find

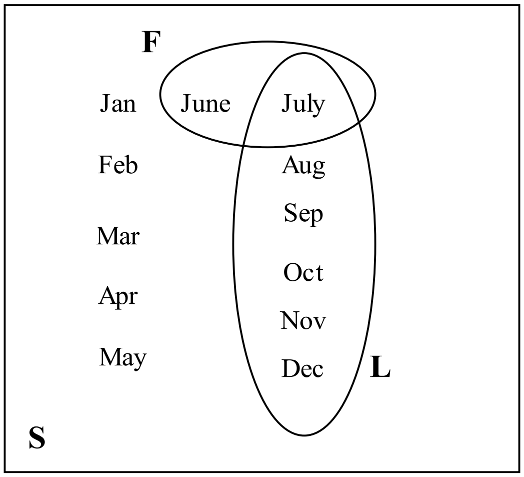
1. P(L) = P(July, August, September, October, November, December)
   = .088 +.091+.088+.098+.082+.085 = .521.
2. P(F) = P(June, July) = .081+.088 = .169.

Now we are ready to illustrate the two probability rules.

What is the probability the student is *not* born during the last half of the year? We could find this by summing the probabilities of the first six months of the year. It is easier to note that we wish to find the probability of the complement of the event L, and we apply the complement rule to find the probability.

$$P(\bar{L}) = 1 - P(L) = 1 - .521 = .479.$$

What is the probability the student is either born during the last six months of the year *or* a month four letters long? In the below figure, we show the sample space S consisting of the twelve possible birth months and the sets F and L are shown by circling the relevant outcomes. We wish to find the probability of the set $F \cup L$ which is the union of the two circled sets.



Applying the addition rule, we find the probability of $F \cup L$ by adding the probabilities of *F* and *L* and subtracting the probability of the intersection event $F \cap L = \{July\}$:

$$\begin{aligned} P(F \cup L) &= P(F) + P(L) - P(F \cap L) \\ &= .521 + .169 - .088 = .602 \end{aligned}$$

Looking at the figure, the formula should make sense. When we add the probabilities of the events *F* and *L*, we add the probability of the month July twice, and to get the correct answer, we need to subtract the outcome (July) that is common to both *F* and *L*.

SPECIAL NOTE: Is it possible to simply add the probabilities of two events, say *A* and *B*, to get the probability of the union $A \cup B$? Suppose the sets *A* and *B* are mutually exclusive which means they have no outcomes in common. In this special case, $A \cap B = \varnothing$, $P(A \cap B) = 0$, and $P(A \cup B) = P(A) + P(B)$. For example, suppose you are interested that your student is born in the last half the year (event L) or in May (event M). Here it is not possible to be born in the last half of the year and in May so $L \cap M = \varnothing$. In this case, $P(L \cup M) = P(L) + P(M) = .521 + .082 = .603$.

PRACTICE: THE ADDITION AND COMPLEMENT RULES

Consider again the simple lottery game where you choose two numbers (without replacement} from the set {1, 2, 3, 4, 5}. There are 10 possible outcomes shown below.

$$\{1, 2\}, \{1, 3\}, \{1, 4\}, \{1, 5\}, \{2, 3\}, \{2, 4\}, \{2, 5\}, \{3, 4\}, \{3, 5\}, \{4, 5\}$$

Here each outcome is equally likely, so we can assign a probability of 1/10 to each. As before, let O denote the event that the number 1 is chosen, and S denote the event that the two numbers chosen sum to 6.

1. Find P(O) and P(S).
2. Using the complement rule, find the probability $P(\bar{O})$.
3. Are O and S mutually exclusive events? Why or why not?
4. Use the addition rule to find $P(O \cup S)$.
5. Find an event C such that events C and S are mutually exclusive.

WRAP-UP

In this topic, we started talking about assigning probabilities. First one writes down a list of all possible outcomes, the sample space, and then one assigns probabilities to the different outcomes. There are basic rules, called axioms, that all probabilities must follow. The probability of any event must be nonnegative, the probability of the sample space is equal to 1, and the probability of a union of mutually exclusive events is the sum of the probabilities of the events. There are additional facts about probabilities, called properties that can be derived from the axioms. Generally, probabilities are difficult to specify, but one can assign reasonable sets of probabilities given information for a particular problem.

EXERCISES

1. **Writing Sample Spaces**

For the following random experiments, give an appropriate sample space for the random experiment. You can use any method (a list, a tree diagram, a two-way table) to represent the possible outcomes.

a. You simultaneously toss a coin and roll a die.

b. Construct a word from the five letters a, a, e, e, s

c. Suppose a person lives at point 0 and each second she randomly takes a step to the right or a step to the left. You observe the person's location after four steps.

d. In the first round of next year's baseball playoff, the two teams, say the Phillies and the Diamondbacks play in a best-of-five series where the first team to win three games wins the playoff.

e. A couple decides to have children until a boy is born.

f. A roulette game is played with a wheel with 38 slots numbered 0, 00, 1, …, 36. Suppose you place a $10 bet that an even number (not 0) will come up in the wheel. The wheel is spun.

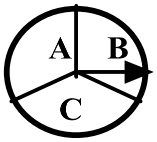
g. Suppose three batters, Marlon, Jimmy, and Bobby, come to bat during one inning of a baseball game. Each batter can either get a hit, walk, or get out.

2. **Writing Sample Spaces**

For the following random experiments, give an appropriate sample space for the random experiment. You can use any method (a list, a tree diagram, a two-way table) to represent the possible outcomes.

a. You toss three coins.

b. You spin the spinner (shown to the right) three times.



c. When you are buying a car, you have a choice of three colors, two different engine sizes, and whether or not to have a CD player. You make each choice completely at random and go to the dealership to pick up your new car.

d. Five horses, Lucky, Best Girl, Stripes, Solid, and Jokester compete in a race. You record the horses that win, place, and show (finish first, second, and third) in the race.

e. You and a friend each think of a whole number between 0 and 9.

f. On your computer, you have a playlist of 4 songs denoted by a, b, c, d. You play them in a random order.

g. Suppose a basketball player takes a "one-and-one" foul shot. (This means that he attempts one shot and if the first shot is successful, he gets to attempt a second shot.)

3. **Writing Sample Spaces**

For the following random experiments, give an appropriate sample space for the random experiment. You can use any method (a list, a tree diagram, a two-way table) to represent the possible outcomes.

a. Your school plays four football games in a month.

b. You call a "random" household in your city and record the number of hours that the TV was on that day.

c. You talk to an Ohio resident who has recently received her college degree. How many years did she go to college?

d. The political party of our next elected U.S. President.

e. The age of our next President when he/she is inaugurated.

f. The year a human will next land on the moon.

4. **Writing Sample Spaces**

For the following random experiments, give an appropriate sample space for the random experiment. You can use any method (a list, a tree diagram, a two-way table) to represent the possible outcomes.

a. The time you arrive at your first class on Monday that begins at 8:30 AM.

b. You throw a ball in the air and record how high it is thrown (in feet).

c. Your cost of textbooks next semester.

d. The number of children you will have.

e. You take a five question true/false test.

f. You drive on the major street in your town and pass through four traffic lights.

5. **Probability Assignments**

Give reasonable assignments of probabilities based on the given information.

a. In the United States, there were 4058 thousand babies born in the year 2000 and 1980 thousand were girls. Assign probabilities to the possible genders of your next child.

Gender	Boy	Girl
Probability		

b. Next year, your school will be playing your neighboring school in football. Your neighboring school is a strong favorite to win the game.

Winner of game	Your school	Your neighboring school
Probability		

c. You have an unusual die that shows 1 on two sides, 2 on two sides, and 3 and 4 on the remaining two sides.

Roll of die	1	2	3	4	5	6
Probability						

6. **Probability Assignments**

Based on the given information, decide if the stated probabilities are reasonable. If they are not, explain how they should be changed.

a. Suppose you play two games of chess with a chess master. You can either win 0 games, 1 game, or 2 games, so the probability of each outcome is equal to 1/3.

b. Suppose 10% of cars in a car show are Corvettes and you know that red is the most popular Corvette color. So the chance that a randomly chosen car is a red Corvette must be larger than 10%.

c. In a Florida community, you are told that 30% of the residents play golf, 20% play tennis, and 40% of the residents play golf and tennis.

d. Suppose you are told that 10% of the students in a particular class get A, 20% get B, 20% get C, and 20% get D. That means that 30% of the class must fail the class.

7. **Finding the Right Key**

Suppose your key chain has five keys, one of which will open up your front door of your apartment. One night, you randomly try keys until the right one is found. Here are the possible numbers of keys you will try until you get the right one:

1 key, 2 keys, 3 keys, 4 keys, 5 keys

a. Circle the outcome that you think is most likely to occur.

1 key, 2 keys, 3 keys, 4 keys, 5 keys

b. Circle the outcome that you think is least likely to occur.

1 key, 2 keys, 3 keys, 4 keys, 5 keys

c. Based on your answers to parts a and b, assign probabilities to the six possible outcomes.

Outcome	1 key	2 keys	3 keys	4 keys	5 keys
Probability					

8. **Playing Roulette**

One night in Reno, you play roulette five times. Each game you bet $5 – if you win, you win $10; otherwise, you lose your $5. You start the evening with $25. Here are the possible amounts of money you will have after playing the five games.

$0, $10, $20, $30, $40, $50 .

a. Circle the outcome that you think is most likely to occur.

$0, $10, $20, $30, $40, $50

b. Circle the outcome that you think is least likely to occur.

$0, $10, $20, $30, $40, $50

c. Based on your answers to parts a and b, assign probabilities to the six possible outcomes.

Outcome	$0	$10	$20	$30	$40	$50
Probability						

9. **Cost of Your Next Car**

Consider the cost of the next new car you will purchase in the future. There are five possibilities:

cheapest: the car will cost less than $5000

cheaper: the car will cost between $5000 and $10,000.

moderate: the car will cost between $10,000 and $20,000

expensive: the car will cost between $20,000 and $30,000

really expensive: the car will cost over $30,0000

a. Circle the outcome that you think is most likely to occur.

cheapest, cheaper, moderate, expensive, really expensive

b. Circle the outcome that you think is least likely to occur.

cheapest, cheaper, moderate, expensive, really expensive

c. Based on your answers to parts a and b, assign probabilities to the five possible outcomes.

Outcome	cheapest	cheaper	moderate	Expensive	Really expensive
Probability					

10. **Flipping a Coin**

Suppose you flip a coin twice. There are four possible outcomes (H stands for heads and T stands for tails).

HH, HT, TH, TT

a. Circle the outcome that you think is most likely to occur.

HH, HT, TH, TT

b. Circle the outcome that you think is least likely to occur.

HH, HT, TH, TT

c. Based on your answers to parts a and b, assign probabilities to the four possible outcomes.

Outcome	HH	HT	TH	TT
Probability				

11. **Playing Songs in Your IPod**

Suppose you play three songs by Jewell (J), Madonna (M), and Plumb (P) in a random order.

a. Write down all possible ordering of the three songs.

b. Let M = event that the Madonna song is played first and B = event that the Madonna song is played before the Jewell song. Find P(M) and P(B).

c. Write down the outcomes in the event $M \cap B$ and find the probability $P(M \cap B)$

d. By use of the complement rule, find $P(\overline{B})$.

e. By use of the addition rule, find $P(M \cup B)$.

12. **Student of the Day**

Suppose that students at a local high school are distributed by grade level and gender by the following table.

	Freshmen	Sophomores	Juniors	Seniors	TOTAL
Male	25	30	24	19	98
Female	20	32	28	15	95
TOTAL	45	62	52	34	193

Suppose that a student is chosen at random from the school to be the "student of the day". Let F = event that student is a freshmen, J = event that student is a junior, and M = event that student is a male.

a. Find the probability $P(\overline{F})$.

b. Are events F and J mutually exclusive. Why?

c. Find $P(F \cup J)$.

d. Find $P(F \cap M)$

e. Find $P(F \cup M)$

13**. Proving Properties of Probabilities**

Given the three probability axioms and the properties already proved, prove the complement rule $P(\overline{A}) = 1 - P(A)$. An outline of the proof is written below.

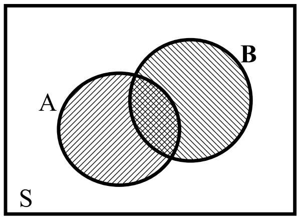
a. Write the sample space S as the union of the sets A and $\overline{A}$.

b. Apply Axiom 3.

c. Apply Axiom 2.

14. **Proving Properties of Probabilities**

Given the three probability axioms and the properties already proved, prove the addition rule $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. A Venn diagram and an outline of the proof are written below.



a. Write the set $A \cup B$ as the union of three sets that are mutually exclusive.

b. Apply Axiom 2 to write $P(A \cup B)$ as the sum of three terms.

c. Write the set A as the union of two mutually exclusive sets.

d. Apply Axiom 2 to write *P(A)* as the sum of two terms.

e. By writing B as the union of two mutually exclusive sets and applying Axiom 2, write *P(B)* as the sum of two terms.

f. By making appropriate substitutions to the expression in part b, one obtains the desired result.

TOPIC P3: LET ME COUNT THE WAYS ...

SPOTLIGHT: ROLLING DICE AND YAHTZEE

Dice are one of the oldest randomization devices known to man. Egyptian tombs, dated from 2000 BC, were found containing dice and there have some evidence of dice in archaeological excavations dating back to 6000 BC. It is interesting to note that dice appeared to be invented independently by many ancient cultures across the world. In ancient times, the result of a die throw was not just considered luck, but determined by gods. So casting dice was often used as a way of making decisions such as choosing rulers or dividing inheritances. The Roman goddess, Fortuna, daughter of Zeus was believed to determine the outcome of a throw.

In the 19th and 20th centuries, standard six-sided dice became a basic component of many commercial board games that were developed. One of the most current popular games is Yahtzee that is played with five dice. The Hasbro game company (http://www.hasbro.com) presents the history of the game. Yahtzee was invented by a wealthy Canadian couple to play aboard their yacht. This "yacht" game was popular among the couple's friends, who wanted copies of the game for themselves. The couple approached Mr. Edwin Lowe, who made a fortune selling bingo games, about marketing the game. Mr. Lowe's initial attempts to sell the game of Yahtzee by placing ads were not successful. Lowe thought that the game had to be played to be appreciated and he hosted a number of Yahtzee parties and the game became very successful. The Milton Bradley company acquired the E. S. Lowe Company and Yahtzee in 1973 and currently more than 50 million games are sold annually.

PREVIEW

In this topic, we consider a situation where probabilities are easy to compute. Suppose we write down all of the outcomes of our random experiment in such a way so that all of the outcomes are equally likely. Then we can compute a probability of interest by counting the number of outcomes in the sample space and counting the number of outcomes in the event of interest. We review some basic rules helpful in counting outcomes.

In this topic, your learning objectives are to:

- Understand and apply the multiplication counting rule.
- Understand the notion of an arrangement and be able to compute the number of arrangements of distinct objects in a given application.
- Understand the use of a combinations rule when objects are selected without regard for order.
- Understand which counting rule is appropriate in a given application.

NCTM Standards

✓ In Grades 6-8, all students should compute probabilities for simple compound events, using such methods as organized lists, and tree diagrams,

✓In Grades 9-12, all students should understand the concepts of sample space and probability distribution and construct sample spaces and distributions in simple cases

EQUALLY LIKELY OUTCOMES

Assume we can write the sample space in such a way that the outcomes are equally likely. Then, applying the classical interpretation, the probability of each outcome will be

$$P(outcome) = \frac{1}{number\ of\ outcomes}.$$

If we are interested in the probability of some event, then the probability is given by

$$P(event) = \frac{number\ of\ outcomes\ in\ event}{total\ number\ of\ outcomes}.$$

This simple formula should be used with caution. To illustrate the use (and misuse) of this formula, suppose you have a box containing five balls of which three are red, one is blue, and one is white. You select three balls without replacement from the box – what is the probability that you choose all red balls?

Let's consider two representations of the sample space of this experiment.

Sample space 1: Suppose we don't distinguish between balls of the same color and don't care about the order in which the balls are selected. Then if we let R, B, W denote choosing a red, blue, and white ball respectively, then there are four possible outcomes:

S1 = {{R, R, R}, {R, R, B}, {R, R, W}, {R, B, W}).

If these outcomes in S1 are assumed equally likely, then the probability of choosing all red balls is

Prob(all reds) = ¼.

Sample space 2: Suppose instead that we distinguish the balls of the same color, so the balls in the box are denoted by R1, R2, R3, B, W. Then we can write down ten possible outcomes

S2 = {{R1, R2, R3}, {R1, R2, B}, {R1, R2, W}, {R1, R3, B}, {R1, R3, W}, {R2, R3, B}, {R2, R3, W}, {R1, B, W}, {R2, B, W}, {R3, B, W}}.

If we assume these outcomes are equally likely, then the probability of choosing all reds is

Prob(all reds) = 1/10.

If we compare our answers, we see an obvious problem since we get two different answers for the probability of choosing all reds. What is going on? The problem is that the outcomes in the first sample space S1 *are not* equally likely. In particular, the chance of choosing three reds (R, R, R) is smaller than the chance of choosing a red, blue and white (R, B, W) -- there is only one way of selecting three reds, but there are three ways of selecting exactly one red. On the other hand, the outcomes in sample space S2 are equally likely since we were careful to distinguish the five balls in the box, and it is reasonable that any three of the five balls has the same chance of being selected.

From this example, we have learned a couple of things. First, when we write down a sample space, we should think carefully about the assumption that outcomes are equally likely. Second, when we have an experiment with duplicate items (like three red balls), it may be preferable to distinguish the items when we write down the sample space and compute probabilities.

PRACTICE: EQUALLY LIKELY OUTCOMES

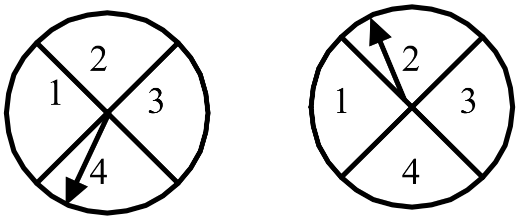
For each of the following experiments, are the outcomes in the given sample space equally likely? If the outcomes are not equally likely, explain why.

1. You will record the weather in your city next Monday. The sample space is S = {sunny, cloudy, rain, snow}.

2. You randomly choose a number from the group of digits {1, 2, 3, 4, 5, 6, 7} and record if the digit is even or odd. The sample space is S = {even, odd}.

3. Three people Ann, Bob, and Jacob are randomly put in a line. Suppose you record Ann's position in the line and the sample space is S={front of line, middle, back of line}.

Consider an experiment using two spinners:



SPINNER 1 SPINNER 2

4. Suppose you spin Spinner 1 and the sample space is S = {1, 2, 3, 4}.

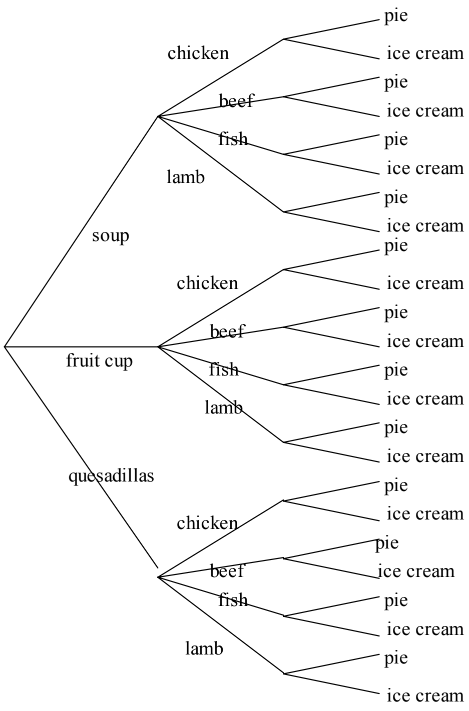
5. Suppose you spin both spinners and record the sum. The sample space is S = {2, 3, 4, 5, 6, 7, 8}.

THE MULTIPLICATION RULE

To apply the equally-likely recipe for computing probabilities, we need some methods for counting the number of outcomes in the sample space and the number of outcomes in the event. Here we illustrate a basic counting rule called the multiplication rule.

Suppose you are dining at your favorite restaurant. Your dinner consists of an appetizer, an entrée, and a dessert. You can either choose soup, fruit cup, or quesadillas for your appetizer, you have the choice of chicken, beef, fish, or lamb for your entrée, and you can have either pie or ice cream for your dessert.

We first use a tree diagram to write down all of your possible dinners. (The first set of branches shows the appetizers, the next set of branches the entrées, and the last set of branches the desserts.)



Note that there are 3 possible appetizers, 4 possible entrées, and 2 possible desserts. For each appetizer, there are 4 possible entrées, and so there are 3 x 4 = 12 possible choices of appetizer and entrée. Using similar reasoning, for each combination of appetizer and entrée, there are 2 possible desserts, and so the total number of complete dinners would be

$$\text{Number of dinners} = 3 \text{ x } 4 \text{ x } 2 = 24.$$

The above dining example illustrates a general counting rule that we call the multiplication rule.

MULTIPLICATION RULE: Suppose you are doing a task that consists of *k* steps. You can do the first step in n_1 ways, the second step in n_2 ways, the third step in n_3 ways, and so on. Then the number of ways of completing the task, which we will denote by *n*, is the product of the different ways of doing the *k* steps, or

$$n = n_1 \times n_2 \times \cdots \times n_k.$$

PRACTICE: THE MULTIPLICATION RULE

1. Suppose you are taking a one-way trip from Toledo to Buffalo to Rochester to New York City. You have two ways of driving from Toledo to Buffalo, three ways of driving from Buffalo to Rochester, and two ways of driving from Rochester to New York City. How many possible routes can you take on your trip?

2. Suppose you flip a penny, flip a quarter, and roll a six-sided die – you observe the side of the penny, the side of the quarter, and the roll on the die. How many possible outcomes are there?

3. Suppose you are ordering a large pizza and you see that the possible toppings are pepperoni, mushroom, extra cheese, peppers, ground beef, and ham. Some possible orders are pepperoni and extra cheese, ham, mushroom, and ground beef, everything (all six toppings), and plain (no toppings).

Suppose the waiter takes this order by asking you the following questions:

Q1. Do you want pepperoni?

Q2. Do you want mushrooms?

Q3. Do you want extra cheese?

Q4. Do you want peppers?

Q5. Do you want ground beef?

Q6. Do you want ham?

(a) How many possible answers are there to question Q1?

(b) How many possible answers are there to questions Q1 and Q2?

(c) How many possible answers are there to all six questions?

(d) How many possible ways can you order your pizza?

PERMUTATIONS

Suppose you load six songs, Song A, Song B, Song C, Song D, Song E, and Song F in your MP3 player. The songs are played in a random order and you listen to the first three songs. How many different selections of three songs can you hear?

In this example, we are assuming that the order that the songs are played is important. So hearing the selections

Song A, Song B, Song C

in that order will be considered different from hearing the selections in the sequence

Song C, Song B, Song A.

We call an outcome such as this a *permutation* or *arrangement* of 3 out of the 6 songs.

We can represent possible permutations by a set of three blanks, where we place songs in the blanks.

________ ________ ________
1st Song 2nd Song 3rd Song

We find the number of permutations as follows:

1. First, we know that 6 possible songs can be played first. We place this number in the first blank above.

____6____ ________ ________
1st Song 2nd Song 3rd Song

2. If we place a particular song, say Song A, in the first slot, there are 5 possible songs in the second position. We put this number in the second blank.

____6____ ___5_____ ________
1st Song 2nd Song 3rd Song

By use of the multiplication rule, there are 6 x 5 = 30 ways of placing two songs in the first two slots.

3. Continuing in the same way, we see that there are 4 ways of putting a song in the 3rd slot and completing the list of three songs.

____6____ ____5____ ___4_____
1st Song 2nd Song 3rd Song

Again using the multiplication rule, we see that the number of possible permutations of six songs in the three positions are

$$6 \text{ x } 5 \text{ x } 4 = 120.$$

We have illustrated a second basic counting rule:

PERMUTATIONS RULE: If we have n objects (all distinguishable), then the number of ways to arrange r of them, called the number of *permutations*, is

$$\#of\ \ permutations{=}_nP_r = n \times (n-1) \times \cdots (n-r+1).$$

In this example, n = 6 and r = 3, and ${}_nP_r = 120$. If three songs are played in your MP3 player, each of the 120 possible permutations will be equally likely to occur. So the probability of any single permutation, say

Song A, Song D, Song B

is equal to 1/120.

Suppose you listen to all six songs on your player. How many possible orders are there? In this case, we are interested in finding the number of ways of arranging the entire set of 6 objects. Here n = 6 and r = 6 and, applying our formula, the number of permutations is

$${}_6P_6 = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720.$$

We use the special symbol n!, pronounced "n factorial", to denote the product of the integers from 1 to n. So the number of ways of arranging n distinct objects is

$${}_nP_n = n! = n \times (n-1) \times (n-2) \times \cdots \times 1.$$

PRACTICE: PERMUTATIONS

Suppose you are interested in constructing a four-letter word from the letters a, b, c, d, e, f, g. Here we use "word" to denote an arrangement of letters – it is very possible that the arrangement is not a real word.

1. In the permutation formula, what are the values of n and r?

2. How many possible four-letter words can you make?

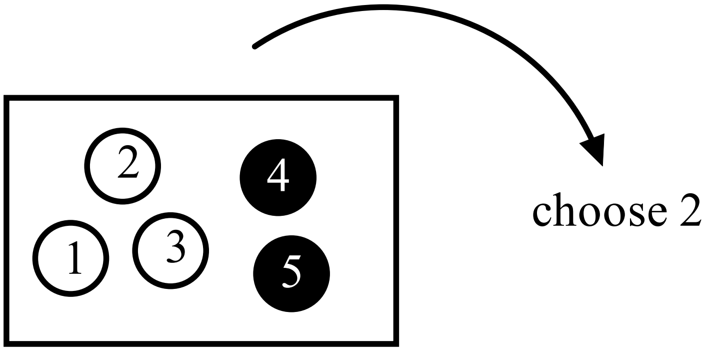
3. What is the probability a randomly arranged word is "bead"?

4. How many possible four-letter words begin with a vowel? (Construct this word in two steps: first choose a vowel for the first letter, and then choose the remaining three letters. Find the number of ways of performing each of the two steps and apply the multiplication rule.)

5. Suppose you wish to form a word using all seven letters. What are the values of n and r? Find the number of possible arrangements.

COMBINATIONS

Suppose you have a box with five balls -- three are white and two are black. You first shake up the box and then you choose two balls out *without replacement*. (This means that once you take a ball out, you do not return it to the box before you take the second ball out.)

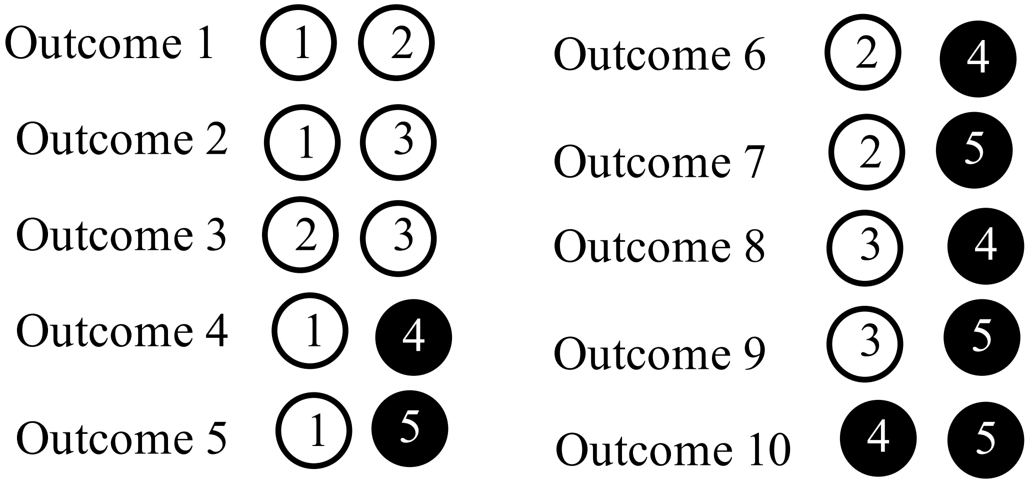


To make it easier to talk about outcomes, we have labeled the five balls from 1 to 5. Remember we are choosing two balls from the box and an outcome would be the numbers of the two balls that we select.

When we list possible outcomes, we should decide if it matters how we order the selection of balls. That is, if we choose ball 1 and then ball 2, is that different than choosing ball 2 and then ball 1?

We could say that order is important -- so choosing ball 1 then ball 2 is a different outcome from ball 2 then ball 1. But in this type of selection problem, it is common practice *not* to consider the order of the selection. Then all that matters is the collection or set of two balls that we select. In this case, we call the resulting outcome a *combination*.

When order doesn't matter, there are 10 possible pairs of balls that we can select. These outcomes or combinations are written below -- this list represents a sample space for this random experiment.



There is a simple formula for counting the number of outcomes in this situation. COMBINATIONS RULE: Suppose we have *n* objects and we wish to take a subset of size *r* from the group of objects without regards to order. Then the number of subsets or combinations, is given by the formula

$$number\ of\ combinations = {}_nC_r = \frac{n!}{r!(n-r)!}.$$

where *k*! stands for *k* factorial $k! = k \times (k-1) \times (k-2) \times \cdots \times 2 \times 1$.

Let's try the formula in our example to see if it agrees with our number. In our setting, we have n = 5 balls and we are selecting a subset of size r = 2 from the box of balls. Using n = 5 and r = 2 in the formula, we get

$$ {}_5C_2 = \frac{5!}{2!(5-2)!} = \frac{5\times 4\times 3\times 2\times 1}{[2\times 1]\times[3\times 2\times 1]} = \frac{120}{12} = 10 $$

that agrees with our earlier answer of 10 outcomes in the sample space.

PRACTICE: COMBINATIONS

PART A: Choosing balls from a box.

We continue with our example of choosing two balls from a box of five where three are white and two are black and there are ${}_5C_2 = 10$ possible combinations.

1. Assuming the outcomes are equally likely, what is the probability of each outcome?

2. Suppose we are interested in the probability that we choose exactly one white ball. To find this probability, we first want to count the number of outcomes that result in one white ball and one black ball.

(a) How many ways are there for choosing exactly one white ball?

(b) How many ways are there for choosing exactly one black ball?

(c) By the multiplication rule, how many ways are there for choosing one white ball *and* one black ball?

(d) Use the work in 1 and part 2 (c) to compute the probability of choosing one white ball.

3. Next consider the probability that we choose two balls of the same color. The first thing to realize is that "balls of the same color" means that either we are choosing two white balls or two black balls.

(a) How many ways can we choose 2 black balls?

(b) How many ways can we choose 2 whites?

(c) How many ways can we choose balls of the same color? (Add the numbers from parts (a) and (b).)

(d) Find the probability of choosing two balls of the same color.

PART B. Ordering a pizza

Let's return to our pizza example that we earlier discussed in this topic. We were interested in ordering a pizza and there were six possible toppings. We use the multiplication rule to compute the total number of possible pizzas that we could order.

How many toppings can there be in our pizza? Since there are six possible toppings, we could either have 0, 1, 2, 3, 4, 5, or 6 toppings on our pizza. Using combination formulas …

1. How many different pizzas can we order that have no toppings?
2. How many different one-topping pizzas can we order?
3. How many different two-topping pizzas can be order?
4. Find the number of possible three topping, four topping, five topping, and six topping pizzas. Record your answers (and the answers to 1, 2, and 3) in the below table.

Number of toppings	Number of ways
None	
One	
Two	
Three	
Four	
Five	
Six	
TOTAL	

5. To compute the total number of different pizzas, add the numbers in the table to get the total that we want. Compare your answer to the total number of pizzas that we computed in the earlier activity.

COMMENT: The above practice exercise demonstrated a general formula. Suppose we have a group of n objects and we are interested in the total number of subsets of this group. Then this total number is

$$2^n = {}_nC_0 + {}_nC_1 + \cdots + {}_nC_n.$$

The formula 2^n is found by noticing there are two possibilities for each object – either the object is in the subset or it is not – and then applying the multiplication rule. The right hand side of the equation is derived by first counting the number of subsets of size 0, of size 1, of size 2, and so on, and then adding all of these subset numbers to get the total number.

ARRANGEMENTS OF NON-DISTINCT OBJECTS

First let's use a simple example to review the two basic counting rules that we have discussed. Suppose you are making up silly words from the letters "a", "b", "c", "d", "e", "f", like

bacedf, decabf, eabcfd

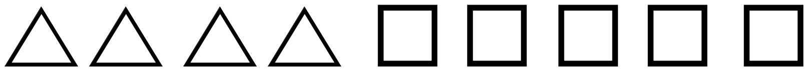
How many silly words can I make up? Here we have n = 6 objects, and so the number of possible permutations is

$$6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720.$$

To illustrate the second counting rule, suppose I have six letters "a", "b", "c", "d", "e", "f", and I am going to choose three of the letters to construct a three-letter word. I can't choose the same letter twice and the order in which I choose the letters is not important. In this case, we are interested in the number of combinations -- applying our combination rule with n = 6 and k = 3, the number of ways of choosing three letters from six is equal to

$${}_6C_3 = \frac{6!}{3!\,3!} = 20.$$

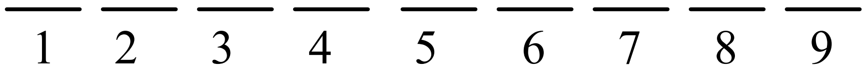
Let's consider a different arrangement problem. Suppose we randomly arrange the four triangles and five squares as shown below.



What is the chance that the first and last locations are occupied by triangles? This is an arrangement problem with one difference -- the objects are not all distinct -- we can't distinguish the four triangles or the five squares. So we can't use our earlier permutation rule that assumes the objects are distinguishable.

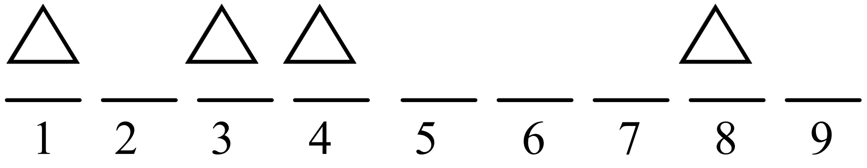
How can we count the number of possible arrangements? It turns out that a combination formula is useful here. (Surprising, but true.)

To think about possible arrangements, suppose we write down a list of nine slots and an arrangement is constructed by placing the triangles and the squares in the nine slots. It is helpful to label the slots with the numbers 1 through 9.

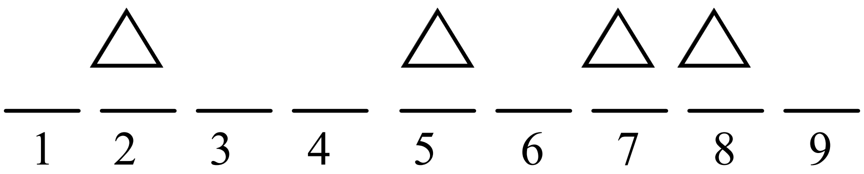


We construct an arrangement in two steps. First, we place the four triangles in four slots, and then we place the squares in the remaining slots.

How many ways can we put the triangles in the slots? First note that we can specify a placement by the numbers of the slots that are used. For example, we could place the triangles in slots 1, 3, 4, and 8.



Or we could place the four triangles in slots 2, 5, 7, and 8.



We can specify an arrangement by choosing four locations from the slot locations {1, 2, 3, 4, 5, 6, 7, 8, 9}. How many ways can we do this? That's easy. We know that the number of ways of selecting four objects (here labels of locations) from a group of nine objects is

$$_9C_4 = \frac{9!}{4!\,(9-4)!} = 126.$$

So there are 126 ways of choosing the four locations for the triangles. Once we have placed the triangles, we can finish the arrangement by putting in the squares. But there is only one way of doing this. For example, if we place triangles in slots 2, 5, 6, 7, then the squares must go in slots 1, 3, 4, 8, 9. So applying the multiplication rule, the number of ways of arranging four triangles and five squares is 126 x 1 = 126.

We have derived a new counting rule:

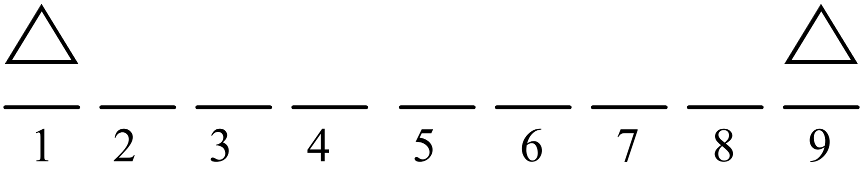
PERMUTATIONS RULE FOR NON-DISTINCT OBJECTS: The number of permutations of n *non-distinct* objects where r are of one type and n - r are of a second type is

$$_nC_r = \frac{n!}{r!\,(n-r)!}$$

Recall our question that we wanted to answer: Suppose we randomly arrange four triangles and five squares. What is the chance that the first and last locations are occupied by triangles?

We have already shown that there are 126 ways of mixing up four triangles and five squares. Each possible arrangement is equally likely and has a chance of 1/126 of occurring.

To find our probability, we need to count the number of ways of arranging the triangles and squares so that the first and last positions are filled with triangles.



If we place triangles in slots 1 and 9 (and there is only one way of doing that), then we are free to arrange the remaining two triangles and five squares in slots 2, 3, 4, 5, 6, 7, 8, 9. By use of our new arrangements formula, the number of ways of doing this is

$${}_7C_2 = 21.$$

and so our probability the first and last slots are filled with triangles is equal to 21/126.

WHICH RULE?

We have described three important counting rules, the permutations rule, the combinations rule, and the permutations rule for non-distinct objects. How can you decide which rule to apply in a given problem? Here are some tips to help you find the right rule. The practice activity that follows will give you some experience in distinguishing between these rules.

1. Do We Care About Order?

If an outcome consists of a collection of objects, does the order that you list the objects matter? If order does matter, then a permutation rule may be appropriate. If the order of the objects doesn't matter, such as choosing a subset from a larger group, then a combination rule is probably more suitable.

2. Are the Objects Distinguishable?

We have two permutation formulas, one that applies when all of the objects are distinguishable, and the second where there are two types of objects and you can't distinguish between the objects of each type.

3. When In Doubt …

If the first two tips don't seem helpful, it may benefit to start writing down a few outcomes in the sample space. When you look at different outcomes, you should recognize if order is important and if the objects are distinguishable.

PRACTICE: THREE COUNTING RULES

For each of the following problems, state the appropriate counting rule (permutations for distinct objects, combinations, permutations for nondistinct objects) and use the rule to answer the problem.

1. Suppose you line up five coins, a penny, a nickel, a dime, a quarter, and a half-dollar, in a row. How many possible lineups can you have?

2. Suppose you line up 3 pennies and 2 nickels. How many possible lineups can you have?

3. Suppose you have five coins, a penny, a nickel, a dime, a quarter, and a half-dollar, in your pocket and you select three coins out. How many possible groups of coins can you have?

4. Suppose you flip a coin 10 times and record the sequence of heads and tails. If you flip exactly six heads, how many possible sequences are there? (One possible sequence is HHHTTTTHHH.)

PLAYING YAHTZEE

Yahtzee is a popular game played with five dice. The game is similar to the card game Poker – in both games, one is trying to achieve desirable patterns in the dice faces or cards, and some types of patterns are similar in the two games. In this section, we describe some of the dice patterns in the first roll in Yahtzee and consider the problem of determining the chances of several of the patterns.

Outcomes of one roll of five dice

When a player rolls five dice in the game Yahtzee, the most valuable result is when all of the five dice show the same number such as

2, 2, 2, 2, 2.

This is called a "Yahtzee" and the player scores 50 points with this pattern. A second valuable pattern is a "four-of-a-kind" where you observe one number appearing four times, such as

3, 4, 3, 3, 3.

The following table gives all of the possible patterns when you roll five dice in Yahtzee. When you play the game, some of these patterns are worth a particular number of points and these points are given in the right column.

Pattern	Sample of pattern	Point value
Yahtzee	4, 4, 4, 4, 4	50
Four of a kind	6, 6, 6, 4, 6	
Large straight	2, 6, 4, 5, 3	40
Small straight	4, 2, 1, 3, 2	30
Full house	5, 1, 1, 5, 1	25
Three of a kind	2, 2, 3, 4, 2	
Two pair	6, 3, 3, 6, 2	
One pair	4, 3, 4, 1, 5	
Nothing	1, 3, 2, 5, 6	
TOTAL		

Total number of outcomes

As in the case of two dice, it is useful to distinguish the five dice when we count outcomes. We can represent an outcome by placing a value of individual die rolls (1 through 6) in the six slots.

—	—	—	—	—
die 1	die 2	die 3	die 4	die 5

So two possible outcomes are

2, 3, 4, 5, 5 and 3, 2, 4, 5, 5

Each die has 6 possibilities and so, applying the multiplication rule, the total number of outcomes in the rolls of five dice is

$$6 \times 6 \times 6 \times 6 \times 6 = 7776.$$

Since all of the outcomes are equally likely, we assign a probability of 1/7776 to each outcome.

Probability of a Yahtzee

We can represent the Yahtzee roll as the outcome

x, x, x, x, x

where x denotes an arbitrary roll of one die. There are six possible choices for x, and so the number of possible Yahtzees is 6.

Since each outcome has probability 1/7776, the probability of a Yahtzee is

$$\text{Prob(Yahtzee)} = 6/7776.$$

Probability of four of a kind

In the pattern "four of a kind", we want to have one number appear four times and a second number appear once. In other words, we are interested in counting outcomes of the form

x, x, x, x, y

where the four x's and the single y can be in different orders.

To apply the multiplication rule, we think of writing down a possible "four of a kind" in three steps.

Step 1: We first choose the number for x (the number that appears four times).

Step 2: We next choose the number for the singleton y.

Step 3: We mix up the orders of the four x's and the one y.

For example, we choose the outcome 5, 5, 5, 3, 5 by (1) choosing 5 to be the number that appears four times, (2) choosing 3 as the number that appears once, and then arranging the digits 5, 5, 5, 5, 3 to get 5, 5, 5, 3, 5.

We next count the number of ways of doing each of the three steps.

Step 1: There are 6 ways of choosing x.

Step 2: Once x has been chosen, there are 5 ways of choosing the value for y.

Step 3: Last, once x and y have been selected, there are ${}_5C_4 = 5$ ways of mixing up the x's and y's.

To find the number of four-of-a-kinds, we use the multiplication rule using the number of ways of doing each of the three steps:

Number of ways = 6 × 5 × 5 = 150.

The corresponding probability of four-of-a-kind is

Prob(four-of-a-kind) = 150/7776.

PRACTICE: YAHTZEE

1. (Probability of a two pair.) We follow the same basic method as described above in counting the number of two-pairs. Represent an outcome by the sequence

x, x, y, y, z

where x and y denote the numbers that will each occur twice, and z is the number that appears one.

(a) How many ways are there for choosing the numbers x and y? (Note that since both x and y each appear twice, it is incorrect to say that the number of ways of choosing x and y is 6 × 5 = 30.)

(b) How many ways are there for choosing z, the identity of the number that appears once?

(c) The number of ways of mixing up the orders of x, x, y, y, z is

$$\frac{5!}{2!2!1!} = 30.$$

(This is a generalization of the earlier counting rule of arrangements of two different types.) Combine this result with the results of parts (a) and (b) to find the number of two-pairs.

(d) Find the probability of a two-pair.

2. (Probability of a large straight.) A large straight is observing five consecutive numbers in our roll. We write down a large straight in two steps:

Step 1: We write down the possible five numbers in the large straight.
Step 2: We arrange the numbers.

(a) How many ways can you do step 1? In other words, how many possible choices are there for the five numbers that make up the straight?

(b) How many ways can you arrange the five numbers in the straight?

(c) Combining the two steps, how many ways can you have a large straight?

(d) What is the probability of a large straight?

ACTIVITY – MOTHERS AND BABIES

DESCRIPTION: One day four babies were born at the local hospital. But for some reason, the nurses forgot to put identification bands on the babies, and decided (believe it or not) to give the babies back to the four mothers in some random fashion and hope for the best. How many babies will be correctly matched with the mothers? We will answer this question two ways, first using a simulation with cards, and then by listing all possible outcomes and using the classical notion of probability.

MATERIALS NEEDED: Sets of playing cards where one set contains eight cards: four red cards of different values, say seven, eight, Jack, and Queen, and four black cards of the same values.

METHOD 1. (Simulation) We will simulate this experiment using four red cards of different types (the moms) and four black cards of the same types.

- Put the four red cards down in a row.
- Mix up the four black cards and place them below the red cards.
- Count the number of matches.

Repeat this experiment 20 times – record your answers in the first table, and summarize your values in the second table.

TRIAL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
# of matches																				

Number of matches	0	1	2	3	4
COUNT					
PROBABILITY					

METHOD 2. (Thinking of all possible outcomes) Suppose that the names of the four babies are Abby Albert, Bobby Brown, Cindy Crawford, and Darren Daulton. In your lab book, write down all ways of arranging or permuting the four first names (ABCD and ABDC are two possible arrangements). For each arrangement, count the number of matches. For example if babies ABDC are assigned to mom's ABCD, the number of matches is 2. Find the probabilities of the number of matches and put your answers in a table.

Number of matches	0	1	2	3	4
PROBABILITY					

ACTIVITY: SAMPLING FROM A BAG

DESCRIPTION: Pick up a lunch bag and 5 blocks – 2 blocks have one color and 3 blocks have a different color. Let's assume the colors are black and white (your colors may vary). Think of black as the darker color of the two colors you have. You put the 5 blocks in the bag, mix them up, and choose two out (without replacement) – how many blocks will be black? We will address this question first by doing a simulation, and then by enumerating all of the outcomes of the experiment.

MATERIALS NEEDED: A number of lunch bags, where each bag contains 5 blocks (these could be balls or dice), where 2 blocks have one color and 3 blocks have a different color.

METHOD 1 (Simulation)

Simulate this process 20 times.

- Put all the blocks in the bag and mix them up.
- Select two out without replacement.
- Record the number of blacks you see in your sample.

Put your answers in the table below.

TRIAL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
# of black selected																				

Find the probability of choosing 0 black, 1 black, etc – put your answers in the table.

Number of blacks	0	1	2
PROBABILITY			

METHOD 2 (Listing outcomes)

Distinguish the blocks in the bag; if you have 2 black and 3 white blocks write them as {B1, B2, W1, W2, W3}. Suppose you keep track of the order of the blocks you select (so choosing B1, W1 is different from choosing W1, B1).

Write down all possible outcomes (selections of two blocks).

Find the probabilities of choosing 0 black, 1 black etc – put your answers in the table.

Number of blacks	0	1	2
PROBABILITY			

METHOD 3 (Using counting arguments)

1. If the five blocks are distinguishable, you select two blocks, and the order in which you select the blocks is important, how many possible outcomes are there?

2. (Continuation of 1.) Count the number of outcomes where you choose a black first and a white second.

3. Count the number of outcomes where you choose a white first and a black second.

4. Count the number of outcomes where you choose exactly one black ball. (Combine answers from parts 1 and 2.)

5. Find the probability of choosing exactly one black.

WRAP-UP

In this topic, we introduced several counting rules helpful for computing probabilities in the case where the outcomes in the sample space are equally likely. The basic rule is the *multiplication rule* that counts the number of outcomes when the experiment consists of several stages. A *permutations* rule is appropriate when one is arranging a set of distinct objects. In contrast, a *combinations* rule is used when one is selecting a subset of a larger group without replacement. We discussed a third permutations rule useful for counting the number of arrangements of nondistinct objects. To decide on the appropriate counting rule, one should carefully consider possible outcomes of the random experiment. One should ask if the order of selecting items is important and if the items selected are all distinguishable.

EXERCISES

1. **Constructing a Word**

Suppose you select three letters at random from {a, b, c, d, e, f} to form a word.

a. How many possible words are there?

b. What is the probability the word you choose is “fad”?

c. What is the probability the word you choose contains the letter “a”?

d. What is the chance that the first letter in the word is “a”?

e. What is the probability that the word contains the letters “d”, “e”, and “f”?

2. **Running a Race**

There are seven runners in a race – three runners are from Team A and four runners are from Team B.

a. Suppose you record which runners finish first, second, and third. Count the number of possible outcomes of this race.

b. If the runners all have the same ability, then each of the outcomes in (a) are equally likely. Find the probability that Team A runners finish first, second, and third.

c. Find the probability that the first runner across the finish line is from Team A.

3. **Rolling Dice**

Suppose you roll three fair dice.

a. How many possible outcomes are there?

b. Find the probability you roll three sixes.

c. Find the probability that all three dice show the same number.

d. Find the probability that the sum of the dice is equal to 10.

4**. Ordering Hash Browns**

When you order Waffle House’s world famous hash browns, you can order them scattered (on the grill), smothered (with onions), chunked (with ham), topped (with chili), diced (with tomatoes), and peppered (with peppers). How many ways can you order hash browns at Waffle House?

5. **Selecting Balls from a Box**

A box contains 5 balls -- 2 are white, 2 are black, and one is green. You choose two balls out of the box at random without replacement.

a. Write down all possible outcomes of this experiment. (Assume that the order in which you select the balls is important.)

b. Find the probability that you choose two white balls.

c. Find the probability you choose two balls of the same color.

d. Find the probability you choose a white ball second.

6. **Dividing into Teams**

Suppose that ten boys are randomly divided into two teams of equal size. Find the probability that the three tallest boys are on the same team.

7. **Choosing Numbers**

Suppose you choose three numbers from the set {1, 2, 3, 4, 5, 6, 7, 8} (without replacement).

a. How many possible choices can you make?

b. What is the probability you choose exactly two even numbers?

c. What is the probability the three numbers add up to 10?

8. **Choosing People**

Suppose you choose two people from three married couples.

a. How many selections can you make?

b. What is the probability the two people you choose are married to each other?

c. What is the probability that the two people are of the same gender?

9. **Football Plays**

Suppose a football team has five basic plays, and they will randomly choose a play on each down.

a. On three downs, find the probability that the team runs the same play on each down.

b. Find the probability the team runs three different plays on the three downs.

10. **Playing the Lottery**

In a lottery game, you make a random guess at the winning three-digit number (each digit can be 0, 1, 2, 3, 4, 5, 6, 7, 8, 9). You win $200 if your guess matches the winning number, $20 if your guess matches in exactly two positions and $2 if your guess matches in exactly one position. Find the probabilities of winning $200, winning $20, and winning $2.

11. **Dining at a Restaurant**

Suppose you are dining at a Chinese restaurant with the menu given below. You decide to order a combination meal where you get to order one soup or appetizer, one entrée (seafood, beef, or poultry), and a side dish (either fried rice or noodles).

a. How many possible combination meals can you order?

b. If you are able to go to this restaurant every day, approximately how many years could you order different combination meals?

c. Suppose that you are allergic to seafood (this includes crab, shrimp, and scallops). How many different combination meals can you order?

d. Suppose your friend orders two different entrées completely at random. How many possible dinners can she order? What is the probability the two entrées chosen contain the same meat?

SOUP
HOT AND SOUR SOUP
WONTON SOUP
EGG DROP SOUP
APPETIZERS
EGG ROLL
BARBECUED SPARERIBS
FRIED CHICKEN STRIPS
BUTTERFLY SHRIMP
CRAB RANGOON
SEAFOOD
SHRIMP WITH GARLIC SAUCE
CURRY SHRIMP
KUNG PAO SCALLOPS
FLOWER SHRIMP
SHRIMP WITH PEA PODS
BEEF
KUNG PAO BEEF
HUNAN BEEF
SZECHUAN STYLE BEEF
ORANGE BEEF (HOT & SPICY)

POULTRY
KUNG PAO CHICKEN
HUNAN CHICKEN
CHICKEN WITH DOUBLE NUTS
CHICKEN WITH GARLIC SAUCE
CURRY CHICKEN
FRIED RICE
CHICKEN FRIED RICE
BEEF FRIED RICE
SHRIMP FRIED RICE
PORK FRIED RICE
THREE DELIGHT FRIED RICE
VEGETABLE FRIED RICE
NOODLES/RICE
PAN FRIED NOODLES
MOO SHU PANCAKE
CHOW MEIN NOODLES
STEAMED RICE .

12. **Ordering Pizza**

If you buy a pizza from Papa John’s, you can you order the following toppings: ham, bacon, pepperoni, Italian sausage, sausage, beef, anchovies, extra cheese, baby portabella mushrooms, onions, black olives, Roma tomatoes, green peppers, jalapeno peppers, banana peppers, pineapple, grilled chicken.

a. If you have the option of choosing two toppings, how many different two topping pizzas can you order?

b. Suppose you want your two toppings to be some meat and some peppers. How many two-topping pizzas are of this type?

c. If you order a "random" two-topping pizza, what is the chance that it will have peppers?

d. If you are able to order at most four toppings, how many different pizzas can you order?

13. **Mixed Letters**

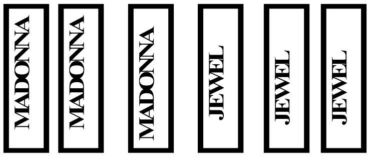
You randomly mix up the letters "s", "t", "a", "t", "s".

a. Find the probability the arrangement spells the word "stats".

b. Find the probability the arrangement starts and ends with "s".

14. **Arranging CDs**

Suppose you have three Madonna cds and three Jewel cds sitting on a shelf as follows. (We assume that you can't distinguish the cds of a given artist.)



The cds are knocked off of the shelf and you place them back on the shelf completely at random.

a. What is the probability that the mixed-up cds remain in the same order?

b. What is the probability that the first and last cds on the shelf are both Jewel music?

c. Which artist do you prefer, Madonna or Jewel?

d. What is the probability that the Jewel cds stay together on the shelf?

15. **Playing a Lottery Game**

The Minnesota State Lottery has a game called Daily 3. A three digit number is chosen randomly from the set {000, 001, …, 999} and you win by guessing correctly certain characteristics of this three digit number. The lottery website lists the following possible plays such as First Digit, Front Pair, etc. Find the probability of winning for each play.

First Digit *Pick one number. To win, match the first number drawn.*

Front Pair *Pick 2 numbers. To win, match the first 2 numbers drawn in exact order*

Straight *Pick 3 numbers. To win, match all 3 numbers drawn in exact order.*

3-Way Box *Pick 3 numbers, 2 that are the same. To win, match all three numbers drawn in any order.*

6-Way Box *Pick 3 different numbers. To win, match all 3 numbers drawn in any order.*

16. **Booking a Flight**

Suppose you are booking a flight to San Francisco on Orbitz. To save money, you agree to either leave Monday, Tuesday, or Wednesday, and return on either Friday, Saturday, or Sunday. Assume that Orbitz randomly assigns you a day to leave and randomly assigns you a day to return.

a. What is the probability you leave on Tuesday and return on Saturday?

b. What is the chance that your trip will be exactly three days long?

c. What is the most likely trip length in days?

d. Do you think that the assumptions about Orbitz are reasonable? Explain.

17. **Assigning Grades**

A math class of ten students takes an exam.

a. If the instructor decides to give exam grades of A to two randomly selected students, how many ways can this be done?

b. Of the remaining eight students, three will receive B's and the remaining will receive C's. How many ways can this be done?

c. If the instructor assigns at random, two A's, three B's and five C's to the ten students, how many ways can this be done?

d. Under this grading method, what is the probability that Jim (the best student in the class) gets an A?

18. **Choosing Officers**

A club consisting of 8 members has to choose three officers.

a. How many ways can this be done?

b. Suppose that the club needs to choose a president, a vice-president, and a treasurer. How many ways can this be done?

c. If the club consists of 4 men and 4 women and the officers are chosen at random, find the probability the three officers are all of the same gender.

d. Find the probability the president and the vice-president are different genders.

19. **Playing Yahtzee**

Find the number of ways and the corresponding probabilities of getting all of the following patterns in Yahtzee. Here are some hints for the different patterns.

Four of a kind: The pattern here is {x, x, x, x, y}, where x is the number that appears four times and y is the number that appears once.

Small straight: This roll will either include the numbers 1, 2, 3, 4, the numbers 2, 3, 4, 5, or the numbers 3, 4, 5, 6. If the numbers 1, 2, 3, 4 are the small straight, then the remaining number can not be 5 (otherwise it would be a large straight).

Full house: The pattern here is {x, x, x, y, y}, where x is the number that appears three times and y is the number that appears twice.

Three of a kind: The pattern here is {x, x, x, y, z}, where x is the number that appears three times, and y and z are the numbers that appear only once.

One pair: The pattern here is {x, x, w, y z}, where x is the number that appears two times, and w, y and z are the numbers that appear only once.

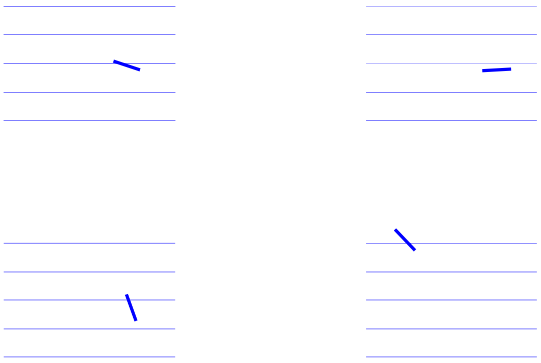
Nothing: This is the most difficult number to count directly. Once the number of each of the remaining patterns is found, then the number of "nothings" can be found by subtracting the total number of other patterns from the total number of rolls (7776).

TOPIC P4: COMPUTING PROBABILITIES BY SIMULATION

SPOTLIGHT: BUFFON’S NEEDLE SIMULATION

Buffon’s Needle is an “old” problem in probability, dating back to the 18th century. A simple version of the problem can be described as follows. Suppose you have a lined sheet with lines spaced one inch apart. You drop a needle of length one inch on the sheet. What is the probability the needle crosses one of the lines on the page?

We can approximation this probability by a simulation experiment. Take a lined sheet and repeatedly drop a needle on the sheet, keeping track of the number of times that the needle crosses a line. The probability of interest is approximately the fraction of “successful” drops. Below we illustrate four drops of the needle – here since three needles cross the line, our estimate of the probability is ¾ = .75. This is a relatively crude estimate of the probability since it is based on four trials and we can obtain a more accurate estimate by repeating the experiment for many trials.



PREVIEW

Suppose that we are interested in computing a probability of some outcome in a random experiment. Suppose that we can repeat the experiment (either by hand or by the use of a computer or a calculator) under the same conditions. Then, using the frequency notion of probability, we can approximate a probability by the fraction of time the event occurs in the many experiments. In this topic, we illustrate the use of these simulation experiments to obtain probabilities for interesting problems.

In this topic, your learning objectives are to:

- Understand the basic components of a simulation experiment.
- Design simulation experiments for simple random experiments.
- Implement a simulation experiment using a software program or a calculator.
- Use simulation output to find probabilities of interest.

NCTM Standards

✓ In Grades 6-8, all students should use proportionality and a basic understanding of probability to make and test conjectures about the results of experiments and simulations.

✓ In Grades 9-12, all students should understand how to use simulations to construct empirical probability distributions.

SIMULATING A LOTTERY GAME

Recall the lottery game we discussed earlier. We choose the three-digit number "123" with the hope that it will be a winner. A winning number will be drawn at random from the list of all possible three-digit numbers. We will win

- $100 if the winning number matches "123"
- $10 if the winning number matches "123" in two positions
- $2 if the winning number matches "123" in exactly one position

We are interested in the probabilities that the winning number will match "123" in one position, in two positions, and in three positions.

We can approximate the probabilities of these three events by use of a simulation experiment. Essentially this experiment consists of two steps:

- [SIMULATE] we simulate the process of choosing a three-digit random number

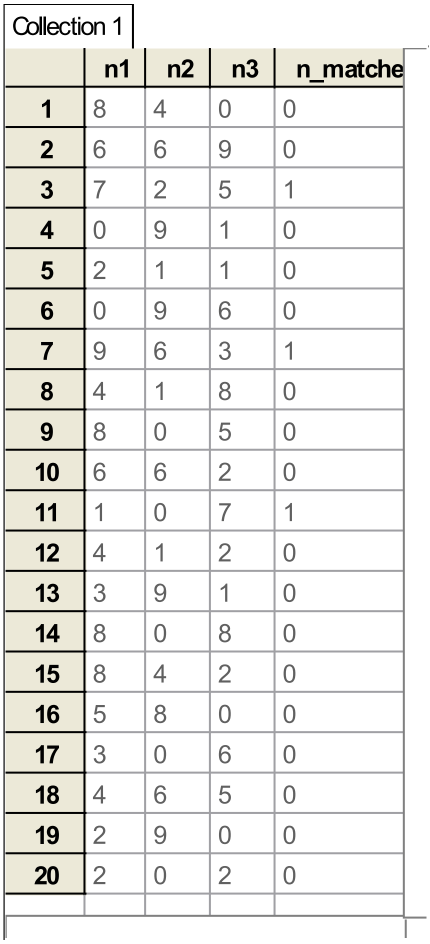
- [RECORD] we record the number of matches of the simulated winning number with our number “123”

By repeating the “simulate” and “record” steps many times and summarizing the simulated data, we can approximate the probabilities of interest.

There are two general ways of performing the simulation step. We can use some physical device for randomly choosing numbers, such as a die or spinner or a deck of cards. This “hands-on” style of simulation is attractive in teaching, since the underlying process behind the simulation experiment is clear and easy to understand. We typically use physical simulations when we can, but there is a limit to the number of trials of the simulation experiment. The computer or calculator is an alternative method of simulation. The use of technology is attractive in that one can perform a simulation experiment many times and get accurate approximations at probabilities of interest. But the process of using the computer or calculator can appear mysterious to students, and an incorrectly programmed simulation on a computer can lead to erroneous results.

Let’s illustrate physical and computer simulations for our lottery example. A ten-sided die is a convenient way of physically simulating random digits. This special die has the digits 0, 1, ..., 9 on its ten sides and each side has the same chance of showing in a roll. You can simulate a winning three-digit number by rolling the die three times -- if the results of the three rolls are 3, 7, 0, then the winning number is "370". An alternative way of simulating winning numbers is by use of a computer program such as *Fathom*. It is easy to have the computer simulate a three-digit number randomly from the possible numbers 000, 001, ..., 999. Although it is not obvious, the computer is actually performing the same simulation experiment (with the same long-term behavior) as the rolls of the ten-sided die.

Using the program *Fathom*, I simulated 20 winning lottery numbers. In each case, I had the program record the number of position matches of the number with "123". The results of the twenty simulations are shown in the table below. The variables n1, n2, and n3 are the three random digits and the variable n_matches records the number of matches of the winning number with our number “123”. Note that the first simulated winning number was "840" which matched "123" in zero positions, and the third winning number was "725" which matched "123" in one position.



Collection 1

	n1	n2	n3	n_matche
1	8	4	0	0
2	6	6	9	0
3	7	2	5	1
4	0	9	1	0
5	2	1	1	0
6	0	9	6	0
7	9	6	3	1
8	4	1	8	0
9	8	0	5	0
10	6	6	2	0
11	1	0	7	1
12	4	1	2	0
13	3	9	1	0
14	8	0	8	0
15	8	4	2	0
16	5	8	0	0
17	3	0	6	0
18	4	6	5	0
19	2	9	0	0
20	2	0	2	0

Suppose I wish to use these twenty simulations to estimate the probability that there will be exactly one match. Looking at the output of 20 simulations, note that we had one match exactly two times. So the probability P(exactly one match) can be approximated by the relative frequency 2 / 20.

We can get a more accurate approximation to the probability by repeating the simulation more times. I had Fathom simulate the process of choosing a winning number 1000 times. In each simulation, I record the number of matches and the table below tabulates the outcomes.

Number of matches	Count	simulation probability
0	718	718/1000 = .002

1	252	252/1000 = .028
2	28	28/1000 = .252
3	2	2/1000 = .718
Total number of simulations	1000	

Note that, out of 1000 simulations, there was one match 252 times, so the probability of one match is approximately 252/1000 = .252. Similarly, the table gives the probabilities of 0, 2, and 3 matching positions.

Now these are not the exact probabilities of 0, 1, 2, 3 matches -- these are approximate ones computed by simulation. If we were able to repeat the simulation an infinite number of times, the simulation probabilities would approach the actual probabilities.

PRACTICE: SIMULATING THE LOTTERY EXAMPLE BY USING TABLES OF RANDOM DIGITS.

If a computer is not available, then one can simulate the lottery game by use of a table of random digits. Suppose that your winning number is “565”. Look at the row of the random digit table corresponding to the day of your birth. (For example, if you were born on November 16, then look at row 16 of the table.) Starting with that row, record the first three digits of 20 rows. In the below table, record the random digits and the number of matching position with “565”.

Random digits	Number of matches

Random digits	Number of matches

Using these results, find the probabilities that the random three-digit number has exactly 1, 2 and 3 matching positions with "565" – put your answers in the below table.

Number of matches	simulation probability
1	
2	
3	

BASIC COMPONENTS OF A SIMULATION EXPERIMENT

Simulation is very useful in observing the inherent variation of outcomes in random experiments and in computing probabilities of interest. But the construction of a simulation experiment, either a hands-on or a computer variety, requires some care. Here we describe the basic components of a simulation experiment. A basic knowledge of these components is necessary in the construction of one's own simulations.

Suppose we wish to simulate a random experiment where we wish to approximate a probability of some outcome of interest. Any simulation should consist of the following elements.

1. HOW ARE WE GOING TO SIMULATE? First, one needs to decide on the randomization device (coin, die, spinner, computer, etc.) that will be used in the simulation. The choice of device often depends on the problem. Simple probability problems may be easily simulated using coins or dice, while others may require a more complicated random mechanism, such as that available on a computer.

2. DETAILS OF ONE SIMULATION TRIAL? Once we have selected a randomization device, then one should describe the details of "one trial" of the simulation experiment. This means we have to describe precisely what we plan to do (flip a coin, roll a die, roll many dice, etc.) and describe how the outcomes of the randomization device correspond to the outcomes in our probability problem.

3. WHAT IS OUR MEASUREMENT? In a simulation experiment, one typically focuses on a few measurements of interest, and these should be described. By collecting these measurements over many trials of the experiment, we will be able to approximate the probability of interest.

4. HOW MANY TRIALS? Finally, we need to decide on the number of trials of the simulation experiment. Since the objective is to compute a probability, one needs to simulate the experiment a sufficient number of times so that one obtains a reasonable estimate at the probability. We purposely won't give a formula for how many trials are needed, but one will get an intuitive sense of "how many trials" through experience with a number of simulation experiments.

THE COLLECTOR'S PROBLEM

We illustrate the components of a simulation experiment by considering a variation of a famous probability problem – the collector's problem.

In a current promotion, a cereal manufacturer packages a poster of a famous women athlete in each box of its most popular cereal for kids. There are six possible posters – Serena Williams (tennis), Michelle Kwan (skating), Mia Hamm (soccer), Marion Jones (track), Annika Sorenstam (golf), and Sheryl Swoopes (basketball) (the posters are shown below). Suppose that each of the six posters is equally likely to be in each cereal box. If you purchase 10 boxes of this brand of cereal, what is the chance that you'll get a complete set of posters?

1. HOW ARE WE GOING TO SIMULATE? First we have to decide on a randomization device. Since there are six possible posters, each equally likely to be found in a box, a convenient device here is the usual six-sided die.

2. DETAILS OF ONE SIMULATION TRIAL? We roll a die – what does it mean? A die roll represents a person unpackaging the poster from a cereal box and noting the athlete on the poster. We will let

- a roll of 1 correspond to a poster of Williams
- a roll of 2 correspond to a poster of Kwan
- a roll of 3 correspond to a poster of Hamm
- a roll of 4 correspond to a poster of Jones
- a roll of 5 correspond to a poster of Sorenstam
- a roll of 6 correspond to a poster of Swoopes

Then by rolling our die 10 times, we simulate the results of observing the women on the 10 posters. So, for example, if we observe the die rolls 4, 2, 4, 3, 6, 5, 1, 4, 6, 1, we have purchased posters with the athletes Jones, Kwan, Jones, Hamm, Swoopes, Sorenstam, Williams, Jones, Swoopes, and Williams.

3. WHAT IS OUR MEASUREMENT? We are interested in the probability of getting a complete set when we purchase ten posters. So, for example, if the result of our 10 die rolls is 4, 2, 4, 3, 6, 5, 1, 4, 6, 1, we are interested if we got a complete set. So our measurement in this case is either YES or NO; for our particular simulation, our result is YES since all six die rolls (or all six posters) are represented.

4. HOW MANY TRIALS? Last, we have to repeat this experiment a "large" number of times. Since we are doing this by hand, there are limitations how many experiments we can run. Suppose we decide to repeat this experiment a total of 50 times, each time recording whether we got a complete set.

Now that the experiment is clearly defined, we can carry it out. A convenient way of doing one trial of our experiment is to record the die outcomes by placing tallies in the boxes below – if all of the boxes are filled after 10 rolls, we have a complete set.

Die roll	1	2	3	4	5	6
Tally						

In our example of ten rolls, 4, 2, 4, 3, 6, 5, 1, 4, 6, 1, we record

Die roll	1	2	3	4	5	6
Tally	||	|	|	|||	|	||

and note that we have a complete set.

Here are the results of our 50 simulations:

Exp.	Comp. Set?	Exp.	Comp. Set?	Exp.	Comp. Set?	Exp.	Comp. Set?	Exp.	Comp. Set?
1	NO	11	YES	21	NO	31	NO	41	NO
2	NO	12	NO	22	YES	32	YES	42	NO
3	YES	13	NO	23	NO	33	NO	43	NO
4	YES	14	NO	24	YES	34	NO	44	NO
5	YES	15	NO	25	YES	35	NO	45	NO
6	YES	16	NO	26	NO	36	NO	46	NO
7	NO	17	NO	27	YES	37	YES	47	NO
8	NO	18	NO	28	YES	38	YES	48	NO
9	YES	19	NO	29	NO	39	NO	49	NO

10	YES	20	YES	30	YES	40	YES	50	YES

We summarize the results using a count table.

Complete set?	Count	Proportion
YES	19	19/50
NO	31	31/50

The probability of getting a complete set with a purchase of 10 boxes is then approximated by

$$\text{Prob(complete set)} = 19/50 = 0.38.$$

This computation suggests that one needs to buy more than 10 boxes to be pretty sure of getting a complete set. In one of the exercises, we will use a simulation experiment to see how many boxes, on average, you will need to purchase to get the poster set.

PRACTICE: SETTING UP SIMULATION EXPERIMENTS

Suppose you have a regular six-sided die, a 10-sided die, and coins in your classroom. For each of the following problems, describe a simulation experiment using one of these randomization devices to compute the probability of interest. Each description should discuss the four components of an experiment described in this topic.

1. [Variation of the collector's problem.] In a current promotion, suppose that a poster of a famous actor is placed in every box of macaroni and cheese. There are ten possible posters, and it is equally likely that each poster is contained in a given box. Suppose that you plan on buying boxes of macaroni and cheese until you have a complete set. How many boxes, on average, do you need to purchase? What is the probability that you need to purchase at least 20 boxes?

(a) How are you going to simulate?

(b) Give the details of one simulation trial.

(c) What is your measurement?

(d) How many trials will you perform?

2. [A random walk by an ant.] Suppose an ant's current location is at "5" on the number line. She will take a step to the right with probability 2/3 and a step to the left with probability 1/3. What is the probability that she will be at location 8 or greater after four steps?

```
+---+---+---+---+---A---+---+---+---+---+
0   1   2   3   4   5   6   7   8   9  10
```

(a) How are you going to simulate?

(b) Give the details of one simulation trial.

(c) What is your measurement?

(d) How many trials will you perform?

3. [Shooting free throws.] Suppose a basketball player makes a free-throw shot with probability 0.7. If he takes ten free-throws during a game, what's the chance that he will make at least 7 shots in the game?

(a) How are you going to simulate?

(b) Give the details of one simulation trial.

(c) What is your measurement?

(d) How many trials will you perform?

ACTIVITY: MIXED UP LETTERS

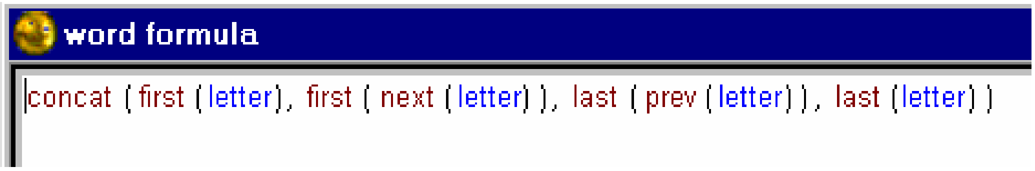
Suppose you randomly mix up the letters “s”, “t”, “a”, and “t”. What is the chance that the mixed-up combination of letters spells “stat”? What is the probability the mixed-up combination of letters spells a word? Each part of this activity describes a different method for computing these probabilities.

Part I: Hands-on simulation

Place the four letters on four blank cards. You can simulate this experiment by (1) mixing up the four cards, (2) placing the cards down in a row, and (3) recording the arrangement that is spelled. Perform this experiment 20 times. Approximate the probability that the arrangement is “stat” and the arrangement spelled is actually a word.

Part II: Simulation using *Fathom*

1. Define a New Collection and a New Data Table. Define an Attribute called "letter" and put the letters s, t, a, t in the collection.
2. Take a Sample of size 4 without replacement from the Collection. This sample represents an arrangement of the four letters.
3. Define a new Measure from your Sample:
   - define a new measure “word”
   - in the formula box, type the following formula



```
concat ( first ( letter ), first ( next ( letter ) ), last ( prev ( letter ) ), last ( letter ) )
```

(All this does is to put the first, second, third, and fourth letters together to form a word. So if your arrangement is "t", "s", "t", "a", the word is "tsta".)

4. Select the Sample collection and select Collect Measures from the Analyze menu. Inspect the Measures from Sample and change it so you collect 1000 measures (words).

5. Run the simulation (select Measures from Sample and hit Control-Y). Find a frequency table of all possible arrangements that can be formed.

Arrangement	Count	Probability	Arrangement	Count	Probability

6. Using your output

(a) What is the probability that the random word is "stat"?

(b) What is the probability the arrangement is actually a word?

PART III: Finding the exact probability

1. Find the number of possible arrangements of the four letters “s”, “t”, “a”, and “t”.

2. Assuming that the possible arrangements are equally likely, find the exact probability that the random arrangement is "stat".

3. Find the exact probability that the random arrangement is actually a word.

ACTIVITY: THE LONGEST RUN

Suppose you have six balls, three are red and three are blue, and you randomly arrange them in a row. Consider the “longest run” which is the length of the longest streak of blues or reds. If your arrangement of balls is

Red, Red, Blue, Blue, Red, Blue,

then the longest run will be equal to 2. What are possible values for the longest run, and what is the probability of each possible value?

PART I: Doing an exact calculation

1. How many possible arrangements are there? (The three reds are identical and also the three blues. This is an example of a counting rule for arrangements when the objects are not all distinguishable.)

2. What is the probability the arrangement will be Red, Red, Blue, Blue, Red, Blue ?

3. Assuming that all of the possible arrangements are equally likely, find the probability that the longest run is equal to 1, 2, and 3. (These are the possible values.) Express your answers as fractions.

Longest run	1	2	3
Probability			

PART II: Simulating this problem on the *TI-84 Plus* Calculator

On a calculator, we first define a list of six elements, where the red balls are represented by three 1's, and the blue balls are represented by three 2's. To simulate this experiment, we

- randomly arrange the elements of the list
- compute the longest run in the random arrangement

There is no simple function on the calculator to find the length of the longest run. But it is possible to compute the number of switches in the sequence which is the number of changes from a 1 to 2 or from a 2 to 1. Then we can convert the number of switches to the length of the longest run. We illustrate this computation below:

1. We start with a random arrangement of 1, 1, 1, 2, 2, 2 – say it is 2, 1, 2, 2, 1, 1.
2. We compute the differences in the values of the list:
   differences = -1, 1, 0, -1, 0

(-1 is the difference between the second and first elements, 1 is the difference between the third and second elements, etc.)

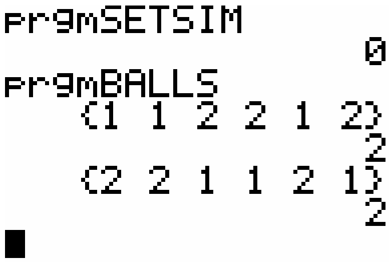
3. The number of switches is the sum of the absolute values of the differences:
   SWITCHES = |-1|+|1|+|0|+|-1|+|0| = 3.
4. Then one can show that
   if SWITCHES = 1 or 2, LONGEST RUN = 3
   if SWITCHES = 3 or 4, LONGEST RUN = 2
   if SWITCHES = 5, LONGEST RUN = 1.

The code for the program BALLS to perform one trial of this experiment is shown below.

What keys you press	What you see on the screen	What it does
2nd (1 , 1 , 1 , 2 , 2 , 2 2nd) STO〉 2nd 1 ENTER	$\{1,1,1,2,2,2\}\rightarrow L_1$	This creates a list with elements 1, 1, 1, 2, 2, 2 and stores the list in L1
MATH 〉 〉 〉 ENTER (6) STO〉 2nd 2 ENTER	$\text{rand}(6)\rightarrow L_2$	This generates six random uniform numbers between 0 and 1 and stores them in list L2
2nd STAT 〉 ENTER 2nd 2 , 2nd 1) ENTER	$\text{SortA}(L_2,L_1)$	Sorts the random numbers in L2, carrying along the elements in list L1
ALPHA x^2 + 1 STO〉 ALPHA x^2 ENTER	$I+1 \rightarrow I$	Updates the counter I by one.
2nd STAT 〉 〉 5 MATH 〉 1 2nd STAT 〉 7 2nd 1))) STO〉 ALPHA LN ENTER	$\text{sum}(\text{abs}(\Delta\text{List}(L_1)))\rightarrow S$	Computes the number of switches in the list L1.

PRGM 〉 3 2nd 1 ENTER	Disp L1	Displays the scrambled values in L1.
(ALPHA LN 2nd MATH 1 5) + 2 (ALPHA LN 2nd MATH 6 4) + (ALPHA LN 2nd MATH 6 2) STO〉 2nd 3 (ALPHA x^2) ENTER	(S=5)+2(S<=4)+(S<=2)→L_3(I)	Converts the number of switches to the length of the longest run and stores the result in list L3.

To run this program, one first runs the program SETSIM that sets the variable I to zero and clears the list L3 that will store the simulated results. Then one runs the program BALLS that will perform one trial of this experiment. The program will display the random arrangement and the length of the longest run. After BALLS is executed many times, the list L3 will contain the longest run values for all trials.



```
prgmSETSIM
                0
prgmBALLS
   {1 1 2 2 1 2}
                2
   {2 2 1 1 2 1}
                2
▮
```

Run this program so you get at least 50 trials of this experiment. In the below table, write down the number of trials where the longest run is of length 1, 2, and 3, and find the associated probabilities.

Longest run	1	2	3
Count			
Simulation Probability			

PART III: Simulating this problem on *Fathom*

- Open a New Collection – call the Attribute "ball" and put in the colors "red", "red", "red", "blue", "blue" and "blue". (There are 6 cases in your collection.)
- Scramble your collection – this will mix up the arrangement.
- From your scrambled collection, Define a Measure called "longest_run" – it is defined as `max(runlength(ball))`
- Select your Scrambled Collection and Collect Measures. Inspect the Collection, turn off animation and take 1000 measures.

Collect more Measures and construct a Count table of the values of longest_run.
Write down your counts from your 1000 simulations:

Longest run	1	2	3
Count			
Simulation Probability			

ACTIVITY: SAMPLING PEOPLE FROM A ROOM

Suppose you have a room of 6 women and 4 men and you select a random sample of 4 people without replacement. How many women will be in your sample?

PART I: Simulate this process on *Fathom*. Each step of this simulation procedure is described in detail below.

First start *Fathom*. Drag a New Collection and a New Case Table from the shelf.

STEP 1 – PUT THE PEOPLE INTO THE COLLECTION

In the Case Table, define two attributes, Name and Gender. In the Name column, type in the first names of 10 people, 6 female and 4 male. In the Gender column, type "Female" or "Male" corresponding to the gender of the person you typed in the Name column.

STEP 2 – TAKE A RANDOM SAMPLE FROM THE COLLECTION

- Select the Collection by clicking on it with the mouse
- Select the item Sample Cases from the Analyze menu.
- You should see a new Collection called Sample of Collection 1. If you look at the Case Table you should see a sample of 10 selected from your original collection.

STEP 3 – TELL *FATHOM* HOW TO TAKE THE RANDOM SAMPLE

When we take a sample, there are two questions:

- How many people are we sampling? (Here we want this to be 4.)
- Are we selecting people *with replacement* or *without replacement*? (Here we want to sample without replacement.)

To tell *Fathom* how to take the sample …

- Select the Sample of Collection.
- Select the menu item Inspect Collection from the Object menu.
- In the Sample portion of the Inspect Sample of Collection, change the number 10 to 4.
- Uncheck the With Replacement box.

STEP 4: RECORD A MEASURE FROM THE SAMPLE

- We want to record a measure, the number of women selected in the sample. With the Inspect Collection still open,
- In the Measures portion of the Inspect Sample, click in the Measure column to create a new measure called n_women

- Double click in the Formula box – you should see a new box open where you define this measure.
- In the formula box, type
  count(gender=”Female”)
- Close the Formula box and the Inspect Collection box.

STEP 5: REPEAT THIS SAMPLING PROCESS MANY TIMES

Now we want to take many samples of size 4, each time recording the number of women. To do this

- Select the Sample of Collection.
- Select the Collect Measures option from the Analyze menu. You should see a new collection called “Measures from Sample of Collection”. If you look at the corresponding Case Table, you should see the number of women observed from 5 samples taken from the room.

Suppose we want to take 100 samples instead of 5. To take more samples

- Select the Measures from Sample of Collection.
- Inspect this collection and look at the Collect Measures tab.
- Turn off the animation and change the number 5 to 100.
- Close the inspector box.

To take the 100 samples, select the Measures from Sample collection and type Apple-Y – you will see 100 samples being generated and the number of women found in all these samples is in the Case Table.

LOOK AT THE MEASURES (THE NUMBER OF WOMEN SAMPLED FOR THE 100 SAMPLES)

Construct a histogram of the Number of Women attribute.

To get a count table of the Number of Women attribute

- Drag a New Summary Table from the shelf.
- With the shift key depressed, drag the n_women attribute name to the blank area in the Summary Table – you should see a count table for this attribute.

Fill in the table below from this Summary Table.

	n_women selected				
	0	1	2	3	4
Count					
Approx. Probability					

PART II: Simulating the experiment on the *TI-84 Plus* calculator.
The code for a program "SAMPLING" to perform one simulation experiment is shown below. We represent the people in the room by a list consisting of 6 ones (women) and 4 two's (men). To take a random sample of size 4 without replacement, we

- randomly arrange the elements of the list
- find the number of women in the first 4 elements of the list

To run this program, one first runs the program SETSIM that sets the variable I to zero and clears the list L3 that will store the simulated results. Then one runs the program SAMPLING that will perform one trial of this experiment. The program will display the number of women in one random sample. After SAMPLING is executed many times, the list L3 will contain the number of women for many trials of this experiment.

What you type	What you see on the screen	What it does
2nd [{] 1, 1, 1, 1, 1, 1, 2, 2, 2, 2 2nd [}] STO> 2nd [L1]	{1,1,1,1,1,1,2,2,2,2}→L_1	This creates a list with elements 1, 1, 1, 1, 1, 1, 2, 2, 2, 2 and stores the list in L1

2nd [{] 1, 1, 1, 1, 0, 0, 0, 0, 0, 0 2nd [}] STO> 2nd [L4]	$\{1,1,1,1,0,0,0,0,0,0\}\to L_4$	This creates a list with elements 1, 1, 1, 1, 0, 0, 0, 0, 0, 0 and stores the list in L4
MATH ▶ PROB 1 (10) STO> 2nd [L2]	$\text{rand}(10)\to L_2$	This generates ten random uniform numbers between 0 and 1 and stores them in list L2
2nd [LIST] ▶ OPS 1 2nd [L2], 2nd [L1])	$\text{SortA}(L_2,L_1)$	Sorts the random numbers in L2, carrying along the elements in list L1
ALPHA [I] +1 STO> ALPHA[I]	$I+1 \to I$	Updates the counter I by one.
2nd [LIST] ▶ MATH 5 (2nd [L1] 2nd [TEST] 1) 2nd [L4]) STO> 2nd [L3](ALPHA[I])	$\text{Sum}((L_1=1)L_4)\to L_3(I)$	Computes the number of 1's in the first four components of L1 and stores the result in L3.

PART III: By using counting arguments …

1. Find the number of ways of selecting four people from the room (assume order is important).
2. Find the number of ways of selecting four women from the room.
3. Compute the exact probability of choosing four women – compare your answer with your simulation results.
4. Count the number of ways of selecting three women and one man.
5. Compute the exact probability of choosing three women – compare your answer with above.
6. Using this method, find the exact probability of choosing 0 women, 1 woman, and two women.

ACTIVITY: THE BIRTHDAY PROBLEM

PART I: Using *Fathom*

1. In the probability survey, I asked you to guess at the probability that at least two people in our class have the same birthday (month and day). Your probability guesses can be found at the webpage
http://personal.bgsu.edu/~albert/247datasets/birthday_problem.htm
Open up a new *Fathom* document. Import this dataset into *Fathom* (using the File > Import from Url menu item). Graph and summarize these probability guesses. Summarize how the class guessed at this probability question.

2. Now we will check if we have two matching birthdays in our class. Are you surprised at this result? Why or why not?

SIMULATING THE BIRTHDAY PROBLEM ON *FATHOM*
Open the *Fathom* document birthday_problem.ftm in the MATH 247 class folder. This document is designed to simulate the process of simulating birthdays from a class of students.

- The **birthdays collection** contains all possible birthdays (365 of them from January 1 through December 31).
- We take a sample of size n (with replacement) from the birthdays collection – the sample is placed in the **Sample of birthdays** collection.
- We can check if there is a matching birthday by doing a bar chart of the **Sample of birthdays** – if any of the heights of the bars are greater than 1, we have observed at least one birthday match.

3. By using the slider, change the value of n to the number of students actually in today's class. Select the **Sample of birthdays** collection and type Apple-Y to take a random sample of birthdays. Do this 10 times – each time record if there was a matching birthday (Y or N) and put your answers in the table.

TRIAL	1	2	3	4	5	6	7	8	9	10	TOTAL Yes
Match? (Y or N)											

What is your new guess at the probability of having matching birthdays? __________

REPEATING THE SIMULATION MANY TIMES ON *FATHOM*

On the right hand side of the *Fathom* page, I constructed a simulation that repeats the above process many times. In each simulation I have *Fathom*

- take a random sample of n birthdays
- count the number of matches between birthdays (could be 0, 1, 2, …)

Then *Fathom* repeats this (take a random sample of birthdays and count the number of matches) 200 times. The histogram and table on the right graph and summarize the number of matches for the 200 experiments.

4. Perform these 200 simulations by

- selecting the Measures from Sample of birthdays collection
- type Apple-Y

Fathom should be running these simulations and displaying the results in the histogram and summary table. Put your answers in the table below.

Number of matches	0	1	2	3	4	5	6 or more
Count							

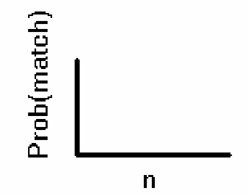
The probability that there will be at least one match is _______________________

5. The probability that there will be at least one birthday match will depend on how many students there are in the class (n).

Change the value of n using the slider. After you change n, rerun the 200 simulations and recompute the probability of getting at least one match. Do this for the different values of n in the table – each time, compute the probability of a match.

n	Probability of at least one match
10	
15	
20	
25	
30	
40	

6. Graph the values of n (horizontal axis) against the probability of a match (vertical axis). Connect the points using a smooth curve. Describe the general pattern that you see. Also, find the class size n so that the probability of getting a match is close to 0.5.



PART II: Using the *TI-84 Plus* Calculator:

The code for a program "BTHDAY" to perform one simulation experiment is shown below. Note that the variable N corresponds to the number of students in the class. You can set N equal to the number of students in your class. The random birthdays of the N students are represented as a list of N random integers from 1 to 365. From this list, the program will compute the number of matching pairs of birthdays.

To run this program, one first runs the program SETSIM that sets the variable I to zero and clears the list L3 that will store the simulated results. Then one runs the program BTHDAY that will perform one trial of this experiment. The program will display the number of matching pairs of birthdays in one random sample. After

BTHDAY is executed many times, the list L3 will contain the number of women for many trials of this experiment.

What you type	What you see on the screen	What it does
25 STO> ALPHA [N]	25→N	You set the number of birthdays to simulate.
MATH ▶ PROB 5 1,365, ALPHA[N]) STO> 2nd [L1]	randInt(1,365,N)→L_1	This generates N random integers between 1 and 365 and stores them in list L1
2nd [LIST] ▶ OPS 1 2nd [L1])	SortA(L_1)	Sorts the random numbers in L1 in ascending order.
ALPHA [I] +1 STO> ALPHA[I]	I+1 → I	Updates the counter I by one.
2nd [LIST] ▶ MATH 5 2nd [LIST] ▶ OPS 7 2nd [L1])2nd [TEST] 1 0) STO> 2nd [L3](ALPHA[I])	sum(ΔList(L_1)=0)→L_3(I)	Computes the number of matching pairs of birthdays and stores the result in L3

PART III: Doing an Exact Calculation:

Suppose there are 25 people in your class. One can represent the birthdays of the 25 students by the slots

____ ____ ____ ____ … ____ ____
1 2 3 4 24 25

where each slot contains a number from 1 to 365.

1. How many collections of birthdays are possible?

2. Assuming that each possible collection is equally likely, then we wish to count the number of birthday collections where there is at least one match. It actually is easier to count the number of collections where we have no match. If we represent outcomes by numbers placed in the 25 slots …

(a) How many possible birthdays can you place in the first slot?
(b) If the second person has a different birthday from the first, how many possible birthdays are there in the second slot?
(c) Continuing in this way, how many possible collections are possible where there is no match?
(d) Using (c), find the probability there is no match.
(e) Find the probability that there is at least one matching birthday.

WRAP-UP

Simulation is an attractive method for computing probabilities. This topic introduced the basic components of a probability simulation that include the choice of randomization device, the details of a simulation trial, the measurement of interest, and the number of trials to run. The use of physical devices such as dice or cards, the computer software package Fathom, and the TI-83 graphing calculators were used to illustrate simulation experiments for a number of problems.

EXERCISES

1. **Collector's Problem**

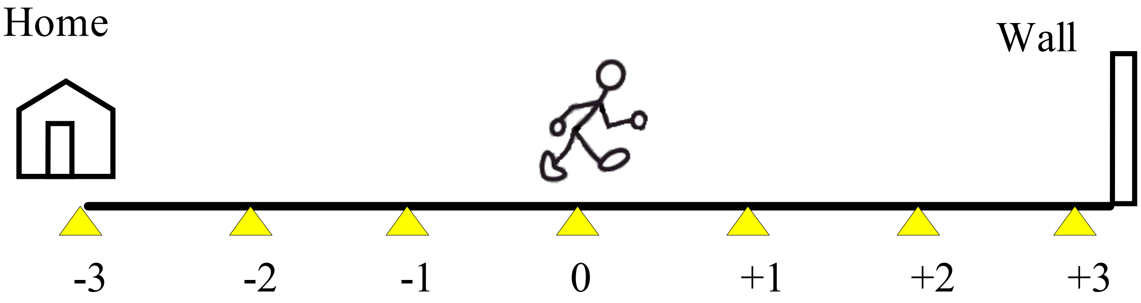
Suppose we decide to buy boxes of *Wheaties* until we obtain a complete set of six posters.

a. Hand simulate the process of purchasing cereal boxes by rolling a die multiple times. You roll repeatedly until you have rolled all numbers from 1 and 6. When you are done, record the number of roll of the die. Repeat this experiment 20 times. When you are done estimate (1) the mean number of boxes you need to get a complete set and (2) the probability that you have to purchase more than 15 boxes to get a set.

b. By using a variation of the simulation method described in the notes, use *Fathom* or some other software program to simulate this experiment 1000 times. Estimate the mean number of boxes to get a complete set and the probability of purchasing more than 15 boxes. Compare your answers with the answers obtained by the hand simulation in (a).

2. **Random Walk**

Suppose that a man is taking a random walk through a town. Each minute, he is equally likely to take a step to the right (to +1) or the left (to -1). He starts at location 0; if he reaches location +3, he hits a wall and can only walk to the left on the next step. If he reaches home (at location -3), then he remains there to take a long nap. If he takes 10 steps, we are interested in (a) the probability his current location is at a positive value (+1, +2, or +3), and (b) the probability he gets home.



Explain how you can use a die to simulate this random walk. Use hand simulation with 20 simulations to approximate the above probabilities.

3. **Matching Birth Months**

Suppose you record the birth months for five people. We are interested in the probability that at least two people have a matching birth month. Using the computer, one can simulate rolling a 12-sided die many times – these rolls can represent the birth months for many people. Here are 100 simulated rolls of the 12-sided die.

```
 12     8     1     1    11     3     6     9    10     2
  3    10     5     9     1     9    11     4    12     1
  8    12    10     6     9     4    10    11     7    11
  6     9     1    12     5     7     8     7    11     3
 11     3     2     6    10     2    10     5     3     4
 10     5     3     6     7     9     8     9    12     8
  6    12     3    11     9     5     5     7     4     4
  1    12     8     7     6    11     4     6     4     6
 10     5     4     3     4    11     5     9    11     1
  6    11     3     9     3     8     7     8     9    12
```

a. Explain how you can use a 12-sided die to simulate the birth months for five people.

b. Using the simulated rolls above, perform the birth month simulation 20 times. Approximate the probability of getting at least one matching birth month.

4. **Shooting One-and-One Shots in Basketball**

Suppose a basketball player has a 70% chance of making a single shot from the free-throw line. He takes a one-and-one shot – he takes one shot and if he makes the first shot, he has the opportunity to take a second shot. What is the probability that he makes exactly one shot (and scores one point) when he has this one-and-one opportunity? Explain how you can simulate this experiment using a 10-sided die. Perform this simulation 20 times and estimate the probability of scoring exactly one point.

5. **Waiting until a Boy**

Suppose that a married couple plans on having children until a boy is born. What is the probability that this couple will have three or more children? Explain how you can simulate this experiment using a coin flip. Perform this simulation 20 times and estimate the probability that the couple has at least three children.

6. **Game of Rock, Paper, and Scissors**

In the well-known rock, paper, scissors game between two players, each person plays a rock, paper, or scissors. Then

Rock: wins against scissors, loses to paper and stalemates against itself

Paper wins against Rock, loses to scissors and stalemates against itself

Scissors wins against paper, loses to rock and stalemates against itself

Suppose that each person plays one of the three things at random.

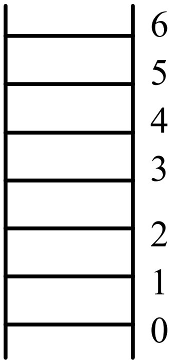
a. Explain how you could simulate this game using two dice.

b. Use the dice to simulate this game 30 times. Approximate the probability that Player I wins, and the probability that the game will be a stalemate.

7. **Moving Up a Ladder**

Suppose you start at the ladder rung marked 0. You then roll a die – if the die roll is even, you move up that number of rungs; if the die roll is odd, you move down that number of rungs. The lowest you can go is rung 0.

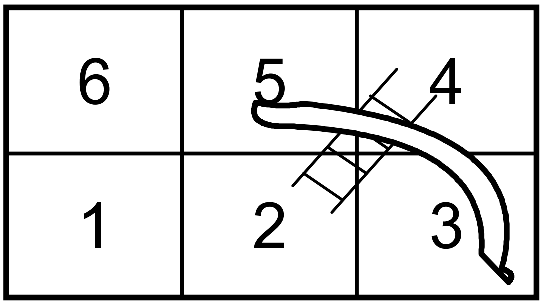
a. Play this game 20 times, recording each time how many rolls it takes to reach the top of the ladder. (You don't need to reach rung 6 on an exact roll.)



b. Find the average number of rolls it takes to reach the top of the ladder. Also, approximate the probability you can reach the top of the ladder in 5 rolls or less.

8. **Game of Snakes and Ladders**

The diagram shows a simple version of the famous snakes and ladders game. You play the game by rolling a die and moving along the board with the goal of reaching the 6 square. If you land on the 2 square, you can immediately jump to square 4 by the ladder. However, if you land on square 5, the snake will take you immediately back to square 3.



a. Play this game 20 times, recording the sequence of moves in each game until you reach square 6. (You don't need to reach square 6 by an exact roll.)

b. Approximate the probability that

- you reach square 6 in exactly 3 moves

- you never land on an odd square
- you complete the game in 3 moves or less

9. **Lottery Game**

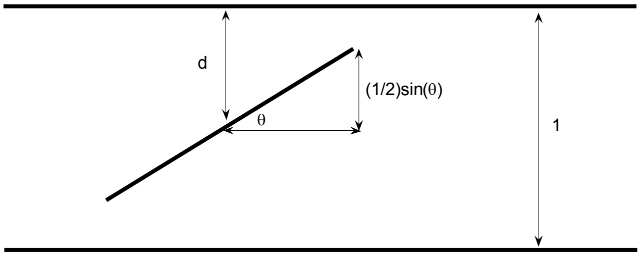
In the CASH 3 game offered by the Florida Lottery, you have the option of selecting a three-digit number with all digits different, and you win if the winning number is any arrangement of the three digits you select. (This is called the 6-Way Combo play.) Choose your favorite three-digit number (with all three digits different), and compute (by simulation) the probability that you win in this game. Given that the probability of winning this game is small, it probably will be better to use *Fathom* to do this simulation.

10. **Buffon's Needle Problem**

Consider Buffon's Needle problem as described in the spotlight. The below figure illustrates the problem, where d is the distance from the center of the needle to the closest line and θ is the angle of the needle from a horizontal line through the center of the needle. The needle will cross the line if

$d \le \frac{1}{2}\sin(\theta)$. To simulate this experiment, you simulate a distance d randomly between 0 and ½ , simulate an angle θ randomly between 0 and π, and record a hit if

$d \le \frac{1}{2}\sin(\theta)$. Using Fathom or the calculator, simulate dropping a needle 1000 times and approximate the probability of crossing the line.



TOPIC P5: CONDITIONAL PROBABILITY

SPOTLIGHT: THE THREE CARD PROBLEM

Suppose you have three cards – one card is blue on both sides, one card is pink on both sides, and one card is blue on one side and pink on the other side. Suppose you choose a card and place it down showing "blue". What is the chance that the other side is also blue?

This is an illustration of a famous conditional probability problem. We are given certain information – here the information is that one side of the card is blue – and we wish to determine the probability that the other side is blue.

Most people think that this probability is 1/2, but actually this is wrong. We can demonstrate the correct answer by simulating this experiment many times. You can do this simulation by hand, but I'll illustrate this using the Fathom program.

Suppose we think of this experiments are first choosing a card, and then choosing a side from the card. There are three possible cards, which we call "blue", "pink" and "mixed". For the blue card, there are two blue sides; for the pink card, there are two pink sides, and the "mixed" card has a blue side and a pink side. I had Fathom randomly choose a card and then a side. I repeat this 1000 times, and the below table categorizes the outcomes by the card and the side observed.

		Side observed		
		blue	pink	TOTAL
Card observed	blue	321	0	321
	pink	0	353	353
	mixed	160	166	326
	TOTAL	481	519	1000

We observed "side is blue" and we are interested in the probability of the event "card is blue". In this experiment, we observed the blue side 481 times – of these, the card was blue 321 times. So the probability the other side is blue is approximately

$$321/481 \approx 2/3.$$

This example illustrates that our intuition can be faulty in figuring out probabilities of the conditional type.

PREVIEW

Probabilities that we assign to events are dependent on what we know about the given experiment. In other words, probabilities are *conditional* on our current knowledge. As we learn more about the experiment, the probabilities we assign can change. In this topic, we begin by showing how new information will change the sample space of possible outcomes, and conditional probabilities will be found using this new sample space. Then we illustrate finding conditional probabilities informally in real-life problems and in the context of two-way count tables. We conclude by formally defining a conditional probability and use this definition to develop new rules (the multiplication rule and Bayes' rules) that will be helpful for solving more sophisticated probability problems.

In this topic, your learning objectives are to:

- Understand how to modify a sample space in the presence of new information.
- Understand how to informally modify your probabilities given new information and to compute conditional probabilities for a two-way table.
- Understand the formal definition of conditional probability and be able to use the multiplication rule for two-stage experiments.
- Be able to check if two events are independent.
- Understand how to compute inverse probabilities using Bayes' rule.

NCTM Standards

✓In Grades 9-12, all students should understand the concepts of conditional probability and independent events.

NEW INFORMATION, REDUCED SAMPLE SPACE, AND CONDITIONAL PROBABILITY

To illustrate the conditional nature of probabilities, suppose you have a box that has 6 slips of paper – the slips are labeled with the numbers 2, 4, 6, 8, 10, and 12. You select two slips at random from the box.

We assume that we're sampling without replacement and the order that we select the slips is not important. Then we can list all of the possible outcomes. (By the way, since we are choosing two numbers from six, the total number of outcomes will be $\binom{6}{2} = 15$.)

S = {(2, 4), (2, 6), (2, 8), (2, 10), (2, 12), (4, 6), (4, 8), (4, 10), (4, 12)
(6, 8), (6, 10), (6, 12), (8, 10), (8, 12), (10, 12)}.

Suppose we are interested in the probability the sum of the numbers on the two slips is 14 or higher. Assuming that the 15 outcomes we listed above are equally likely, we see there are 9 outcomes where the sum is 14 or higher and so

P(sum 14 or higher) = 9/15.

Next, suppose we're given some new information about this experiment -- both of the numbers on the slips are single digits. Given this information, we now have only six possible outcomes. We call this new sample space the *reduced sample space* based on the new information.

S = {(2, 4), (2, 6), (2, 8), (4, 6), (4, 8), (6, 8)}

We can evaluate the probability Prob(sum is 14 or higher) given that both of the slip numbers are single digits. Since there is only one way of obtaining a sum of 14 or higher in our new sample space, we see

P(sum of 14 and higher) = 1/6.

NOTATION: Suppose that E is our event of interest and H is our new information. Then we write the probability of E given the new information H as Prob (E | H), where the vertical line "|" means "conditional on" or "given" the new information. Here we found P(sum is 14 or higher | both slip numbers are single digits).

How does the probability of "14 or higher" change given the new information? Initially, the probability of 14 and higher was pretty high (9/15), but given the new information, the probability dropped to 1/6. Does this make sense? Yes. If we are told that both numbers are single digits, then we have drawn small numbers and that would tend to make the sum of the digits small.

PRACTICE: NEW INFORMATION, REDUCED SAMPLE SPACE, AND CONDITIONAL PROBABILITY

Let's consider the familiar example where we roll a red die and a white die – the 36 possible outcomes of rolling the dice are shown in the table below.

Consider the following three events:

S = the sum of the two rolls is 7

E = the red die is an even number

D = the rolls of the two dice are different

		Roll on red die					
		One	Two	Three	Four	Five	Six
	One	X	X	X	X	X	X

Roll on white die	Two	X	X	X	X	X	X
	Three	X	X	X	X	X	X
	Four	X	X	X	X	X	X
	Five	X	X	X	X	X	X
	Six	X	X	X	X	X	X

1. Find the probability you roll a sum equal to 7, that is, P(S).

2. Suppose we are told that the red die is an even number (event E). Circle the outcomes that correspond to the reduced sample space given this information. How many outcomes are there in this reduced sample space?

3. Find the probability of that the sum is equal to 7 given this information, that is, P(S | E).

4. Compare your answers to 1. and 3. Does knowing the red die is even change your probability of rolling a seven?

5. When the knowledge of one event *does not change* the probability of a second event, then we say that the two events are *independent*. Are events S and E independent?

6. In a similar fashion, compute P(S) and P(S | D) and compare the two probabilities. Are events S and D independent?

CONDITIONAL PROBABILITY IN EVERYDAY LIFE

Generally our beliefs about uncertain events can change when we get new information. Conditional probability provides a way for us to precisely say how our beliefs change. Let's illustrate this with a simple example.

Suppose you are interested in estimating the population of Philadelphia, Pennsylvania in the current year. Consider three possible events:

A = Philadelphia's population is under one million
B = Philadelphia's population is between one and two million
C = Philadelphia's population is over two million

If you know little about Philadelphia, then you probably are not very knowledgeable about its population. So you initially assign the probabilities shown in the table below.

Event	P(Event | I)
under one million	0.3
between one and two million	0.3
over two million	0.4
TOTAL	1.0

You are assigning approximately the same probability to each of the three possibilities, indicating that they are all equally likely in your mind. These can be viewed as conditional probabilities since they are conditional on your initial information – we denote them by P(E | I), where I denotes your Initial information.

Now suppose I give you some new information about Philadelphia's population. I won't tell you its current population, but I tell you that in 1990, Philadelphia was the fifth largest city in the country, and the population of the sixth largest city, San Diego, is 1.1 million in 1990. So this tells us that in 1990, the population of Philadelphia has to exceed 1.1 million. Now you might not be sure about how the population of Philadelphia has changed between 1990 and 2003, but it probably has not changed a significant amount. So you think that

- The population of Philadelphia is most likely to be between 1 and 2 million.
- It is very unlikely that Philadelphia's population is over 2 million.
- There is a small chance that Philadelphia's population is under 1 million.

You then revise your probabilities that reflect these beliefs as shown in the table below. We denote these probabilities as P(E | N), which are probabilities of these population events conditional on our newer information N.

Event	P(Event | N)
under one million	0.20
between one and two million	0.78
over two million	0.02
TOTAL	1.0

Now, I will give you additional information. To find the current population of Philadelphia, I look up the 2000 Census figures and the population of Philly is reported to be 1,517,550. Even though the Census number is three years old, you don't think that the population has changed much – definitely not enough to put in a new category of the table. So your probabilities will change again as shown in the below table. We call these probabilities of events conditional on additional information A.

Event	P(Event | A)
under one million	0
between one and two million	1
over two million	0
TOTAL	1.0

PRACTICE: CONDITIONAL PROBABILITY IN EVERYDAY LIFE

The notion of independence between two events can be understood in the context of probabilities in everyday life. Suppose you enroll in a history class. You are a pretty good student and have heard that this history class is relatively easy. There are three events of interest: you get an A in the class, you don't get an A but still pass, and you fail.

1. Based on your current beliefs at the beginning of the class, you assign the following probabilities to these three events.

Event	Get an A	Get B, C, or D	Fail
Probability			

2. Now you take the first test – it is much harder than you expected and you get a D. Assign new probabilities to these grade events.

Event	Get an A	Get B, C, or D	Fail
Probability			

3. Let X be the event that you get an A as a final grade and Y the event that you got a D in the first test. Are events X and Y independent? (Compare the probability P(X) that you assigned in 1. with the probability P(X|Y) that you assigned in 2.)

4. Let T be the event that a bad thunderstorm went through your town on the second day of class. Are events X and T independent? Why?

We actually make many judgments every day based on uncertainty. For example, we make decisions about the weather based on information such as the weather report, how it looks outside, and advice from friends. We make decisions about who we think will win a sports event based on what we read in the paper, our knowledge of the teams' strengths, and discussion with friends. Conditional probability is simply a way of quantifying our beliefs about uncertain events given information.

CONDITIONAL PROBABILITY IN A TWO-WAY TABLE

It can be easier to think about, and compute conditional probabilities when they are found from observed counts in a two-way table.

In the table below, we have classified high school athletes in 14 sports with respect to their sport and their gender. These numbers are recorded in thousands, so the 454 entry in the Baseball/Softball –Male cell means that 454,000 males played baseball or softball this year.

	Male	Female	TOTAL
Baseball/Softball	454	373	827
Basketball	541	456	997
Cross Country	192	163	355
Football	1048	1	1049
Gymnastics	2	21	23
Golf	163	62	225
Ice Hockey	35	7	42
Lacrosse	50	39	89
Soccer	345	301	646
Swimming	95	141	236
Tennis	145	163	308
Track and Field	550	462	1012
Volleyball	39	397	436
Wrestling	240	4	244
TOTAL	3899	2590	6489

Suppose we choose a high school athlete at random who is involved in one of these 14 sports. Consider several events

F – athlete chosen is female
S – athlete is a swimmer
V – athlete plays volleyball

What is the probability that the athlete is female? Of the 6489 (thousand) athletes, we see that 2590 were female, so the probability is

Prob(F) = 2590 / 6489 = 0.3991

Likewise, the probability that the randomly chosen athlete is a swimmer is

Prob(S) = 236 / 6489 = 0.0364.

and the probability he or she plays volleyball is

Prob(V) = 436 / 6489 = 0.0672.

Next, let's consider the computation of some conditional probabilities. What is the probability a volleyball player is female? In other words, conditional on the fact that the athlete plays volleyball, what is the chance that the athlete is female:

Prob(F | V)

To find this probability, we restrict attention only to the volleyball players in the table.

	Male	Female	TOTAL
Volleyball	39	397	436

Of the 436 (thousand) volleyball players, 397 are female, so

Prob(F | V) = 397/436 = 0.9106.

What is the probability a woman athlete is a swimmer? In other words, if we know that the athlete is female, what is the (conditional) probability she is a swimmer, or Prob(S | F)?

Here since we are given the information that the athlete is female, we restrict attention to the "Female" column of counts. There are a total of 2590 (thousand) women who play one of these sports; of these, 141 are swimmers. So

Prob(S | F) = 141 / 2590 = 0.0544.

Are events F and V independent? We can check this several ways. Above we found that the probability a randomly chosen athlete is a volleyball player is P(V) = 0.0672. Suppose we are told that the athlete is a female (F). Will that change the probability that she is a volleyball player? Of the 2590 women, 397 are volleyball players, and so P(V | F) = 397/2590 = 0.1533, Note that P(V) is different from P(V | F), that means that the knowledge the athlete is female has increased our probability that the athlete is a volleyball player. So the two events are not independent.

PRACTICE: CONDITIONAL PROBABILITY IN A TWO-WAY TABLE

The table below classifies all of the 2002 motor vehicle crashes by the time of day and the day of week. (The source is *Traffic Safety Facts 2002* from the National Highway Traffic Safety Administration of the U.S. Department of Transportation.) The numbers are in thousands of crashes.

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	**TOTAL**
Midnight to 3 am	83	29	30	27	35	37	65	**306**
3 am to 6 am	42	27	19	28	24	30	45	**215**
6 am to 9 am	38	129	141	148	139	128	52	**775**
9 am to Noon	66	135	126	130	133	141	127	**858**
Noon to 3 pm	126	178	168	175	169	227	160	**1203**
3 pm to 6 pm	119	256	251	243	235	288	156	**1548**
6 pm to 9 pm	105	117	125	126	135	159	118	**885**
9 pm to Midnight	69	68	67	62	65	106	93	**530**
TOTAL	**648**	**939**	**927**	**939**	**935**	**1116**	**816**	**6320**

Let S = event that a crash occurs on a Saturday, W = event that a crash occurs on Wednesday, and E = event that a crash occurs early in the morning (defined between midnight to 6 am).

1. Compute P(E).

2. Compute P(E | S).

3. Compute P(E | W).

4. Does the probability of an early morning crash depend on the day of the week? Explain.

DEFINTION OF CONDITIONAL PROBABILITY AND MULTIPLYING PROBABILITIES

In this topic, we have computed conditional probabilities by considering a reduced sample space. There is a formal definition of conditional probability that we will find useful in computing probabilities of complicated events.

Suppose we have two events A and B where the probability of event B is positive, that is P(B) > 0. Then the probability of A given B is defined as the quotient

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Example (How many boys?) To illustrate this conditional probability formula, suppose a couple has four children. You are told that this couple has at least one boy. What is the chance that they have exactly two boys?

If we let L be the event "at least one boy" and B be the event "have two boys", we wish to find P(B | L).

Suppose we represent the genders of the four children (from youngest to oldest) as a sequence of four letters. For example, the sequence BBGG means that the first two children were boys and the last two were girls. If we represent outcomes this way, there are 16 possible outcomes of four births:

BBBB
BBBG
BBGB
BGBB
BGBG
BGGB
GBBB
GBBG
GBGB
GGBB
GGBG
GGGB

BBGG BGGG GBGG GGGG

If we assume that boys and girls are equally likely (is this really true?), then each of the outcomes are equally likely and so we assign each outcome a probability of 1/16. Applying the definition of conditional probability, we have

$$P(B \mid L) = \frac{P(B \cap L)}{P(L)}.$$

There are 15 outcomes in the set L, and 6 outcomes where both events B and L occur. So using the definition

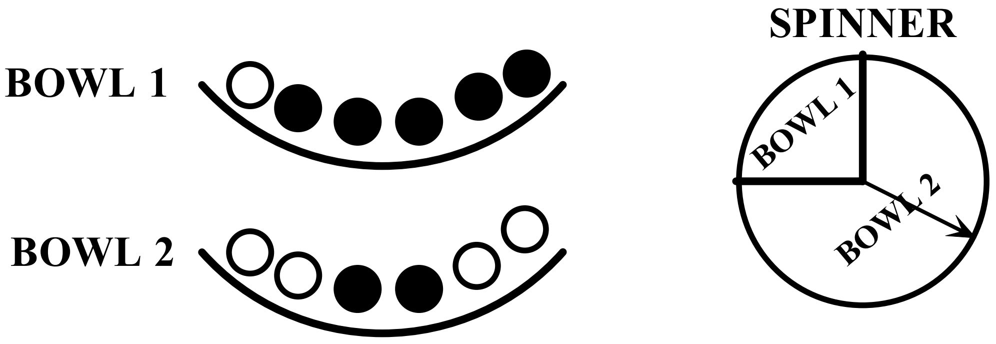
$$P(B \mid L) = \frac{P(B \cap L)}{P(L)} = \frac{6/16}{15/16} = \frac{6}{15}.$$

The Multiplication Rule

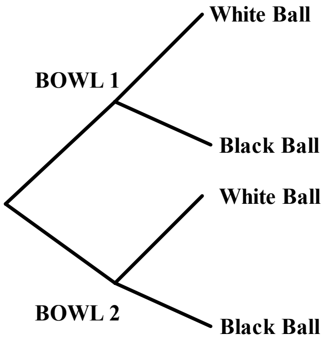
If we take the conditional probability formula and multiply both sides of the equation by P(B), we get the multiplication rule

$$P(A \cap B) = P(B)P(A \mid B)$$

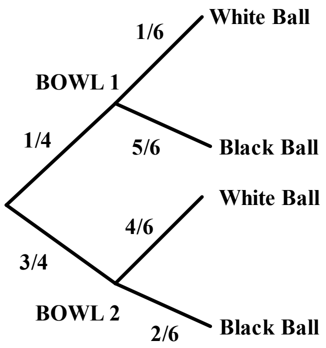
Example: Choosing balls from a random bowl. The multiplication rule is especially useful for experiments that can be divided into stages. Suppose we have two bowls -- Bowl 1 is filled with one white and 5 black balls, and Bowl 2 has 4 white and 2 black balls. We first spin the spinner below that determines which bowl to select, and then select one ball from the bowl. What the chance that the ball I select is white?



We can implement the multiplication rule by the tree diagram shown below. The first set of branches corresponds to the spinner result (choose Bowl 1 or choose Bowl 2) and the second set of branches corresponds to the ball selection.



We place numbers on the diagram corresponding to the probabilities that are given in the problem. Since one quarter of the spinner region is “Bowl 1”, the chance of choosing Bowl 1 is ¼ and so the chance of choosing Bowl 2 is ¾ -- we place these probabilities at the first set of branches. Also we know that if we select Bowl 1, the chances of choosing a white ball and a black ball are respectively 1/6 and 5/6. We place these conditional probabilities, P(white | Bowl 1) and P(black | Bowl 2) at the top set of branches at the second level. Also, if we select Bowl 2, the conditional probabilities of selecting a white ball and a black ball are given by P(white | Bowl 2) = 4/6 and P(black | Bowl 2) = 2/6 – these probabilities are placed at the bottom set of branches.



Now that we have the probabilities assigned on the tree, we can use the multiplication rule to compute the probabilities of interest:

- What is the probability of selecting Bowl 1 and selecting a white ball? By the multiplication rule

  $$\begin{aligned} P(Bowl\ 1 \cap White) &= P(Bowl\ 1) \times P(White \mid Bowl\ 1) \\ &= (1/4) \times (1/6) = 1/24 \end{aligned}$$

  (We are just multiplying probabilities along the top branch of the tree.)
- What is the probability of selecting a white ball? We see from the tree that there are two ways of selecting a white depending on which bowl is selected. We can either (1) select Bowl 1 and choose a white ball or (2) select Bowl 2 and choose a white ball. We find the probability of each of the two outcomes and add the probabilities to get the answer.

  $$\begin{aligned} P(white) &= P(Bowl\ 1 \cap white) + P(Bowl\ 2 \cap white) \\ &= P(Bowl\ 1)P(white \mid Bowl\ 1) + P(Bowl\ 2)P(white \mid Bowl\ 2) \\ &= (1/4)(1/6) + (3/4)(4/6) = 13/24 \end{aligned}$$

PRACTICE: DEFINITION OF CONDITIONAL PROBABILITY AND MULTIPLYING PROBABILITIES

Suppose that a restaurant keeps a gift bag for children; if the child cleans his/her plate, he or she is allowed to randomly choose one toy from the bag. Currently the bag consists of 10 yo-yos and 20 gliders – of the yo-yos, 3 are red, 5 are blue, and 2 are yellow, and of the 20 gliders, half are blue and half are yellow.

Let Y denote the event that a child selects a yo-yo, B the event that the toy is blue, and R the event that the toy is red.

1. Find the probability that a child selects a blue yo-yo, $P(B \cap Y)$.
2. Find the probability that a child selects a yo-yo, $P(Y)$.
3. Using the definition of conditional probability and your answers to 1 and 2, find $P(B \mid Y)$, the probability the toy is blue given that it is a yo-yo.
4. Using the definition, find $P(Y \mid B)$, the probability that a blue toy is a yo-yo.

For the remaining problems, suppose a child has the opportunity to choose two toys.

5. Find the probability that he/she chooses two gliders. (You wish to find the probability $P(choose\ glider\ 1st \cap choose\ glider\ 2nd)$ and remember that the toys are selected without replacement.)
6. Find the probability that the first toy is blue and the second toy is yellow.
7. Find the probability the child chooses two different toys.

THE MULTIPLICATION RULE UNDER INDEPENDENCE

When two events A and B are independent, then the multiplication rule takes the simple form

$$P(A \cap B) = P(A)P(B).$$

Moreover, if you have a sequence of independent events, say A, B, C, D and E, then the probability that all events happen simultaneously is the product of the probabilities of the individual events

$$P(A \cap B \cap C \cap D \cap E) = P(A)P(B)P(C)P(D)P(E).$$

By use of the assumption of independent events and multiplying, one can find probabilities of sophisticated events. We illustrate this in several examples.

Blood Types of Couples. White Americans have the blood types O, A, B, AB with respectively proportions .45, .40, .11, 04. Suppose two people in this group are married.

1. What is the probability that the man has blood type O and the woman has blood type A?

Let O_M denote the event that the man has "O" blood type and A_W the event that the woman has "A" blood type. Since these two people are not related, it is reasonable to assume that O_M and A_W are independent events. Applying the multiplication rule, the probability the couple have these two specific blood types is

$$\begin{aligned} P(O_M \cap A_W) &= P(O_M)P(A_W) \\ &= (.45)\times(.40) = .18. \end{aligned}$$

2. What is the probability the couple have O and A blood types?

This is a different question from the first one since we haven't specified who has the two blood types. Either the man can have blood type "O" and the woman have blood type "A" or the other way around. So the probability of interest is

$$\begin{aligned} P(two\ have\ A, O\ types) &= P(O_M \cap A_W\ or\ O_W \cap A_M) \\ &= P(O_M \cap A_W) + P(O_W \cap A_M). \end{aligned}$$

We can add the probabilities since $O_M \cap A_W$ and $O_W \cap A_M$ are different outcomes. We use the multiplication rule with our independence assumption to find the probability:

$$\begin{aligned} P(two\ have\ A, O\ types) &= P(O_M \cap A_W\ or\ O_W \cap A_M) \\ &= P(O_M)P(A_W) + P(O_W)P(A_M) \\ &= (.45)\times(.40) + (.45)\times(.40) \\ &= .36. \end{aligned}$$

3. What is the probability the man and the woman have the same blood types?

This is a more general question than the earlier parts since we haven't specified the blood types – we just are interested in the event that the two people have the *same* type. There are four possible ways for this to happen: they can both have type O, they both have type A, they have type B, or they have type AB. We first find the probability of each possible outcome and then sum the outcome probabilities to obtain the probability of interest. We obtain

$$\begin{aligned} P(same\ type) &= P(O_M \cap O_W \ or\ A_W \cap A_M \ or\ B_W \cap B_M \ or\ AB_W \cap AB_M) \\ &= (.45)^2 \times (.40)^2 + (.11)^2 \times (.04)^2 \\ &= .3762. \end{aligned}$$

4. What is the probability the couple have *different* blood tests?

One way of doing this problem is to consider all of the ways to have different blood types – the two people could have blood types O and A, types O and B, and so on, and add the probabilities of the different outcomes. But it is simpler to note that the event "having different blood tests" is the complement of the event "have the same blood". Then using the complement property of probability,

$$\begin{aligned} P(different\ type) &= 1 - P(same\ type) \\ &= 1 - .3762 \\ &= .6238. \end{aligned}$$

A Five-game Playoff. Suppose two baseball teams play in a "best of five" playoff series, where the first team to win three games wins the series. Suppose the Yankees play the Angels and one believes that the probability the Yankees will win a single game is .6. If the results of the games are assumed independent, what is the probability the Yankees win the series?

This is a more sophisticated problem than the first example, since there are numerous outcomes of this series of games. The first thing to note is that the playoff can last three games, four games, or five games. In listing outcomes, we let "Y" and "A" denote respectively the single-game outcomes "Yankees win" and "Angels win". Then a series result can be represented by a sequence of letters. For example, "YYAY" means that the Yankees won the first two games, the Angels won the third game, and the Yankees won the fourth game and the series. Using this notation, we write down below all of the possible outcomes of the five-game series.

Three games	Four games	Five games
YYY	YYAY, AAYA	YYAAY, AAYYA
AAA	YAYY, AYAA	YAYAY, AYAYA
	AYYY, YAAA	YAAYY, AYYAA
		AYYAY, YAAYA
		AYAYY, YAYAA
		AAYYY, YYAAA

We are interested in the probability the Yankees win the series. We underline all of the outcomes above where the Yankees win. By the assumption of independence, we can find the probability of a specific outcome – for example, the probability of the outcome "YYAY" is

$$\begin{aligned} P(YYAY) &= (.6)\times(.6)\times(.4)\times(.6) \\ &= .0864. \end{aligned}$$

We find the probability that the Yankees win the series by finding the probabilities of each type of Yankees win and adding the outcome probabilities. Below we write down the probability of each outcome.

Three games	Four games	Five games
P(YYY) = .216	P(YYAY) = .0864	P(YYAAY) = .0346
	P(YAYY) = .0864	P(YAYAY) = .0346
	P(AYYY) = .0864	P(YAAYY) = .0346
		P(AYYAY) = .0346
		P(AYAYY) = .0346
		P(AAYYY) = .0346

So the probability of interest is given by

$$
\begin{aligned}
P(Yankees\ win\ series) &= P(YYY, YYA, YAY, \ldots, AAYYY) \\
&= .216 + 3(.0864) + 6(.0346) \\
&= .683.
\end{aligned}
$$

PRACTICE: PLAYING CRAPS

One of the most popular casino games is craps. Here we describe a basic version of the game, and we will use the multiplication rule together with the use of conditional probabilities to find the probability of winnings.

This game is based on the roll of two dice. One begins by rolling the dice: if the sum of the dice is 7 or 11, the player wins, and if the sum is 2, 3, or 12, the player loses. If any other sum of dice is rolled (that is, 4, 5, 6, 8, 9, 10), this sum is called the "point". The player continues rolling two dice until either his point or a 7 are observed – he wins if he sees his point and loses if he observes a 7.

What is the probability of winning at this game?

1. On the first roll, the player can win by rolling 7 or 11, or lose by rolling 2, 3, or 12. Find the probabilities of these five outcomes and put the answers in the below table.

Initial roll	Probability	Outcome
7		Win
11		Win
2		Lose
3		Lose
12		Lose

2. If the player rolls initially a 4, 5, 6, 8, 9 or ten, he keeps rolling. First find the probabilities of rolling these sums (of two dice) and put your answer in the Prob(Roll) column of the table.

First Roll	Prob(Roll)	Secondary	Outcome	Prob(Win|Roll)	Product
		4	Win		

4		7	Lose		
5		5	Win		
		7	Lose		
6		6	Win		
		7	Lose		
8		8	Win		
		7	Lose		
9		9	Win		
		7	Lose		
10		10	Win		
		7	Lose		

3. Suppose you initially roll 4 and this becomes your point. Now you keep rolling until you get your point of 4 (you win) or you get a 7 (you lose). All of the other sums of two dice are not important. Write down all possible rolls of (first die, second die) that either gives you a sum of 4 or a sum of 7. Using this reduced sample space, find the conditional probability P(Win | First Roll is 4) – put this value in the Prob(Win | Roll) column.

4. Using a similar method to 3, find the P(Win | First Roll) if the first roll is 5, if the first roll is 6, …, the first roll is 10. Place these conditional probabilities in the Prob(Win | Roll) column.

5. Using the multiplication rule, the probability of rolling a 4 first and then winning is given by

$$P(Roll = 4 \cap Win) = P(Roll = 4)P(Win \mid Roll = 4)$$

Use this rule to find the probability of rolling a 4 and winning and place it in the "Product" column of the table.

6. Using a similar calculation, find the probabilities

$$P(Roll = 5 \cap Win),\ P(Roll = 6 \cap Win),\ P(Roll = 8 \cap Win),\ P(Roll = 9 \cap Win),\ P(Roll = 10 \cap Win)$$

and place them in the "Product" column of the table.

7. The probability you win at craps is the sum

$$\begin{aligned} P(Win) = P(Roll = 7) + P(Roll = 11) + P(Roll = 4 \cap Win) + P(Roll = 5 \cap Win) \\ + P(Roll = 6 \cap Win) + P(Roll = 8 \cap Win) + P(Roll = 9 \cap Win) + P(Roll = 10 \cap Win) \end{aligned}$$

Use the calculations from the table to find the probability you win at craps.

8. Is craps a fair game? Who has the advantage in this game: you or the casino? Is it a large advantage? Explain.

LEARNING USING BAYES' RULE

We have seen that probabilities are conditional in that one's opinion about an event is dependent on our current state of knowledge. As we gain new information, our probabilities can change. Bayes' rule provides a mechanism for changing our probabilities when we obtain new data.

Suppose that you are given a blood test for a rare disease. The proportion of people who currently have this disease is .1. The blood test comes back with two results: positive, which is some indication that you may have the disease, or negative. It is possible that the test will give the wrong result. If you have the disease, it will give a negative reading with probability .2. Likewise, it will give a false positive result with probability .2. Suppose that you have a blood test and the result is positive. Should you be concerned that you have the disease?

In this example, you are uncertain if you have the rare disease. There are two possible alternatives: you have the disease, or you don't have the disease. Before you have a blood test, you can assign probabilities to “have disease" and “don't have disease" that reflect the plausibility of these two models. You think that your chance of having the disease is similar to the chance of a randomly selected person from the population. Thus you assign the event “have disease" a probability of .1 By a property of probabilities, this implies that the event “don't have disease" has a probability of 1- .1 = .9.

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. Here the two possible data results are a positive result (+) or a negative result (-). We are given the probabilities of the observations for each model. If we “have the disease," the probability of a + observation is .8 and the probability of a - observation is .2. Since these are conditional probabilities, we can write

$$P(+\,|\,disease) = .8, \quad P(-\,|\,disease) = .2.$$

Likewise, if we “don't have the disease," the probabilities of the outcomes + and - are .2 and .8, respectively. Using symbols, we have

$$P(+\,|\,no\ disease) = .2, \quad P(-\,|\,no\ disease) = .8.$$

Suppose you take the blood test and the result is positive (+) – what is the chance you really have the disease? We are interested in computing the conditional probability

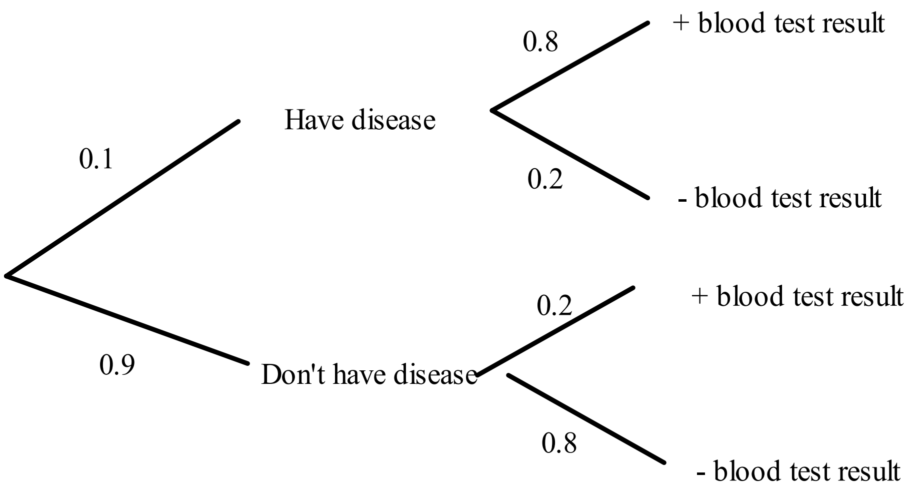
$$P(disease\,|\,+).$$

This should not be confused with the earlier probability $P(+\,|\,disease)$ that is the probability of getting a positive result if you have the disease. Here the focus is on the so-called *inverse probability* $P(disease\,|\,+)$ -- the probability of having the disease given a positive blood test result.

We describe the computation of this inverse probability using two methods. They are essentially two ways of viewing the same calculation.

METHOD 1: Using a tree diagram.

A person either has or does not have the disease, and given the person's disease state, he or she either gets a positive or negative test result. We can represent the outcomes by a tree diagram where the first set of branches corresponds to the disease states and the second set of branches corresponds to the blood test results. We label the branches of the tree by the given probabilities.



By the definition of conditional probability,

$$P(disease \mid +) = \frac{P(disease \cap +)}{P(+)}.$$

We can find the numerator $P(disease \cap +)$ by use of the multiplication rule:

$$\begin{aligned} P(disease \cap +) &= P(disease)P(+ \mid disease) \\ &= 0.1 \times 0.8 = 0.08 \end{aligned}.$$

In the tree diagram, we are multiplying probabilities along the disease/+ branch to find this probability.

To find the denominator $P(+)$, we first note that there are two ways of getting a positive blood test result – either the person has the disease and gets a positive blood test result, or the person doesn't have the disease and gets a positive result. These two outcomes are the disease/+ and no disease/+ branches of the tree. We find the probability $P(+)$ by using the multiplication rule to find the probability of each outcome, and then summing the outcome probabilities:

$$\begin{aligned} P(+) &= P(disease \cap +) + P(no\ disease \cap +) \\ &= P(disease)P(+ \mid disease) + P(no\ disease)P(+ \mid no\ disease) \\ &= 0.1 \times 0.8 + 0.9 \times 0.2 \\ &= 0.26 \end{aligned}$$

So the probability of having the disease, given a positive blood test result is

$$P(disease \mid +) = \frac{P(disease \cap +)}{P(+)} = \frac{0.08}{0.26} = 0.31\,.$$

As one would expect, the new probability of having the disease (.31) is larger than the initial probability of having the disease (.1) since a positive blood test was observed.

METHOD 2: Using a Bayes' box.

There is an alternative way of computing the inverse probably based on a two-way table that classifies people by the disease status and the blood test result. This is an attractive method since it based on expected counts rather than probabilities.

Suppose we have 1000 people in our community – we place "1000" in the lower right corner of the table.

		Blood test result		
		+	-	TOTAL
Disease	Have disease			

status	Don't have disease			
TOTAL				1000

We know that the chance of getting the disease is 10% -- so we expect 10% of the 1000 = 100 people to have the disease and the remaining 900 people to be disease-free. We place these numbers in the right column corresponding to "Disease status".

		Blood test result		
		+	-	TOTAL
Disease status	Have disease			100
	Don't have disease			900
TOTAL				1000

We know the test will err with probability .2. So if 100 people have the disease, we expect 20% of 100 = 20 to have a negative test result and 80 will have a positive result – we place these counts in the first row of the table. Likewise, if 900 people are disease-free, then 20% of 900 = 180 will have an incorrect positive result and the remaining 720 will have a negative result – we place these in the second row of the table.

		Blood test result		
		+	-	TOTAL
Disease status	Have disease	80	20	100
	Don't have disease	180	720	900
TOTAL				1000

Now we are ready to compute the probability of interest $P(disease \mid +)$ from the table of counts. Since we are conditioning on the event +, we restrict attention to the + column of the table – we see that 260 people had a positive test result. Of these 260, 80 actually had the disease, so

$$P(disease\,|\,+) = \frac{80}{260} = 0.31.$$

Note that, as expected, we get the same answer for the inverse probability.

PRACTICE: LEARNING USING BAYES' RULE

Suppose a friend is flipping a coin. Either the coin is the usual fair variety, or it is a special coin with heads on both sides. You are 80% sure that the coin is fair. Suppose your friend flips this coin three times and obtains three heads. In the following, we will calculate the new probability that the coin is fair.

1. There are two possible states of the coin, fair or two-headed, and assume that there are two possible outcomes --- HHH or "not HHH". Construct a tree diagram where the first set of branches corresponds to the states of the coin and the second set of branches corresponds to the two outcomes.

2. In this problem, identify the following probabilities:

P(fair coin) = ______________ P(two-headed coin) = _______________

P(HHH | fair coin) = ____________ P(HHH | two-headed coin) = ___________

Label the branches of the tree diagram with these probabilities.

3. Find the probability of interest P(fair coin | HHH).

4. Use the following Bayes' box to compute the probability P(fair coin | HHH). Note that we start with 1000 coins like the one that your friend is holding.

		Coin result		
		HHH	Not HHH	TOTAL
Coin Type	Fair			
	Two-headed			

TOTAL				1000

TECHNOLOGY ACTIVITY: ROLLING TWO DICE

Suppose you roll two fair dice and you keep track of two measures: the sum of the rolls and the maximum of the two rolls. We will use Fathom to simulate this process 1000 times.

1. Open a new Collection and Data Table. Define an attribute called "Die" and put the numbers 1, 2, 3, 4, 5, 6 in the table.

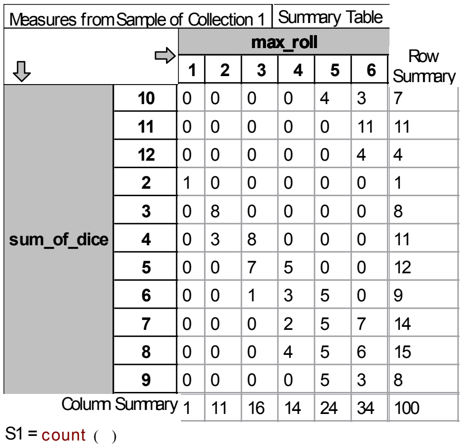
2. Take a Sample of size 2 with replacement from your collection. This sample represents the rolls of the two dice. (It may help to turn off the animation of this sampling.)

3. Define two Measures from your sample:

- sum_of_dice = sum(die)
- max_roll = max(die)

4. Now you want to repeat rolling 2 dice 1000 times. Collect Measures from your Sample of Collection. When you inspect the Collect Measures collection, you want to collect 1000 measures and, to speed things up, turn off the animation.

5. After you do your simulation, construct a two-way frequency table of sum_of_dice and max_rolls like the one shown to the right. When you drag the attribute names to the table, make sure the Shift key is depressed.



		max_roll						
		1	2	3	4	5	6	Row Summary
sum_of_dice	10	0	0	0	0	4	3	7
	11	0	0	0	0	0	11	11
	12	0	0	0	0	0	4	4
	2	1	0	0	0	0	0	1
	3	0	8	0	0	0	0	8
	4	0	3	8	0	0	0	11
	5	0	0	7	5	0	0	12
	6	0	0	1	3	5	0	9
	7	0	0	0	2	5	7	14
	8	0	0	0	4	5	6	15
	9	0	0	0	0	5	3	8
Column Summary		1	11	16	14	24	34	100

S1 = count ( )

6. Using this table, find the following probabilities:

(a) P(sum = 7 and max roll = 6)

(b) P(sum = 7)

(c) P(sum = 7 | max roll = 6)

(d) P(max roll = 6 | sum = 7)

(e) Suppose you are given the information that the maximum roll is 5. What is the most likely value of the sum? What is its probability?

TECHNOLOGY ACTIVITY: HOW MANY DEFECTIVES?

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.

Initially, the worker believes, based on past experience, that there is a 60% chance of no defectives in the box and if there any defectives, it is equally likely to be 1, 2, 3, or 4 defectives. 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?

We'll simulate this process using Fathom and answer the above question by the use of Bayes' rule.

1. Define a new collection and data table. Define two Attributes

- The first Attribute "n_defectives" is defined by the formula
  randompick(0,0,0,0,0,0,1,2,3,4)
  This attribute represents the unknown number of defectives in the box.
- The second Attribute "observation" represents the process of choosing a part from the box and observing if the selected part is defective or acceptable.
  if randominteger(1,4)<=n_defectives then "defective", otherwise "acceptable"

2. Place 1000 cases in your collection. Construct a two-way count table of the attributes n_defectives and observation. From this table, find the following probabilities.

(a) Find the probability that there is actually one defective in the box and an acceptable item is chosen.

(b) Find the probability an acceptable item is chosen.

(c) If you find an acceptable item, find the probability there are no defectives in the box.

(d) Assuming you find an acceptable item, find the probability of exactly 0, 1, 2, 3, and 4 defectives in the box from the Fathom output. Place your probabilities in the "probability after observing acceptable part" column of the table. Also place your initial probabilities in the "probability before sampling" column.

Number of defectives in box	Probability before sampling	Probability after observing acceptable part
0		
1		
2		
3		
4		

(e) After finding an acceptable part, what is the most likely number of defectives in the box?

WRAP-UP

Probabilities are always conditional on current information or assumptions. In the case where one can specify a sample space, new information will result in a **reduced sample space**, and conditional probabilities can be found using this new sample space. Examples were given that illustrated the use of conditional probability in everyday life. One states beliefs about events by means of probabilities and these probabilities can change in the light of new information.   If knowledge about an event has no effect on the probability of a second event, then we say that the two events are **independent**. In a two-way of counts, conditional probabilities can be found by finding proportions restricted on certain rows or columns of the table. The conditional probability of one event given a second event was defined formally. This definition gives the **multiplication rule** that is useful for finding probabilities in many-stage experiments. The topic concluded with **Bayes' rule** that can be used to compute probabilities of events in the presence of new information.

EXERCISES

1. **Flipping Coins**

Suppose you flip a fair coin four times. The 16 possible outcomes of this experiment are shown below.

HHHH	HHHT	HHTT	HHTH
HTHH	HTHT	HTTT	HTTH
THHH	THHT	THTT	THTH
TTHH	TTHT	TTTT	TTTH

a. Let A denote the event that you flip exactly three heads. Find the probability of A.

b. Suppose you are given the information N that at least two heads are flipped. Circle the possible outcomes in the reduced sample space based on knowing that event N is true.

c. Using the reduced sample space, find the conditional probability P(A | N).

d. Compare P(A) computed in part a with P(A | N) computed in part (c). Based on this comparison, are events A and N independent? Why?

2. **Choosing a Committee**

Suppose you randomly choose three people from the group

{Sue, Ellen, Jill, Bob, Joe, John}

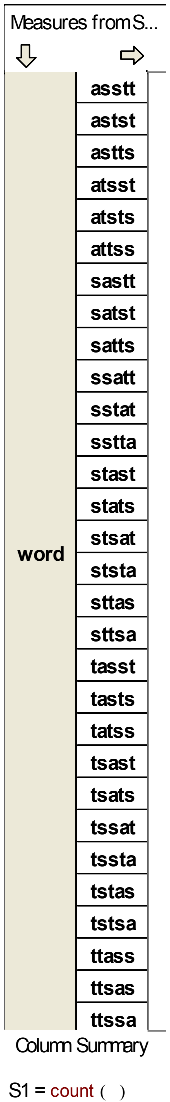
to be on a committee. Below I have listed all possible committees of size three:

{Sue, Ellen, Jill}, {Sue, Ellen, Bob}, {Sue, Ellen, Joe}, {Sue, Ellen, John}
{Sue, Jill, Bob}, {Sue, Jill, Joe}, {Sue, Jill, John}.
{Sue, Bob, Joe}
{Sue, Bob, John}, {Sue, Joe, John}, {Ellen, Jill, Bob},
{Ellen, Jill, Joe}
{Ellen, Jill, John}, {Ellen, Bob, Joe} {Ellen, Bob, John},
{Ellen, Joe, John}
{Jill, Bob, Joe}, {Jill, Bob, John}, {Jill, Joe, John},
{Bob, Joe, John}

a. Find the probability of the event A that exactly two women are in the committee.

b. Suppose you are told that Jill is on the committee – call this event J. Circle the possible outcomes in the reduced sample space if we know that J is true.

c. Compute the conditional probability P(A | J).



Measures from S...

⇩ ⇨

word	asstt
	astst
	astts
	atsst
	atsts
	attss
	sastt
	satst
	satts
	ssatt
	sstat
	sstta
	stast
	stats
	stsat
	ststa
	sttas
	sttsa
	tasst
	tasts
	tatss
	tsast
	tsats
	tssat
	tssta
	tstas
	tstsa
	ttass
	ttsas
	ttssa

Column Summary

S1 = count ( )

d. Based on your computations in parts a and c, are events A and J independent?

e. Let F denote the event that more women are on the committee than men. Find P(F).

f. Suppose you are given the information S that all three people on the committee are of the same gender. Find P(F | S).

g. Based on your computations in e and f, are events F and S independent?

3. **Arranging Letters**

Suppose you randomly arrange the letters a, s, s, t, t. Your author had Fathom do this arranging 200 times and the table on the right lists all of the possible "words" that came up. (There were 30 distinct arrangements.)

a. Assuming each possible arrangement is equally likely, what is the probability that the word formed is "stats"?

b. What is the probability that the word formed begins and ends with an "s"?

c. Suppose you are told that the word formed starts with "s" – write down all of the possible words in the reduced sample space.

d. Given that the word begins with "s", what is the probability the word is "stats"?

4. **Rolling Two Dice**

Suppose two dice are rolled.

a. Suppose you are told that the sum of the dice is equal to 7. Write down the six possible outcomes.

b. Given the sum of the dice is equal to 7, find the probability the largest die roll is 6.

c. Suppose you are told that the two dice have different numbers. Write down the possible outcomes.

d. If the two dice have different numbers, what is the probability the largest die roll is 6?

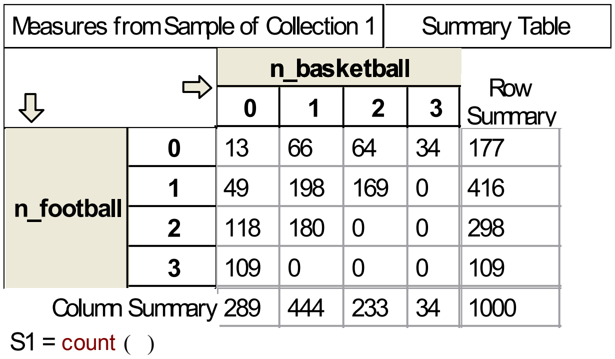
5. **Choosing Sport Balls**

Suppose you have a bin in your garage with three sports balls – four are footballs, three are basketballs, and two are tennis balls. Suppose you take three balls from the bin – you count the number of footballs and the number of basketballs. The first time this is done, the following balls were selected:

basketball, basketball, football

so the number of footballs selected was 1 and the number of basketballs selected was 2.

We repeat this sampling experiment 1000 times, each time recording the number of footballs and basketballs we select. The table below summarizes the results of the 1000 experiments:



Measures from Sample of Collection 1 | Summary Table

⇩	⇨	n_basketball				Row Summary
		0	1	2	3	
n_football	0	13	66	64	34	177
	1	49	198	169	0	416
	2	118	180	0	0	298
	3	109	0	0	0	109
Column Summary		289	444	233	34	1000

S1 = count ( )

Let F1 denote that event that you have chosen exactly one football and B1 the event that you chose exactly one basketball from the bin.

a. Find P(F1) and P(B1).

b. Find P(F1 | B1).

c. Find P(B1 | F1).

d. From your calculations above, explain why F1 and B1 are not independent events.

6. **Rating Movies**

On the Internet Movie Database (www.imdb.com), people are given the opportunity to rate movies on a scale from 1 to 10. The table below shows the ratings of the movie "Sleepless in Seattle" for men and women who visited the website.

a. Suppose you choose at random a person who is interested in rating this movie on the website. Find the probability that the person gives this movie a high rating between 8 and 10 – that is, P(H).

	8,9,10	5,6,7	1,2,3,4	
Rating	(High)	(Median)	(Low)	TOTAL
Males	2217	3649	754	6620
Females	1059	835	178	2072
TOTAL	3276	4484	932	8692

b. Find the conditional probabilities P(H | M) and P(H | F), where M and F are the events that a man and a woman rated the movie, respectively.

c. Interpret the conditional probabilities in part b – does this particular movie appeal to one gender?

d. The table below shows the ratings of the movie "Die Hard" for men and women who visited the website. Answer questions a, b, and c for this movie.

	8,9,10	5,6,7	1,2,3,4	
Rating	(High)	(Median)	(Low)	TOTAL
Males	16197	6937	882	24016
Females	1720	1243	258	3221
TOTAL	17917	8180	1140	27237

7. **Rating Movies**

The Internet Movie Database also breaks down the movie ratings by the age of the reviewer. For the movie "Sleepless in Seattle", the table below classifies the reviewers by age and their rating.

		5,6,7	1,2,3,4	
	8,9,10 (High)	(Median)	(Low)	TOTAL
under 18	74	76	16	166
18-29	1793	2623	555	4971

30-44	886	1280	272	2438
45+	438	300	60	798
TOTAL	3191	4279	903	8373

a. Find the probability that a reviewer gives this movie a high rating – that is, find P(H).

b. Define a "young adult" (YA) as a person between the ages of 18 and 29, and a "senior" (S) as a person 45 or older. Compute P(H | YA) and P(H | S).

c. Based on your computations in parts a and c, is "giving a high rating" and age independent events? If not, explain how the probability of giving a high rating depends on age.

8. **Family Planning**

Suppose a family plans to have children until they have two boys. Suppose there are two events of interest, A = event that they have at least five children and B = event that the first child born is male. Assuming that each child is equally likely to be a boy or girl, and genders of different children born are independent, then this process of building a family was simulated 1000 times. The results of the simulation are displayed in the following table.

		Gender of First Born		
		Female	Male	TOTAL
	2	0	247	247
Number of Children	3	125	138	263
	4	126	58	184
	5 or more	250	56	306
	TOTAL	501	499	1000

a. Use the table to find P(A).

b. Find P(A | B) and decide if events A and B are independent.

c. Suppose another family plans to continue to have children until they have at least one of each gender. The table of simulated results of 1000 families of this type is shown below. Again find P(A), P(A | B) and decide if events A and B are independent.

		Gender of First Born		
		Female	Male	TOTAL
	2	235	261	496
Number of Children	3	106	152	258
	4	71	63	134
	5 or more	50	62	112
	TOTAL	462	538	1000

9. **Conditional Nature of Probability**

For each of the following problems

(i) Make a guess at the probabilities of the three events based on your current knowledge.

(ii) Ask a friend about this problem. Based on his or her opinion about the event, make new probability assignments.

(iii) Do some research on the Internet to learn about the right answer to the question. Make new probability assignments based on your new information.

a. What is the area of Pennsylvania?

Initial probabilities:

Event	under 30,000 sq miles	between 30,000 and 50,000 sq miles	over 50,000 sq miles
Probability			

Probabilities after talking with a friend:

Event	under 30,000 sq miles	between 30,000 and 50,000 sq miles	over 50,000 sq miles
Probability			

Probabilities after doing research on the Internet.

Event	under 30,000 sq miles	between 30,000 and 50,000 sq miles	over 50,000 sq miles
Probability			

a. Robin Williams has appeared in how many movies?

Initial probabilities:

Event	Under 15	Between 16 and 30	Over 30
Probability			

Probabilities after talking with a friend:

Event	Under 15	Between 16 and 30	Over 30
Probability			

Probabilities after doing research on the Internet.

Event	Under 15	Between 16 and 30	Over 30
Probability			

10. **Conditional Nature of Probability**

For each of the following problems

(i) Make a guess at the probabilities of the three events based on your current knowledge.

(ii) Ask a friend about this problem. Based on his or her opinion about the event, make new probability assignments.

(iii) Do some research on the Internet to learn about the right answer to the question. Make new probability assignments based on your new information

a. How many plays did Shakespeare write?

Initial probabilities:

Event	Under 30	Between 31 and 50	Over 50
Probability			

Probabilities after talking with a friend:

Event	Under 30	Between 31 and 50	Over 50

Probability			

Probabilities after doing research on the Internet.

Event	Under 30	Between 31 and 50	Over 50
Probability			

b. What is the average temperature in Melbourne, Australia in June?

Initial probabilities:

Event	Under 40° F	Between 40° and 60° F	Over 60° F.
Probability			

Probabilities after talking with a friend:

Event	Under 40° F	Between 40° and 60° F	Over 60° F
Probability			

Probabilities after doing research on the Internet.

Event	Under 40° F	Between 40° and 60° F	Over 60° F
Probability			

11. **Picnic Misery**

Twenty boys went on a picnic. Five got sunburned, 8 got bitten by mosquitoes, and 10 got home without mishap. What is the probability that the mosquitoes ignored a sunburned boy? What is the probability that a bitten boy was also burned?

12. **A Mall Survey**

Suppose 30 people are surveyed at a local mall. Half of the 10 men surveyed approve the upcoming school levy and a total of 17 people don't approve of the levy. Based on the survey data,

a. What is the probability a woman is in favor of the levy?

b. If the person is in favor of the levy, what is the probability the person is a woman?

13. **Drawing Tickets**

Have 12 tickets numbered from 1 to 12. Two tickets are drawn, one after the other, without replacement.

a. Find the probability that both numbers are even.

b. Find the probability both numbers are odd.

c. Find the probability one number is even and one is odd.

14. **Testing for Steroids**

Suppose that 20% of all baseball players are currently on steroids. You plan on giving a random player a test, but the test is not perfectly reliable. If the player is truly on steroids, he will test negative (for steroids) with probability 0.1. Likewise, if the player is not on steroids, he will get a positive test result with probability 0.1.

a. What is the probability the player is on steroids and will test negative?

b. If you give a player a test, what is the probability he will test positive?

c. If the test result is positive, what is the probability the player is on steroids?

15. **Preparing for the SAT**

Suppose a student has a choice of enrolling (or not) in an expensive program to prepare for taking the SAT exam. The chance that she enrolls in this class is 0.3. If she takes the program, the chance that she'll do well on the SAT exam is 0.8. On the other hand, if she doesn't take the prep program, the chance that she'll do well on the SAT is only 0.4. Let E denote the event "enrolls in the class" and W denote the event "does well on the SAT exam".

a. Find $P(W \mid E)$

b. Find $P(E \cap W)$

c. Find $P(E \mid W)$, that is, the probability that she took the class given that she did well on the test.

16. **Working Off-Campus**

At my college campus, 33% of the students are freshman and 25% are seniors. 13% of the freshman works over 10 hours off-campus, and 37% of the seniors work over 10 hours off-campus.

a. Suppose you sample a student who is either a freshman and senior. Find the probability he works over 10 hours off-campus.

b. If this person does work over 10 hours off campus, find the probability she is a senior.

17. **Flipping Coins**

You flip a coin three times. Let A be the event that a head occurs on the first flip and B is the event that (exactly) one head occurs. Are A and B independent?

18. **A Two-headed Coin?**

One coin in a collection of 65 has two heads. Suppose you choose a coin at random from the collection – you toss it 6 times and observe all heads. What is the probability it was the two-headed coin?

19. **Smoking and Gender**

Suppose the proportion of female students at your school is 60%. Also you know that 26% of the male students smoke and only 16% of the female students smoke. Suppose you randomly select a student.

a. Find the probability the student is a male smoker.

b. Find the probability the student smokes.

c. the student smokes, what is the probability the student is female?

20. **Choosing Until One Selects a Red**

Suppose you have a box with 4 green and 2 red balls. You select balls from the box one at a time until you get a red, or until you select three balls. If you don't select a red on the first draw, find the probability that you will select three balls.

21. **Mutually Exclusive and Independence**

Suppose that two events A and B are mutually exclusive. Are they independent events?

22. **Blood Type of Couples**

Consider the example in the book where white Americans have the blood types O, A, B, AB with proportions .45, .40, .11, 04. If two people are married

a. Find the probability both people have blood type A.

b. Find the probability the couple have A and B blood types.

c. Find the probability neither person has an A type.

23. **Five-Game Playoffs**

Consider the “best of five” playoff series between the Yankees and the Indians described in the example. We assume the probability the Yankees win a single game is .6.

a. Find the probability the Yankees win in three games.

b. Find the probability the series lasts exactly three games.

c. Find the probability the series lasts five games and no team wins more than one game in a row.

24. **Computer and Video Games**

The Entertainment Software Association reports that of all computer and video games sold, 53% are rated E (Everyone), 30% are rated T (Teen), and 16% are rated M (Mature). Suppose three customers each purchase a game at a local store. Assume that the software choices for the customers can be regarded as independent events.

a. Find the probability that all three customers buy games that are rated E.

b. Find the probability that exactly one customer purchases a M rated game.

c. Find the probability that the customers purchase games with the same rating.

25. **Washer and Dryer Repair**

Suppose you purchase a washer and dryer from a particular manufacturer. From reading a consumer magazine, you know that 20% of the washers and 10% of the dryers will need some repair during the warranty period.

a. Find the probability that both the dryer and washer will need repair during the warranty period.

b. Find the probability that exactly one of the machines will need repair.

c. Find the probability the neither machine will need repair.

26. **Basketball Shooting**

In a basketball game, a player has a "one and one" opportunity at the free-throw line. If she misses the first shot, she is done. If instead she makes the first shot, she will have an opportunity to make a second shot. From past data, you know that the probability this player will make a single free-throw shot is .7.

a. Find the probability the player only takes a single free-throw.

b. Find the probability the player makes two shots.

c. Find the probability the player makes the first shot and misses the second.

27. **Playing Roulette**

You play the game of roulette in Reno. Each game you always bet on "red" and the chance that you win is 18/38. If you play the game four times …

a. Find the probability you win in all games.

b. Find the probability you win in the first and third games, and lose in the second and fourth games.

c. Find the probability you win in exactly two of the four games.

28. **Is a Die Fair?**

Suppose a friend is about to roll a die. The die either is the usual "fair" type or it is a "special" type that has two sides showing 1, two sides showing 2, and two sides showing 3. You believe that the die is the fair type with probability .9. Your friend rolls the die and you observe a 1.

a. Find the probability that a 1 is rolled.

b. If you observe a 1, what is the probability your friend was rolling the fair die?

29. **How Many Fish?**

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 in the pond is equally likely to be 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.)

a. There are two stages of this experiment. At the first stage you have 1, 2, 3, or 4 fish in the pond, and at the second stage, you observe either a tagged or not-tagged fish. Draw a tree diagram to represent this experiment, and label the branches of the tree with the given probabilities.

b. Find the probability of getting a tagged fish.

c. If you find a tagged fish, find the probability there was exactly 1 fish in the pond. Also find the probabilities of exactly 2 fish, 3 fish, and 4 fish in the pond.

30. **Shopping at the Mall**

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 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.

a. If you let T = event that shopper lives downtown, S = event that shopper lives in the suburbs and D = event that shopper is a Democrat, write down the probabilities given in the above paragraph.

b. Suppose you interview a random shopper. Find the probability that the shopper is a Democrat.

c. If your shopper is a Democrat, find the probability he/she lives in the suburbs.

31. **What Bag?**

Suppose that you have two bags in your closet. The "white bag" contains four white balls and the "mixed bag" contains two white and two black balls. The closet is dark and you just grab one bag out at random and select a ball. The ball you choose can either be white or black.

a. Suppose there are 1000 hypothetical bags in your closet. By use the Bayes' box shown below, classify the 1000 bags by the type "white" and "mixed" and the ball color observed.

		Ball color observed		
		White	Black	TOTAL
Bag Type	White			
	Mixed			
TOTAL				1000

b. Using the Bayes' box, find the probability that you observe a white ball.

c. If you observe a white ball, find the probability that you were selecting from the white bag.

TOPIC P6: PROBABILITY DISTRIBUTIONS

SPOTLIGHT: THE HAT CHECK PROBLEM

Some time ago, it was common for men to wear hats when they went out for dinner. When one entered a restaurant, each man would give his hat to an attendant who would keep the hat in a room until his departure. Suppose the attendant gets confused and returns hats in some random fashion to the departing men. What is the chance that no man receives his personal hat? How many hats, on average, will be returned to the right owners?

This is a famous "matching" probability problem given different names. The "Mothers and Babies" activity in a previous topic was the same problem – there we were trying to match mothers with their babies instead of men with their hats.

To start thinking about this problem, it is helpful to start with some simple cases. Suppose only one man checks his hat at the restaurant. Then obviously this man will get his hat back. Then the probability of "no one receives the right hat" is 0, and the average number of hats returned will be equal to 1.

Let n denote the number of men who enter the restaurant. We considered the case n = 1 above. What if n = 2? If the two men are Barry and Bobby, then there are two possibilities shown below.

	Barry receives	Bobby receives	# of matching hats
1.	Barry's hat	Bobby's hat	2
2.	Bobby's hat	Barry's hat	0

These two outcomes are equally likely, so the probability of no match is ½. Half the time there will be 2 matches and half the time there will be 0 matches, and so the average number of matches will be 1.

What if we have n = 3 men that we'll call Barry, Bobby, and Jack. Then there are 3! = 6 ways of returning hats to men:

	Barry receives	Bobby receives	Jack receives	# of matching hats
1.	Barry's hat	Bobby's hat	Jack's hat	3
2.	Barry's hat	Jack's hat	Bobby's hat	1
3.	Bobby's hat	Barry's hat	Jack's hat	1
4.	Bobby's hat	Jack's hat	Barry's hat	0
5.	Jack's hat	Barry's hat	Bobby's hat	0
6.	Jack's hat	Bobby's hat	Barry's hat	1

Again these outcomes are equally likely, so the probability of no match is 2/6. One can show that the average number of matches is again 1.

What happens if you have a large number of hats checked? It turns out that the probability of no matches is given by

Prob(no matches) = 1/e,

where e is the special irrational number 2.718… . Also it is interesting that the average number of matches for any value of n is given by

Average number of matches = 1.

You will get the opportunity of exploring this famous problem by simulation in the exercises.

PREVIEW

In the previous topics, you have computed probabilities for a variety of random experiments. In many experiments, the outcome of interest is a particular number. One can summarize this type of experiment by a probability distribution that is a list of all possible number outcomes and the corresponding probabilities. In this topic, you will get experience in constructing, using, and summarizing probability distributions.

In this topic, your learning objectives are to:

- Construct probability distributions for simple random experiments.
- Use a probability distribution to find different probabilities of interest.
- Be able to summarize a probability distribution by a mean and standard deviation.
- Be able to construct a probability distribution by the use of a simulation experiment.

NCTM Standards

✓In Grades 9-12, all students should understand the concepts of sample space and probability distribution and construct sample spaces and distributions in simple cases

✓In Grades 9-12, all students should compute and interpret the expected value of random variables in simple cases.

A RANDOM VARIABLE

Suppose that Peter and Paul play a simple coin game. A coin is tossed. If the coin lands heads, then Peter receives $2 from Paul; otherwise Peter has to pay $2 to Paul. The game is played for a total of five coin flips.

After the five flips, what is Peter's net gain (in dollars)?

Well, it depends on the results of the coin flips. There are two possible outcomes of each coin flip (heads or tails) and, by applying the multiplication rule, there are $2^5 = 32$ possibilities for the five flips. We write down the 32 possible outcomes below.

HHHHH
HHHHT
HHHTH
HHHTT
HHTHH
HHTHT
HTHHH
HTHHT
HTHTH
HTHTT
HTTHH
HTTHT
THHHH
THHHT
THHTH
THHTT
THTHH
THTHT
TTHHH
TTHHT
TTHTH
TTHTT
TTTHH
TTTHT

```
HHTTH     HTTTH     THTTH     TTTTH
HHTTT     HTTTT     THTTT     TTTTT
```

For each possible outcome of the flips, say HTHHT, there will be a corresponding net gain for Peter. For this outcome, Peter won three times and lost twice, so his net gain is 3(2) – 2(2) = 2 dollars. The net gain is an example of a *random variable* – this is simply a number that is assigned to each outcome of the random experiment.

Let G denote Peter's gain in this experiment. For each of the 32 outcomes, we can assign a value of G – we do this below.

```
HHHHH G=10  HTHHH  G=6   THHHH  G=6TTHHH  G=2
HHHHT  G=6  HTHHT  G=2   THHHT  G=2TTHHT  G=-2
HHHTH  G=6  HTHTH  G=2   THHTH  G=2TTHTH  G=-2
HHHTT  G=2  HTHTT  G=-2  THHTT  G=-2    TTHTT  G=-6
HHTHH  G=6  HTTHH  G=2   THTHH  G=2TTTHH  G=-2
HHTHT  G=2  HTTHT  G=-2  THTHT  G=-2    TTTHT  G=-6
HHTTH  G=2  HTTTH  G=-2  THTTH  G=-2    TTTTH  G=-6
HHTTT  G=-2 HTTTT  G=-6  THTTT  G=-6    TTTTT  G=-10
```

We see from the table that the possible gains for Peter are –10, -6, -2, 2, 6, and 10 dollars. We are interested in the probability that Peter will get each possible gain. To do this, we put all of the possible values of the random variable in a table.

Gain G (dollars)	Number of outcomes	Probability
-10		
-6		
-2		
2		
6		
10		

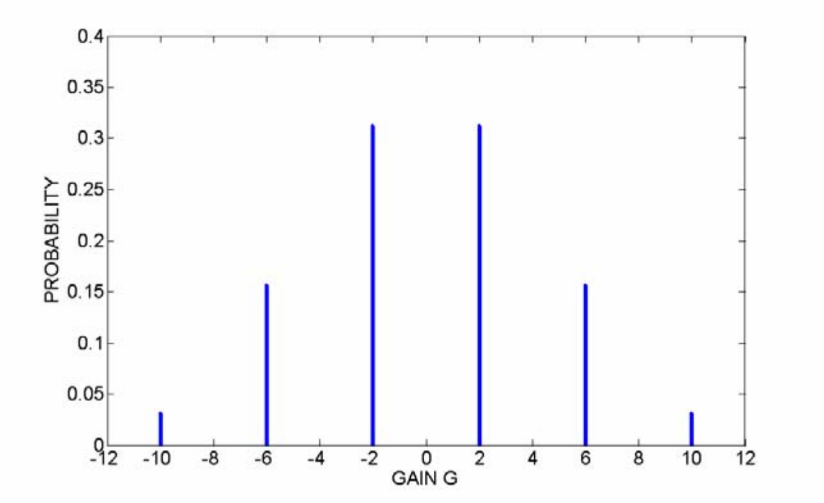
What is the probability that Peter gains $6 in this game? Looking at the table of outcomes, we see that Peter won $6 in five of the outcomes. Since there are 32 possible outcomes of the five flips, and each outcome has the same probability, we see that the probability of Peter winning $6 is 5/32.

We continue this for all of the possible values of G – we place the number of outcomes for each value and the corresponding probability in the table.

Gain G (dollars)	Number of outcomes	Probability
-10	1	1/32
-6	5	5/32
-2	10	10/32
2	10	10/32
6	5	5/32
10	1	1/32
SUM	32	1

This is an example of a *probability distribution* for G – this is simply a list of all possible values for a random variable together with the associated probabilities.

We can graphically display this probability distribution with a line graph. We place all of the values of G on the horizontal axis, mark off a probability scale on the vertical scale, and then draw vertical lines on the graph corresponding to the probability values.



This graph visually shows that it is most likely for Peter to finish with a net gain of +2 or –2 dollars. Also we notice the symmetry of the graph – the graph looks the same way on either side of 0. This symmetry about 0 indicates that this game is fair. We will shortly discuss a way of summarizing this probability distribution that confirms that this is indeed a fair game.

PRACTICE: RANDOM VARIABLES

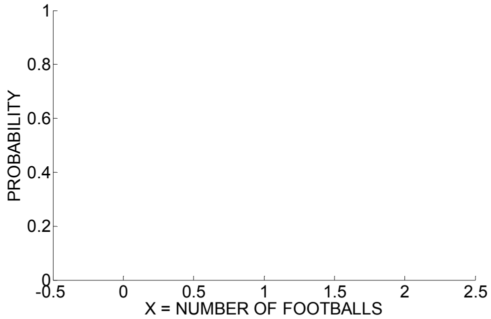
Suppose you have a toy box containing two footballs and three basketballs. You choose two toys at random from the box and record X, the number of footballs you select.

1. If the balls in the box are denoted by F1, F2, B1, B2, B3, and you don't care about the order in which the balls are selected, then two possible outcomes are {F1, F2} and {F1, B2}. Write down the ten possible outcomes.

2. Assign a value of the random variable X to each outcome in part 1.

3. Find the probability distribution for X – place the values of X and the probabilities in the below table. Draw a graph of the probabilities in the below figure.

X	P(X)



SUMMARIZING A PROBABILITY DISTRIBUTION

Once we have constructed a probability distribution – like we did above – it is convenient to use this to find probabilities.

What is the chance that Peter will win at least $5 in this game? Looking at the probability table, we see that winning "at least $5" includes the possible values

G = 6 and G = 10

We find the probability of interest by adding the probabilities of the individual values.

P(G is at least 5) = P(G = 6 or G = 10) = P(G = 6) + P(G = 10) = (5 + 1)/32.

What is the probability Peter wins money in this game? Peter wins money if the gain G is positive and this corresponds to the values G = 2, 6, 10. By adding up the probabilities of these three values, we see the probability that Peter wins money is

P(Peter wins) = P(G > 0) = P(G = 2) + P(G = 6) + P(G = 10) = (10 + 5 + 1)/32 = 1/2.

It is easy to also compute the probability Peter loses money is also 1/2. Since the probability Peter wins in the game is the same as the probability he loses, the game is clearly fair.

When we have a distribution of data, it is helpful to summarize the data with a single number, such as median or mean, to get some understanding about a typical data value. In a similar fashion, it is helpful to compute an "average" of a probability distribution – this will give us some feeling about typical or representative values of the random variable when we observe it repeated times.

A common measure of "average" is the mean or expected value of X, denoted μ or E(X). We find the mean (or expected value) by

1. Computing the product of a value of x and the corresponding probability for all values of X.
2. Summing the products.

In other words, we find the mean by the formula

$$\mu = \sum x \, P(x) \, .$$

We illustrate the computation of the mean for the Peter and Paul game in the table below. For each value of the gain G, we multiply the value by the associated probability – the products are given in the rightmost column of the table. Then we sum the products – we see that the mean of G is μ = 0.

Gain G (dollars)	Probability	G x Probability
-10	1/32	-10/32
-6	5/32	-30/32
-2	10/32	-20/32
2	10/32	20/32
6	5/32	30/32

10	1/32	10/32
SUM	1	0

How do we interpret a mean value of 0? Actually it is interesting to note that G = 0 is not a possible outcome of the game – that is, Peter cannot break even when this game is played. But if Peter and Paul play this game a large number of times, then the value μ = 0 represents (approximately) the mean winnings in all of these games.

Let's illustrate this interpretation of μ by simulating this game on a computer. I had Fathom simulate this experiment 100 times – here are Peter's winnings in these 100 games:

```
6     -2    2     -6    2     -2    -6    2     -2    -2
2     2     -6    -2    2     -2    2     -2    2     2
-2    10    -2    -2    -2    10    2     2     -2    2
-6    2     2     -2    -2    6     -10   -2    2     2
-6    -6    -2    -6    2     2     6     6     10    -6
2     6     2     2     6     -2    -2    -2    6     -2
6     6     -2    -2    2     -2    2     -6    -2    -2
2     -2    -2    -2    -6    -2    6     -2    2     -6
6     2     -6    10    6     -6    -2    -2    2     2
-2    10    2     -2    2     -6    -2    -2    -2    -2
```

If we add these winnings for these 100 games, we get that Peter's total winning was

$$\sum G = 12 \text{ dollars}$$

and so Peter's mean winning in these 100 games was

$$\bar{G} = \frac{12}{100} = .12 \text{ dollars.}$$

This value of $\bar{G}$ is approximately equal to the mean of G, μ = 0. If Peter was able to play this game for a much larger number of games, then we would see that his average winning $\bar{G}$ would be very close to μ = 0.

PRACTICE: SUMMARIZING A PROBABILITY DISTRIBUTION

Consider again the problem where you are selecting two toys from a box with two footballs and three basketballs and X is the number of footballs you select.

1. Copy the probability distribution you found earlier for X in the first two columns of the table.

X	P(X)	X ×P(X)

2. Find the probability you select at least one football from the box.

3. Compute the products of the values of X and the probabilities, X ×P(X), and place your answers in the third column of the table. Find the mean of X.

4. Suppose that you randomly select two toys from the box many days, each time recording a value for X. In this context, give an interpretation to the mean of X.

STANDARD DEVIATION OF A PROBABILITY DISTRIBUTION

Consider two dice – one we will call the “fair die” and the other one will be called the “loaded die”. The fair die is the familiar one where each possible number (1 through 6) has the same chance of being rolled. The loaded die is designed in a special way that 3’s or 4’s are relatively likely to occur, and the remaining numbers (1, 2, 5, and 6) are unlikely to occur. The table below gives the probabilities of the possible rolls for both dice.

Fair Die	
Roll	Probability
1	1/6

Loaded Die	
Roll	Probability
1	1/12

2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

2	1/12
3	1/3
4	1/3
5	1/12
6	1/12

How can we distinguish the fair and loaded dice? An obvious way is to roll each a number of times and see if we can distinguish the patterns of rolls that we get. We first rolled the fair die 20 times with the results

3, 3, 5, 6, 6, 1, 2, 1, 4, 3, 2, 5, 6, 4, 2, 5, 6, 1, 2, 3 (mean 3.5)

We then rolled the loaded die 20 times with the results

3, 2, 1, 4, 4, 1, 4, 3, 3, 3, 1, 3, 3, 5, 3, 3, 3, 6, 3, 4 (mean 3.1)

Below we have displayed back to back stemplots of the rolls from the two dice.

```
FAIR DIE   LOADED DIE
    000 | 1 | 000
   0000 | 2 | 0
   0000 | 3 | 0000000000
     00 | 4 | 0000
    000 | 5 | 0
   0000 | 6 | 0
```

What do we see? For the fair die, the rolls appear to be evenly spread out among the six possible numbers. In contrast, the rolls for the loaded die tend to concentrate on the values and 3 and 4, and the remaining numbers were less likely to occur.

Can we compute a summary number to contrast the probability distributions for the fair and loaded dice?

We have already discussed one summary number for a random variable, the mean μ. This number represents the average outcome for the random variable when one performs the experiment many times.

Let's compute the mean for the two probability distributions. For the fair die, the mean is given by

$$\mu \text{ (Fair Die)} = (1)\,(1/6) + (2)\,(1/6) + (3)\,(1/6) + (4)\,(1/6) + (5)\,(1/6) + (6)\,(1/6) = 3.5,$$

and for the loaded die the mean is given by

$$\mu \text{ (Loaded Die)} = (1)\,(1/12) + (2)\,(1/12) + (3)\,(1/3) + (4)\,(1/3) + (5)\,(1/12) + (6)\,(1/12) = 3.5.$$

The means of the two probability distributions are the same – this means that we'll tend to get the same average roll when we roll the fair die and roll the loaded die many times.

But we know from our rolling data that the two probability distributions are different. For the loaded die, we're more likely to roll 3's or 4's. In other words, for the loaded die, it is more likely to roll a number close to the mean value $\mu = 3.5$.

The standard deviation of a random variable X, denoted by the Greek letter σ, measures how close the random variable is to the mean μ. It is called a standard deviation since it represents an "average" (or standard) distance (or deviation) from the mean μ.

To find the standard deviation σ for a random variable, we

1. (Compute deviations.) For each value of the random variable X, we compute the difference (or deviation) of X from the mean value μ.
2. (Square deviations.) We then square each of the differences that we computed in step 1.
3. (Compute the average squared deviation.) We find the average squared deviation, by multiplying each squared deviation by the corresponding probability, and summing the products.

4. (Take square root.) The standard deviation σ is the square root of the average squared deviation.

We illustrate in the following table the computation of the standard deviation for the roll of the fair die and for the roll of the loaded die.

Computation of the standard deviation σ for the Fair Die.

Roll	R - μ	$(R - \mu)^2$	Probability	$(R - \mu)^2$ x Probability
1	1 – 3.5 = -2.5	$(-2.5)^2 = 6.25$	1/6	6.25 x (1/6)
2	2 – 3.5 = -1.5	$(-1.5)^2 = 2.25$	1/6	2.25 x (1/6)
3	3 – 3.5 = -0.5	$(-0.5)^2 = 0.25$	1/6	0.25 x (1/6)
4	4 – 3.5 = 0.5	$(0.5)^2 = 0.25$	1/6	0.25 x (1/6)
5	5 – 3.5 = 1.5	$(1.5)^2 = 2.25$	1/6	2.25 x (1/6)
6	6 – 3.5 = 2.5	$(2.5)^2 = 6.25$	1/6	6.25 x (1/6)
SUM				2.917
				Sqrt(2.917) = 1.71

Computation of the standard deviation σ for the Loaded Die.

Roll	R - μ	$(R - \mu)^2$	Probability	$(R - \mu)^2$ x Probability
1	1 – 3.5 = -2.5	$(-2.5)^2 = 6.25$	1/12	6.25 x (1/12)
2	2 – 3.5 = -1.5	$(-1.5)^2 = 2.25$	1/12	2.25 x (1/12)
3	3 – 3.5 = -0.5	$(-0.5)^2 = 0.25$	1/3	0.25 x (1/3)
4	4 – 3.5 = 0.5	$(0.5)^2 = 0.25$	1/3	0.25 x (1/3)
5	5 – 3.5 = 1.5	$(1.5)^2 = 2.25$	1/12	2.25 x (1/12)
6	6 – 3.5 = 2.5	$(2.5)^2 = 6.25$	1/12	6.25 x (1/12)
SUM				Sqrt(1.583) = 1.26

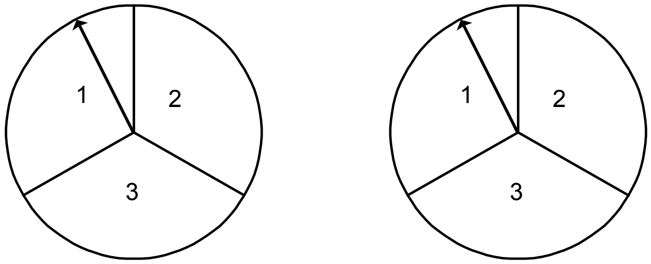
We see from our calculations that

σ (Fair Die) = 1.71, σ (Loaded Die) = 1.26

What does this mean? Since the loaded die roll has a smaller standard deviation, this means that the roll of the loaded die tends to be closer to the mean (3.5) than for the fair die. When we roll the loaded die many times, we will notice a smaller spread or variation in the rolls than when we roll the fair die many times.

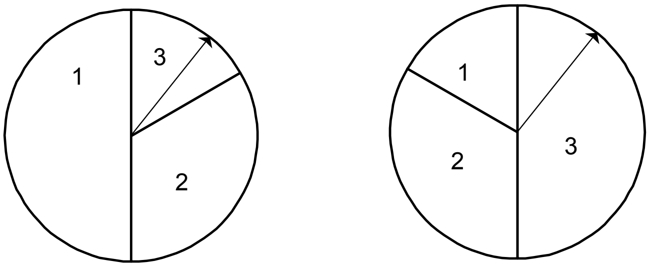
PRACTICE: THE STANDARD DEVIATION

1. Suppose you spin two spinners of the type shown below where each spinner is equally likely to land 1, 2, or 3. Find the probability distribution for the sum S of the two spins.



2. Find the mean and standard deviation of S.

3. Suppose you spin the two different spinners shown below. Find the probability distribution for the sum T of the two spins.



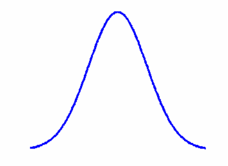
4. Graph the probability distributions for S and T. By looking at the graphs, how do the two distributions differ?

5. Compute the means and standard deviations for S and T and place these values next to the graphs you drew in part 4. Do S and T differ with respect to their average value? Do the two random variables differ with respect to their spread?

INTERPRETING THE STANDARD DEVIATION FOR A BELL-SHAPED DISTRIBUTION

Once we have computed a standard deviation σ for a random variable, how can we use this summary measure? We illustrated one use of σ in the dice example above. The probabilities for the roll of the loaded die were more concentrated about the mean than the probabilities for the roll of the fair die, and that resulted in a smaller value of σ for the roll of the loaded die.

The standard deviation has an attractive interpretation when the probability distribution of the random variable is bell-shaped. When the probability distribution has the following shape



then approximately

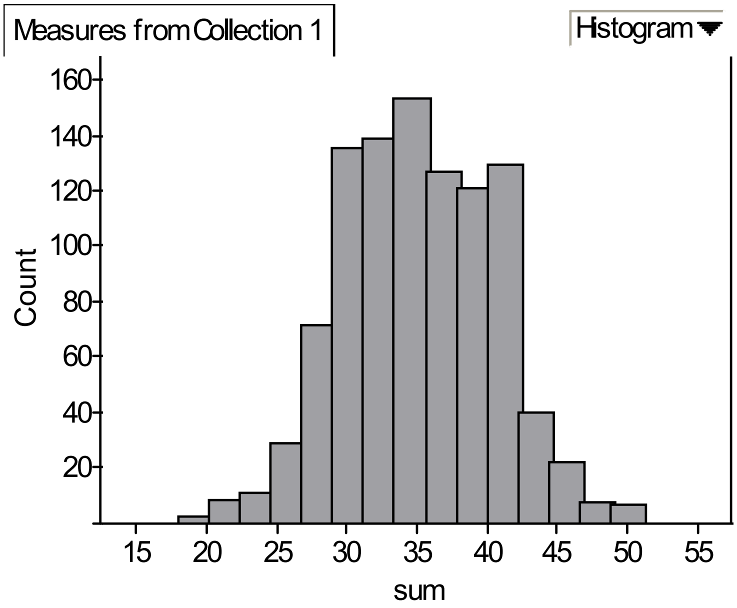
- the probability that X falls within one standard deviation of the mean is 0.68.
- the probability that X falls within two standard deviations of the mean is 0.95.

In more mathematical jargon,

$$\text{Prob}(\mu - \sigma < X < \mu + \sigma) \text{ is approximately } 0.68.$$

$$\text{Prob}(\mu - 2\,\sigma < X < \mu + 2\,\sigma) \text{ is approximately } 0.95.$$

To illustrate this interpretation of the standard deviation, suppose we roll ten fair dice and record the sum of the numbers appearing on the dice. It is easy to simulate this experiment on Fathom. A histogram of the results from 1000 trials of this experiment is shown below.



Note that the shape of this histogram is approximately bell shaped about the value 35. Since this histogram is a reflection of the probability distribution of the sum of the rolls of ten dice, this means that the shape of the probability distribution for the sum will also be bell-shaped.

For this problem, it can be shown (this will be an exercise) that the mean and standard deviation for the sum of the rolls of ten fair dice are respectively

$$\mu = 35 \text{ and } \sigma = 5.4.$$

Applying our rule, the probability that the sum falls between

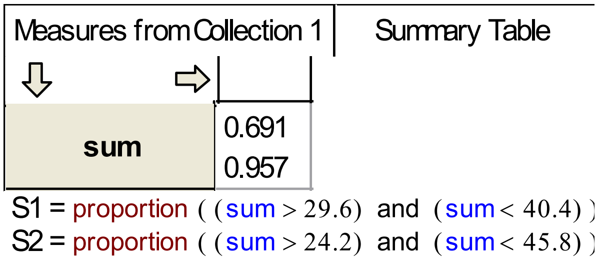
$$\mu - \sigma \text{ and } \mu + \sigma\text{, or } 35 - 5.4 = 29.6 \text{ and } 35 + 5.4 = 40.4$$

is approximately 0.68. and the probability that the sum of the rolls falls between

$$\mu - 2\ \sigma \text{ and } \mu + 2\ \sigma\text{, or } 35 - 2(5.4) = 24.2 \text{ and } 35 + 2\ (5.4) = 45.8 \text{ is}$$

approximately 0.95.

To see if these are accurate probability computations, we return to our simulation of this experiment and see how often the sum of the ten rolls fell within the above limits. In the Fathom output below, we compute the proportion of sums of ten rolls that fell between 29.6 and 40.4, and between 24.2 and 45.8.



Measures from Collection 1 | Summary Table

sum	
	0.691
	0.957

S1 = proportion ((sum > 29.6) and (sum < 40.4))
S2 = proportion ((sum > 24.2) and (sum < 45.8))

We see that the proportions of values that fall within these limits are 0.691 and 0.957, respectively. Since these proportions are close to the numbers 0.68 and 0.95, we see that this rule is pretty accurate.

PRACTICE: INTERPRETING THE STANDARD DEVIATION FOR A BELL-SHAPED DISTRIBUTION

Suppose you flip 10 coins and record the number of heads X. We will see in a future topic that X has the following probability distribution.

X	P(X)		X	P(X)
0	.001		6	.205
1	.010		7	.117
2	.044		8	.044
3	.117		9	.010
4	.205		10	.001
5	.246			

1. Construct a graph of the probability distribution and comment on its shape.

2. The mean and standard deviation of X are given by $\mu=5$ and $\sigma=1.58$. Find the probability that X falls within one standard deviation of the mean.

3. Find the probability that X falls within two standard deviations of the mean.

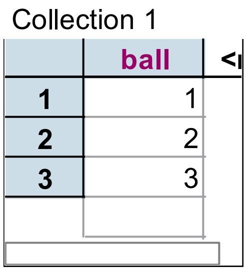
4. Since the probability distribution is bell-shaped, we expect about 68% and 95% of the probability to fall within one and two standard deviations of the mean. Are your computed probabilities close to what we expect?

TECHNOLOGY LAB – CONSTRUCTING PROBABILITY DISTRIBUTIONS BY SIMULATION

Suppose we have a bag with three types of balls – numbered 1, 2, and 3 – and there are an equal number of balls of each type. We repeatedly sample balls from the bag with replacement.

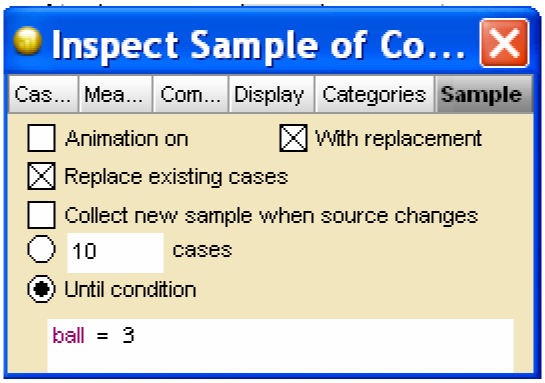
Suppose we keep sampling until we choose a ball numbered 3. Let X denote the number of balls we choose.

1. On Fathom, I first created a new Collection and named a new Attribute called "ball". I placed the numbers 1, 2, and 3 in the collection.



Collection 1

	ball	<
1	1	
2	2	
3	3	



□ Animation on ☒ With replacement

☒ Replace existing cases

□ Collect new sample when source changes

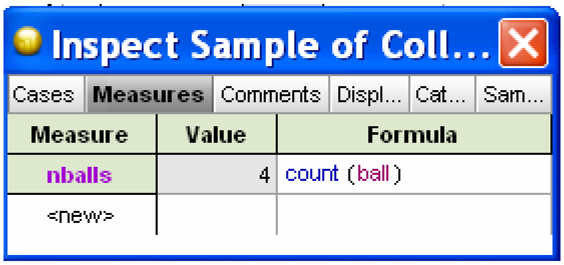
○ 10 cases

● Until condition

ball = 3

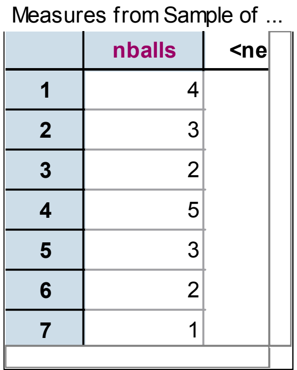
2. Next I took a Sample from the Collection. When inspecting the sample collection , I modified the sampling procedure so that I sample until ball = 3.

Also I defined a new Measure called nballs and define nballs to be count(balls). (This measure will correspond to the value of X.)



Inspect Sample of Coll...
Cases Measures Comments Displ... Cat... Sam...

Measure	Value	Formula
nballs	4	count (ball)
<new>		

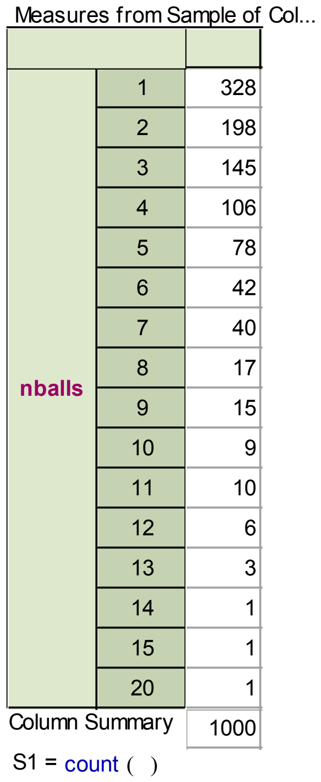


Measures from Sample of ...

	nballs	<ne
1	4	
2	3	
3	2	
4	5	
5	3	
6	2	
7	1	

3. Now I used the Collect Measures command to repeat this sampling procedure many times (specifically 1000 times), and put the values of X in the Measures from Sample collection.

4. I tabulated the different values of X (or nballs) in a Summary Table:



Measures from Sample of Col...

nballs	1	328
	2	198
	3	145
	4	106
	5	78
	6	42
	7	40
	8	17
	9	15
	10	9
	11	10
	12	6
	13	3
	14	1
	15	1
	20	1
Column Summary		1000

S1 = count ( )

5. Looking at the output,

(a) What is the most likely value of X?

(b) Find the probability that it will take you at most 4 draws to choose a ball numbered 3.

(c) Find the mean of X.

6. Consider a slight variation of the above experiment. As before, you sample from a bag with an equal number of "1", "2", and "3" balls. But you continue sampling until you get balls of all three types. Let Y denote the number of selections. Adjust the above Fathom simulation to approximate the probability distribution of Y. (You can define your new measure to be uniquevalues(ball).) From the output, find (a) the most likely value of Y, (b) the probability that it takes you at most 6 balls to choose balls of all three types, and (c) the mean of Y.

TECHNOLOGY LAB – PLAYING ROULETTE

You have decided to take on a new job. You're going to make money (you hope) playing roulette full-time in Reno. Each day next year, you will play the roulette wheel 20 times, betting $5 each game. You will do this all 365 days and keep track of your winnings (or losing) each day.

To learn about roulette and play on-line, go to the web site www.roulette.com.

1. First, you have to decide what numbers you are going to bet on. The table below gives some possible bets.

Type of bet	Description	Payoff odds (if you bet $1 and your number comes up, you win the left number plus your $1 bet; otherwise you lose $1)
Straight bet	bet on one number	35 to 1
Split bet	bet on 2 consecutive numbers	17 to 1
Trio bet	bet on 3 consecutive numbers	11 to 1
Corner bet	bet on 4 consecutive numbers	8 to 1
Five number bet	bet on 5 consecutive numbers	6 to 1
Six number bet	bet on 6 consecutive numbers	5 to 1
Dozens bet	bet on either 1-12, 13-24, or 25-36	2 to 1
High or low	bet on low numbers (1-18) or high numbers (19-36)	1 to 1

Write down the consecutive numbers you plan to always bet on and the payoff (remember you are betting $5 each time and if you win, you also keep your $5 bet)

Numbers: ________________________

Payoff (amount you win): _________________________

2. In Fathom, open a New Collection and Case Table. Call the attribute "slot" and put the numbers 0, 0, 1, 2, 3, …, 36 into this collection. This collection represents the 38 different outcomes of the roulette wheel.

3. Next, define a new attribute called “PAYOFF”. Next to each value in slot, put in the payoff. For example, if you decide to bet on numbers 5, 6, 7 (a trio bet that pays off at 11 to 1), you would put in 11 × 5+5 = 60 in the payoff column next to numbers 5, 6, 7, and 0 in the remaining rows.

4. To represent your 20 plays on a single day, take a Sample of size 20 with replacement from your collection. Write down the results of these 20 games.

Play	SLOT	PAYOFF	Play	SLOT	PAYOFF	Play	SLOT	PAYOFF
1			8			15		
2			9			16		
3			10			17		
4			11			18		
5			12			19		
6			13			20		
7			14					

5. Next, let’s define a measure that computes your total winnings for the day. Select the Sample Collection, Inspect this collection, and click on the Measures tab. Define a measure called “winnings”, double-click in the Formula box, and type the formula

sum(PAYOFF)

6. You have computed your payoff for a single day. To play for 365 days (every day next year)

- select your Sample Collection and then Collect Measures from the Analyze menu
- select your Measures Collection -- inspect it to change to 365 measures and turn off animation
- collect more measures

7. Your Measures Collection contains your winnings for each day next year.

(a) Construct a histogram of your winnings. From this histogram, write a few sentences about this distribution of winnings (what was an average winning, what were your best and worst days, etc).

(b) Construct a summary table that lists all of the possible winnings and how often each winnings happened. Fill in the table below. (You find probabilities by dividing each count by the total count.)

Winnings	Count	Probability

Winnings	Count	Probability

(c) How did you do for the year? Did you actually win money? If the answer is "no", how much money did you lose, on average, each day?

8. Repeat all of the above work using a different bet. Compare the two bets – would you prefer one bet? Why?

ACTIVITY: HOW MANY KEYS?

It is a dark and stormy night and you are trying to open the door of your apartment. On your key ring, you have 6 keys, 2 of which will open the door and 4 that won't work. Since it is dark, you randomly select keys (without replacement) until you find a key that will work.

(a) If B represents trying a bad key and G represents trying a good key, list all of the possible outcomes of this experiment. (There are five possible outcomes.)

(b) Find the probability of each outcome (use the multiplication rule and conditional probability). Compute each probability to the nearest hundredth.

(c) If X = number of keys that you need to try before opening the door, find the probability distribution for X. Put your answer in the table to the right.

X	P(X)

(d) Simulate this experiment using Fathom. (You first define a collection with 4 bad and 2 good keys, you sample until you find a key that is good, and your measure is the number of keys that you select.) Simulate this experiment 1000 times and place your Fathom probabilities in the below table – check if your Fathom probabilities are close to the values you found in part (c).

X	Fathom approx. probs

ACTIVITY: INVESTING MONEY – COMPARING SAFE AND RISKY INVESTMENTS.

Suppose you have $1000 and wish to invest your money. There are three different investments (called YELLOW, RED, WHITE) that you can make. We describe each investment in terms of the percentage return R in one year – if you start with $1000, the value of your money after one year is $1000 (1+R/100). So, for example, if your percentage return is R = +10 %, the value of $1000 after one year will be $1000 (1 + 10/100) = $1100. If your return is –20 %, the value of $1000 will be $1000 (1 – 20/100) = $800.

Each of the investments is a probability distribution for R. The mean and standard deviation of each probability distribution is also shown. Note that the average returns of YELLOW, RED, and WHITE are 6 %, 71%, and 7.5 %, respectively, and the standard deviations of the returns are quite different.

YELLOW	
R	Prob
-10 %	1/6
0	1/6
0	1/6
0	1/6
0	1/6
+10 %	1/6
Mean	0 %
Stand dev	6 %

RED	
R	Prob
- 94 %	1/6
- 80 %	1/6
0	1/6
+ 200 %	1/6
+ 200 %	1/6
+ 200 %	1/6
Mean	71 %
Stand dev	132 %

WHITE	
R	Prob
- 20 %	1/6
- 10 %	1/6
+ 5 %	1/6
+ 10 %	1/6
+ 20 %	1/6
+ 40 %	1/6
Mean	7.5 %
Stand dev	20 %

1. Verify the computation of the mean and standard deviation for one of the investments.

2. Suppose you want to follow a single investment strategy for 20 years. Which would you prefer? Explain.

3. Now we will simulate trying out all investment strategies for 20 years. Work in pairs, where one person is the dice roller and rolls the yellow, red, and white dice. (The colors of your dice might be different depending on the availability of dice colors from your instructor.) The second person records the value multipliers in the below table. When

you are finished entering in the value multipliers for all 20 years, then compute the investment values.

When we are done, we will collect the results of the investments after 20 years.

Value multipliers for Dice Simulation

OUTCOME	YELLOW	RED	WHITE
1	0.9	0.06	0.8
2	1	0.2	0.9
3	1	1	1.05
4	1	3	1.1
5	1	3	1.2
6	1.1	3	1.4

	Value Multipliers				Investment values		
Round	YELLOW	RED	WHITE		YELLOW	RED	WHITE
Start	1	1	1		1000	1000	1000
Year 1							
Year 2							
Year 3							
Year 4							
Year 5							
Year 6							
Year 7							
Year 8							

Year 9							
Year 10							
Year 11							
Year 12							
Year 13							
Year 14							
Year 15							
Year 16							
Year 17							
Year 18							
Year 19							
Year 20							

4. For each investment type (Yellow, White, and Red), graph the investment values. Describe each distribution of values, including statements about shape, average, and spread.

5. Based on your work, what is the best investment? Explain.

6. If you were providing advice to a young couple, what investment type would you recommend? Why?

WRAP-UP

In many random experiments, the outcome of interest is a number called a **random variable**. A **probability distribution** is a table that lists all possible values of the random variable together with the associated probabilities. In this topic, we constructed probability distributions for simple random experiments. A probability distribution can be summarized by a mean μ and a standard deviation σ. The mean μ is approximately equal to the sample mean of the random variable when the experiment is repeated many times. The standard deviation is informative about the spread of values of the random variable. When the probability distribution has a bell-shape, then we can use μ and σ to predict the probability of falling within one and two standard deviations from the mean. One convenient way of constructing a probability distribution is by repeated simulations of the random experiment.

EXERCISES

1. **Coin-tossing Game**

In the Peter/Paul coin-tossing game described in the text, let the random variable X be the number of times Paul is in the lead. For example, if the coin tosses are HTHHT, Paul's running winnings are $2, 0, $2, $4, $2, and the number of times he is in the lead is X = 4.

HHHHH	HTHHH	THHHH	TTHHH
HHHHT	HTHHT	THHHT	TTHHT
HHHTH	HTHTH	THHTH	TTHTH
HHHTT	HTHTT	THHTT	TTHTT
HHTHH	HTTHH	THTHH	TTTHH
HHTHT	HTTHT	THTHT	TTTHT
HHTTH	HTTTH	THTTH	TTTTH
HHTTT	HTTTT	THTTT	TTTTT

a. Find the probability distribution for X.

b. Construct a graph of the probabilities for X.

c. What is the most likely value of X?

d. Find the probability that X > 2.

2. **Sampling Without Replacement**

Suppose you choose two coins from a box with two nickels and three quarters. Let X denote the number of nickels you draw.

a. Write out all possible 10 outcomes of this experiment.

b. Find the probability distribution for X.

c. What is the most likely value of X?

d. Find the probability that X ≥ 1.

3. **Shooting Free Throws**

Suppose you watch your favorite basketball player attempt five free throw shots during a game. You know that the chance that he is successful on a single shot is 0.5, so that the possible sequences of successes (S) and misses (M) shown below are equally likely. Suppose you measure the number of runs X where a run is defined to be a streak of S's or M's. For example, in the sequence MMSSM, there are three runs (one run of two misses, one run of two successes, and one run of one miss).

SSSSS
SSSSM
SSSMS
SSSMM
SSMSS
SSMSM
SSMMS
SSMMM
SMSSS
SMSSM
SMSMS
SMSMM
SMMSS
SMMSM
SMMMS
SMMMM
MSSSS
MSSSM
MSSMS
MSSMM
MSMSS
MSMSM
MSMMS
MSMMM
MMSSS
MMSSM
MMSMS
MMSMM
MMMSS
MMMSM
MMMMS
MMMMM

a. Find the probability distribution for X.

b. Construct a graph of the probabilities for X.

c. What is the most likely number of runs in the sequence?

d. Find the probability that you have at most 2 runs in the sequence.

4. **Rolling Two Dice**

Suppose you roll two dice and you keep track of the larger of the two rolls which we denote by X. (For example, if you roll a 4 and a 5, then X = 5.)

a. Find the probability distribution for X.

b. Construct a graph of the probabilities for X.

c. What is the most likely value of X?

d. Find the probability that X is either 5 or 6.

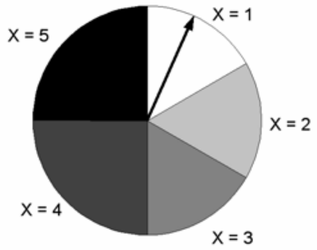
5. **Spinning a Spinner**

Let X denote the number you get when you spin the spinner shown to the right.

a. Find the probability distribution for X.

b. Find the probability that $X \geq 2$.

c. Find the mean and standard deviation of X.



6. **Rolling Four Dice**

Suppose you are asked to roll four dice and record the sum X. A lazy student thinks this is too much work. As a shortcut, he decides to roll only two dice, record the sum of the dice, and then double the result – call this random variable Y.

The probability distributions of X and Y are shown below:

X = sum of rolls of four dice

X	4	5	6	7	8	9	10	11	12	13	14
Prob	.0008	.0031	.0077	.0154	.0270	.0432	.0617	.0802	.0965	.1080	.1127

X	15	16	17	18	19	20	21	22	23	24
Prob	.1080	.0965	.0802	.0617	.0432	.0270	.0154	.0077	.0031	.0008

Y = 2 × (sum of rolls of two dice)

Y	4	6	8	10	12	14	16	18	20	22	24
Prob	.0278	.0556	.0833	.1111	.1389	.1667	.1389	.1111	.0833	.0556	.0278

a. Compute the mean and standard deviation of the probability distributions of X and Y.

b. Plot the probability distributions of X and Y on the same graph.

c. Compare and contrast the two probability distributions. How are the distributions similar? How are they different? How would you respond to the lazy student who thinks that doubling a two-dice result is equivalent to finding the sum of four fair dice?

7. **Running a Marathon Race**

Suppose three runners from college A and four runners from college B are participating in a marathon race. Suppose that all seven runners have equal abilities and so all possible orders of finish of the seven runners are equally likely. (One possible order of finish is AAABBBB where the three A runners finish first, second, and third.) Let X denote the finish position of the best runner from college A.

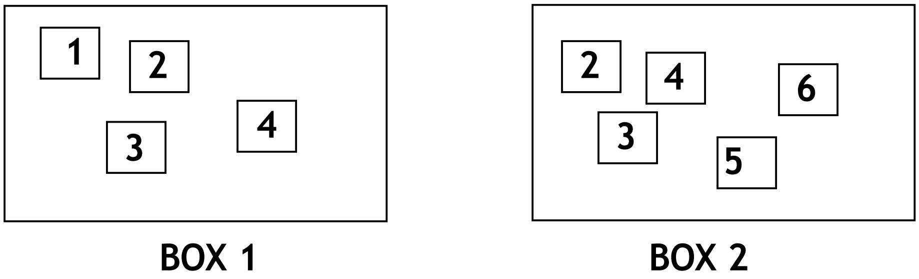
a. Find the probability distribution of X.

b. Find the probability that X is at most 2.

c. Find the average finish of the best runner from college A.

8. **Choosing a Slip from a Random Box**

Suppose you roll a die. If the die roll is 1 or 2, you choose a slip from box 1; otherwise you choose a slip from box 2. Let Y denote the number on the slip.

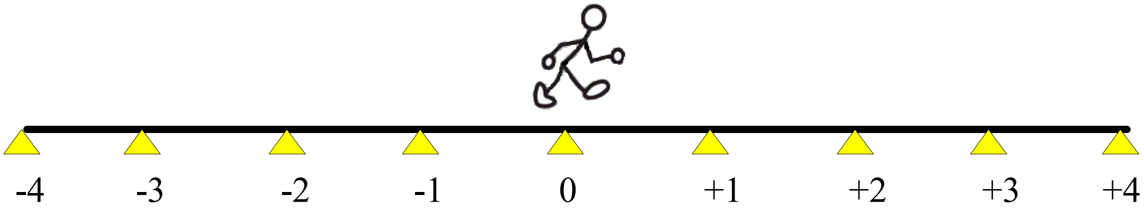


a. Find the probability distribution for Y.

b. Find the probability that Y is between 2 to 4.

9. **A Random Walk**

Suppose that a person starts at location 0 on the number line and each minute he is equally likely to take a step to the left and to the right. Let Y denote the person's location after four steps.



a. Find the probability distribution for Y.

b. Find the probability that he is at least two steps away from his start after four steps.

c. Suppose there is some gravitational pull towards the 0 (home) location. Then if he is currently at a negative location, the probability he will take a positive step is .7, and likewise if he is at a positive location, the probability he takes a negative step is .7. (If he is at point 0, he is equally likely to take a negative or positive step.) Find the probability distribution of Y.

d. Compare the two probability distributions in (a) and (c) using the mean and standard deviation.

10. **Selecting a Prize from a Bag**

Suppose you select a prize (with replacement) from a bag that contains three prizes – one worth \$1, one worth \$5, and one worth \$10. You have three opportunities to select a prize and you get to keep the largest prize of the three you select. Let X denote the value of the prize you keep.

a. Find the probability distribution of X.

b. Find the probability you win more than $1.

c. Find your expected winning.

11. **Playing Roulette**

Suppose you place a single $5 bet on three numbers (the Trio Bet) in roulette that has a payoff odds of 11 to 1. Let X denote your payoff. Recall that if you win you receive 11 times your betting amount plus your $5 bet; if you lose, your payoff is nothing.

a. Find the probability distribution for X.

b. Find the mean of X. On average, how much money do you lose in a single $5 bet?

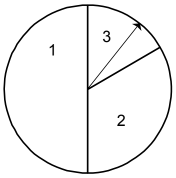
c. Consider placing $5 instead on a Five Number Bet that pays at 6 to 1. Find the probability distribution for the payoff Y for this bet. Compute the mean of Y. How does this average payoff compare with the average payoff for the Trio Bet?

d. Find the standard deviation of the payoffs for X and Y. Which bet has the larger standard deviation? Interpret what it means to have a large standard deviation.

12. **Sum of Independent Random Variables**

Suppose you have *k* random variables $X_1, \ldots, X_k$. Each random variable has a mean μ and a standard deviation σ. Suppose the random variables are independent – this means that the probability that one variable, say X_1 takes a value will not be affected by the values of the other random variables. In this case, it can be shown that the mean and standard deviation of the sum $S = X_1 + \cdots + X_k$ will have mean E(S)=$k\mu$ and standard deviation SD(S)=$\sqrt{k}\sigma$.

a. It has been shown that if X denotes the roll of a single die, then the mean and standard deviation of X are given by μ=3.5 and σ=1.71. Suppose you roll 10 dice and the outcomes of these dice are represented by $X_1, \ldots, X_{10}$ .Using the above result, find the mean and standard deviation of the sum of these 10 rolls.



b. Suppose you spin the spinner pictured here five times and record the sum of the five spins S. Find the mean and standard deviation of S. (HINT: First you need to find the mean and standard deviation of X_1, a single spin of the spinner. Then you can apply the above result.)

13. **Selecting a Coin from a Box**

Suppose you select a coin from a box containing 3 nickels, 2 dimes and one quarter. Let X represent the value of the coin.

a. Find the probability distribution of X.

b. Find the mean and standard deviation of X.

c. Suppose that your instructor will give you twice the value of the coin that you select, so your profit is Y = 2 X. Make intelligent guesses at the mean and standard deviation of Y.

d. Check your guesses by actually computing the mean and standard deviation of Y.

e. This is an illustration of a general result. If X has mean μ and standard deviation σ and Y = c X where c is a positive constant, then the mean of Y is equal to _______ and the standard deviation of Y is equal to _________.

14. **How Many Tries to Open the Door?**

You have a ring with four keys, one of which will open your door. Suppose you try the keys in a random order until you open the door. Let X denote the number of wrong keys you try before you find the right one. It can be shown that X has the following distribution.

X	P(X)
0	1/4
1	1/4
2	1/4
3	1/4

a. Find the mean and standard deviation of X.

b. Suppose you record instead Y, the total number of keys you try. Note that Y = X + 1. Find the probability distribution for Y and the mean and standard deviation.

c. This is an illustration of a general result. If X has mean μ and standard deviation σ and Y = X + c for some constant c, then the mean of Y is equal to _______ and the standard deviation of Y is equal to _________.

15. **The Hat Check Problem**

Consider the hat check problem described in the Spotlight.

a. Consider the special case where n = 4 men are checking their hats. If the names of the four men are represented by the initials A, B, C, D, then you can represent the hats given to these four men by the arrangements ABCD, ABDC, and so on.

a. Write down the 24 possible arrangements and find the probability distribution for X = the number of matches.

b. Find the probability of no matches.

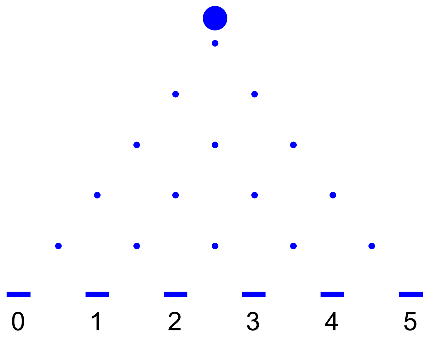
c. Find the expected number of matches.

d. Suppose instead that n = 10 men are checking their hats. It would be too tedious to write down all 10! = 3,628,800 possible arrangements of hats, but it is straightforward to design a simulation experiment for this problem. Simulate this experiment (using a computer program or the calculator) 1000 times. Approximate the probability of no matches and the expected number of matches. Compare your answers with the "large sample" answers given in the Spotlight.

TOPIC P7: COIN-TOSSING DISTRIBUTIONS

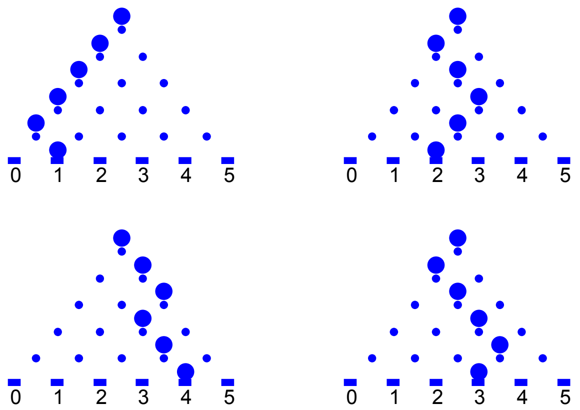
SPOTLIGHT: A GALTON BOARD

A Galton board is a physical device for simulating a special type of random experiment that we describe in this chapter. It was named after the famous scientist Sir Francis Galton who lived from 1822 to 1911. Galton is noted for a wide range of achievements in the areas of meteorology, genetics, psychology, and statistics. The Galton board consists of a set of pegs laid out in the configuration shown in the below figure – one peg is in the top row, two pegs are in the second row, three pegs in the third row, and so on. A ball is placed above the top peg. When the ball is dropped and hits a peg, it is equally likely to fall left or right. We are interested in the location of the ball after striking five pegs – as shown in the diagram, the ball can land in locations 0, 1, 2, 3, 4, or 5.



The below figure shows the path of four balls that fall through a Galton board. The chances of falling in the locations follow a special probability distribution that has a strong connection with a simple coin-tossing experiment.

Topic P7: Coin-Tossing Distributions



PREVIEW

Consider the following random experiment. You take a quarter and flip it ten times, recording the number of heads you get. There are four special characteristics of this simple coin-tossing experiment.

1. You are doing the same thing (flip the coin) ten times. We will call an individual coin flip a *trial*, and so our experiment consists of ten identical trials.
2. On each trial, there are two possible outcomes, heads or tails.
3. In addition, the probability of flipping heads *on any trial* is ½.
4. The results of different trials are independent. This means that the probability of heads, say, on the fourth flip, does not depend on what happened on the first three flips.

We are interested in the number of heads we get – we will refer to this number as X. In particular, we are interested in the probability of getting five heads, or Prob(X = 5).

In this topic, we will see that this binomial probability model applies to many different random phenomena in the real world. We discuss probability computations for the binomial and closely related negative binomial models and illustrate the usefulness of these models in representing the variation in real-life experiments.

In this topic, the learning objectives are to:

- Understand and recognize binomial and negative binomial experiments.

- Understand how to compute and apply binomial and negative binomial probabilities to real-life problems.
- Understand how real coin tossing can be differentiated from fake coin tossing where students are imagining flips of heads and tails.

NCTM Standards

✓In Grades 9-12, all students should understand the concepts of sample space and probability distribution and construct sample spaces and distributions in simple cases

✓In Grades 9-12, all students should compute and interpret the expected value of random variables in simple cases.

PROBABILITIES OF A COIN-TOSSING EXPERIMENT

Let's return to our experiment where a quarter is flipped ten times, recording X, the number of heads. We are interested in the probability of flipping exactly five heads, that is, Prob(X = 5). To compute this probability, we first have to think of possible outcomes in this experiment. Suppose we record if each flip is heads (H) or tails (T). Then one possible outcome when we make ten flips is

Trial	1	2	3	4	5	6	7	8	9	10
Result	H	H	T	T	H	T	T	H	H	T

Another possible outcome is TTHHTHTHHH. The sample space consists of all possible ordered listings of ten letters, where each letter is either an H or a T.

Next, consider computing the probability of a single outcome of ten flips such as the HHTTHHTHHT sequence shown above. We can write the probability of this outcome as

Prob("H on toss 1" AND "H on toss 2" AND "T on toss 3" AND … AND "T on toss 10")

Using the fact that outcomes on different trials are independent, this probability can be written as the product

Prob(H on toss 1) × Prob(H on toss 2) × Prob(T on toss 3) × … × Prob(T on toss 10).

Since the probability of heads (or tails) on a given trial is ½, we have

$$\text{Prob(HHTTHHTTHT)} = \left(\frac{1}{2}\right)\times\left(\frac{1}{2}\right)\times\cdots\times\left(\frac{1}{2}\right)=\left(\frac{1}{2}\right)^{10}.$$

Actually, the probability of any outcome (sequence of ten letters with H's or T's) in this experiment is equal to $\left(\frac{1}{2}\right)^{10}$.

Let's return to our original question – what is the probability that we get exactly five heads? If we think of the individual outcomes of the ten trials, then we'll see that there are many ways to get five heads. For example, we could observe

HHHHHTTTTT or HHHHTTTTTH or HHHTTTTTHH

In each of the three outcomes, note that the number of heads is five.

How many outcomes (like the ones shown above) will result in exactly five heads? As before, we label the outcomes of the individual flips by the trial number:

Trial	1	2	3	4	5	6	7	8	9	10
Outcome	_	_	_	_	_	_	_	_	_	_

If we observe five heads, then we wish to place five H's in the ten slots above. In the outcome HHHHHTTTTT, the heads occur in trials 1, 2, 3, 4, 5, and in the outcome HHHTTTTTHH, the heads occur in trials 1, 2, 3, 9, and 10. If we observe exactly 5 heads, then we must choose five numbers from the possible trial numbers 1, 2, …, 10 to place the five H's. There are ${}_{10}C_5$ ways of choosing these trial numbers. (The order in which we choose the trial numbers is not important.) Since there are ${}_{10}C_5$ ways of getting exactly five heads, and each outcome has probability $\left(\frac{1}{2}\right)^{10}$, we see that

$$\text{Prob}(X = 5) = {}_{10}C_5\left(\frac{1}{2}\right)^{10} = 0.246.$$

From a basic property of probabilities, we see that the Prob(five heads *are not* tossed) = 1- 0.246 = 0.754. It is interesting to note that although we *expect* to get five heads when we flip a coin ten times, it is actually much more likely *not* to flip five heads than to flip five heads.

PRACTICE: COIN-TOSSING EXPERIMENTS

Suppose you flip a coin five times.

1. Write down all possible sequences of five coin flips.

2. Next to each outcome of five flips, write down the value of X = the number of heads observed.

3. Using the work in 1 and 2, find the probability distribution for X. Put your probabilities in the below table.

X	0	1	2	3	4	5
P(X)						

4. By using counting arguments, how many sequences of five flips will contain exactly 2 heads? Check that this number agrees with your computation in part 3.

5. If you flip a coin 20 times, how many outcomes will result in exactly 10 heads? What is the probability of 10 heads?

BINOMIAL EXPERIMENTS

Although the coin tossing experiment described above seems pretty artificial, many random experiments share the same basic properties as coin tossing. Consider the following *binomial experiment*:

1. We are repeating the same basic task or trial many times – let the number of trials be denoted by n.

2. On each trial, there are two possible outcomes, which we will call "success" or "failure". (We could call the two outcomes "black" and "white", or "0" or "1", but they are usually called success and failure.)

3. The probability of a success, denoted by p, is the same for each trial.

4. The results of outcomes from different trials are independent.

Here are some examples of binomial experiments.

A sample survey. Suppose the Gallup organization is interested in estimating the proportion of adults in the United States who use the popular auction web site EBay. They take a random sample of 100 adults and 45 say that they use EBay. In this story, we see that

1. The results of this survey can be considered to be a sequence of 100 trials where one trial is asking a particular adult if he or she uses EBay.
2. There are two possible responses to the survey question – either the adult says "yes" (he or she uses EBay) or "no" (he or she doesn't use EBay).
3. Suppose the proportion of all adults that use EBay is p. Then the probability that the adult says "yes" will be p.
4. If the sampling is done randomly, then the chance that one person says "yes" will not depend on the answers of the people who were previously asked. This means that the responses of different adults to the question can be regarded as independent events.

A baseball hitter's performance during a game. Suppose you are going to a baseball game and your favorite player comes to bat five times during the game. This particular player is a pretty good hitter and his batting average is about .300. You are interested in the number of hits he will get in the game. This can also be considered a binomial experiment:

1. The player will come to bat five times – these five at-bats can be considered the five trials of the experiment (n = 5).
2. At each at-bat, there are two outcomes of interest – either the player gets a hit or he doesn't get a hit.
3. Since the player's batting average is .300, the probability that he will get a hit in a single at-bat is p = .300.
4. It is reasonable to assume that the results of the different at-bats are independent. That means that the chance that the player will get a hit in his fifth at-bat will be unrelated to his performance in the first four at-bats. (This is a debatable assumption, especially if you believe that a player can have a hot-hand.)

Sampling without replacement. Suppose a committee of four will be chosen at random from a group of five women and five men. You are interested in the number of women that will be in the committee. Is this a binomial experiment?

1. If we think of selecting this committee one person at a time, then we can think this experiment as four trials (corresponding to selecting the four people).
2. On each trial, there are two possible outcomes – either we select a woman or a man.

At this point, things are looking good – this may be a binomial experiment. But …

3. Is the probability of choosing a woman the same for each trial? For the first pick, the chance of picking a woman is 5/10. But once this first person has been chosen, the probability of choosing a woman is not 5/10 – it will be either 4/9 or 5/9 depending on the outcome of the first trial. So the probability of a "success" is not the same for all trials, so this violates the third property of a binomial experiment.
4. Likewise, in this experiment, the outcomes of the trials are not independent. The probability of choosing a woman on the fourth trial is dependent on who was selected in the first three trials, so again the binomial assumption is violated.

PRACTICE: BINOMIAL EXPERIMENTS

Are each of the following binomial experiments? If so, indicate what is a "success" and give values of n and p.

1. From weather records, you know that 60% of the days in your town will be sunny. You record the number of sunny days for ten randomly selected days in the year.

2. You keep flipping a coin until you observe two heads.

3. You have a box of 10 mittens of which six are red. You select four mittens from the box and count the number of red.

4. A company knows from experience that 10% of the products they sell will need some repair during the 90-day warranty period. They survey 12 consumers who have purchased this product and record the number who need repair within the warranty period.

5. A car dealer passes out of consumer satisfaction survey. Twenty people are surveyed of which 10 are satisfied, 5 are not satisfied, and 5 have no opinion.

ACTIVITY: COIN FLIPPING: IS IT REAL OR FAKE?

1. (Fake coin tossing.) Pretend to flip a coin 200 times – put your results (H or T for each toss) in the boxes below.

Pretend Coin Flips																			

X	Y	Z

2. (Real coin tossing) Now flip a quarter 200 times – put your results (H or T) in the boxes below.

Real Coin Flips																			

X	Y	Z

3. For a given sequence of coin tosses, we define a *run* as a consecutive sequence of heads or tails. So for example, if we observe the sequence

TTHHTHHHTTTTTHHHTTT

we observe a run of two tails, a run of two heads, a run of one tail, a run of three heads, and so on. Here the length of the longest run is 5, since the longest run is TTTTT and the length of this run is 5. We define the *number of switches* as the number of changes from H to T, or from T to H. I count a total number of six switches in the above sequence.

For each row of 20 tosses in the above two tables, compute X = the number of heads, Y = the length of the longest run of heads or tails, Z = the number of switches.

4. Collect the values of X, Y, and Z from all students in the class for the fake coin flips and the real coin flips. Place the data in the boxes below

Fake coins – values of X	Fake coins – values of Y	Fake coins – values of Z

(number of heads)	(longest run)	(number of switches)
Real coins – values of X (number of heads)	Real coins – values of Y (longest run)	Real coins – values of Z (number of switches)

5. Compare the number of heads for the fake coin flips and the real coin flips by constructing parallel dotplots. By looking at the two histograms and calculating suitable summary statistics, explain how the numbers of heads for the real coins look different from the number of heads for the fake coins.

6. Do the same comparison using the longest run variable. Which dataset tends to have "long" runs – the fake coins or the real coins?

7. Repeat the comparison using the number of switches variable. Do you notice any differences between the histogram for the number of switches for the real coins and the number of switches for the fake coins?

TECHNOLOGY ACTIVITY: SIMULATED COIN FLIPPING

This activity simulates coin flipping on Fathom. We'll consider an imaginary coin where the chance of heads on a single flip is p, which could be different from ½. We'll toss this coin 20 times – each time we will keep track of the number of heads and the length of the longest run of heads.

1. Open up the Fathom document "coin_tossing.ftm". You'll see a

- slider where you can change p, the probability of getting heads on a single flip

- the results of 20 random coin flips

2. Select the Collection and simulate 20 flips by typing Apple (or Control) – Y. For each simulation, record the number of heads, and the length of the longest run.

Simulation	Number of heads	Length of longest run
1		
2		
3		
4		
5		

3. This Fathom program has been set up to repeat this experiment (of flipping 20 coins) 1000 times – for each experiment, we record the number of heads and the length of the longest run.

To do this, select the Measures from Collection and type Apple-Y. This collection contains the number of heads and length of longest run for these 1000 experiments.

4. First look at the Number of Heads for the 1000 experiments.

(a) Construct a histogram of the number of heads.

(b) Construct a count or frequency table of the number of heads.

(c) What is the most likely number of heads you will get?

(d) What is the probability you will flip exactly 10 heads?

(e) What is the probability you will flip 15 or more heads?

5. Now look at the Length of the Longest Run for the 1000 experiments.

(a) Construct a histogram and count table.

(b) What is the most likely length of the longest run?

(c) What is the probability the longest run is 6 or more?

6. Above, we assumed that p = .5 (we were flipping a fair coin). Now using the slider, change p to .3. Redo the simulation of 1000 experiments.

Answer the same questions as in 4 and 5.

ACTIVITY: IS A PROFESSIONAL ATHLETE STREAKY?

Sports fans are often interested in streaky performances of athletes during a game. For example, the following table gives the results of 40 shots taken by Kobe Bryant (basketball player who plays for the Lakers) during a 2002 professional basketball game.

ALL OF KOBE BRYANT SHOTS for Lakers 96, Warriors 89
11/15/2002 STAPLES Center, Los Angeles, CA

1st Period
(8:06) Bryant Jump Shot: MADE
(7:33) Bryant Jump Shot: MADE
(6:46) Bryant Jump Shot: MADE
(5:50) Bryant Jump Shot: MADE
(5:16) Bryant Jump Shot: MISSED
(4:55) Bryant Dunk Shot: MADE
(4:21) Bryant Turnaround Jump: MISSED
(3:15) Bryant Jump Shot: MISSED
(3:05) Bryant Jump Shot: MADE
(0:39) Bryant Jump Shot: MISSED
(0:01) Bryant Jump Shot: MISSED

2nd Period
(10:24) Bryant Jump Shot: MADE
(9:41) Bryant Jump Shot: MISSED
(5:20) Bryant Jump Shot: MADE
(4:38) Bryant Jump Shot: MADE
(4:03) Bryant Fade Away: MISSED
(1:09) Bryant Driving Layup: MADE
(0:34) Bryant Jump Shot: MISSED
(0:02) Bryant Turnaround Jump: MISSED

3rd Period
(11:15) Bryant Turnaround Jump: MADE
(10:31) Bryant Jump Shot: MISSED
(8:31) Bryant Layup Shot: MISSED
(7:04) Bryant Jump Shot: MISSED
(6:33) Bryant Jump Shot: MISSED
(4:23) Bryant Driving Finger Roll: MADE
(2:06) Bryant Dunk Shot: MADE
(0:57) Bryant Jump Shot: MISSED

4th Period
(8:19) Bryant Turnaround Jump: MADE
(6:54) Bryant Slam Dunk Shot: MADE
(6:24) Bryant Jump Shot: MADE
(5:48) Bryant Jump Shot: MISSED
(5:10) Bryant Jump Shot: MISSED
(4:12) Bryant Jump Shot: MISSED
(1:57) Bryant Jump Shot: MADE
(0:47) Bryant Fade Away: MISSED
(0:18) Bryant Jump Shot: MISSED

1st Overtime (4)
(4:26) Bryant Turnaround Jump: MISSED
(2:08) Bryant Jump Shot: MISSED
(1:31) Bryant Turnaround Jump: MADE
(0:34) Bryant Turnaround Jump: MISSED

1. For Kobe's data, compute the length of the longest run of makes or misses.

2. Do you think this value is unusually small or large? Why?

3. One way of deciding if Kobe's longest run of makes or misses is unusual is to compare this value with the longest run of heads or tails in 40 flips of a fair coin. (We use a fair coin since Kobe's probability of making a particular shot is approximately .5.) Below we have simulated 20 sequences of forty coin flips. For each sequence, compute Y = the longest run of heads or tails and record this value on the right.

Y = longest run

```
seq 1   HTHTHHHTTTTTHHTHTHTTHHHTTTHTHTTHHTTHTTTH
seq 2   TTHTHHTTTTTTTHTHHTHTTHTHHTHTTHTTTTHHHTHT
seq 3   TTHHTHHTHTHHTHTHTTTHTTTHHHHTHHHHHTHTHTTH
seq 4   THHTTHHHHTTTTHHTTTTHHTTTTHHTHTHHHHHTHTHH
seq 5   HHTTTTHHHHHHTHTHHHTHTHHHTTHHTTHHHHTTHHTT
seq 6   THTHHHTTHHTTTTTTHHHHHHTHHTHTHTTHTHTHHHHH
seq 7   HHHHHTHHHHTTTHHTHTHTHHTHTTHHHHTTHTHTHHTT
seq 8   HTHHTTTHHTTTHTTTHTHHHHTTHTHTHTTHHTTHTTHH
seq 9   TTHHHHHHHTTHHTTTTTHHTHHTHHHTTHHTHTTHHHHT
seq 10  HTHTTTTHTHTTHTHTHTHTTHHHTHTTTHHHHHHTTTTT
seq 11  HTTTTTHTHHTTTHTHHHHHHHHHTHHTTTHTHTTTHHTH
seq 12  TTHHHTHTHHHTHHHTHTHHHTHHTTTHHHHHHTTTTTHT
seq 13  THHTTHTHTHHTHHHTHHHTTHTTHTTTHTTHTTHHHTHH
seq 14  TTTTHHHHTTHTTTHTHTHTHTTTHTTHTTHTTHTTHTTH
seq 15  THHHTHTTTTHTTTTTTHHHHHHTHTTHTHHHHHTTTHTH
seq 16  THHHTTTHHTHHHTHTHTTHHHHHHHTTHTHTHHHHHTTH
seq 17  HHTHHTTTTHHHHTTHTTTTHHHHHTTTHHTHHTHHHHHT
seq 18  THHTTTTTTHHHTTHTTHHHTTTHHHTHTTTHTTHTHHTH
seq 19  THHTTTHTTHTTHHTHHTTTTHHHHTTHTTTTHHHTTHHT
seq 20  HTTTHHTHTTTTTHTHHHTHTHTTTHTHHTTTHTTHTTTH
```

4. Construct a dotplot of the 20 values of the longest run Y. Indicate the length of Kobe's longest run by a vertical line placed on the dotplot. Based on this graph, would you say that Kobe's longest run is unusual relative to the distribution of the longest run for 40 coin flips? Explain.

BINOMIAL COMPUTATIONS

A binomial experiment is defined by two numbers

n = the number of trials, and p = probability of a "success" on a single trial.

If we recognize an experiment as being binomial, then all we need to know is n and p to determine probabilities for the number of successes X.

Using the same argument as we made in the coin-tossing example, one can show that the probability of k successes in a binomial experiment is given by

$$P(X = k) = {}_nC_k\, p^k (1-p)^{n-k}, \quad k = 0,\ldots,n.$$

Let's illustrate using this formula for a few examples.

Baseball example (revisited). Remember our baseball player with a true batting average of .300 is coming to bat five times during a game. What is the probability that he gets exactly two hits?

We showed earlier that this was a binomial experiment. Since the player has five opportunities, the number of trials is n = 5. If we regard a success as getting a hit, the probability of success on a single trial is p = 0.3. The random variable X is the number of hits of the player during this game.

Using the formula, the probability of exactly two hits is

$$P(X = 2) = {}_5C_2 (0.3)^2 (1-0.3)^{5-2} = 0.3087$$

What is the probability that the player gets at least one hit? To do this problem, we first construct the collection of binomial probabilities for n = 5 trials and probability of success p = 0.3. The table below shows all possible values of X (0, 1, 2, 3, 4, 5) and the associated probability that can be found using the binomial formula.

X	P(X)
0	0.168
1	0.360
2	0.309
3	0.132
4	0.029
5	0.002

We are interested in the probability that the player gets at least one hit or Prob(X >= 1). "At least one hit" means that X can be 1, 2, 3, 4, or 5. To find this we simply sum the probabilities of X between 1 and 5:

Prob(X>=1) = P(X = 1, 2, 3, 4, 5) = 0.360 + 0.309 + 0.132 + 0.029 + 0.002 = 0.832.

There is a simpler way of doing this computation using the complement property of probabilities. We note that if the player doesn't get at least one hit, then he was hitless in the game (that is, X = 0). Using the complement property

Prob(X>=1) = 1 – Prob(X = 0) = 1 – 0.168 = 0.832.

PRACTICE: BINOMIAL COMPUTATIONS

Suppose a student takes a six-question true/false test. He or she guesses at each question.

1. Explain why this is a binomial experiment and give values of n and p.
2. Write down the formula expression for the probability that the student gets exactly three correct.
3. The probability distribution for X = the number correct is shown below. Using this table, find the probability the student gets at least 3 correct.

X	P(X)
0	0.016
1	0.094
2	0.234
3	0.312
4	0.234
5	0.094
6	0.016

4. Find the probability the student gets fewer than 2 correct.
5. Suppose the student gets all of the questions correct. Is it reasonable to assume that the student is really guessing at all the questions?

MEAN AND STANDARD DEVIATION OF A BINOMIAL

There are simple formula for the mean and variance for a binomial random variable. First let X_1 denote the result of the first binomial trial where

$$X_1 = \begin{cases} 1, & \textit{if we observe a success} \\ 0, & \textit{if we observe a failure} \end{cases}$$

In the exercises, you will be asked to show that the mean and variance of X_1 are given by

$$E(X_1) = p, \quad Var(X_1) = p(1-p).$$

If X_1, …, X_n represent the results of the n binomial trials, then the binomial random variable X can be written as

$$X = X_1 + \cdots + X_n$$

Using this representation, the mean and variance of X are given by

$$E(X) = E(X_1) + \cdots + E(X_n), \quad Var(X) = Var(X_1) + \cdots + Var(X_n).$$

(the result about the variance is a consequence of the fact that the results of different trials of a binomial experiment are independent). Using this result and the previous result on the mean and variance of an individual trial outcome, we obtain

$$E(X) = p + \cdots + p = np$$

$$Var(X) = p(1-p) + \cdots + p(1-p) = np(1-p).$$

To illustrate these formulas, recall the first example where X denoted the number of heads when a fair coin is flipped 10 times. Here the number of trials and probability of success are given by n = 10 and p = .5. The expected number of heads would be

E(X) = 10 (.5) = 5

and the variance of the number of heads would be

$$Var(X) = 10 (.5) (1- .5) = 2.5.$$

PRACTICE: MEAN AND STANDARD DEVIATION OF A BINOMIAL

1. Suppose you have a box with 15 black beads and 5 white beads. You select six beads from the box with replacement and count the number of white beads. Explain why this is a binomial experiment and give values of n and p.
2. If X is the number of white beads you select, find the mean and standard deviation of X.
3. You should notice that E(X) is not a whole number. Is it reasonable to say that E(X) is the most likely value of X? If not, give an alternative interpretation to E(X).
4. Suppose you flip a fair coin 10 times and count the number of heads X. In a second experiment, you flip the coin 20 times and count the number of heads Y. In which experiment, do you expect to get the larger number of heads? In which experiment will you see the greater spread of values? Explain.

NEGATIVE BINOMIAL EXPERIMENTS

The 2004 baseball season was exciting since particular players had the opportunity to break single-season records. Let’s focus on Ichiro Susuki of the Seattle Mariners who had the opportunity to break the season record for the most hits that was set by George Sisler in 1920. Sisler’s record was 257 hits and Susuki had 255 hits before the Mariners’ game on September 30. Was it likely that Susuki would tie Sisler’s record during this particular game?

We can approximate this process as a coin-tossing experiment. When Susuki comes to bat, there are two relevant outcomes: either he will get a hit, or he will get an out. (We are ignoring other batting plays such as a walk or sacrifice bunt that don’t result in a hit or an out.) Assume the probability that he gets a hit on a single at-bat is p = .372

(his 2004 batting average) and we can assume (for simplicity) that the outcomes on different at-bats are independent.

Susuki needs two more hits to tie the record. How many at-bats will it take him to get two hits?

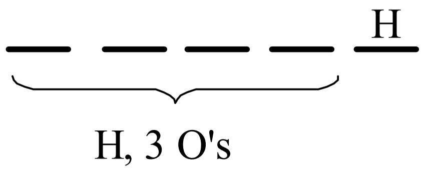
This is not a binomial experiment since the number of trials is not fixed. Instead the number of successes (hits) is fixed in advanced and the number of trials to achieve this is random. Consider

Y = number of at-bats to get two hits.

We are interested in probabilities about the number of bats Y.

It should be obvious that Y has be at least 2 (he needs at least 2 at-bats to get 2 hits), but Y could be 3, 4, 5, etc. Let's find the probability that Y = 5.

First we know that the 2nd hit must have occurred in the fifth trial (since Y=5). Also we know that there must have been one hit and three outs in the first four trials – there are ${}_4C_1$ ways of arranging the H's and the O's in these trials.



Also the probability of each possible outcome is $p^2(1-p)^3$, where p is the probability of a hit. So the probability that it takes 5 trials to observe 2 hits is

$$P(Y=5)={}_4C_1\, p^2(1-p)^3 .$$

Since p = .372 in this case, we get

$$P(Y=5)={}_4C_1\, .372^2(1-.372)^3 = .1371$$

A general *negative binomial* experiment can be described as follows:

- We have a sequence of independent trials where each trial can be a success (S) or a failure.
- The probability of a success on a single trial is p.
- We continue the experiment until we observe r successes, and Y = number of trials we observe.

The probability that it takes us y trials to observe r successes is

$$P(Y = y) = {}_{(y-1)}C_{(r-1)}\, p^{r}(1-p)^{y-r},\ y = r, r+1, r+1, \ldots$$

Let's use this formula in our baseball example where r = 2 and p = .372. The table below gives the probabilities for the number of at-bats y = 2, 3, …, 9.

Y	P(y)
2	.1384
3	.1738
4	.1637
5	.1371
6	.1076
7	.0811
8	.0594
9	.0426

Note that it is most likely that Ichiro will only need three at-bats to get his two additional hits, but the probability of three at-bats is only 17%. Actually each of the values 2, 3, 4, 5, and 6 have probabilities exceeding 10%. There is a significant probability that Ichiro will take a large number of bats – by adding the probabilities in the table, we see that the probability that Y is at most 9 is .904, so the probability that Y exceeds 9 is 1 - .904 = .096.

For a negative binomial experiment where Y is the number of trials needed to observe r successes, one can show that the mean value is

$$E(Y) = \frac{r}{p}.$$

For our baseball example, r = 2 and p = .372, so the expected number of at-bats to get two hits would be E(Y) = 2/.372 = 5.4. It is interesting to note that although Y = 3 is the most probable value, Ichiro would average over 5 at-bats to get 2 hits in many repetitions of this random experiment.

PRACTICE: NEGATIVE BINOMIAL EXPERIMENTS

Suppose a candy box contains a large number of candies of which 30% are peppermint. You select candies from the box with replacement until you choose a peppermint.

1. Explain why this is a negative binomial experiment and give values of r and p.
2. If Y represents the number of draws until you choose a peppermint, find the probability that Y is equal to 2.
3. Find P(Y <= 3).
4. How many draws do you expect to make before you choose a peppermint?
5. I simulated this experiment 20 times and I've listed the values of Y for these twenty experiments. Use this data to approximate P(Y <=3) and E(Y) and compare your answers with your answers to parts 3 and 4.

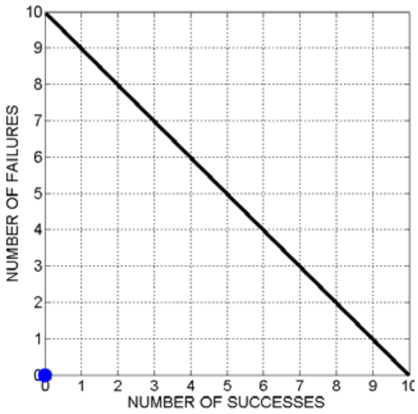
```
3  3  2  3  3 10  1 11  1  1  2  5  1  1  4  4  1  4  3  1
```

ACTIVITY – GRAPHING BINOMIAL AND NEGATIVE EXPERIMENTS

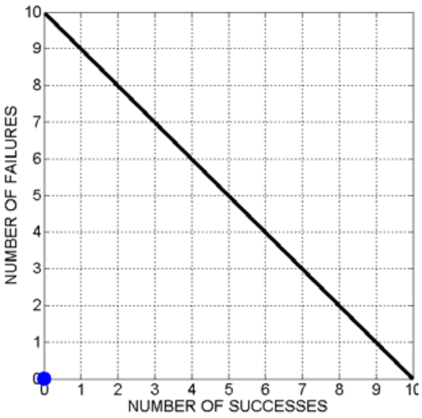
Suppose a basketball player makes a free throw shot with probability 0.7. We can simulate this shot using a ten-sided die – **if the die roll is between 1 and 7, she makes the shot; otherwise she misses the shot.** (If a ten-sided die is not available, then a table of random digits can be used.)

Using the die, simulate the shooting of 10 free throws for four games. Record your data on the graph. Start at the (0, 0) point (where there is a big dot). If the player makes the shot, draw a line one unit to the right; if she misses, draw a line one unit up. When you are done shooting 10 shots, your line should be on the dark diagonal line.

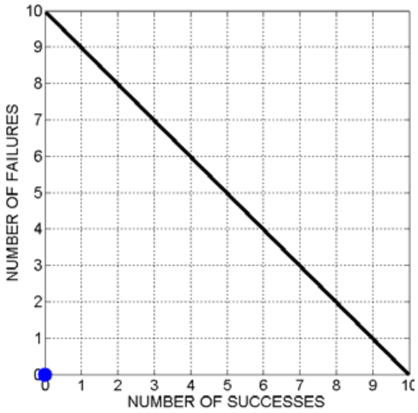
Game 1: Number of successes: _____



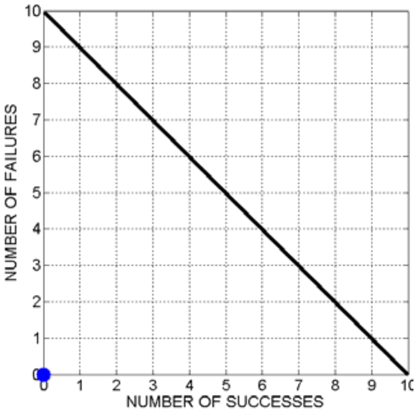
Game 2: Number of successes: _____



Game 3: Number of successes: _____



Game 4: Number of successes: _____



1. Record the number of makes (successes) for each game – put your answers in the blanks.

2. Collect the number of makes from all students in. Construct a suitable graph of these data.

3. What is the most likely number of successes during a game of 10 shots?

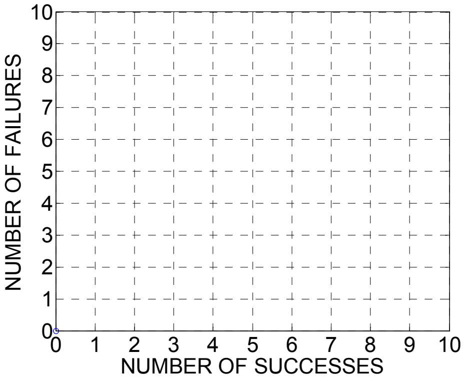
4. Find the probability she makes at least half her shots.

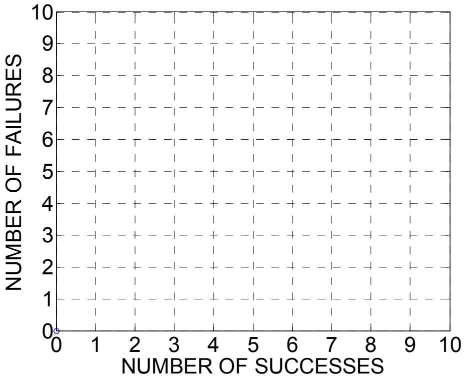
5. Find the probability she makes all of her shots.

6. Suppose that the shooter continues to shoot free throws until she misses three shots. Use this same diagram to record the results of the individual shots until the experiment is completed. For each experiment, record Y = the total shots taken.

Game 1: Total shots = ______

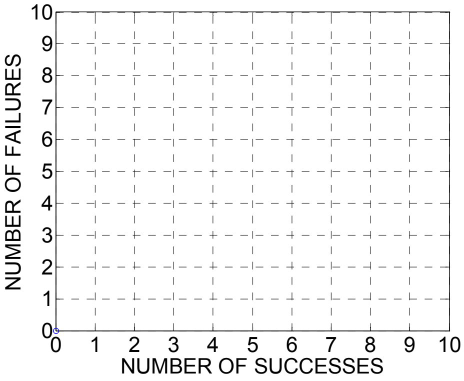
Game 2: Total shots = ______

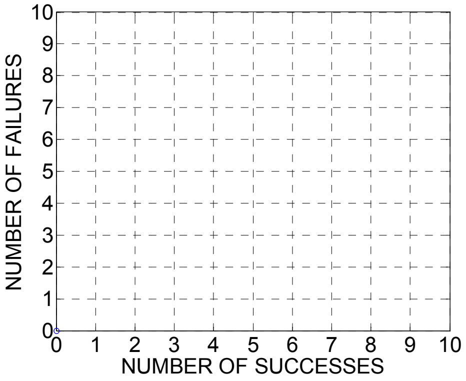




Game 1: Total shots = ______

Game 2: Total shots = ______





7. Collect the number of makes from all students in. Construct a suitable graph of these data.

8. What is the most likely number of shots taken?

9. What is the probability the woman will take at least 8 shots?

WRAP-UP

In this topic, we were introduced to the binomial experiment which represents a popular type of random experiment that resembles coin tossing. The experiment is a sequence of trials where there are two possible outcomes on each trial, the probability of a success is the same for each trial, and outcomes from different trials are independent. The focus is on the number of successes X that has a binomial distribution with parameters n, the number of trials, and p, the probability of success. A probability formula to compute P(X = k) was derived, and simple expressions were presented for the mean and standard deviation of X. The negative binomial experiment is a similar coin-tossing experiment where one continues sampling until one observes r successes and the random variable is Y, the number of trials.

EXERCISES

1. **Binomial Experiments**

Is each random process described below a binomial experiment? If it is, give values of n and p. Otherwise, explain why it is not binomial.

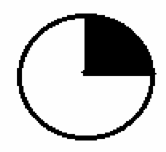
a. Roll a die 20 times and count the number of sixes you roll.

b. There is a room of 10 women and 10 men – you choose five people from the room without replacement and count the number of women you choose.

c. Same process as (b) but you sample with replacement instead of without replacement.

d. You flip a coin repeatedly until you observe 3 heads.

e. The spinner shown to the right is spun 50 times – you count the number of spins in the black region.



2. **Binomial and Negative Binomial Experiments**

Each of the random processes below is a binomial experiment, a negative binomial experiment, or neither. If the process is binomial, give values of n and p, and if the process is negative binomial, give values of r and p.

a. Suppose that 30% of students at a college regularly commute to school. You sample 15 students and record the number of commuters.

b. Same scenario as part a. You continue to sample students until you find two commuters and record the number of students sampled.

c. Suppose that a restaurant offers apple and orange juice. From past experience, the restaurant knows that 30% of the breakfast customers order apple juice, 50% order orange juice, and 20% order no juice. One morning, the restaurant has 30 customers and the numbers ordering apple juice, orange juice, and no juice are recorded.

d. Same scenario as part b. The restaurant only records the number ordering orange juice out of the first 30 customers.

e. Same scenario as part b. The restaurant counts the number of customers that order breakfast until exactly three order apple juice.

f. Same scenario as part b. Suppose that from past experience, the restaurant knows that 40% of the breakfast bills will exceed $10. Of the first 30 breakfast bills, the number of bills exceeding $10 is observed.

3. **Shooting Free Throws**

Suppose that Michael Jordan makes 80% of his free throws. Assume he takes 10 free shots during one game.

a. What is the most likely number of shots he will make?

b. Find the probability that he makes at least 8 shots.

c. Find the probability he makes more than 5 shots.

4. **Purchasing Audio CDs**

Suppose you know that 20% of the audio cd’s sold in China are defective. You travel to China and you purchase 20 cd's on your trip.

a. What is the probability that at least one cd in your purchase is defective?

b. What is the probability that between 4 and 7 cd’s are defective?

c. Compute the “average” number of defectives in your purchase.

5. **Rolling Five Dice**

Suppose you roll five dice and count the number of 1’s you get.

a. Find the probability you roll exactly two 1’s. [Do an exact calculation.]

b. Find the probability all the dice are 1’s. [Do an exact calculation.]

c. Find the probability you roll at least two 1’s. [Do an exact calculation.]

6. **Choosing Socks from a Drawer**

Suppose a drawer contains 10 socks, of which 4 are brown. I select 5 socks from the drawer with replacement.

a. Find the probability two of the five selected are brown.

b. Find the probability I choose more brown than non-brown.

c. How many brown socks do I expect to select?

d. Does the answer to part a change if we select socks from the drawer without replacement? Explain.

7. **Choosing Socks from a Drawer**

Suppose that I select socks from the drawer with replacement until I see two that are brown.

a. Find the probability that it takes me four selections.

b. Find the probability it takes more than 2 selections.

c. How many selections do I expect to make?

8. **Sampling Voters**

In your local town, suppose that 60% of the residents are supportive of a school levy that will be on the ballot in the next election. You take a random sample of 15 residents

a. Find the probability that a majority of the sample support the levy.

b. How many residents in the sample do you expect will support the levy?

c. If you sample the residents one at a time, find the probability that it will take you five residents to find three that support the levy.

9. **Taking a True/False Test**

Suppose you take a true/false test with twenty questions and you guess at the answers.

a. Find the probability you pass the test assuming that passing is 60% or higher correct.

b. Find the probability you get a B or higher where B is 80% correct.

c. If you get an 80% on this test, is it reasonable to assume that you were guessing? Explain.

10. **Bernoulli Experiment**

Let X_1 denote the result of one binomial trial, where X_1=1 if you observe a success and X_1= 0 if you observe a failure. Find the mean and variance of X_1.

11. **Rolling a Die**

Suppose we roll a die until we observe a 6. This is a special case of a negative binomial experiment where r = 1 and p = 1/6. When we are interested in the number of trials until the first success, this is a geometric experiment and Y is a geometric random variable.

a. Find the probability that it takes you 4 rolls to get a 6.

b. Find the probability that it takes you more than 2 rolls to get a 6.

c. How many rolls do you need, on average, to get a 6?

12. **Heights of Male Freshmen**

Suppose that one third of male freshmen entering a college are over 6 feet tall. Four men are randomly assigned to a dorm room. Let X denote the number of men in this room that are under 6 feet tall. (You can ignore the fact that the actual sampling of men is done without replacement.)

a. Assuming X has a binomial distribution, what is a "success" and give values of n and p.

b. What is the most likely value of X? What is the probability of this value?

c. Find the probability that at least three men in this room will be under 6 feet tall.

13. **Basketball Shooting**

Suppose a basketball player is practicing shots from the free-throw line. She hasn't been playing for a while and she becomes more skillful in making shots as she is practicing. Let X represent the number of shots she makes in 50 attempts. Explain why the binomial distribution should not be used in finding probabilities about X.

14. **Collecting Posters from Cereal Boxes**

Suppose that a cereal box contains one of four posters and you are interested in collecting a complete set. You first purchase one box of cereal and find poster #1.

a. Let X_2 denote the number of boxes you need to purchase to find a different poster than #1. Find the expected value of X_2.

b. Once you have found your second poster, say #2, let X_3 denote the number of boxes you need to find a different poster than #1 or #2. Find the expected value of X_3.

c. Once you have collected posters #1, #2, #3, let X_4 denote the number of boxes you need to purchase to get poster #4. Find the expected value of X_4.

d. How many posters do you need, on average, to get a complete set of four?

15. **Baseball Hitting**

In baseball, it is important for a batter to get "on-base" and batters are rated in terms of their on-base percentage. In the 2004 baseball season, Bobby Abreu of the Philadelphia Phillies had 705 "plate appearances" or opportunities to bat. Suppose we divide his plate appearances into groups of five – we record the number of times Abreu was on-base for plate appearances 1 through 5, for 6 through 10, for 11 through 15, and so on. If we let X denote the number of times on-base for five plate appearances, then we observe the following counts for X:

X	0	1	2	3	4	5	TOTAL
Count	10	29	44	40	15	3	141

To help understand this table, note that the count for X=1 is 29 – this means there were 29 periods where Abreu was on-base exactly one time. The count for X=2 is 44 – this means that for 44 periods Abreu was on-base two times.

Since each outcome is either a success or failure, where success is getting on-base, one wonders if the variation in these data can be explained by a binomial distribution.

X	0	1	2	3	4	5	TOTAL
P(X)							

Expected Count							

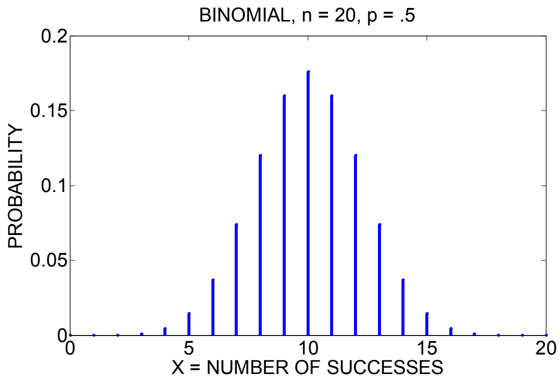
a. Find the probabilities for a binomial distribution with n=5 and p=.443. (This value of p is Abreu's on-base rate for the entire 2004 baseball season.) Place these probabilities in the P(X) row of the table.

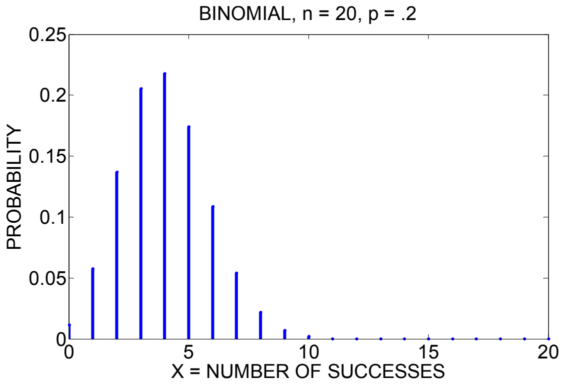
b. Multiply the probabilities you found in part (a) by 141, the number of periods in the 2004 season. Place these numbers in the Expected Count row of the table. These represent the expected number of times Abreu would have 0, 1, 2, .., 5 times on-base if the probabilities followed a binomial distribution.

c. Compare the expected counts with the actual observed counts in the first table. Does a binomial distribution provide a good description of these data?

16. **Graphs of Binomial Distributions**

The below figures show the binomial distributions with n = 20 and p = .5 (left) and n = 20 and p = .2 (right).





Recall the 68 rule from Topic P6 that said that if a probability distribution is approximately bell-shaped, then approximately 68% of the probability falls within one standard deviation of the mean.

a. For the binomial distribution with n = 20 and p = .5, find the mean μ and standard deviation σ and compute the interval (μ – σ, μ + σ).

b. Find the exact probability that X falls in the interval (μ – σ, μ + σ).

c. Repeat parts a and b for the binomial distribution n = 20 and p = .2.

d. For which distribution was the 68% rule more accurate? Does that make sense based on the shapes of the two distributions?

17. **Guessing on a Test**

Students in a statistics class were given a five-question baseball trivia quiz. On each question, the students had to choose one of two possible answers. The number correct X was recorded for each student – a count table of the values of X is shown below.

X = number correct	Count	Probability	Expected
0	0		
1	3		
2	4		
3	7		
4	6		
5	1		

a. Suppose the students know little about baseball and so they are guessing on each question. If this is true, find the probability distribution of the number correct X.

b. Using this distribution, find the probability of each value of X and place these probabilities in the above table.

c. By multiplying these probabilities by the number of students (21), find the expected number of students for each value of X.

d. Compare your expected counts with the actual counts – does a binomial distribution seem like a reasonable assumption in this example?

18. **Playing Roulette**

Suppose you play the game roulette 20 times. Each game, you place a Trio Bet on three numbers and you win with probability 3/38.

a. Find the probability you win the game exactly two times.

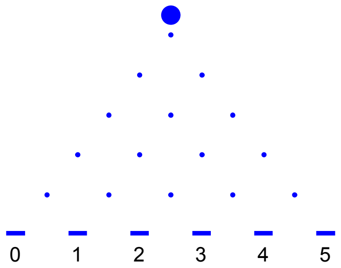
b. Find the probability that you are winless in the 20 games.

c. Find the probability you win at least once.

d. How many games do you expect to win?

19. **The Galton Board**

Consider the Galton board described in the Spotlight at the beginning of this topic. A ball is placed above the first peg and dropped. When it strikes a peg, it is equally likely to fall left or right. The location at the bottom X is equal to the number of times that the ball falls right.



a. Explain why X has a binomial distribution and give the values of n and p.

b. Find P(X = 2).

c. Find the probability the ball falls to the right of the location "1".

d. Suppose that we change the experiment so that the probability of falling right is equal to 1/4. Explain how this changes the binomial experiment and find P(X = 2).

20. **Drug Testing**

In a *New York Times* article "Facing Questions, Rodriguez Raises More" (February 21, 2008), Major League Baseball is said to have a drug-testing policy where 600 tests are randomly given to a group of 1200 professional ballplayers. Alex Rodriguez claimed one season that he received five random tests.

a. If every player is equally likely to receive a single random blood test, what is the probability that Rodriguez gets tested?

b. If X represents the number of tests administered to Rodriguez among the 600 test, then explain why X has a binomial distribution and give the values of n and p.

c. Compute the probability that Rodriguez receives exactly one test.

d. Recall Rodriguez's claim that he received five random tests. Compute the probability of this event.

e. You should find the probability computed in part d to be very small. If Rodriguez is indeed telling the truth, what do you think about the randomness of the drug-testing policy?

TOPIC P8: CONTINUOUS DISTRIBUTIONS

SPOTLIGHT: A SPINNER BASEBALL GAME

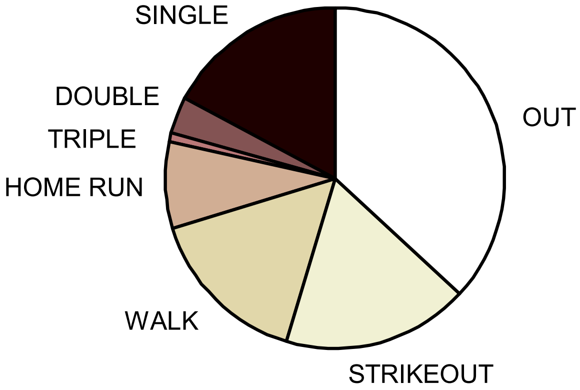
The baseball board game *All-Star Baseball* has been honored as one of the fifty most influential board games of all time according to the Wikipedia Encyclopedia (http://en.wikipedia.org). This game is based on a collection of spinner cards, where one card represents the possible batting accomplishments for a single player. The game is played by placing a card on a spinner and a spin determines the batting result for that player.

A spinner card is constructed by use of the statistics collected for a player during a particular season. To illustrate this process, the below table shows the batting statistics for the famous player Mickey Mantle for the 1956 baseball season. When Mantle comes to bat, that is called a plate appearance (PA) – we see from the table that he had 632 plate appearances this season. There are several different events possible when Mantle came to bat – he could get a single (1B), a double (2B), a triple (3B), or a home run (HR). Also he could walk (BB), strike out (SO), or get other type of out.

PA	1B	2B	3B	HR	BB	SO	Other OUTS
632	109	22	5	52	99	112	233

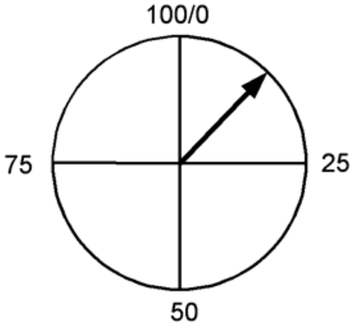
We can find the probability of each type of event by dividing each count by the number of plate appearances. We convert each probability to an angle on the spinner by multiplying each probability by the total number of degrees (360). From these degree measurements, we can construct a spinner, displayed below, where the area of each wedge of the circle is proportional to the probability of that event occurring. We can now simulate a single plate appearance of Mickey Mantle by spinning the spinner and observing the batting event.

PA	1B	2B	3B	HR	BB	SO	Other OUTS
632	109	22	5	52	99	112	233
Probability	.172	.035	.008	.082	.157	.177	.369
Degrees in spinner	62	13	3	30	57	64	133



PREVIEW

The binomial described in the last topic is an example of a discrete random variable which takes on only values in a list, such as 0, 1, …, 10. How can we think about probabilities where the random variable is not discrete? As a simple example, consider the experiment of spinning the spinner shown below where the random variable X is the recorded location. Here X is a continuous random variable that can take on any value between 0 and 100.



We will show that probabilities for a continuous random variable are represented by means of a smooth curve where the probability that X falls in a given interval is equal to an area under the curve. Through a series of examples, we illustrate probability calculations and summarization for continuous random variables.

In this topic, your learning objectives are to:

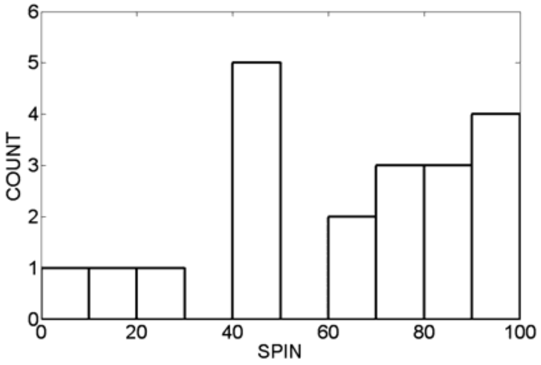
- Understand basic properties of a density curve for a continuous random variable.
- Be able to compute probabilities given a density curve.
- Be able to summarize a continuous random variable by the computation of a mean and a standard deviation.
- Compute a cumulative density function for a continuous random variable.
- Be able to compute percentiles for a continuous random variable.

THE UNIFORM DISTRIBUTION

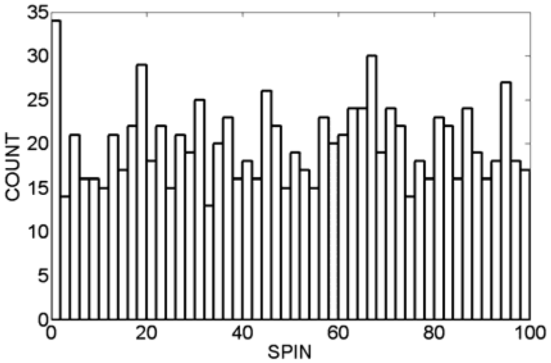
Consider the spinner experiment described in the Preview where the location of the spinner X can be any number between 0 and 100. I had my computer simulate spinning this spinner 20 times with the following results (rounded to the nearest tenth):

95.0	23.1	60.7	48.6	89.1	76.2	45.6	1.9	93.5	91.7
82.1	44.5	61.5	79.2	92.2	73.8	17.6	40.6	41.0	89.4

A histogram of these values of X is shown in the below figure.

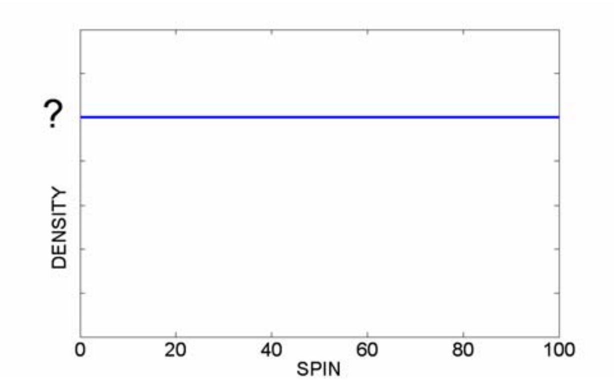


Although we think that any spin between 0 and 100 is equally likely to occur, I don't see any obvious shape of this histogram. But we only spun the spinner 20 times. Let's try spinning 1000 times – a histogram of the spins is shown below.



Note that since we have a large sample of values, we chose a small interval width for each bin in the histogram. Now we are seeing a clearer shape in the histogram – although there is variation in the bar heights, the general shape of the histogram seems to be pretty flat or uniform over the entire interval of possible values of X between 0 and 100.

Suppose we were able to spin the spinner a LARGE number of times. If we did this, then the shape of the histogram would look close to the *uniform curve* shown below.



When the random variable X is continuous, such as the case of the spinner result here, then we represent probabilities by means of a smooth curve that we call a *density curve* (or more formally, a probability density curve). How do we find probabilities? When X is continuous, then probabilities are represented by areas under the density curve.

As a simple example, what is the chance that the spinner result falls between 0 and 100? Since the scale of the spinner is from 0 to 100, we know that all spins must fall in this interval, so the probability of X landing in (0, 100) is 1. This probability is represented by the total area under the flat line between 0 and 100. Since the area of this rectangle is given by height TIMES base, and the base is equal to 100, the height of this density curve must by 1/100 = 0.01. (This is the value that should replace the "?" in the figure.)

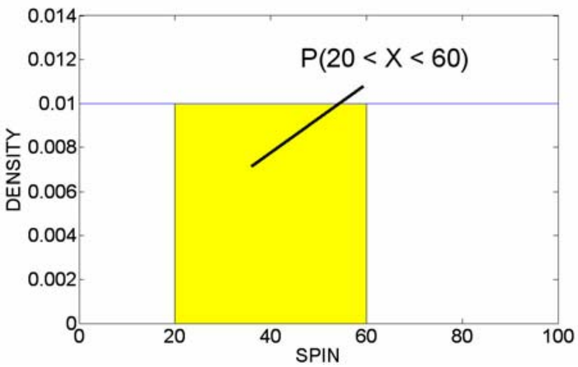
By means of similar area computations, we can find other probabilities about the spinner location X.

1. What is the probability the spin falls between 20 and 60? That is, what is

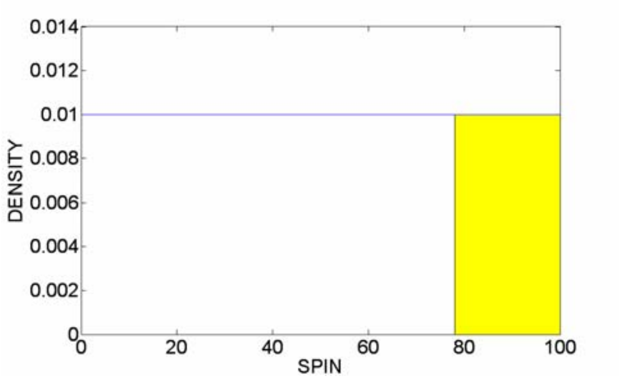
$$P(20 < X < 60)?$$

This probability is equal to the shaded area under the uniform density between 20 and 60. (See the figure below.) Using again the formula for the area of a rectangle, the base is 60 – 20 = 40 and the height is 0.01, so

$$P(20 < X < 60) = 40\ (0.01) = 0.4.$$



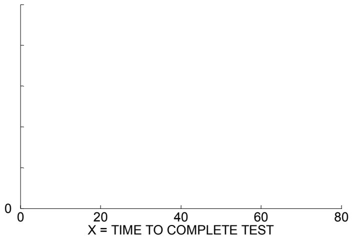
2. What is the probability the spin is greater than 80? That is, what is P(X > 80)?
The figure below shows the area that needs to be computed to find this probability. Note that the area under the curve only between the values 80 and 100 is shaded, since X cannot be larger than 100. Again by finding the area of the shaded rectangle, we see that P(X > 80) = 20 (0.01) = 0.2.



PRACTICE: THE UNIFORM DENSITY

Suppose that the time X that a student will take to complete a test is uniformly distributed between 20 and 60 minutes.

1. On the below figure, graph the density function of X. What is the height of this density curve?



2. Find the probability the student takes less than 30 minutes to complete the test.

3. Find the probability the student takes between 50 and 60 minutes.

4. What would be an average time for a student to complete this test? How did you compute this average?

PROBABILITY DENSITY / WAITING FOR A BUS

Suppose we have a random experiment and we observe a continuous random variable X such as the location of the spinner in the previous example. To describe probabilities about X, we define a *density* function denoted by f(x). Any function f won't work – we require that f satisfy two properties:

Property 1. The probability density f must be ***nonnegative*** which means that

$$f(x) \geq 0 \textit{ for all } x.$$

Property 2. The total area under the probability density curve f must be equal to 1 – using symbols

$$\int_{-\infty}^{\infty} f(x) = 1.$$

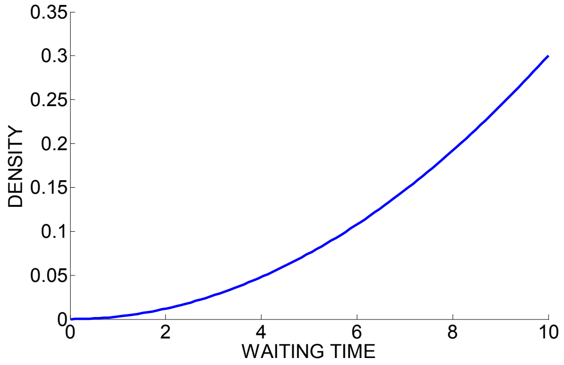
To illustrate a probability density, suppose that I have a class that meets three times a week. To get to class, I walk to a bus stop and wait for a bus to take me to school. From past experience, I know that I can wait any time between 0 and 10 minutes for my bus, and I know that each waiting time between 0 and 10 minutes is equally likely.

For a given week, what's the chance that my longest wait will be under 7 minutes?

Let W denote my longest waiting time for the week. One can show that the density for W is given by

$$f(w) = \frac{3\,w^2}{1000}, \quad 0 < w < 10.$$

This density for this longest waiting time is shown below.



Before we go any further, we should check if this is indeed a legitimate probability density:

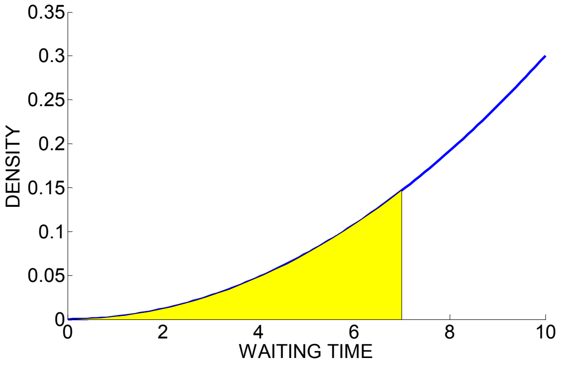
1. We note from the graph that the density does not take on negative values, so the first property is satisfied.

2. Second, for it to be a probability density, the entire area under the curve must be equal to 1. Let's check this by finding the integral of the density between 0 and 10 (the region where the density if positive):

$$\int_0^{10} \frac{3\,w^2}{1000}\,dw = \left.\frac{w^3}{1000}\right|_0^{10} = \frac{10^3}{1000} - \frac{0^3}{1000} = 1.$$

The entire area under the curve is indeed equal to 1, so f is a legitimate probability density.

Now that we know f is a probability density, we can use it to find probabilities. To find the probability that this longest waiting time is less than 7 minutes, P(W < 7), we wish to compute the area under the density curve between 0 and 7.



This is equivalent to the integral

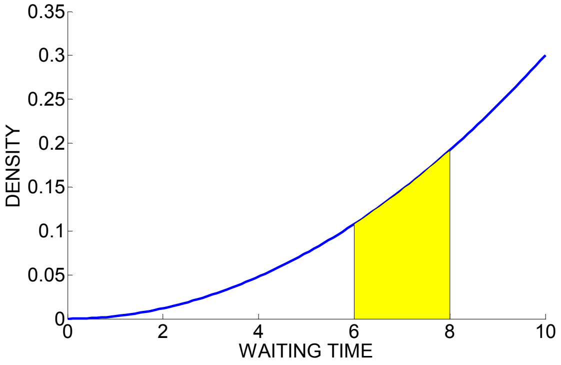
$$P(W < 7) = \int_0^7 \frac{3\,w^2}{1000}\,dw\,,$$

and, by evaluating this, we obtain the probability

$$P(W < 7) = \int_0^7 \frac{3\,w^2}{1000}\,dw = \left.\frac{w^3}{1000}\right|_0^7 = \frac{7^3}{1000} - \frac{0^3}{1000} = 0.343.$$

.

Suppose we are interested in the probability that our longest waiting time is between 6 and 8 minutes. We can represent this by the shaded area below.



To compute this area, we find the integral of the density between 6 and 8:

$$P(6 < W < 8) = \int_6^8 \frac{3\,w^2}{1000}\,dw = \left.\frac{w^3}{1000}\right|_6^8 = \frac{8^3}{1000} - \frac{6^3}{1000} = 0.296.$$

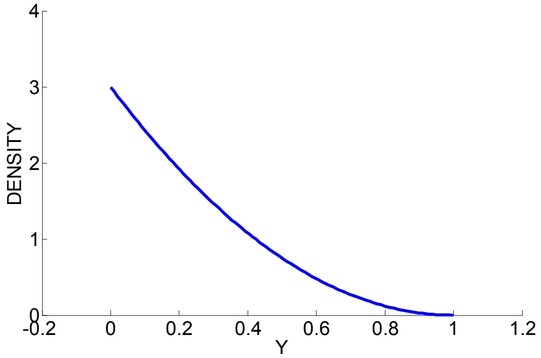
.

PRACTICE: PROBABILITY DENSITY

Suppose you generate three random numbers X_1, X_2, X_3 between 0 and 1 from your calculator and you store the smallest value Y = min$\{X_1, X_2, X_3\}$. It can be shown that the density function for Y is given by

$$f(y) = 3(1-y)^2, \quad 0 < y < 1.$$

A graph of this density function is shown below.



1. Show that f satisfies the two properties of a density function.

2. Find the probability that Y is between 0.2 and 0.4.

3. Find the probability Y is larger than .5.

4. Based on looking the graph, estimate the average value of Y.

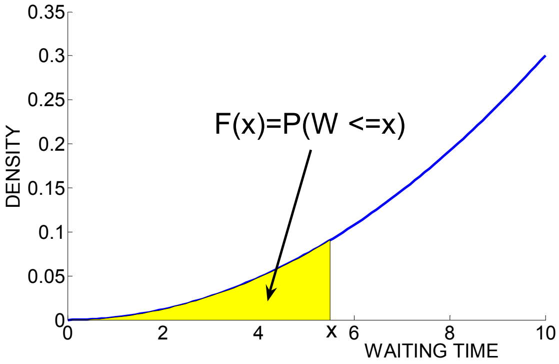
THE PROBABILITY FUNCTION (THE CDF F(X)).

To find any probability about the maximum waiting time, we can compute an area under the curve that is equivalent to integrating the density curve over a region. But there is a basic function that we can compute at the beginning that will simplify these probability computations.

Choose an arbitrary point x – the cumulative distribution function at x, or cdf for short, is the probability that W is less than or equal to x:

$$F(x) = P(W \le x) = \int_{-\infty}^{x} f(w)dw.$$

Here suppose we choose a value of x in the interval (0, 10). Then F(x) would be the area under the density curve between 0 and x shown in the below figure.



Writing this area as an integral, we compute F(x) as

$$F(x) = P(W \le x) = \int_{0}^{x} \frac{3\,w^2}{1000}\,dw = \left.\frac{w^3}{1000}\right|_0^x = \frac{x^3}{1000}.$$

.

This formula is valid for any value of x in the interval (0, 10).

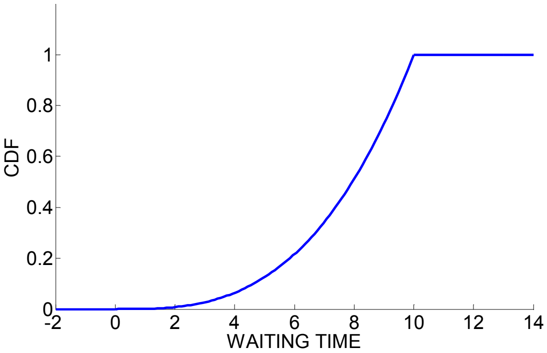
Actually F(x) is defined for all values of x on the real line.

- If x is a value smaller or equal to 0, then we see from the figure that the probability that W is smaller than x is equal to 0. So F(x) = 0 for x <= 0.

- On the other hand, if x is greater or equal to 10, then the probability that W is smaller than x is 1. So F(x) = 1 for x >= 10.

Putting all of our work together, we see that the cdf F is given by

$$F(x) = \begin{cases} 0, & x \le 0 \\ \dfrac{x^3}{1000}, & 0 < x < 10 \\ 1, & x > 10 \end{cases}$$



PRACTICE: THE PROBABILITY FUNCTION

Recall the earlier practice problem where you generate three random numbers X_1, X_2, X_3 between 0 and 1 from your calculator and you store the smallest value Y = min$\{X_1, X_2, X_3\}$. The density function for Y is given by

$$f(y) = 3(1-y)^2, \quad 0 < y < 1.$$

Consider the cdf function $F(y) = P(Y \le y)$.

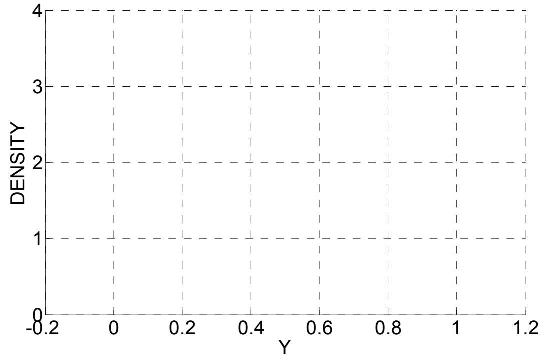
1. Find $F(.3) = P(Y \le .3)$.

2. If y is a value smaller than 0, find F(y).

3. If y is a value larger than 1, find F(y).

4. For a value y between 0 and 1, find F(y).

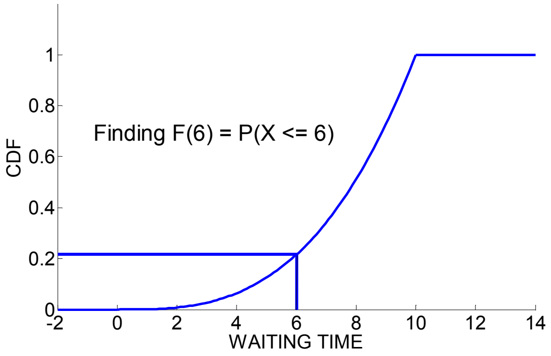
5. On the grid below, graph the function F(y).



FINDING PROBABILITIES USING F.

Once we have computed the cdf function F, probabilities can be found simply by evaluating F at different points. Fortunately, no additional integration is needed.

For example, to find the probability that the maximum waiting time W is less than equal to 6 minutes, we just compute $F(6) = P(W \le 6) = \frac{6^3}{1000} = 0.216,$ which is shown in the below figure.



To compute the probability that the *maximum waiting time exceeds 10 minutes*, we first note that "exceeding 10 minutes" is the complement event to "less than or equal to 10 minutes", and so

$$P(W > 8) = 1 - P(W \le 8) = 1 - F(8) = 1 - \frac{8^3}{1000} = 0.488.$$

Likewise, if we are interested in the chance that the *waiting time W falls between 2 and 4*, we represent the probability as the difference of two "less-than" probabilities, and then subtract the two values of F.

$$P(2 < W < 4) = P(W \le 4) - P(W \le 2) = F(4) - F(2) = \frac{4^3}{1000} - \frac{2^3}{1000} = 0.056.$$

PRACTICE: FINDING PROBABILITIES USING F

Recall the earlier practice problem where you generate three random numbers X_1, X_2, X_3 between 0 and 1 from your calculator and you store the smallest value Y = min$\{X_1, X_2, X_3\}$. The cdf function for Y is given by

$$F(y) = \begin{cases} 0, & y < 0 \\ 1-(1-y)^3, & 0 \le y \le 1 \\ 1, & y > 1 \end{cases} .$$

Using this formula, find

1. The probability the smallest value Y is less than .5.

2. P(Y > .8)

3. P(.3 < Y < 1.2)

4. P(Y < -.3)

SUMMARIZING A CONTINUOUS RANDOM VARIABLE

We are interested in summarizing a continuous random variable. Natural summaries are given by the mean μ and the standard deviation σ, where these quantities are defined in a similar manner as for a discrete random variable, with the exception that summations are replaced by integrals.

The mean μ, or equivalently the expected value of X, is given by

$$\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx\,.$$

Just as in the discrete random variable case, there is an attractive interpretation of μ. If we are able to observe a large number of values of X, then μ will be approximately equal to the sample mean $\bar{X}$ of these random values of X.

To define the spread of the values of X, we first compute the average squared deviation about the mean, the variance,

$$\sigma^2 = Var(X) = E(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx\,.$$

The standard deviation of X, σ, is defined to be the square root of the variance.

Let's illustrate the computation of the mean and standard deviation for our waiting time problem. Using the definition of f, we get that the mean is equal to

$$\mu = \int_{0}^{10} x\left(\frac{3x^2}{1000}\right) dx\,.$$

Performing the integration, we get

$$\mu = \int_0^{10} x\left(\frac{3x^2}{1000}\right)dx = \left.\frac{3x^4}{4000}\right|_0^{10} = \frac{3\,(10)^4}{4000} = 7.5.$$

On, the average, we expect our longest wait in a week to be 7.5 minutes.

The computation of the variance is a bit more tedious, but straightforward.

$$\begin{aligned}
\sigma^2 &= \int_0^{10} (x-7.5)^2\left(\frac{3x^2}{1000}\right)dx \\
&= \int_0^{10} x^2\left(\frac{3x^2}{1000}\right)dx - \int_0^{10} 15x\left(\frac{3x^2}{1000}\right)dx + \int_0^{10} 7.5^2\left(\frac{3x^2}{1000}\right)dx \\
&= \left.\frac{3x^5}{5000}\right|_0^{10} - \left.\frac{45x^4}{4000}\right|_0^{10} + \left.\frac{168.75x^3}{3000}\right|_0^{10} \\
&= 3.75
\end{aligned}$$

So the standard deviation of X is $\sigma = \sqrt{3.75} = 1.94$.

PRACTICE: SUMMARIZING A CONTINUOUS RANDOM VARIABLE

In the earlier practice problem, you generated three random numbers X_1, X_2, X_3 between 0 and 1 from your calculator and you store the smallest value Y = min$\{X_1, X_2, X_3\}$. The density function is given by

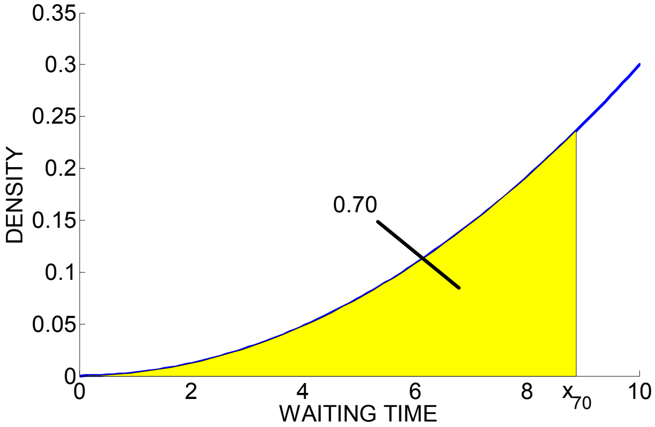
$$f(y) = 3(1-y)^2, \quad 0 < y < 1.$$

Find the mean and standard deviation for Y.

PERCENTILES

Another useful summary of a continuous random variable is a percentile. The 70th percentile, for example, is the value of X, call it x, such that 70% of the probability is to the left. (See the figure below.) That is, the 70th percentile, call it x_{70}, satisfies the equation

$$P(X \le x_{70}) = .70\,.$$



Since we recognize the left hand side of the equation as equivalent to the cdf F (which we have already computed), we can write the equation as

$$F(x_{70}) = .70$$
$$\frac{x_{70}^3}{1000} = .70$$

To find the 70th percentile, we solve the above equation for x_{70} -- after some algebra, we get that

$$x_{70} = \sqrt[3]{700} = 8.88\,.$$

This means that if we wait many weeks for this bus, approximately 70% of the longest waiting times will be shorter than 8.88 minutes.

PRACTICE: PERCENTILES

For our practice problem, the random variable Y has a cdf defined by

$$F(y) = \begin{cases} 0, & y < 0 \\ 1-(1-y)^3, & 0 \le y \le 1 \\ 1, & y > 1 \end{cases} \quad .$$

Use this cdf to find

1. the median of Y

2. the 20th percentile

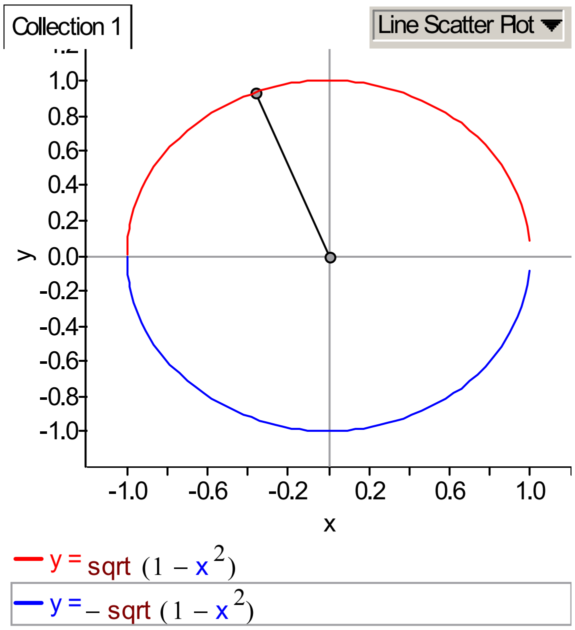
3. the 90th percentile

TECHNOLOGY ACTIVITY: SPINNING AWAY

You are playing a game that uses a spinner. The arrow on the spinner lands randomly on the edge of the circle. Equivalently, the angle of the arrow lands uniformly between 0 and 2 π . We use Fathom to simulate this process.

1. Drag out a New Collection and New Case Table. Define two attributes, called "angle" and "r". Put the values 0 and 1 in the "r" column. Select the "angle" attribute and select Crt-E to bring up the formula box. Type random(0, 2 pi) and select OK. "Angle" will be selected randomly between 0 and 2 π.

2. Define two new attributes "x" and "y". Both are defined by formula:



- definition for “x” – you type in box: r cos(angle)
- definition for “y” – you type in box: r sin(angle)

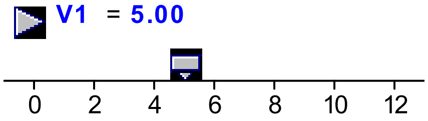
3. Drag out a New Graph.
   - Draw a Line Scatter Plot of “x” and “y”. (This is the arrow.)
   - From Graph menu, select Show Graph Info.
     - Make the x axis horizontal from –1.2 to 1.2.
     - Make the y axis vertical from –1.2 to 1.2.
     - From Edit menu, select Delete Control Text to remove graph information.
   - To draw a circle around the arrow.
     - Select the Graph.
     - Select Plot Function from the Graph menu.
     - In the function expression box, type
       - sqrt(1-x^2)
     - Again select Plot Function from the Graph menu.
     - In the function expression box, type
       - -sqrt(1-x^2)
     - If you did it right, you should see a circular spinner.

4. To get a random spin, select your collection and hit control-Y. Try this a few times to see random spins.

5. Let’s spin this spinner 100 times and draw a histogram of the results.
   - Inspect your collection. Define a new Measure called “spin” – define it to last(Angle)
   - Select the collection and Collect Measures. Inspect the Measures from Collection. Change the number of measures collected from 5 to 100. **(Keep the animation turned on!)**
   - Before you rerun this, <u>construct a histogram of the values of “spin”.</u>
   - Now recollect your Measures. Describe using a few sentences what the histogram looks like.

6. We use Fathom to compute values of the cumulative distribution function F(x) of the spin.

- Drag out a Slider you should see a variable V1 and a number line below it from 0 to 12.



- Select the Measures collection and drag out a Summary Table. Add a new formula and drag "spin" to the table. Write down mean(spin) that is shown:

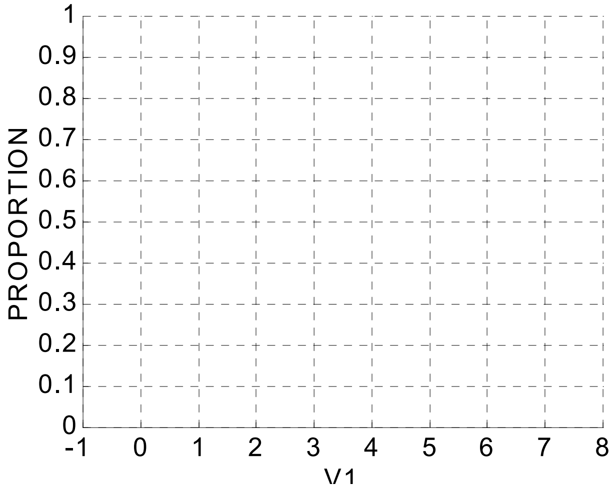
- Double-click on the mean(spin) formula and enter this new formula:
    - Proportion(spin<v1) [write down this answer]

      (this finds the proportion of spin values that are smaller than v1)

- By dragging the slider for v1, the Summary Table will compute different values for Proportion(spin < v1).

Fill out the table below using 8 different values for v1. Graph the points to the right using a line graph.

V1	Proportion(spin < v1)



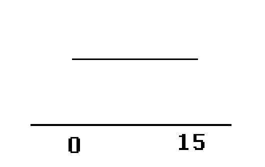
TECHNOLOGY ACTIVITY: WAITING FOR THE SHUTTLE

Suppose you commute to school. You park in commuter parking lot and wait for the shuttle bus to take you to campus. Your first class is probability and statistics at 11:30. This is your favorite class (I hope) and you certainly don't want to be late! (Your instructor gets really mad when students are late.) In any event, you get to the shuttle bus wait area at 11:10. If the bus doesn't arrive by 11:20, you will be late to class. (Your heart beats fast just with the thought of being late.)

From past experience, you know that the time X (in minutes) that you wait for the shuttle bus has a uniform distribution from 0 to 15. That is, the density function for X has the form

$$f(x) = 1/15 \ , \ 0 < x < 15.$$

This density looks like the graph below



What we will do on Fathom is to simulate the random times you wait for the bus on Monday, Wednesday, and Friday.

1. Open a new Collection and Data Table. Create two new Attributes, called "day" and "wait_time". Put the values 'Mon', 'Wed', and 'Fri' in the "day" column. For the attribute "wait_time", use the formula

randomuniform(0, 15)

[This puts 3 random wait times in the column "wait_time".]

2. Define three Measures from this Collection:

Measure	Formula

avg_wait	mean(wait_time)
mon_wait	first(wait_time)
max_wait	max(wait_time)

[Note "avg_wait" is the average time you wait on the 3 days, "mon_wait" is how long you wait on Monday, and "max_wait" is the longest wait during the week.]

3. Collect Measures -- as usual, turn off animation and repeat this process 1000 times. [I guess this means that you are taking this course for 1000 weeks -- actually we only have 15 weeks, but it may seem like 1000 weeks for a few of you.]

HERE ARE THE QUESTIONS TO ANSWER:

4. Construct histograms for each of the three measures ("mon_wait", "avg_wait" and "max_wait"). Fill out this table.

	DRAW WHAT THE HISTOGRAM LOOKS LIKE	DESCRIBE IN WORDS THE SHAPE OF THE HISTOGRAM	WHAT IS THE MOST LIKELY VALUE?
how long you wait on Monday ("mon_wait)			
the average wait of the 3 days ("avg_wait)			
the longest time you wait ("max_wait")			

5. **Using Fathom**, what is the average (mean) time that you wait for the bus on Monday?

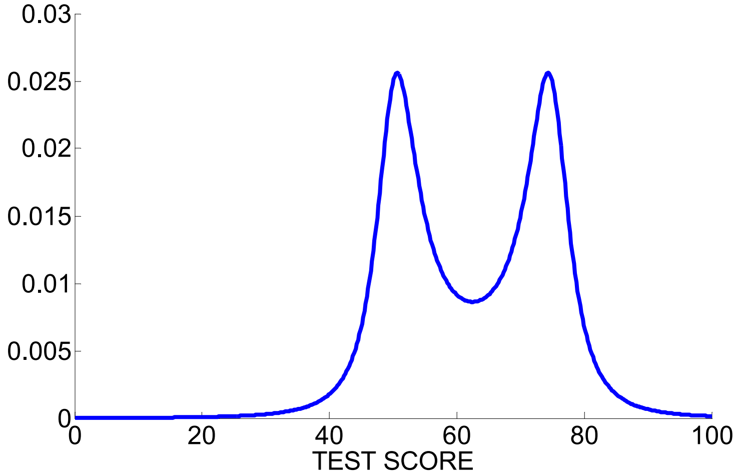
6. Find the expected value of X (not using Fathom) --compare this answer with the answer to question 5.

7. **Using Fathom**, find the standard deviation of the Monday wait time. [The formula to use in the summary table is stddev(mon_wait).]

TECHNOLOGY ACTIVITY: A TEST WITH A BIMODAL DISTRIBUTION

Suppose that grades on a test are well-described by the probability density

$$f(x) = \frac{1}{\left(1+\left(\frac{x-50}{4}\right)^2\right)\left(1+\left(\frac{x-75}{4}\right)^2\right)}, \quad 0 < x < 100$$



Use a computer program (such as Fathom) or a calculator to answer the following questions.

1. Is really a probability density? If not, can we fix it so it is a legitimate density?
2. If X represents a score on the test, find the probability a student scores below 60.
3. Find the probability X is between 70 and 90.
4. Find the median grade on the test.
5. Find F(80).

WRAP-UP

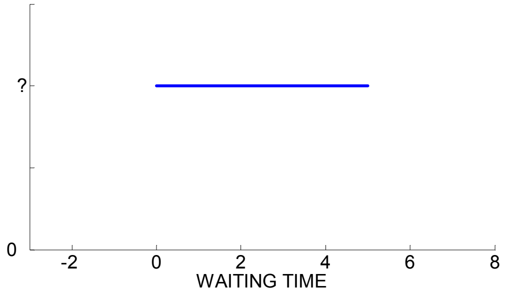
When the outcome X of a random experiment is continuous, probabilities are represented by a **density curve**. To be a legitimate density curve, the values of the curve must be nonnegative and the entire area under the curve is equal to one. A probability such as P(a < X < b) is given by the area under the density curve for values of X between a and b. One can summarize probabilities by the computation of a **cumulative density function** F(x) that finds P(X <= x) for all values of x. A continuous random variable can be summarized by the computation of a **mean** μ and a **standard deviation** σ that are expressible as integrals. Also, percentiles such as the median and quartiles are helpful in summarizing values of X.

EXERCISES

1. **Waiting at a ATM Machine**

You are waiting at your local ATM machine and as usual, you are waiting in a line. Suppose you know that your waiting time can be between 0 to 5 minutes and any value between 0 and 5 minutes is equally likely.

a. The graph to the right shows the density function for X, the waiting time. What is the height of this function?



b. Find the probability you wait more than 2 minutes.

c. Find the probability you wait between 2 and 3 minutes.

2. **Morning Wake-Up**

Suppose you wake up at a random time in the morning between 6 AM and 12 AM.

a. Find the probability you wake up before 11 am.

b. Find the probability you wake up between 8 and 10 am.

c. What is an "average" or typical time you will wake up? Explain how you computed this number.

d. Find the standard deviation of the time.

3. **The Median Waiting Time**

In the "waiting for a bus" example, suppose that you record the median time T (in minutes) that you wait for the bus on the three days. The density function for this median time is given by

$$f(t) = \frac{6\,t\,(10 - t)}{1000}, \quad 0 < t < 10.$$

a. Draw a graph of this density function

b. Find the probability that the median time is between 5 and 7 minutes.

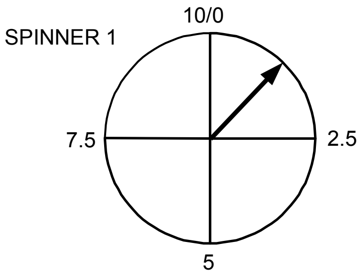
c. Find the cdf F(t) for all values of t.

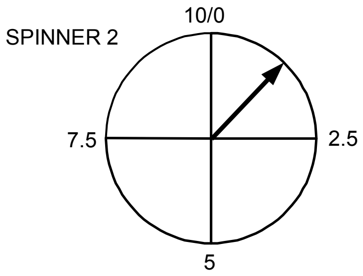
d. Using the cdf you found in part c, find the probability the median time is over 6 minutes.

e. Find the 75% percentile of your median waiting time.

4. **The Sum of Two Spins**

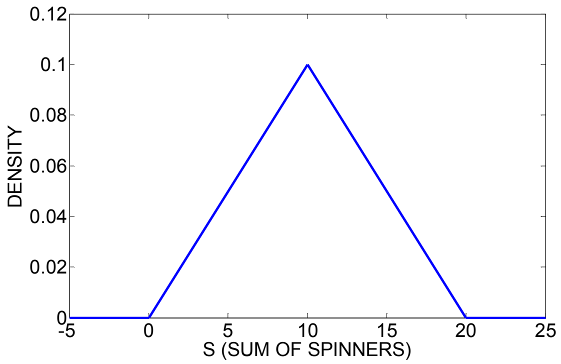
Suppose you spin two spinners, where the location of the arrow for each spinner is equally likely to fall between 0 and 10.





If you let S be the sum of the two spins, it can be shown that the density function of S is given by

$$f(s)=\begin{cases}\dfrac{s}{100}, & 0<s\leq 10\\ \dfrac{20-s}{100}, & 10<s<20\end{cases}.$$



a. Check that this is a proper density function.

b. Find the probability the sum of the two spins is smaller than 5.

c. Find the cdf function F.

d. Using the cdf function, find the probability the sum of spins falls between 8 and 12.

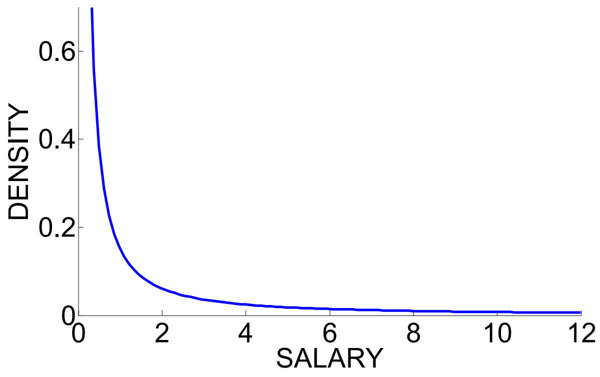
e. Using the cdf function, find the probability the sum of spins exceeds 12.

5. **Salaries for Professional Basketball Players**

Let X denote the salary (in millions of dollars) of a professional basketball player. A reasonable density function for X is given by

$$f(x)=\frac{0.15}{x^{1.3}}, \quad x\geq 0.1,$$

shown by the figure to the right.



a. What proportion of basketball players earn more than 1 million dollars?

b. What proportion of players earn between 1 and 2 million dollars?

c. Find the cdf function.

d. Using the cdf function, find the probability a player earns less than one-half a million dollars.

e. Find the "average" salary of a NBA player.

6. **Grading on a Curve**

Suppose the grades on a math test are distributed according to the curve.

$$f(x) = \frac{x}{5000}, \quad 0 < x < 100 .$$

a. Draw a graph of this density curve.

b. Find the mean grade on this test.

c. What proportion of students who take this test get a grade of 90 or higher?

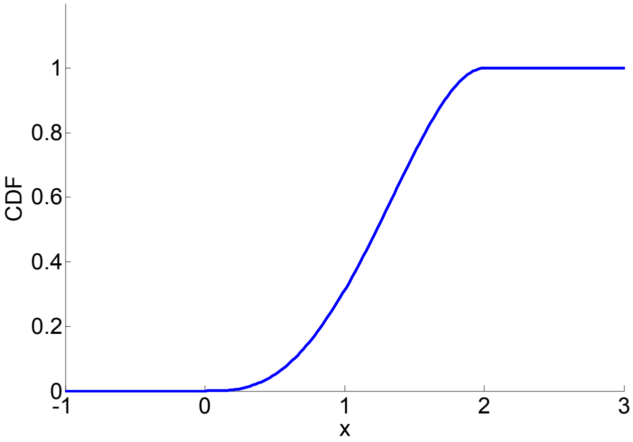
d. What proportion of students get a C grade, where C is defined to be between 70 and 80?

e. Is this test harder or easier than the test grades in your statistics class? Explain.

7. **Time to Clean Your Room**

Suppose the time that it takes you to clean your room (in hours) is a random variable X with the cdf function given below. A graph of the cdf is also shown.

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{3}{4}\left[\frac{2x^3}{3} - \frac{x^4}{4}\right], & 0 \le x \le 2 \\ 1, & x > 2 \end{cases}$$



a. Find the probability you can clean your room in under one hour.

b. Find the probability it takes you over one and a half hour to clean your room?

c. Using the graph, find a value M such that it is equally likely that X is smaller than M and X is larger than M. (M is the 50th percentile of X.)

8. **Time to Complete a Race**

Suppose a group of children are running a race. The times (in minutes) that the children complete the race can be described by the density function

$$f(x) = \frac{4+(x-3)^2}{21}, \quad 3 < x < 6.$$

a. Graph this density function.

b. Looking at your graph, is it more common to have a slow time (near 6 minutes) or a fast time (near 3 minutes)?

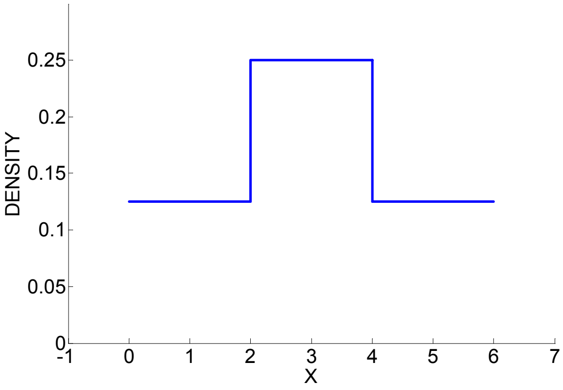
c. Find the probability a child completes the race in under 4 minutes.

d. Find the probability that a child's time exceeds 5 ½ minutes.

e. Find the median running time.

9. **Spinning a Random Spinner**

Suppose you flip a coin. If the coin lands heads, you spin a spinner that is equally likely to fall at any point in the interval (0, 4). If the coin lands tails, you spin a different spinner that lands at any point in the interval (2, 6). If X denotes your spin, the density function for X is graphed below.



a. Check that this graphed function is indeed a probability density.

b. Find the probability that X is greater than 5.

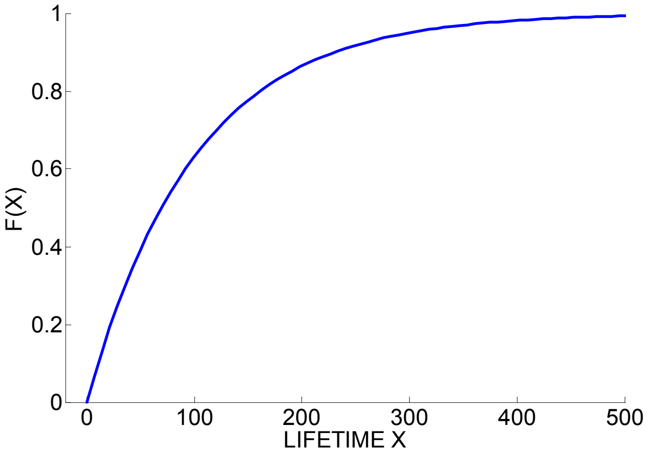
c. Find the probability that X falls between 1 and 3.

10. **Lifetimes of Light Bulbs**

Suppose that a company is interested in the amount of time that a particular type of light bulb will last until it burns out. After sampling the lifetimes for a large group of light bulbs, it is decided that the lifetime X (in hours) is well-described by the exponential distribution of the form

$$f(x) = \frac{1}{100} e^{-x/100}, \quad x > 0 .$$

The cdf for X is drawn below. In addition, the cdf is computed for some values of X in the following table.



x	F(x)	X	F(x)
0	0	180	0.8347
30	0.2592	210	0.8775
60	0.4512	240	0.9093
90	0.5934	270	0.9328

120	0.6988	300	0.9502
150	0.7769		

a. Find the probability that a lifetime of a bulb will be less than 90 hours.

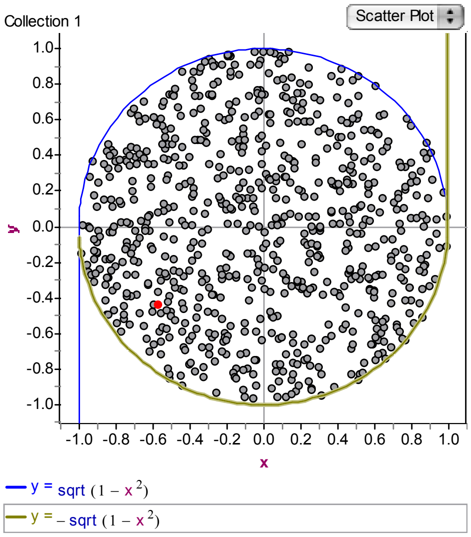
b. Find the probability the lifetime is between 120 and 180 hours.

c. From the table, approximate the median lifetime.

d. Approximate the 95th percentile.

11. **Locations of Dart Throws**

Suppose you throw a dart at a circular target such that the dart is equally likely to land in any location on the target. The locations for a large number of dart throws are shown in the below figure.



Let X denote the distance of a throw from the bulls eye. It can be shown that the density function of X has the form $f(x) = \dfrac{x}{2}, \quad 0 < x < 2.$

a. Find the probability your throw lands within a distance of 1 unit from the target.

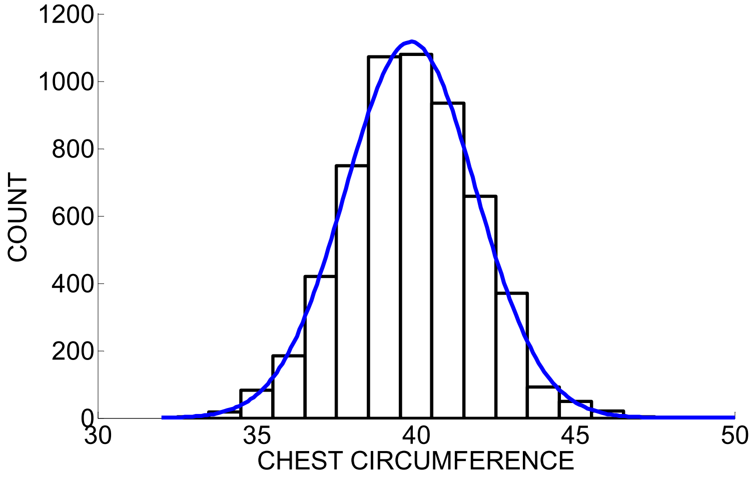
b. Find the probability your throw lands between .5 and 1.5 units from the target.

c. If you threw the dart many times at the target, find your average distance from the target.

TOPIC P9: THE NORMAL DISTRIBUTION

SPOTLIGHT: EARLY USE OF THE NORMAL CURVE

The famous normal curve was independently discovered by several scientists. Abraham De Moivre in the 18th century showed that a binomial probability for a large number of trials n could be approximated by a normal curve. Pierre Simon Laplace and Carl Friedrich Gauss also made important discoveries about this curve. By the 19th century, it was believed by some scientists such as Adolphe Quetelet that the normal curve would represent the distribution of any group of homogeneous measurements. To illustrate his thinking, Quetelet considered the frequency measurements for the chest measurements (in inches) for 5738 Scottish soldiers taken from the Edinburgh Medical and Surgical Journal (1817). A histogram of the chest measurements together with a matching normal curve are shown in the below measure. Quetelet's beliefs were a bit incorrect – *any* group of measurements will not necessarily be normal-shaped. However, it is generally true that a distribution of physical measurements from a homogeneous group, say heights of American women or foot lengths of Chinese men will generally have this bell shape.



PREVIEW

In the previous topic, we were introduced to the notion of a continuous random variable. Here we introduce the normal curve that is a popular model for representing the distribution of a measurement random variable. Also we will see that the normal curve is helpful for computing binomial probabilities and for representing the distributions of means taken from a random sample.

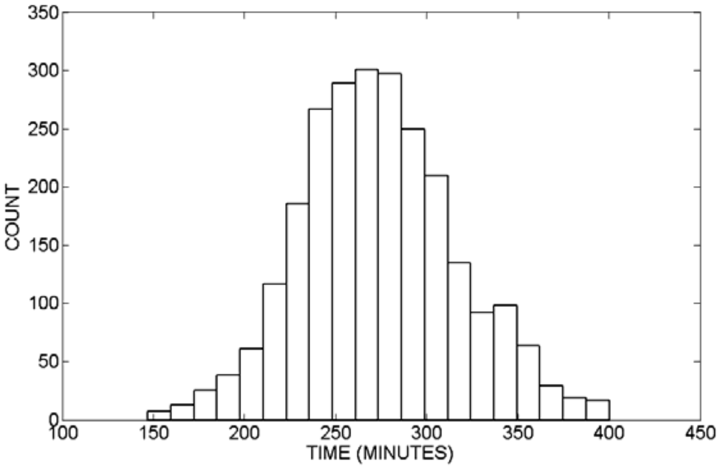
In this topic, your learning objectives are to:

- Be able to compute probabilities and percentiles for a normal distribution.
- Be able to approximate binomial probabilities using a normal curve.
- Understand the statement of the Central Limit Theorem regarding the sampling distribution of the mean.
- Be able to apply the Central Limit Theorem to find approximate probabilities regarding the sample mean of observations.

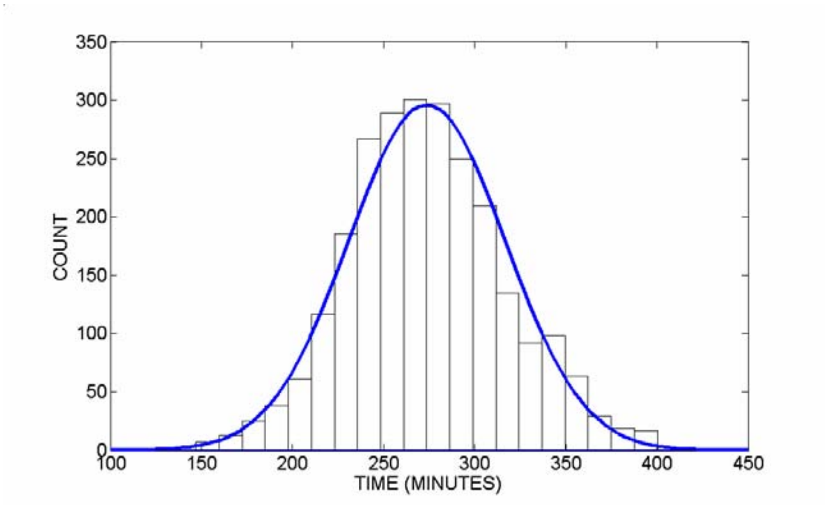
MODELING DATA BY A NORMAL CURVE

One of the most popular races in the United States is a marathon, a grueling 26-mile run. Most people are familiar with the Boston Marathon that is held in Boston, Massachusetts every April. But other cities in the U.S. hold yearly marathons. Here we look at data collected from Grandma's Marathon that is held in Duluth, Minnesota every June.

In the year 2003, there were 2515 women who completed Grandma's Marathon. The completion times in minutes for all of these women can be downloaded from the marathon's website. A histogram of these times, measured in minutes, is shown in the below figure.



Note that these measured times have a bell or mound shape. The figure below superimposes a normal curve on top of this histogram. Note that this curve is a pretty good match to the histogram. In fact, data like this marathon time data that are measurements are often well approximated by a normal curve.



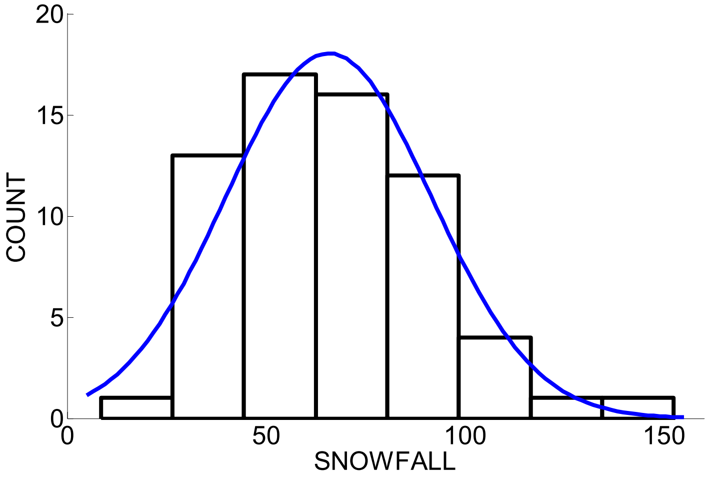
A normal probability curve has the general form

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2} \right), \quad -\infty < x < \infty .$$

This probability curve is described by two numbers -- the mean μ and the standard deviation σ. The mean μ is the center of the curve. Looking at the normal curve above, we see that the curve is centered about 270 minutes -- actually the mean of the normal curve is μ = 274. The number σ, the standard deviation, describes the spread of the curve. Here the normal curve standard deviation is σ = 43. If we know the mean and standard deviation of the normal curve, we can make reasonable predictions where the majority of times of the women runners will fall.

PRACTICE: MODELING DATA BY A NORMAL CURVE

Buffalo is an American city that receives a lot of snow. For each of the years 1941 through 2005, one observes the number of inches of snowfall during the winter months of December, January, and February. A histogram of the snowfall amounts is displayed below. Note that the data is approximately bell-shaped and a normal probability curve with mean μ = 65.8 inches and standard deviation σ = 25.9 inches is drawn on top of the histogram.



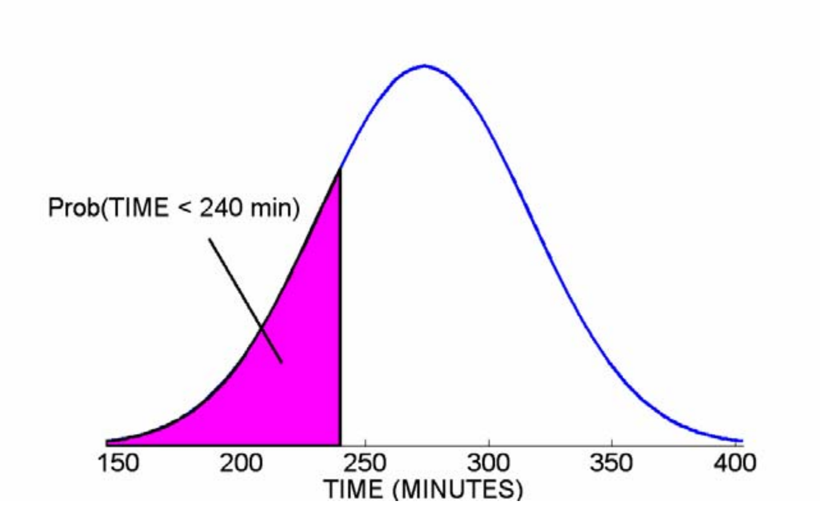
1. What is a typical snowfall amount for Buffalo for the months of December, January, and February?

2. One fact about the normal curve, discussed in Topic D3 is that 68% of the probability falls between the values μ – σ and μ + σ. Using this fact, find an interval that contains approximately 68% of the yearly snowfall amounts.

3. Another fact from Topic D3 is that 95% of the probability falls between the values μ – 2 σ and μ + 2 σ. Find an interval that contains approximately 95% of the snowfall amounts.

COMPUTING NORMAL PROBABILITIES

Suppose that the normal curve with μ = 274 minutes and σ = 43 minutes represents the distribution of women racing times. Say we are interested in the probability that a runner completes the race less than 4 hours or 240 minutes. We can compute this probability by finding an area under the normal curve. Specifically, as indicated in the figure below, this probability is the area under the curve for all times less than 240 minutes.



We can express this area as the integral

$$P(X \le 240) = \int_{-\infty}^{240} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2} \right) dx ,$$

but unfortunately we cannot integrate this function analytically (like we did in the examples of Topic P8) to find the probability. Instead we find this area by expressing it in terms of one special normal distribution, called the standard normal curve, and then use tables for this special normal curve to find probabilities.

Let X denote a normal random variable with mean μ and standard deviation σ. Suppose we wish to find the probability that X is smaller than a specific value, say x. By using some algebra, we can write this probability as

$$P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right).$$

Here is a useful fact. If X is a normal variable with mean μ and standard deviation σ, and we standardize it by computing

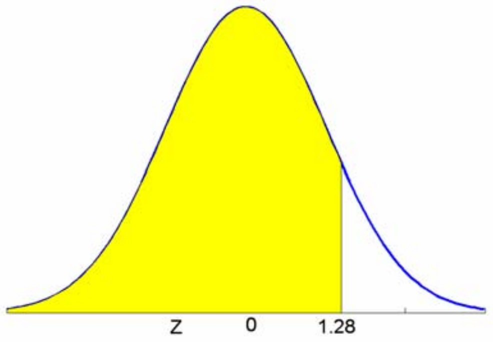
$$Z = \frac{X-\mu}{\sigma},$$

then the random variable Z has a normal distribution with mean 0 and standard deviation 1. So, continuing our work from above, we can write

$$P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P\left(Z \le \frac{x-\mu}{\sigma}\right),$$

and so the area under the normal(μ, σ) to the left of x will be the same as the area under the Z curve (mean 0, standard deviation 1) to the left of (x – μ)/σ.

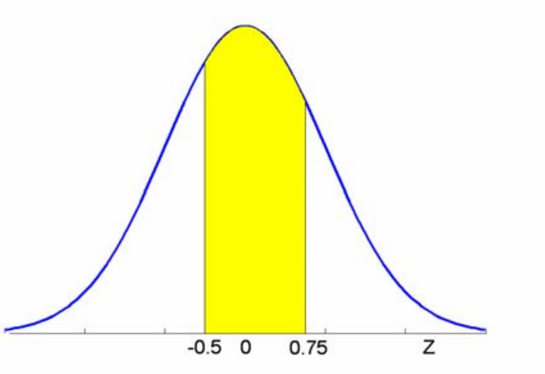
Tables xx and xx give probabilities under the standard normal (Z) curve to the left of specified values. Let's illustrate the use of these tables for three examples.



1. (Finding a "less than" area.) Suppose we wish to find P(Z < 1.28) which is the area under the Z curve to the left of 1.28. This is the type of area that is directly given in the table. We look up the value Z = 1.28 (the first two digits are contained on the right hand side of the table and the last decimal place is shown along the top of the table), and the corresponding area is shown in the body of the table. We see that

$$P(Z < 1.28) = 0.8997.$$

2. (Finding a "between two values" area.) Suppose we are interested in computing the probability that Z falls between two points, such as $P(-0.5 < Z < 0.75)$. One can write this probability as the difference of two "less than" probabilities:

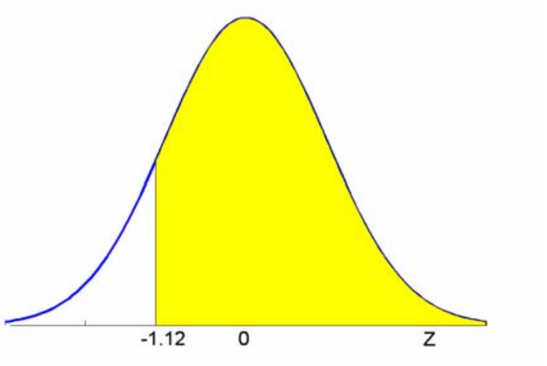


$$P(-0.5 < Z < 0.75) = P(Z < 0.75) – P(Z < -0.5).$$

So we look up the values Z = 0.75 and Z = -0.5 in the table – we find that the areas to the left of these two values are 0.7734 and 0.3085, respectively. So the probability of interest is

$$P(-0.5 < Z < 0.75) = 0.7734 – 0.3085 = 0.4649.$$

3. (Finding a "greater than" area.) Last, sometimes we will be interested in the probability that Z is greater than some value, such as $P(Z > -1.12)$. This probability can be found by noting (by the complement property of probabilities) that



$$P(Z > -1.12) = 1 - P(Z < -1.12).$$

We look up the value Z = -1.12 and find that the area to the left is equal to 0.1314. So

$$P(Z > -1.12) = 1 – 0.1314 = 0.8686.$$

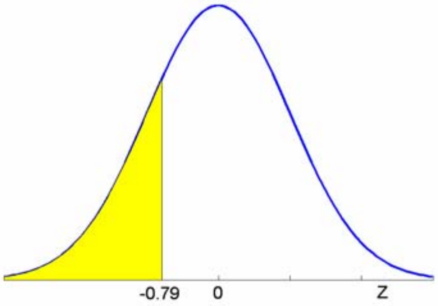
Let's return to our problem of determining the probability that a woman marathon runner complete the race in under 240 minutes. Recall the distribution of marathon times is approximately normally distributed with mean $\mu = 274$ and standard deviation $\sigma = 43$. We write the probability of interest as

$$P(X \le x) = P\left(\frac{X - 274}{43} \le \frac{240 - 274}{43}\right) = P\left(Z \le \frac{240 - 274}{43}\right) = P(Z \le -0.79) .$$

Equivalently, we find the probability that X is smaller than 240, by

(1) converting the time 240 to a z value by subtracting the mean and dividing by the standard deviation (that is, z = (240 – 273)/43 = -0.79)

(2) finding the area to the left of the z value under a standard normal curve. Looking up the value -0.79, we find the area to the left to be 0.2148. So P(X < 240) = 0.2148.



PRACTICE: COMPUTING NORMAL PROBABILITIES

Suppose the snowfall X of Buffalo for the three winter months is approximately normally distributed with mean $\mu = 65.8$ inches and standard deviation $\sigma = 25.9$ inches.

1. Find the Z score for the snowfall amount of 50 inches.

2. Find the Z score for the snowfall of 70 inches.

For each of the following, draw the normal curve and shade the probability (area) to be found.

3. Find the probability a snowfall is smaller than 50 inches.

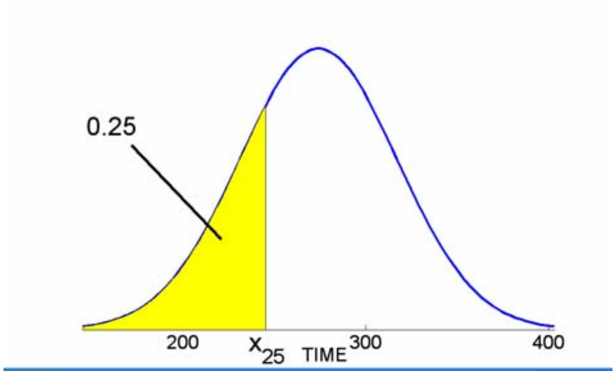
4. Find the probability it snows between 50 and 70 inches.

5. Find the probability it snows more than 70 inches in the winter months.

COMPUTING NORMAL PERCENTILES

In the above example, we were interested in computing a probability that was equivalent to finding an area under the normal curve. A different problem is to compute a percentile of the distribution. In our marathon running example, suppose that t-shirts will be given away to the runners who get the 25% fastest times. How fast does a runner need to run the race to get a t-shirt?

Here we wish to compute the 25th percentile of the distribution of times. This is a time, call it x_{25}, such that 25 percent of all times are smaller than x_{25}. This is shown graphically in the figure below.



Equivalently, we wish to find the value x_{25} such that

$$P(X < x_{25}) = 0.25.$$

As before, we find this percentile by finding the same percentile under the standard normal curve. We can reexpress the normal(μ, σ) as a standard normal probability as

$$P(X < x_{25}) = P\left(Z < \frac{x_{25} - \mu}{\sigma} \right) = 0.25.$$

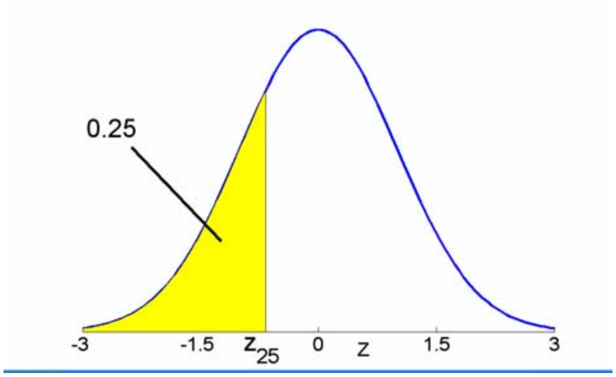
Using this equation, one sees that one can find our percentile in two steps. First, we find the same percentile, call it z_{25} of the Z curve, and then, by solving

$$z_{25} = \frac{x_{25} - \mu}{\sigma}$$

to get the percentile of interest

$$x_{25} = \sigma \; z_{25} + \mu \;\; ,$$

Here, to find the 25th percentile of the running times, we first find the 25th percentile under the standard normal curve as shown in the below figure.



Here we use our normal area table differently. We look in the body in the table for an area close to the target value 0.25, and then note the corresponding value of z. From the table, we see that the closest area to 0.25 is 0.2514, corresponding to z = -0.67. So

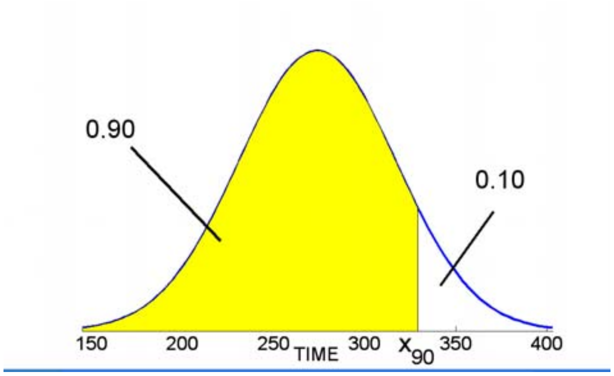
$$z_{25} = -0.67$$

and the 25th percentile of the running times is

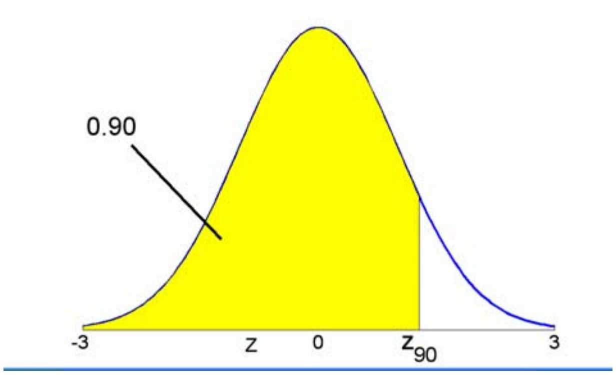
$$x_{25} = \sigma \; z_{25} + \mu = 43(-0.67) + 274 = 245.2$$

This means you need to run faster (lower) than 245.2 minutes to get a t-shirt in this competition.

Suppose you need to complete the race faster than 10% of the runners to be invited to run in the race the following year. How fast do you need to run? If you wish to have a 10% of the times to be larger than your time, this means that 90% of the times will be smaller than your time. That is, you wish to find the 90th percentile, x_{90} of the normal distribution (see the figure below).



We find this percentile the same way as the first example. We find the same percentile (the 90th) for the standard normal curve, and then convert the z percentile to the percentile of the running times distribution. Looking at the table, we look for a "less than" area equal to 0.90 (look at the below figure) – we see that the 90th percentile is equal to z_{90}= 1.28.



$$x_{25} = \sigma\, z_{90} + \mu = 43(1.28) + 274 = 329.0$$

So 329 minutes is the time to beat if you wish to be invited to participate in next year's race.

PRACTICE: COMPUTING NORMAL PERCENTILES

Consider again the snowfall X of Buffalo for the three winter months is approximately normally distributed with mean μ = 65.8 inches and standard deviation σ = 25.9 inches. For each of the following, draw a normal curve and label the probability that is given.

1. Find the 30th percentile of snowfall amounts.

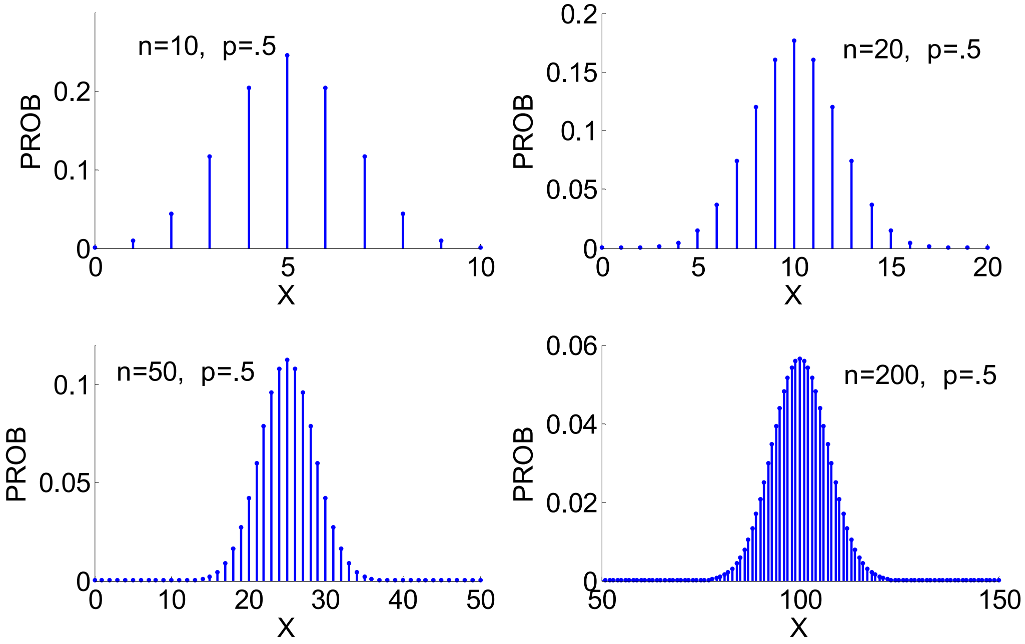
2. Find the 90th percentile.

3. Suppose it snows 100 inches during the three winter months and the meteorologist says that this amount is unusually large. Find the probability the snowfall is 100 inches or greater. Based on this calculation, do you agree that 100 inches is a usually large amount? Explain.

BINOMIAL PROBABILITIES AND THE NORMAL CURVE

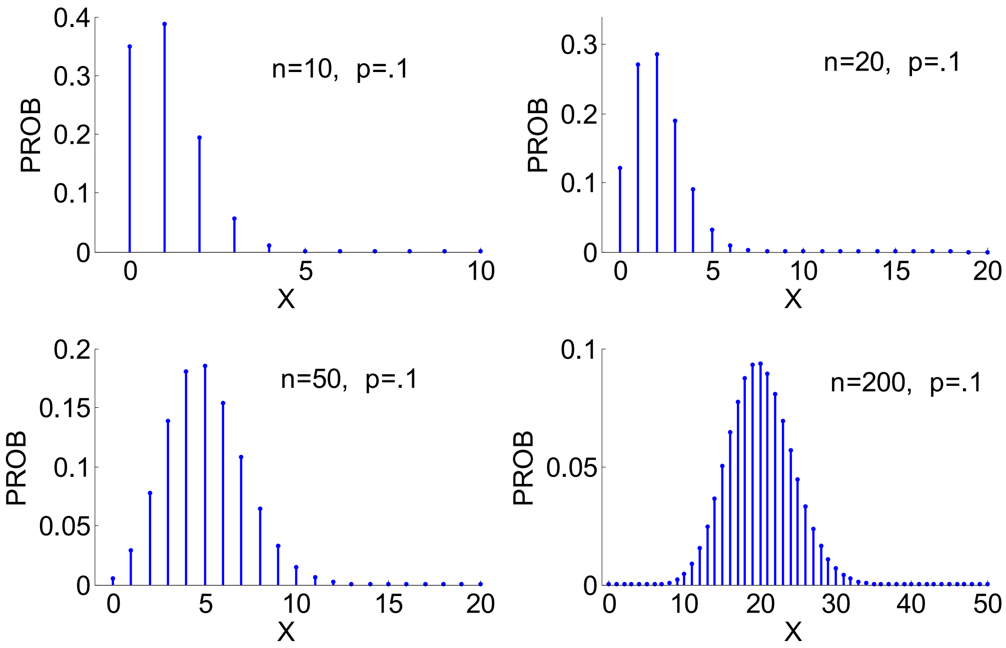
The normal curve is useful for modeling batches of data, especially when we are collecting measurements of some process. But the normal curve actually has a more important justification. We'll explore several important results about the pattern of binomial probabilities and sample means and we'll find these results useful in our introduction to statistical inference.

First, we explore different shapes of binomial distributions. Suppose that half of your student body is female and you plan on taking a sample survey of n students to learn if they are interested in using a new recreational sports complex that is proposed. Let X denote the number of females in your sample. We know (assuming a random sample is chosen) that X will be distributed binomial with parameters n and p=1/2. What is the shape of the binomial probabilities? The below figure displays the binomial probabilities for sample sizes n = 10, 20, 50, and 200.



What do we notice about this probability graphs? First, note that each distribution is symmetric about the mean $\mu = n/2$. But, more interesting, the shape of the distribution seems to resemble a normal curve as the number of trials n increases.

Perhaps this pattern happens since we started with a binomial distribution with p = .5 and you wouldn't see this behavior if you used a different value of p. Let's suppose that only 10% of all students would use the new facility and you let X denote the number of students in your sample who say they would use the facility. The random variable X would be distributed binomial with parameters n and p = .1. The following figure shows the probability distributions again for the sample sizes n = 10, 20, 50, and 200. As we might expect the shape of the probabilities for n=10 aren't very normal-shaped – the distribution is skewed right. But, note that as n increases, the probabilities become more normal-shaped and the normal curve seems to be a good match for n=200.



These graphs illustrate a basic result: if we have a binomial random variable X with n trials and probability of success p, then, as the number of trials n approaches infinity, the distribution of the standardized score

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

approaches a standard normal random variable. This is a very useful result. It means, that for a large number of trials, we can approximate a binomial random variable by a normal random variable with mean and standard deviation

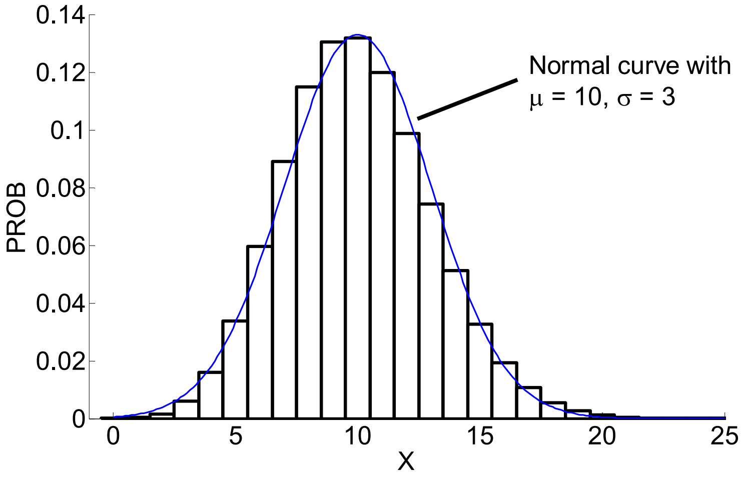
$$\mu = np, \quad \sigma = \sqrt{np(1-p)}\ .$$

We illustrate this result with our student survey example. Suppose that 10% of the student body would use the new recreational sports complex. You take a random sample of 100 students – what's the probability that 5 or fewer students in the sample would use the new facility?

The random variable X in this problem is the number of students in the sample that would use the facility. This random variable has a binomial distribution with n = 100 and p = .1 that is pictured as a histogram in the below figure. By the above result, this distribution can be approximated by a normal curve with

$$\mu = 100 \times .1 = 10, \quad \sigma = \sqrt{100 \times .1 \times .9} = 3\,.$$

This normal curve is placed on top of the probability histogram – note that it is a pretty good fit to the histogram.



We are interested in the probability that at most 5 students use the facility, that is, X ≤ 5. This probability can be approximated by the area under a normal curve ($\mu = 10, \sigma = 3$) between X=0 and X=5. Using the TI-83 calculator, we compute this normal curve area to be

```
normalcdf(0, 5, 10, 3) = .0474
```

In this case, one can also find this probability exactly by a calculator or computer program that computes binomial probabilities. Using the TI-83 calculator, we find the probability that X is at most 5 is

```
binomcdf(100, .1, 5) = .0576 .
```

Here we see that the normal approximation gives a similar answer to the exact binomial computation.

PRACTICE: BINOMIAL PROBABILITIES AND THE NORMAL CURVE

1. Suppose you roll 20 dice and you are interested in the random variable X that is equal to the number of sixes you observe. Explain why this is binomial experiment and define a "success" and give values of n and p.

2. By a direct calculation using the binomial formula, find the probability you observe at most three sixes, that is, P(X <= 2).

3. Use the normal approximation to compute the probability P(X <=2). How close is your approximation to the exact answer?

4. Suppose you instead roll 50 dice and you are interested in the number of sixes denoted by X. Find the probability P(X < = 5) using the binomial formula.

5. Use the normal approximation to compute the probability P(X < = 5). Again compare the exact and approximate answers to the probability.

6. In which case (n = 20 or n = 50) did you obtain a more accurate approximation using the normal curve? Can you explain why?

SAMPLING DISTRIBUTION OF THE MEAN

We have seen that binomial probabilities are well-approximated by a normal curve when the number of trials is large. There is a more general result about the shape of sample means that are taken from any population.

To begin our discussion about the sampling behavior of means, suppose we have a jar filled with a variety of candies of different weights. We are interested in learning about the mean weight of a candy in the jar. We could obtain the mean weight by measuring the weight for every single candy in the jar, and then finding the mean of these measurements. But that could be a lot of work.

Instead of weighing all of the candies, suppose we selected a random sample of 10 candies from the jar and found the mean $\bar{x}$ of the weights of these 10 candies. What have we learned about the mean weight of all candies from this sample information?

To answer this type of question, we

- Assume that we know about the weights of all candies in the jar.
- Look at the pattern of means that we get when we take random samples from the jar.

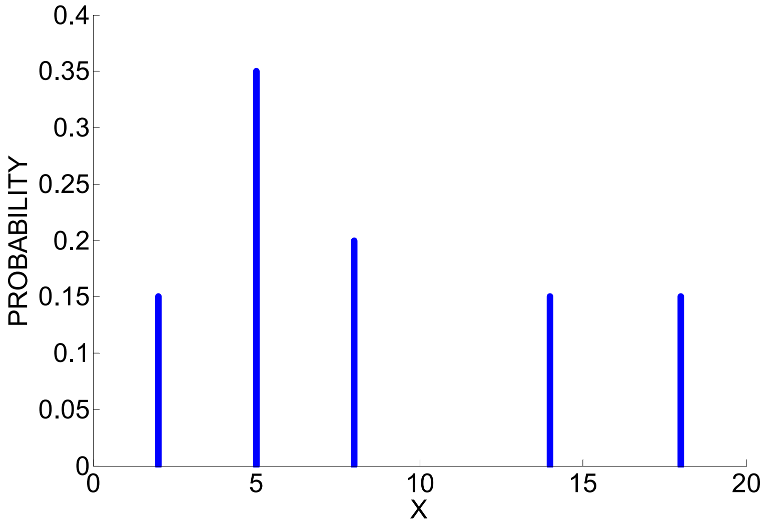
The group of items (here, candies) of interest is called the *population*. We assume first that we know the population – that is, we know exactly the weights of all candies in the jar. There are five types of candies – the table below gives the weight of each type of candy (in grams) and the proportion of candies of that type.

	X = Weight	Proportion
fruity square	2	.15
milk maid	5	.35
jelly nougat	8	.20
caramel	14	.15
candy bars	18	.15

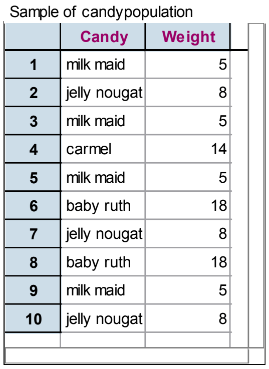
Let X denote the weight of a randomly selected candy from the jar. X is a discrete random variable with the probability distribution given in the table and a graph of the probabilities is shown in the following figure. We can summarize this distribution by computing a mean μ and a standard deviation σ. You will be asked in the exercises to verify that

$$\mu = 8.4500 \text{ and } \sigma = 5.3617.$$

So if we were really able to weigh each candy in the jar, we would find the mean weight to be $\mu = 8.4500$ grams.



Suppose we take a random sample of 10 candies with replacement from the jar and compute the mean $\bar{x}$ (we call this the *sample* mean to distinguish it from the *population* mean μ). We can simulate this sampling on a computer – I get the following candies:

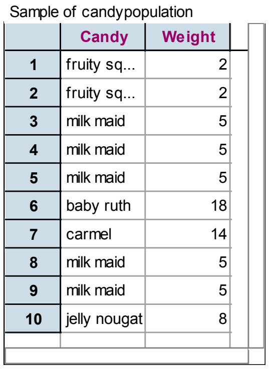


Sample of candypopulation

	Candy	Weight
1	milk maid	5
2	jelly nougat	8
3	milk maid	5
4	carmel	14
5	milk maid	5
6	baby ruth	18
7	jelly nougat	8
8	baby ruth	18
9	milk maid	5
10	jelly nougat	8

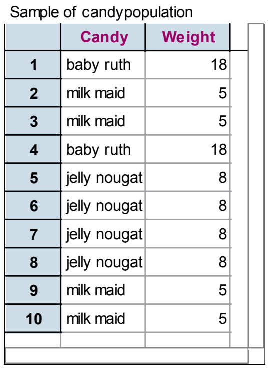
We compute the sample mean $\bar{x} = (5+8+5+14+5+18+8+18+5+8)/10 = 9.4$ gm.

Let's do this two more times – in the second sample, we obtain $\bar{x} = 6.9$ gm and in the third sample, we get $\bar{x} = 8.8$ gm.



Sample of candypopulation

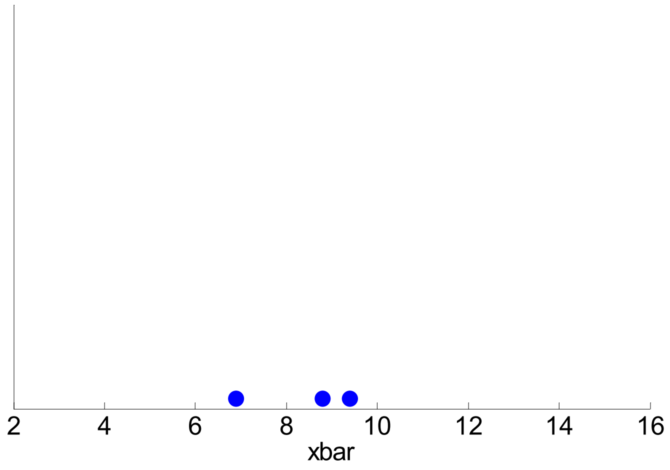
	Candy	Weight
1	fruity sq...	2
2	fruity sq...	2
3	milk maid	5
4	milk maid	5
5	milk maid	5
6	baby ruth	18
7	carmel	14
8	milk maid	5
9	milk maid	5
10	jelly nougat	8



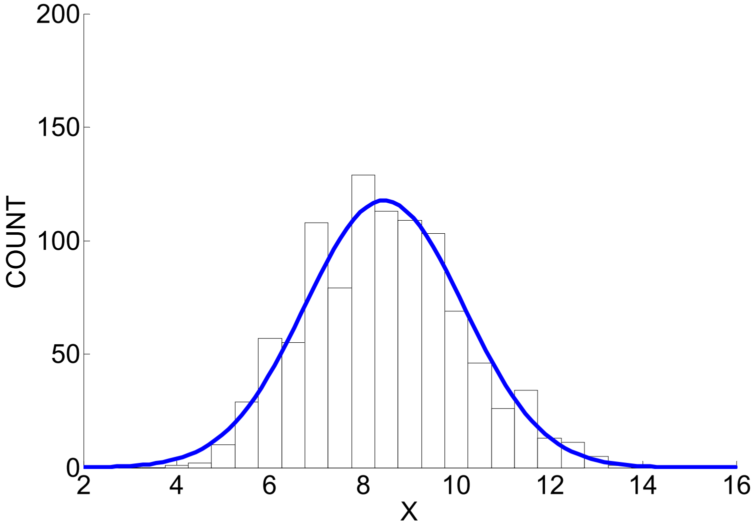
Sample of candypopulation

	Candy	Weight
1	baby ruth	18
2	milk maid	5
3	milk maid	5
4	baby ruth	18
5	jelly nougat	8
6	jelly nougat	8
7	jelly nougat	8
8	jelly nougat	8
9	milk maid	5
10	milk maid	5

We plot the three sample mean values on the below graph.



Suppose that we continue to take random samples of 10 candies from the jar and plot the values of the sample means on a graph – we obtain the *sampling distribution* of the mean $\overline{x}$.



Note that we get an interesting pattern of these sample means – they appear to have a normal shape.

This motivates an amazing result, called the Central Limit Theorem, about the pattern of sample means. If we take sample means from *any* population with mean μ and standard deviation σ, then the sampling distribution of the means (for large enough sample size) will be approximately normally distributed with mean and standard deviation

$$E(\bar{x}) = \mu, \quad SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$

Let's illustrate this result for our candy example. Recall that the population of candy weights had a mean and standard deviation given by μ = 8.4500 and σ = 5.3617, respectively. If we take samples of size n = 10, then, by this result, the sample means $\bar{x}$ will be approximately normally distributed where

$$E(\bar{x}) = 8.45, \quad SD(\bar{x}) = \frac{5.36}{\sqrt{10}} = 1.69$$

We have drawn this normal curve on top of the histogram of sample means above.

There are two important points to mention about this result.

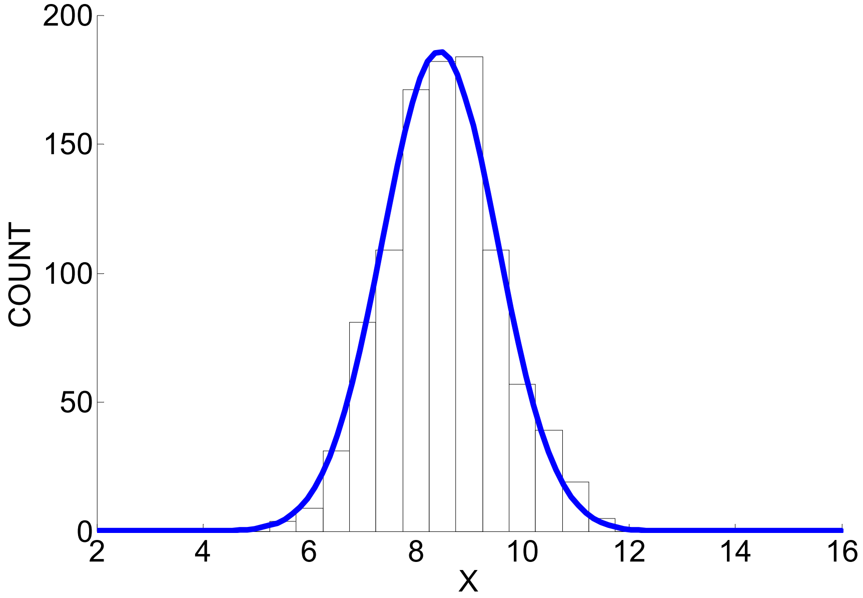
- First the expected value of the sample means, $E(\bar{x})$, is equal to the population mean μ. When we take a random sample, it is possible that the sample mean $\bar{x}$ is far away from the population mean μ. But, if we take many random samples, then, *on the average*, the sample mean will be close to the population mean.
- Second, note that the spread of the sample means, as measured by the standard deviation, is equal to $\sigma / \sqrt{n}$. Since the spread of the population is σ, note that the spread of the sample means will be *smaller* than the spread of the population. Moreover, if we take random samples of a *larger* size, then the spread of the sample means will *decrease*.

We can illustrate the last point in the context of our candy example. Above, we took random samples of size n = 10 and computed the sample means. Suppose instead we took repeated samples of size n = 25 from the candy jar – how does the sampling distribution of means change?

On a computer, we simulated the process of taking samples of size 25 – a histogram of the sample means is shown below. By the Central Limit Theorem, the sample means will be approximately normal-shaped with mean and standard deviation

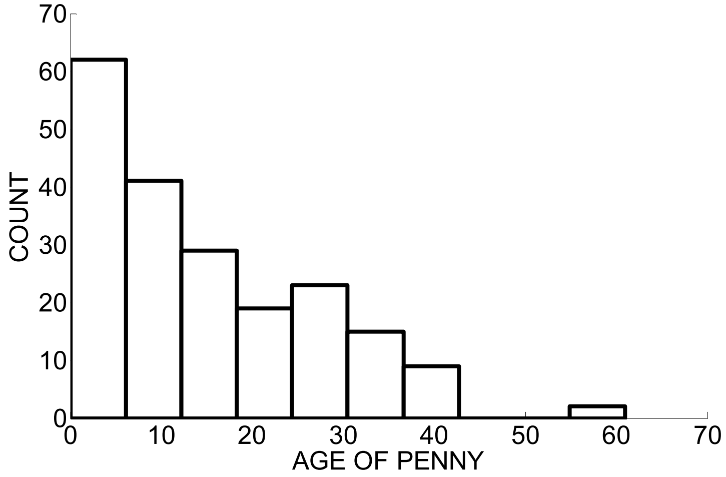
$$E(\bar{x}) = 8.45, \quad SD(\bar{x}) = \frac{5.36}{\sqrt{25}} = 1.07 .$$

Comparing the n = 10 sample means with the n = 25 sample means, what's the difference? Both sets of sample means are normally distributed with an average equal to the population mean. But the n = 25 sample means have a smaller spread – this means that as you take bigger samples, the sample mean $\bar{x}$ is more likely to be close to the population mean μ.



PRACTICE: THE CENTRAL LIMIT THEOREM

In my dresser drawer, I keep a lot of loose change. One night I recorded the ages in years of a large number of pennies from the drawer. A histogram of the population of penny ages is shown below. The mean and standard deviation of the ages are given by μ = 15.0 years and σ =12.1 years, respectively.



1. Describe the shape of the population. Would it be accurate to call this a normal-shaped distribution?

2. Suppose we take samples of ten pennies and for each sample compute the sample mean $\bar{x}$. By the Central Limit Theorem, what is the approximate shape of the distribution of $\bar{x}$'s, and find the mean and standard deviation of this sampling distribution.

3. By looking at the histogram, make an intelligent guess at the proportion of coins that have ages 3 years or less.

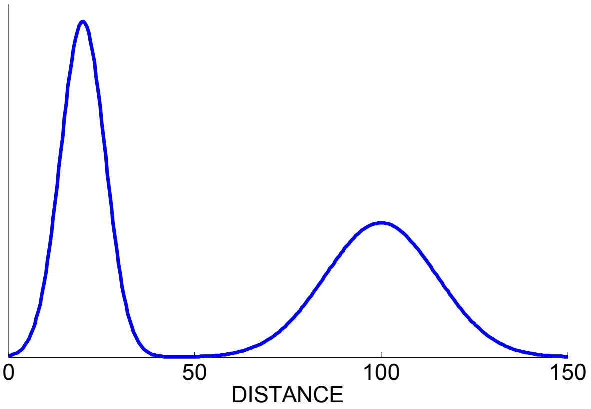
4. If you take a sample of 10 pennies, find the probability the mean age of the sample is 3 years or less.

5. Suppose that you took samples of size 25 instead of 10. How does that change the distribution of the sample mean? Would there be any change in the mean of the sampling distribution? Would there be any change in the standard deviation?

THE CENTRAL LIMIT THEOREM WORKS FOR ANY POPULATION

Let's illustrate the Central Limit Theorem for a second example where the population has a distinctive nonnormal shape. At my university, many of the students' hometowns are within 40 miles of the school. There also are a large number of students whose homes are between 80-120 miles of the university. Given the population of "distances of home" of all students, it is interesting to see what happens when we take random samples from this population.

If we let X denote "distance from home", imagine that the population of distances can be described by the continuous density curve below. We see two humps in this density – these correspond to the large number of students whose homes are in the ranges 0 to 40 miles and 70 to 130 miles. Although we won't describe the computations, one can show that the mean and standard deviation of this population are given by μ = 60 miles and σ = 41.6 miles, respectively.

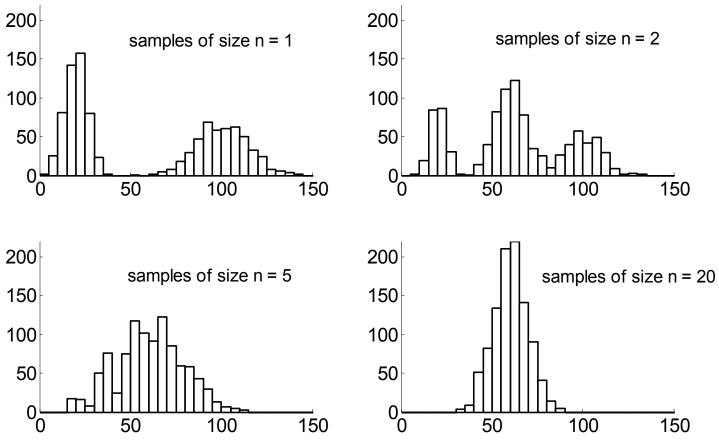


Now imagine that we take a random sample of n students from this population and compute the sample mean $\bar{x}$ from this sample. For example, suppose we take random samples of 20 students and collect the distances from home from these students – once we have collected the 20 distances, we compute the sample mean $\bar{x}$. Here are two samples and the values of $\bar{x}$:

```
Sample 1:
  102    22    23    24   114   102   114   102    22    19
   88    31    30   100   111   105    105   17   100    21     x̄ =67.6 mi.
Sample 2:
   12   127    33    34    73    19   111    99    16    20
   22    16    24    62    22    76    91   115   117    93     x̄ =59.1 mi.
```

If we repeat this sampling process many times, what will the distribution of sample means look like?

To answer this question, we have the computer simulate taking repeated samples of sizes n = 1, n = 2, n = 5, and n = 20. The histograms below show the distributions of sample means for the four sample sizes.



As we might expect, if we take samples of size 1, then our sample means look just like the original population. If we take samples of size 2, then the sample means have a fumy three-hump distribution. But, note as we take samples of larger sizes, then the sampling distribution of means looks more like a normal curve. Well, this is what we expect from the Central Limit Theorem result – no matter what the population shape, the distribution of the sample means will be approximately normal if the sample size is large enough.

What is the distribution of the sample means when we take samples of size n = 20? We just apply the Central Limit Theorem result. The sample means will be approximately normal with mean and standard deviation

$$E(\bar{x}) = \mu, \quad SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$

Since we know the mean and standard deviation of the population and the sample size, we just substitute these quantities and get

$$E(\bar{x}) = 60, \quad SD(\bar{x}) = \frac{41.6}{\sqrt{20}} = 9.3.$$

Let's use these results to answer some questions.

1. What is the probability that a student's distance from home is between 40 and 60 miles?

Actually this is a difficult question to answer exactly, since we don't know the exact shape of the population. But, looking at the graph of the population, we see that the curve takes on very small values between 40 and 60 miles. So this probability is close to zero – very few students live between 40 and 60 miles from our school.

2. What is the probability that, if we take a sample of 20 students, the mean distance from home for these twenty students is between 40 and 60 miles?

This is a different question than 1. We are asking about the chance that the sample mean $\bar{x}$ falls between 40 and 60 miles. Since the sampling distribution of $\bar{x}$ is approximately normal with mean 60 and standard deviation 9.3, we can compute this by tables or a calculator. Using the TI-83 Plus calculator, this probability $P(40 \le \bar{x} \le 60)$ is found by using the command

```
normalcdf(40, 60, 60, 9.3) = .484.
```

It is interesting to note that although it is unlikely for students to live between 40 and 60 miles from my school, it is pretty likely for the sample mean for a group of 20 students to fall between 40 and 60 miles.

3. What is the probability that the mean distance exceeds 100 miles?

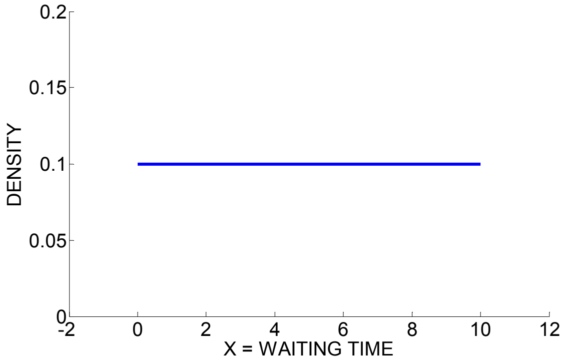
Here we want to find the probability that $\bar{x}$ is greater than 100, that is $P(\bar{x} > 100)$. On the calculator, we can find this by finding the probability of $\bar{x}$ between 100 and some very large value:

```
normalcdf(100, 200, 60, 9.3) = 8.5 10^(-6).
```

This probability is essentially zero, which means that it is highly unlikely that a sample mean of 20 student distances will exceed 100 miles.

PRACTICE: THE CENTRAL LIMIT THEOREM WORKS FOR ANY POPULATION

Suppose that the time that I wait in a line at a bank is uniformly distributed on the interval from 0 to 10 minutes.



1. If X denotes the time that I wait, find the formula for the probability density function for X.

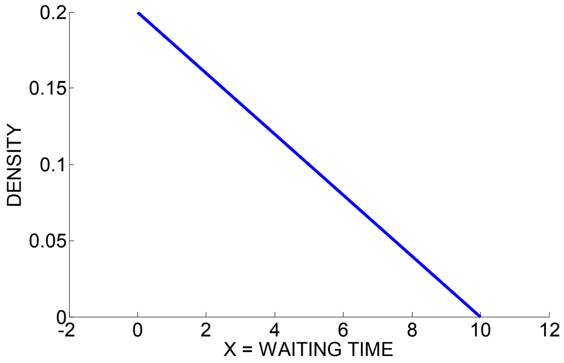
2. Find the mean and standard deviation of X.

3. Suppose I go to the bank 20 times in the fall and each time I record my waiting time. I compute the mean waiting time for my 20 visits $\overline{x}$. What is the (approximate) shape of the probability distribution for $\overline{x}$?

4. Find the mean and standard deviation of the sampling distribution of $\overline{x}$.

5. Find the probability that my average waiting time for the 20 visits exceeds 6 minutes.

6. Suppose that my waiting time X was not uniformly distributed but had the shape pictured below. Looking at the density function, we see that it is more likely to wait a short time. What would be the shape of the sampling distribution of means (for 20 visits) from this distribution?



TECHNOLOGY ACTIVITY: SAMPLING HEIGHTS

In this lab, we explore a dataset that contains the heights (in inches) of all the women who played in the WNBA (professional basketball) league in 2000. We'll take repeated samples of size 25 from this dataset. For each sample, we'll compute the mean, and look at the distribution of sample means.

1. First load the dataset of heights wba_heights.txt into Fathom.

2. Construct a dotplot of the heights. Write three sentences about this dataset, commenting on the general shape of the distribution, a typical height, and any unusual features that you see.

3. Find the names and heights of the tallest and shortest players in the WNBA. (An easy way to identify these players is by selecting the Height attribute, and Selecting Sort Ascending from the Data menu.)

4. Find the mean and standard deviation of the heights. We'll denote these values by μ and σ, respectively.

$\mu =$ $\sigma =$

5. Take a sample of size 25 without replacement from this dataset. (Select the Collection, select Sample Cases from the Analyze menu, inspect the Sample Collection –

turn off animation, make sure the With Replacement box is not checked, and sample 25 cases.)

6. Define a Measure from this Sample Collection (this will be the mean of the sample). Call the measure “sample_mean” – the formula for this measure is mean(height).

Write down the value of the sample mean for one sample. ___________

7. Now let’s repeat this process (taking a sample of 25 and computing the sample mean) 1000 times. Select the Sample Collection and select Collect Measures from the Analyze menu.

For the remainder of this assignment, we’ll focus on the Measures Collection. This collection contains the sample means from the 1000 samples you took from the WNBA heights.

8. Construct a histogram of the sample means. On the number line below, sketch the histogram that you see.

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
66   67    68    69    70    71    72    73    74    75    76 (inches)
```

9. Find the mean and standard deviation of the 1000 sample means.

μ (sample means) = σ (sample means) =

10. Find the probability (on Fathom) that the sample mean is larger than 73 inches.

11. Find the mean and standard deviation of the distribution of sample means. [Check your answer by comparing with your answer in 9.]

12. Using the Central Limit Theorem, find the probability that you will find a sample mean larger than 73. [Check your answer by comparing with your answer to 10.]

13. For a normal curve with mean μ and standard deviation σ, 95% of the probability falls in the interval (μ – 2 σ, μ + 2 σ). Find an interval where you expect 95% of the sample means to fall.

TECHNOLOGY ACTIVITY: ROLLING BIASED DICE

Suppose we are rolling a collection of biased dice. In this activity, we'll decide how the dice are biased, and use Fathom to roll the biased dice and find probabilities of the sum of the dice.

1. First construct your biased dice. In the boxes below, put whole number weights in the boxes so that two numbers are more likely to be rolled than the remaining numbers. (For example, if you assign the weights 5, 1, 1, 1, 1, 5 to the numbers, this means that 1 and 6 are most likely to be rolled.) Also compute the probabilities of each roll.

Roll	⚀	⚁	⚂	⚃	⚄	⚅
Weight						
Probability						

2. If X denotes the die roll, use the above probability distribution to find the mean μ and standard deviation σ of X.

To roll 5 of your biased dice, open the Fathom program
`rolling_biased_dice.ftm`

Here are the steps for rolling your biased dice on Fathom:

Step 1: To set up your biased dice, you put the numbers 1, 2, 3, 4, 5, 6 in Collection 1 where the number of 1's, 2's, etc. correspond to your weights. (For example, if my weights are 5, 1, 1, 1, 1, 5, then I would place five 1's, one 2, one 3, one 4, one 5, and five 6's in Collection 1.)

Step 2: Set n in the Fathom Slider to be 5 since you'll be rolling five dice.

Step 3: To roll your five dice 500 times, push the "Collect More Measures" button. Each time five dice are rolled, Fathom will find the sum of the rolls, and the Measures Collection will contain 500 values of the sum of rolls.

3. Look at the histogram of the sum of rolls. Describe this sampling distribution, including comments about its shape, average value, and spread.

4. Let S denote the sum of your five rolls. Using Fathom, find

(a) E(S) and SD(S)

(b) $P(15 \le S \le 20)$

(c) $P(S \ge 25)$

5. Your five rolls can be considered to be a random sample of size n = 5 from your population defined by the probability distribution of X. By the Central Limit Theorem, the sum of the five rolls $S = \sum_{i=1}^{n} X_i$ can be approximated by a normal curve with mean $n\mu$ and standard deviation $\sqrt{n}\ \sigma$. Using this approximation, compute parts (b) and (c) of part 3. Compare your answers to the Fathom simulation answers (they should be close).

6. The accuracy of the Central Limit Theorem approximation should improve for larger values of the sample size n. To check this, rerun your simulation for n = 10 rolls of your biased dice. Compute the probability $P(S \geq 40)$ two ways (using the Fathom output and the normal approximation) and comment on the accuracy of the approximation.

WRAP-UP

This topic introduced the normal distribution that is the most popular distribution for modeling continuous measurements. This distribution is described by two parameters: the mean μ and the standard deviation σ. By use of normal tables, we can find probabilities or areas under an arbitrary normal curve. Also the tables can be used to find percentiles of a normal distribution. Binomial distributions for a large number of trials n resemble normal curves, and one can accurately approximate binomial probabilities by the normal curve. One of the most important applications of the normal distribution is the Central Limit Theorem that states that the sampling distribution of means from any population tends to be approximately normal. By use of this result, one can approximate probabilities about the sample mean of measurements if one is given the mean and standard deviation of the population.

EXERCISES

1. **Heights of Men**

Suppose heights of American men are approximately normally distributed with mean 70 inches and standard deviation 4 inches.

a. What proportion of men is between 68 and 74 inches?

b. What proportion of men is taller than 6 feet?

c. Find the 90th percentile of heights.

2. **Test Scores**

Test scores in a precalculus test are approximately normally distributed with mean 75 and standard deviation 10. If you choose a student at random from this class

a. What is the probability he or she gets an A (over 90)?

b. What is the probability he or she gets a C (between 70 and 80)?

c. What is the letter grade of the lower quartile of the scores?

3. **Body Temperatures**

The normal body temperature was measured for 130 subjects in an article published in the Journal of the American Medical Association. These body temperatures are approximately normally distributed with mean $\mu = 98.2$ degrees and standard deviation $\sigma = 0.73$.

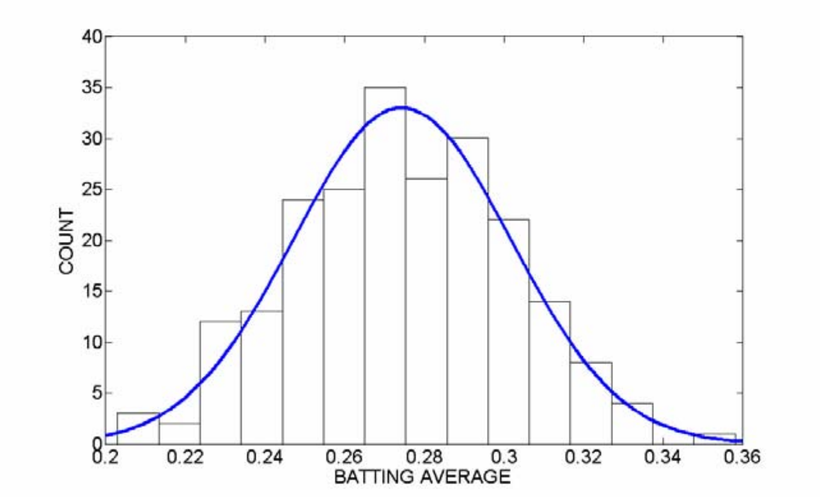
a. Most people believe that the mean body temperature of healthy individuals is 98.6 degrees, but actually the mean body temperature is smaller than 98.6. What proportion of healthy individuals have body temperatures smaller than 98.6?

b. Suppose a person has a body temperature of 96 degrees. What is the probability of having a temperature less than or equal to 96 degrees? Based on this computation, would you say that a temperature of 96 degrees is unusual? Why?

c. Suppose that a doctor diagnoses a person as sick if his or her body temperature is above the 95th percentile of the temperature of "healthy" individuals. Find this body temperature that will give a sick diagnosis.

4. **Baseball Batting Averages**

Batting averages of baseball players can be well approximated by a normal curve. The figure below displays the batting averages of players during the 2003 baseball season with at least 300 at-bats (opportunities to hit). The mean and standard deviation of the matching normal curve shown in the figure are $\mu = 0.274$ and $\sigma = 0.027$, respectively.



a. If you choose a baseball player at random, find the probability his batting average is over .300. (This is a useful benchmark for a "good" batting average.)

b. Find the probability this player has a batting average between .200 and .250.

c. A baseball player is said to hit below the Mendoza line (named for weak-hitting baseball player Minnie Mendoza) if his batting average is under .200. Given our model, find the probability that a player hits below the Mendoza line.

d. Suppose that a player has an incentive clause in his contract that states that he will earn an additional $1 million if his batting average is in the top 15%. How well does the player have to hit to get this additional salary?

5. **Emergency Calls**

Suppose that the AAA reports that the average time it takes to respond to an emergency call on the highway is 25 minutes. Assume that the times to respond to emergency calls are approximately normally distributed with mean 25 minutes and standard deviation 4 minutes.

a. If your car gets stuck on a highway and you call the AAA for help, find the probability that it will take longer than 30 minutes to get help.

b. Find the probability that you'll wait between 20 and 30 minutes for help.

c. Find a time such that you are 90% sure that the wait will be smaller than this number.

6. **Buying a Battery for your iPod**

Suppose you need to buy a new battery for your iPod. Brand A lasts an average of 11 hours and Brand B lasts an average of 12 hours. You plan on using your iPod for eight hours on a trip and you want to choose the battery that is most likely to last 8 hours (that is, have a life that is least as long as 8 hours).

a. Based on this information, can you decide which battery to purchase? Why or why not?

b. Suppose that the battery lives for Brand A are normally distributed with mean 11 hours and standard deviation 1.5 hours, and the battery lives for Brand B are normally distributed with mean 12 hours and standard 2 hours. Compute the probability that each battery will last at least 8 hours.

c. On the basis of this calculation in part b, what battery should you purchase?

7. **Lengths of Pregnancies**

It is known that the lengths of completed pregnancies are approximately normally distributed with mean 266 days and standard deviation 16 days.

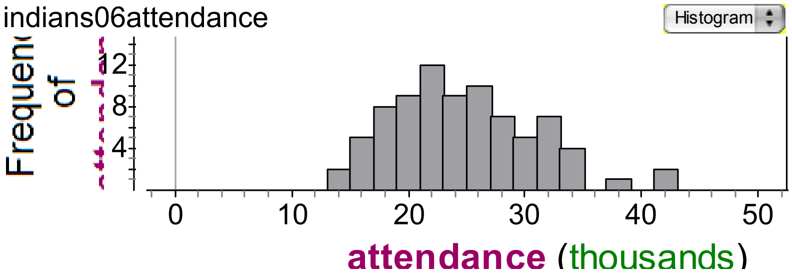
a. What is the probability a pregnancy will last more than 270 days?

b. Find an interval that will contain the middle 50% of the pregnancy lengths.

c. Suppose a doctor wishes to tell a mother that he is 90% confident that the pregnancy will be shorter than x days. Find the value of x.

8. **Attendances at Baseball Games**

The following figure displays the attendance for all of the home games of the Cleveland Indians for the 2006 baseball season. This distribution can be approximated by a normal curve with mean m = 24,667 and standard deviation s = 6144.



Consider the attendance for one randomly selected game during the 2006 season.

a. Find the probability the attendance exceeds 30,000.

b. Find the probability the attendance is between 20,000 and 30,000.

c. Suppose that the attendance at one game in the following season is 12,000. Based on the normal curve, compute the probability that the attendance is at most 12,000. Based on this computation, is this attendance unusual? Why?

9. **Coin Flipping**

Suppose you flip a fair coin 1000 times.

a. How many heads do you expect to get?

b. Find the probability that the number of heads is between 480 and 520.

c. Suppose your friend gets 550 heads. What is the probability of getting at least 550 heads? Do you believe that your friend's coin really was fair? Explain.

10. **Use of On-Line Banking Services**

Suppose that a newspaper article claims that 80% of adults currently use on-line banking services. You wonder if the proportion of adults who use on-line banking services in your community, p, is actually this large. You take a sample of 100 adults and 70 tell you they use on-line banking.

a. If the newspaper article is accurate, find the probability that 70 or fewer of your sample would use on-line banking.

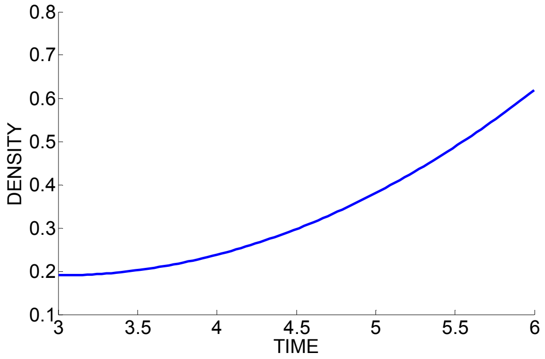
b. Based on your computation, is there sufficient evidence to suggest that less than 80% of your community use on-line banking services? Explain.

11. **Time to Complete a Race**

Suppose a group of children are running a race. The times (in minutes) that the children complete the race can be described by the density function

$$f(x) = \frac{4+(x-3)^2}{21}, \quad 3 < x < 6.$$

A graph of this density is shown below. The mean and standard deviation of this density are given by 4.83 and .84 minutes, respectively.



a. Suppose 25 students run this race and you find the mean completion time. Find the probability that the mean time exceeds 5 minutes.

b. Find an interval that you are 90% confident contains the mean completion time for the 25 students.

12. **Snowfall Accumulation**

Your local meteorologist has collected data on snowfall for the past 100 years. Based on these data, you are told that the amount of snowfall in January is approximately normally distributed with mean 15 inches and standard deviation 4 inches.

a. Find the probability you get more than 20 inches of snow this year.

b. In the next ten years, find the probability that the average snowfall (for these ten years) will exceed 20 inches.

13. **Total Waiting Time at a Bank**

You are waiting to be served at your bank. From past experience, you know that your time to be served is uniformly distributed between 0 and 10 minutes.

a. Find the mean and standard deviation of your waiting time.

b. The Central Limit Theorem can be also stated in terms of the sum of random variables. If the random variables $X_1, \ldots, X_n$ represent a random sample drawn from a population with mean μ and standard deviation σ, then the sum of random variables $S = \sum_{i=1}^{n} X_i$, for large sample size n, will be approximately normally distributed with mean $n\mu$ and standard deviation $\sqrt{n}\,\sigma$. Suppose you wait at the bank for 30 days.

Use this version of the Central Limit Theorem to find the probability that your total waiting time will exceed three hours.

14. **Total Errors in Check Recording**

Suppose you record the amount of a written check to the nearest dollar. It is reasonable to assume that the error between the actual check amount and the written amount is uniformly distributed between -$0.50 and +$0.50.

a. Find the mean and standard deviation of one error.

b. Suppose you write 100 checks in a single month and S denotes the total error in recording these checks. Find the probability that S is smaller than $5. (Use the version of the Central Limit Theorem described in Exercise 5.)

c. Find an interval of the form (-c, c) so that P(-c < S < c) = .95.

15. **Distribution of Measurements**

Suppose that a group of measurements is approximately normally distributed with mean μ and standard deviation σ.

a. Find the probability that a measurement falls within one standard deviation of the mean.

b. Is it likely that you collect a measurement that is larger than $\mu+3\sigma$? Explain.

c. Find an interval that contains the middle 50% of the measurements.

16. **Salaries of Professional Football Players**

Suppose you learn that the mean salary of all professional football players this season is 7 million dollars with a standard deviation of 2 million dollars.

a. Do you believe that the distribution of salaries is approximately normally distributed? If your answer is no, sketch a plausible distribution for the salaries.

b. From your graph, find an approximate probability that a salary is smaller than $6 million.

c. Suppose you take a random sample of 30 salaries. Find the probability that the mean salary for this sample is smaller than $6 million.

17. **Weights of Candies**

In the candy bowl example, the probability distribution of the candy weight X is given in the following table.

	X = Weight	Proportion
fruity square	2	.15
milk maid	5	.35
jelly nougat	8	.20
caramel	14	.15
candy bars	18	.15

Verify by calculation that the mean and standard deviation of X are given by μ = 8.4500 and σ = 5.3617, respectively.

18. **Sleeping Times**

Suppose sleeping times of college students are approximately normally distributed. You are told that 25% of students sleep less than 6.5 hours and 25% of students sleep longer than 8 hours. Given this information, determine the mean and standard deviation of the normal distribution.

TOPIC I1: INTRODUCTION TO STATISTICAL INFERENCE

A CLASSROOM SURVEY

Currently digital audio players are very popular in the United States. People can easily download music from their computers and play music at home, on the go, or in their cars. Among the digital audio players, the most popular models are variations of the iPod manufactured by Apple Computer.

That raises the interesting question: How popular is the iPod among college students at my school? To learn about the iPod's popularity, I plan to conduct a survey of a small group of students and ask each student if they own an iPod.

I have just described a simple illustration of statistical inference. We are interested in learning about the characteristics of a large group of college students, and we gain information about this large group by means of information collected in the survey.

NCTM Standards

✓In Grades 9-12, all students should understand the distinction between a statistic and a parameter.

✓In Grades 9-12, all students should use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions.

✓In Grades 9-12, all students should understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference.

POPULATION, PARAMETER, SAMPLE, AND STATISTIC

Statistical inference is essentially the process of learning about a large group of people or objects on the basis of a small group selected from the large group. The large group that we wish to learn about is called the *population*. In our example, the population would be the group of college students from my college. Precisely defining the population is important. I am not interested in the buying habits of all Americans or even all college students – I'm only interested in the habits of the students at my school and so that is the relevant population.

A *parameter* is a numerical summary of the population. If the population is the group of college students at my school, then possible parameters might be their average age, the proportion of students that are education majors, or the proportion that have visited Hawaii in their lifetime. In this example, since I'm interested in learning about the popularity of iPods, a suitable parameter would be

p = proportion of students that own iPods

What is the value of p? We don't know since a parameter is a characteristic of a population and we wouldn't know p unless we were able to ask every student if he or she owned an iPod. There are approximately 20,000 students at my school, so it would be almost impossible for me to survey every student and find the exact value of p.

A *sample* is the subset of the population that we sample to learn about the value of p. In this example, I may survey the students in one of my classes and the survey results would be my sample. There are actually many ways for me to take my sample. I could call 20 "random" students on the phone; I could interview 20 students in front of the one of the buildings on campus, or perhaps I could ask 20 students who visit me for advising help. All of these methods would give me a sample of students that would help me learn about the characteristics of the population, all students who attend my school.

A *statistic* is a number that we compute from the sample. Obviously there are many possible numbers that we could compute from the sample – we focus on the statistic that we think will help us learn about the population parameter. In my class of 20 students I ask each the question: Do you own an iPod? Of my students, 3 said "yes" and 17 said "no". One statistic would be the proportion of students in my sample that owned iPods that we denote by $\hat{p}$:

$$\hat{p} = \frac{3}{20} = .15.$$

Does that mean that the population proportion p is equal to .15? No – we only know that 15% of our sample of students owns an iPod. We will never know the value of p since we can't collect data from the entire population. But we hope that the value of p will be close to the value of $\hat{p}$. One goal of this topic is to construct an interval of values that we are confident contains the unknown value of p.

PRACTICE: POPULATION, PARAMETER, SAMPLE, STATISTIC

In each of the following situations, describe the population of interest and define the parameter. In addition, what is the sample and the statistic that is collected from the sample?

1. Researchers from McGill University conduced a web-based survey of 277 single adults, including 196 women and 81 men. They found that women were 50% more likely than men to feel guilty about indulging in "comfort" foods containing high amounts of sugar and fat.

2. Americans were asked in a Gallup survey in November 2005 how much money they plan to spend for Christmas gifts. They found that 30% of the adults intended to spend $1000 or more on gifts.

3. The Drive for Life poll, a survey on drivers' attitudes conducted by Mason-Dixon Polling and Research, conducted a telephone survey of 1100 licensed drivers in America. One result of the survey was that 31 percent thought that talking on a cell phone while driving was the most annoying behavior of other drivers.

SAMPLE ESTIMATES – BIAS AND VARIANCE

In statistical inference, we are interested in learning about a parameter from the population of interest. We take a random sample, and compute a quantity, called an *estimate*, that is a good guess at the value of the parameter. Suppose we have several possible estimates – how can we say that one estimate is better than other estimate? We illustrate two important properties of estimators by use of a simple example.

Returning to our iPod example, suppose that 30% of the students at my school own iPods – that is, we know p = .3. We are going to select a sample at random and ask each student "Do you own a iPod?" Here are the results from one hypothetical sample:

STUDENT	1	2	3	4	5	6	7	8	9	10
	NO	NO	NO	YES	YES	YES	NO	NO	YES	YES
STUDENT	11	12	13	14	15	16	17	18	19	20
	YES	NO	NO	NO	YES	YES	NO	NO	NO	YES

Here are three possible estimators of the parameter p that we will call *all data*, *some data*, or *no data* that we define as follows.

1. The *all-data estimator*, denoted by $\hat{p}_{ALLDATA}$, is simply the proportion of students who own iPods in the complete sample. Here we note that 9 out of the 20 students answered yes to the question, so $\hat{p}_{ALLDATA}$=9/20 = .45.

2. The *some-data estimator*, denoted by $\hat{p}_{SOMEDATA}$, is defined to be the proportion of the first ten students who own iPods. Here 5 of the first ten students answered yes, so $\hat{p}_{SOMEDATA}$ = 5/10 = .5. One might use this estimator thinking that a sample of ten students is large-enough to learn about the actual value of p.

3. The *no-data estimator*, denoted by $\hat{p}_{NODATA}$, ignores the data completely and is based on one's prior opinion about the value of p. In this example, suppose that the investigator strongly believes that the proportion of iPod users on campus is 20%, so $\hat{p}_{NODATA}$ = .2.

Which is the best estimator of p? If we knew that the actual proportion of iPod users was p = .3, then we see which estimator is closest to p. Here, we computed $\hat{p}_{ALLDATA}$=.45, $\hat{p}_{SOMEDATA}$ = .5, and $\hat{p}_{NODATA}$ = .2, and we see that $\hat{p}_{NODATA}$ is the best estimator in the sense that it is closest to the actual value of p.

But this conclusion about "best estimator" depends on the particular sample we observe. Suppose we take a new sample of 20 students with the following results:

STUDENT	1	2	3	4	5	6	7	8	9	10

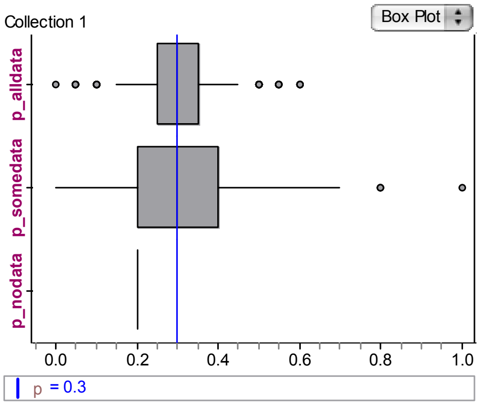
```
        NO   NO   NO YES YES   NO   NO   NO   NO   NO
STUDENT 11   12   13   14   15   16   17   18   19   20
       YES   NO   NO   NO YES   NO   NO YES YES   NO
```

For this sample $\hat{p}_{ALLDATA}$=.3, $\hat{p}_{SOMEDATA}$ = .2, and $\hat{p}_{NODATA}$ = .2, and in this case $\hat{p}_{ALLDATA}$ is closest to the actual value of p = .3.

We see that the *best* estimator for p will depends on the particular sample. Since the result of our sample is unpredictable, we can't be confident one estimator such as $\hat{p}_{ALLDATA}$ will be the best estimator for a single sample.

For this reason, we judge the goodness of estimators by looking at their behavior for a large number of samples. We look at many random samples and look at the batches of $\hat{p}_{ALLDATA}$, $\hat{p}_{SOMEDATA}$, and $\hat{p}_{NODATA}$ we obtain. Our criteria for best estimator will depend on the distributions of these estimators over many samples.

We perform a simple simulation on Fathom to understand the sampling variation of these three estimators. We still assume that the actual value of p is equal to .3 and simulate many samples of size 20 from a population where the proportion of iPod users is equal to .3. For each sample, we compute our three estimators. We take a total of 1000 samples, and collect 1000 values of $\hat{p}_{ALLDATA}$, 1000 values of $\hat{p}_{SOMEDATA}$, and 1000 values of $\hat{p}_{NODATA}$. The figure below displays parallel boxplots of the batches of the three estimators.



The first important property of an estimator is *bias*. If $\hat{p}$ denotes an arbitrary estimator, then the bias of $\hat{p}$ is the difference between the average value of $\hat{p}$ over many samples and the actual value of p. In other words,

$$\text{BIAS}(\hat{p}) = \text{E}(\hat{p}) - \text{p}.$$

It is desirable for an estimator to be *unbiased* where BIAS($\hat{p}$) = 0. Over many samples, an unbiased estimator will, on average, be equal to the actual value of the parameter. That is, an unbiased estimator will tend to be on-target, and will not generally overestimate or underestimate the parameter.

How do our three estimators do with respect to bias? In the above figure, the vertical line shown is located at the actual parameter value p = 3. It seems that the boxplots of $\hat{p}_{ALLDATA}$ and $\hat{p}_{SOMEDATA}$ are centered about the line. Out of the 1000 samples, the mean value of $\hat{p}_{ALLDATA}$ is equal to .29985 and the mean value of $\hat{p}_{SOMEDATA}$ .3018. In fact, it can be shown that both $\hat{p}_{ALLDATA}$ and $\hat{p}_{SOMEDATA}$ are unbiased, that is, BIAS($\hat{p}_{ALLDATA}$) = 0 and BIAS($\hat{p}_{SOMEDATA}$) = 0. In contrast, since the value of $\hat{p}_{NODATA}$ is always equal to .2, this no-data estimator is biased and one can compute BIAS($\hat{p}_{NODATA}$) = E($\hat{p}_{NODATA}$) – p = .2 - .3 = .1.

Generally, we prefer unbiased estimators such as $\hat{p}_{ALLDATA}$ and $\hat{p}_{SOMEDATA}$ that are equal to the actual parameter value p on average in repeated sampling. But there is a second important property in estimation. Given two unbiased estimators, we prefer the one that tends to be closer to the parameter p in repeated sampling. Looking again at the parallel boxplot display, we see that 50% of the values of $\hat{p}_{SOMEDATA}$ are between .2 and .4 and 50% of the values of $\hat{p}_{ALLDATA}$ are between .25 and .35. The values of the all-data estimator are less variable, which means that $\hat{p}_{ALLDATA}$ is more likely than $\hat{p}_{SOMEDATA}$ to be closer to the parameter value p.

We can measure the second property by the use of a standard deviation or another measure of spread of the values of the estimator in repeated sampling. From Fathom, we compute

$$\text{SD}(\hat{p}_{ALLDATA}) = .1024,\ \text{SD}(\hat{p}_{SOMEDATA}) = .1464,$$

which confirms that the all-data estimator has less variation than the some-data estimator.

To summarize, we evaluate estimators by the pattern of their performance across repeated sampling. We prefer unbiased estimators that tend to be equal to the parameter value on average. Both the $\hat{p}_{ALLDATA}$ and $\hat{p}_{SOMEDATA}$ estimators were unbiased. Our no-data estimator will be biased whenever the actual parameter value is different from $\hat{p}_{NODATA}$ = .2. If we wish to compare unbiased estimators, then we prefer the estimator with the smaller variability in repeated sampling. On the surface, the some-data estimator didn't make any sense – why ignore half of the sample to compute the estimator? By taking a proportion of yeses among 10 students instead of 20 students, one increases the variability of the estimator $\hat{p}$. Generally, we will see that the variation of the sample proportion $\hat{p}$ over many samples will decrease as a function of the sample size n.

TECHNOLOGY ACTIVITY: THE TAXI PROBLEM

Suppose you are wandering the streets of a city. You notice that a number identifies each taxi passing by. You observe the five numbers

345, 167, 91, 125, 92

From this information, can you make an intelligent guess at the number of taxis in this city?

This is an illustration of a simple problem in statistical inference. Here the population is the collection of taxis that drive in this city. We are interested in a particular parameter – specifically the number N of taxis. We can't directly observe N, but we get information about this parameter by taking a sample of taxis and recording their numbers.

Let's make several assumptions that will simplify this problem. First, we assume that the taxis are numbered from 1 to N; that is 1, 2, 3, … , N. Second, we assume that you are equally likely to observe any one of the N taxis at a given time. We are making the probably unreasonable assumption that each taxi drives through the area where you are walking. Last, we assume that your observations are independent. This means that if you first observe, say taxi number 5, then you are still equally likely to observe any one of the N taxi numbers in the next sighting.

How can we use our five observed numbers (345, 167, 91, 125, 92) to make an intelligent guess at the total number of taxis N? I describe three possible guesses, or more formally estimates of N based on these data.

1. I can estimate N by the maximum or largest of these five numbers. Here I would estimate N by the largest number 345.
2. The first method for estimating N might seem too small. Although 345 is a possible value of N, it does seem likely that N is larger than 345 since we observed the number on only five taxis. We decide to estimate N using the formula

$$N = (6/5) \text{ (maximum observation).}$$

Here our estimate would be (6/5) (345) = 414.
3. Our third method is based on computing the sample mean of our five observed numbers and then estimating N by twice this mean. Here the mean is (345 + 167 + 91 + 125+92) = 165 and so our estimate would be 2 (165) = 330.

We have described three methods for estimating the total number of taxis in the city. How can we decide on the best method? In this activity, we will use all three methods to estimate N using many simulated datasets, graph the different estimates to understand their sampling distributions, and use these graphs to illustrate several important properties of estimators.

In the following we will assume that we know that the number of taxis in the city is N = 100. We will take random samples of size 5 from the population of 100 cities. For each sample, we will calculate each of the three estimates. We'll repeat this process 1000 times, obtaining collections of estimates for 1000 samples. By graphing each collection of estimates, we will illustrate the concepts of bias and variability.

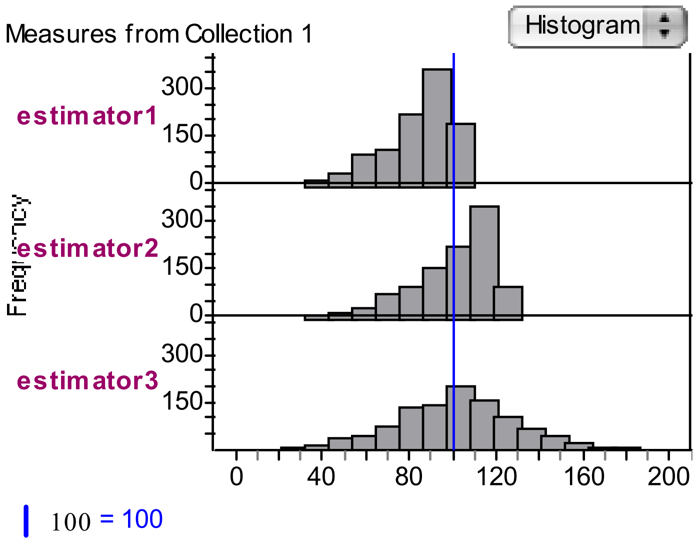
1. Open up a new Collection and define a new Attribute called number. Edit the formula for number and type the formula randominteger(1,1000). Then add 5 new cases to number. You should see five random taxi numbers between 1 and the maximum value N = 1000.

2. Inspect the Collection and define the following three Attributes. These are the three methods for estimating N from the sample.

Name of Attribute	Formula
estimator1	max(number)
estimator2	6/5max(number)
estimator3	2mean(number)

3. Repeat the process of sampling taxis and computing the three estimators by
-- selecting the Collection and choosing Collect Measures from the Collection menu
-- when you inspect the Measures collection, Replace Existing Cases and collect 1000 measures.
When you are done you should have a Measures Collection that contains the values of the three estimators for 1000 random samples.

4. Graph each collection of estimators on the same graph using histograms. You should see a figure like the one below:



5. Describe the shape of the sampling distribution for each estimator.

6. One desirable property of an estimator is that the average or mean value of the estimator over repeated samples should be equal to the true value of N. If the average value of the estimator is equal to N, then we say that the estimator is unbiased. Use Fathom to find the mean value of each estimator. By comparing the mean value with the true value of N (100), decide if each estimator is unbiased.

7. You should have found in part 6. that one of the estimators is not unbiased. In this case, one can measure the bias of the estimator by

$$\text{bias} = \text{mean value of estimator} - \text{true value of N}.$$

Find the bias of this one estimator.

8. Another desirable property of an estimator is that it should, on average, be close to the true value of N. One can measure the closeness of an estimator to N by the quantity

$$\text{MAD} = \text{mean of } |\text{estimator} - \text{N}|.$$

This is a slight difference of our definition of MAD from a previous topic. MAD measures the mean distance of an estimator from the true value N. Find the MAD for each one of the three estimators and decide which is best from this criterion.

9. Based on your work, which is the best estimator of the total number of taxis N? Justify why you believe one estimator is the best.

CONSTRUCTION OF A CONFIDENCE INTERVAL

The basic problem that we want to address is this: Once we compute a value of the proportion $\hat{p}$ from a random sample, what does this proportion value tell us about the population proportion p? To answer this question, we review some results from our earlier topic on coin-tossing patterns.

Suppose we know the value of the population proportion p. For example, suppose that we know that 5% of the college students from my school own iPods – that is, p = .05. Let X represent the number of students that own iPods in my class of 20 students. If p = .05, then we know from Topic P7 that the number of iPod owners in my sample will have a binomial distribution with n = 20 and probability of success p = .05, where we define "success" as owning an iPod. So the chance that x students own iPods is given by the binomial formula

$$p(x) = \binom{20}{x} .05^x (1-.05)^{20-x} .$$

These binomial probabilities are displayed in the below table.

X	0	1	2	3	4	5	6	7	8	9	10
P(X)	0.358	0.377	0.189	0.060	0.013	0.002	0.000	0.000	0.000	0.000	0.000
X	11	12	13	14	15	16	17	18	19	20	
P(X)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	

We see that if the proportion of all students owning iPods is p = .05, then the most likely values of X are 0, 1 and 2. So we would expect 0, 1, or 2 students in my class to own iPods.

What if instead the population proportion was equal to p = .15? Then the number of iPoders in my class would be binomial with parameters n = 20 and p = .15. These binomial probabilities are given below.

X	0	1	2	3	4	5	6	7	8	9	10
P(X)	0.039	0.137	0.229	0.243	0.182	0.103	0.045	0.016	0.005	0.001	0.000
X	11	12	13	14	15	16	17	18	19	20	
P(X)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	

In this case, when 15% of the population own iPods, we see that it is most likely for 1 to 5 students to own iPods in my class.

Suppose that we summarize these two binomial distributions by 5th and 95th percentiles. In particular, suppose that the population proportion p = .05, and we took many samples of size 20 and observed the number of successes X. Then we would find that if p = .05, then

$$P(0 \le X \le 3) \ge .90 .$$

To say it in a different way, if the proportion of iPod owners is really .05, then the chance of observing between 0 and 3 iPod owners in my sample is at least .90. By similar reasoning, one can show that if the population proportion is p = .15, then the probability of X falling between 1 and 6 is at least .90:

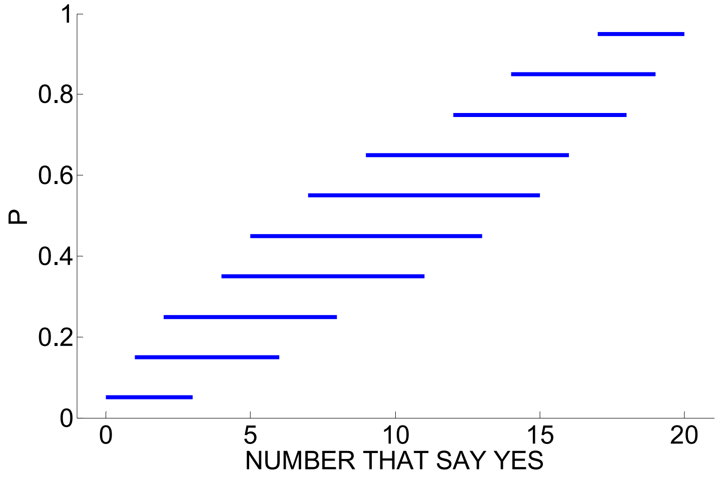
$$P(1 \le X \le 6) \ge .90$$

We summarize other binomial distributions for other values of the parameter p. For each of the values p = .05, .15, …, .95, we find an interval of values of the number of successes X that captures at least 90% of the probability. We find the intervals and place them in the below table.

Value of the population proportion	90% of the values of X fall between	Value of the population proportion	90% of the values of X fall between	Value of the population proportion	90% of the values of X fall between

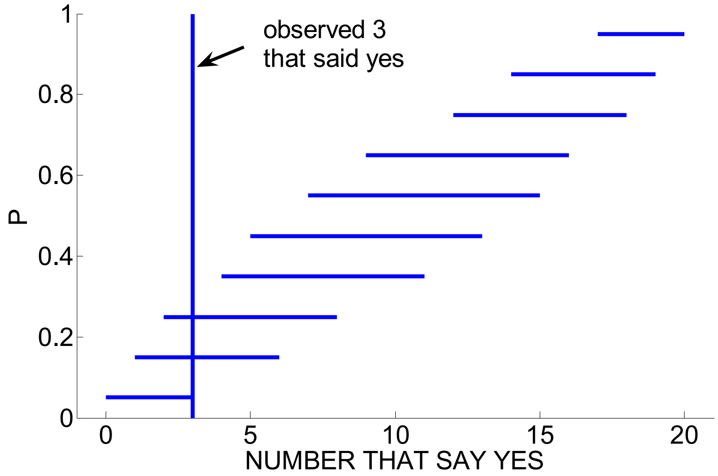
p=.05	0 and 3	p=.45	5 and 13	p=.85	14 and 19
p=.15	1 and 6	p=.55	7 and 15	p=.95	17 and 20
p=.25	2 and 8	p=.65	9 and 16		
p=.35	4 and 11	p=.75	12 and 18		

We graph these intervals on the below figure. The values of the population proportion are shown on the vertical axis and the intervals of likely values of X are drawn as solid bars.



By use of the binomial formula, we now have a good idea how many iPod owners to expect in my sample of students for different values of the population parameter p. But the statistical inference problem asks the reverse question: What have we learned about the parameter p on the basis of the data collected in the sample?

Remember that we observed three "yes's" – three people that said they owned an iPod. In the figure, we draw a vertical line on the figure at the observed value X = 3. We now ask: For which values of the population proportion p is X = 3 a plausible result? We see from the figure that the vertical line hits the horizontal lines for the values p = .05, p = .15, and p = .25. For these three values of p, getting 3 iPod users in my sample is a likely result. So we say that (.05, .25) is a 90% *confidence interval* for p.

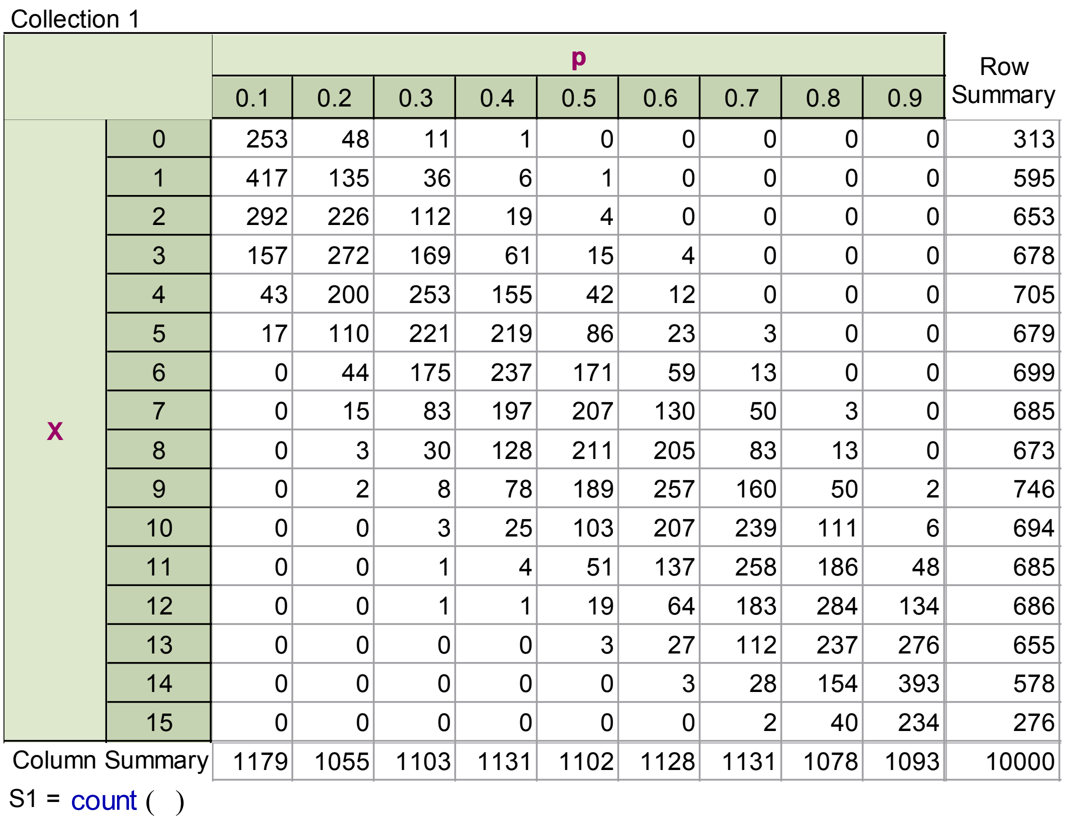


PRACTICE: CONSTRUCTION OF A CONFIDENCE INTERVAL

Suppose there is going to be an election for the major of our city and we are interested in learning about the proportion p of the registered voters who currently prefer the Republican candidate Jones. Suppose there are nine possible values for p – either 10% of the voters prefer Jones (that is, p = .1), or 20% prefer Jones, or 30% prefer Jones, , …, or 90% prefer Jones.

We are going to take a random sample of 15 voters and observe X, the number of voters who prefer Jones in my sample.

To see what the value X tells us about the unknown population proportion p, we do this simple simulation experiment. We first choose a value of p at random from the possible nine values {.1, .2, .3, …, .9}. Then given this value of p, we randomly simulate a number of "favorable" voters X out of a sample of 15. We continue this process – simulate a parameter p and simulate a statistic X – a total of 10,000 times. We summarize the outcomes by the following table that classifies the simulation results by the value of p and the value of X.



Collection 1

		p									Row Summary
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
X	0	253	48	11	1	0	0	0	0	0	313
	1	417	135	36	6	1	0	0	0	0	595
	2	292	226	112	19	4	0	0	0	0	653
	3	157	272	169	61	15	4	0	0	0	678
	4	43	200	253	155	42	12	0	0	0	705
	5	17	110	221	219	86	23	3	0	0	679
	6	0	44	175	237	171	59	13	0	0	699
	7	0	15	83	197	207	130	50	3	0	685
	8	0	3	30	128	211	205	83	13	0	673
	9	0	2	8	78	189	257	160	50	2	746
	10	0	0	3	25	103	207	239	111	6	694
	11	0	0	1	4	51	137	258	186	48	685
	12	0	0	1	1	19	64	183	284	134	686
	13	0	0	0	0	3	27	112	237	276	655
	14	0	0	0	0	0	3	28	154	393	578
	15	0	0	0	0	0	0	2	40	234	276
Column Summary		1179	1055	1103	1131	1102	1128	1131	1078	1093	10000

S1 = count ( )

1. We say that a value of X is "likely" if it occurs at least 20 times in the simulation. If the proportion of all voters in favor of Jones is really 10%, that is p = .1, circle the likely values of the number of favorable voters in our sample X.

2. In a similar fashion, circle the likely values of X if p = .2. Continue in this way and circle the likely values of X for p = .3, .4, .5, .6, .7, .8, .9.

3. Suppose we take our sample of 15 voters and we observe X = 9 voters in favor of Jones. Looking at the table, for what values of the population proportion p is the result X = 9 a likely outcome? (Your answer is a confidence interval for p.)

4. Based on your work, do you believe, on the basis of your sample data X = 9, that the proportion of voters in the population is larger than 50%? Explain.

A LARGE SAMPLE CONFIDENCE INTERVAL FOR A PROPORTION

Although the basic logic behind a confidence interval is described above, the resulting intervals obtained are a bit inexact since we only considered ten possible values for the population proportion p. Obviously, the proportion can take on any possible value between 0 and 1 and so we wish to make more precise statements about the location of p.

A standard method for constructing a confidence interval is based on a large sample result about the sampling distribution of proportions. This is a slight variation of the normal approximation to the binomial distribution discussed in Topic P7.

Suppose we take a large sample from a population where the proportion of successes is equal to p. From our sample, we count the number of successes X and compute the sample proportion $\hat{p} = X/n$. Then the sampling distribution of $\hat{p}$ will be approximately normal-shaped with mean p and standard deviation $\sqrt{p(1-p)/n}$.

Let's interpret this result. Recall the fact that 90% of the probability falls within 1.645 standard deviations from the mean for a normal curve. Applying this fact in this case, if we take many samples from the population, then 90% of the sample proportions $\hat{p}$ will fall between

$$p - 1.645\sqrt{\frac{p(1-p)}{n}} \quad \text{and} \quad p + 1.645\sqrt{\frac{p(1-p)}{n}}.$$

We are interested in learning about the location of the population proportion p. Note that if $\hat{p}$ falls in the interval ($p - 1.645\sqrt{\frac{p(1-p)}{n}}$, $p + 1.645\sqrt{\frac{p(1-p)}{n}}$), then by a simple algebraic manipulation, p will fall in the interval ($\hat{p} - 1.645\sqrt{\frac{p(1-p)}{n}}$, $\hat{p} + 1.645\sqrt{\frac{p(1-p)}{n}}$). We can restate this result as follows. If we take many samples from the population, then 90% of the intervals of the form

$$(\hat{p} - 1.645\sqrt{\frac{p(1-p)}{n}}, \hat{p} + 1.645\sqrt{\frac{p(1-p)}{n}})$$

will cover the proportion p. This is not a very useful interval, since both the left and right endpoints contain the term $\sqrt{p(1-p)/n}$ including the population proportion p that we don't know. But the 90% probability statement is still approximately accurate if we replace $\sqrt{p(1-p)/n}$ by the estimated standard deviation $\sqrt{\hat{p}(1-\hat{p})/n}$. With this replacement, we obtain a 90% confidence interval for the proportion of the form

$$(\hat{p} - 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}).$$

To illustrate the computation of a confidence interval, a February 2005 Harris poll surveyed 1012 adult Americans by phone about their use of seat belts in their cars. Of this sample, 870 said that they always wear seatbelts. What can we say about the proportion of all adult Americans who use seatbelts?

Assume that the phone survey represents a random sample from the population of all American adults. The sample size is n = 1012 and we observe X = 870 who wear seatbelts. The sample proportion is $\hat{p}$ = 870/1012 = .860 and a 90% confidence interval is given by

$$(.860 - 1.645\sqrt{\frac{.860(1-.860)}{1012}}, .860 + 1.645\sqrt{\frac{.860(1-.860)}{1012}})$$

= (.860 - .018, .860 + .018) = (.842, .878).

What does it mean to say that (.842, .878) is a 90% confidence interval for the population proportion p? Is it correct to say that there is a 90% chance that p falls in the interval (.842,.878)? No. The proportion p is just a fixed number and either p falls in the interval (.842,.878) or it doesn't fall in the interval. A correct interpretation is that we are stating a confidence *in our procedure* of estimating p by an interval. We don't know if a particular computed interval would contain p. But we know that if we take repeated samples from the population, then approximately 90% of the intervals we compute will contain the unknown proportion of seatbelt users p.

Let us illustrate the meaning of 90% confidence in this setting. Suppose the proportion of seatbelt users in our population is p = .9. We are going to take a phone survey of 1012 randomly selected adults and observe X, the number of seatbelt users in my sample. After I collect X, I find a 90% confidence interval for p using our formula. I'll repeat this procedure of taking a sample and computing a confidence interval 20 times – the results are shown in the table below.

	X	$\hat{p}$ **=X/1012**	**90% interval**	**Cover p=.9?**
Sample 1	924	0.913	(0.898 0.928)	yes
Sample 2	928	0.917	(0.903 0.931)	no
Sample 3	916	0.905	(0.890 0.920)	yes
Sample 4	890	0.879	(0.863 0.896)	no
Sample 5	914	0.903	(0.888 0.918)	yes
Sample 6	904	0.893	(0.877 0.909)	yes
Sample 7	922	0.911	(0.896 0.926)	yes
Sample 8	908	0.897	(0.882 0.913)	yes
Sample 9	904	0.893	(0.877 0.909)	yes
Sample 10	903	0.892	(0.876 0.908)	yes
Sample 11	900	0.889	(0.873 0.906)	yes
Sample 12	890	0.879	(0.863 0.896)	no
Sample 13	909	0.898	(0.883 0.914)	yes
Sample 14	908	0.897	(0.882 0.913)	yes
Sample 15	924	0.913	(0.898 0.928)	yes
Sample 16	910	0.899	(0.884 0.915)	yes
Sample 17	894	0.883	(0.867 0.900)	yes
Sample 18	902	0.891	(0.875 0.907)	yes
Sample 19	890	0.879	(0.863 0.896)	no
Sample 20	913	0.902	(0.887 0.918)	yes

Note that the value of X and therefore our confidence interval changes from sample to sample – this is expected since the sample result is random. We have indicated in the table by a "yes" or "no" if the confidence interval contains the known value of p = .9.

Here we see that 16 out of 20 = 80% of the intervals do contain p. If we were able to take many more random samples and compute confidence intervals, we would see that 90% of the intervals would cover p. So 90% doesn't mean that we are confident that one particular interval will be successful in covering p – rather it means that 90% of the intervals in repeated sampling will cover the unknown value of p.

PRACTICE: A LARGE-SAMPLE CONFIDENCE INTERVAL FOR A PROPORTION

A poll was taken of 1100 licensed drivers by the Mason-Dixon Polling and Research Company. Of this sample, 367 said that they have driven through a red light at an intersection.

1. Find a 90% confidence interval for the proportion of all licensed drivers who have driven through a red light.

2. Suppose that a new poll of drivers would be taken next week. Would your confidence interval change? Explain.

3. Is it correct to say that the probability your interval computed in part 1. includes p is .90? Explain.

4. Give a correct interpretation of 90% confidence based on repeated sampling.

UNDERSTANDING A CONFIDENCE INTERVAL

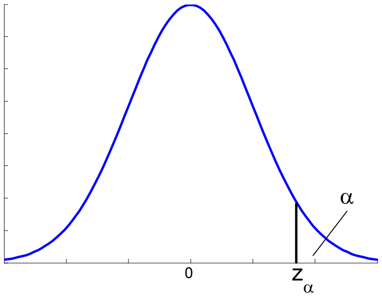
There are two important aspects of a confidence interval:

- the level of confidence
- the margin of error of the interval

The confidence level comes from the statement about the sampling distribution of proportions. When we had 90% confidence, we found an interval that contained the middle 90% of the probability of the sample proportion $\hat{p}$. If we want instead a 95% confidence interval, we want to bound the middle 95% of the probability. Generally, we specify a confidence level $1-\alpha$ and the middle $1-\alpha$ probability for $\hat{p}$ falls between

$$p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \quad \text{and} \quad p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}},$$

where $z_{\alpha/2}$ is the value of a standard normal variable with α/2 in the right tail region. (See the below figure.)



The corresponding $1-\alpha$ confidence interval for p has the form

$$(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}).$$

The value of the percentile $z_{\alpha/2}$ and the corresponding confidence interval are shown in the following table for several popular choices for confidence.

Confidence	$z_{\alpha/2}$	Confidence interval
90%	1.645	$(\hat{p}-1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p}+1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$
95%	1.96	$(\hat{p}-1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p}+1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$
99%	2.58	$(\hat{p}-2.58\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p}+2.58\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$

How does one in practice decide on a confidence level? Generally, one wants to be pretty confident that the population proportion p is in our interval, so typically large confidence levels of .9, .95, and .99 are chosen. But there is a cost to choosing a larger confidence interval – this brings us to the second important aspect of an interval.

We are also interested in the length of our interval. The shorter the interval, the more information we have regarding the location of the population proportion p. Define the margin of error

$$ME = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

The ME simply tells us how close we believe the sample proportion $\hat{p}$ is to the population proportion p. The length of the confidence interval is given by twice the margin of error, or

$$LENGTH\ OF\ INTERVAL = 2\times ME = 2z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Looking at the expressions for the margin of error and the length, we can see the impact of changing the confidence level. If we change the confidence from 90% to a larger value, say 95%, the value of $z_{\alpha/2}$ will increase, and the margin of error and length of the interval will increase. So the cost of increasing confidence is a longer interval.

Another thing to notice in the margin of error expression is the role of the sample size n. Suppose we increase our sample size from n = 30 to n=100? How will this change our sample size? We see that the sample size is in the denominator of the margin of error expression. This means that the margin of error will decrease for a larger sample size. So a confidence interval for n=100 will be shorter than an interval with a sample size of n=30. We will see shortly that we can in practice take a sufficiently large sample so that the margin of error will be smaller than any value of interest.

PRACTICE: UNDERSTANDING A CONFIDENCE INTERVAL

Consider again the poll taken of 1100 licensed drivers by the Mason-Dixon Polling and Research Company where 367 said that they have driven through a red light at an intersection.

1. Find a 90% confidence interval for the proportion of all drivers who have driven through a red light. (You can copy your answer from the previous practice problem.)

2. Find 95% and 98% confidence intervals for the population proportion p.

3. Looking at your answers from parts 1 and 2, how does increasing the confidence level change the confidence interval?

4. Suppose that 2200 drivers are polled and 734 drivers admit to driving through a red light. Compute the sample proportion $\hat{p}$ and a 90% confidence interval for p.

5. Compare your 90% intervals from parts 1 and 4. How does increasing the sample size from 1100 to 2200 change the confidence interval?

CHOOSING A SAMPLE SIZE

We saw above that the margin of error of a confidence interval decreases as we take a larger sample. This means that by taking a sample of a sufficient size, we are able to estimate the population parameter p to a given level of accuracy.

To illustrate this point, suppose that the president of a company is concerned about absenteeism during a particular season and so she is interested in estimating p, the proportion of workers who have missed at least 5 days this month. She wishes to estimate p to within .02 with a confidence level of 95%. How many workers should she sample?

Recall the margin of the confidence interval is given by

$$ME = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\,.$$

Since we are interested in learning about the sample size n, we solve this equation for n, obtaining

$$n = \frac{z_{\alpha/2}^2\; p(1-p)}{ME^2}\,.$$

We call this a *sample size formula* for determining the necessary sample size to achieve a given accuracy or margin in error.

To use this formula in this example, the president wishes to have 95% confidence, so α = .05 and $z_{\alpha/2}$=1.96. In addition, she wishes to estimate the proportion p to within

.02, so the margin of error ME = .02. To use this formula, we also need to specify a value for the proportion p. How can we do this when the proportion of all workers who miss at least five days this month is unknown? There are two options here. If the president has some knowledge about a plausible value of p, then she would use that value in the formula. If she has little knowledge about p, then a conservative choice would be the value of p = .5.

Suppose in this case that she believes that the value of this proportion is close to .1. Applying this formula with p = .1, $z_{\alpha/2}$=1.96, and ME = .02, we obtain the sample size

$$n = \frac{(1.96)^2(.1)(.9)}{(.02)^2} = 864.4\,.$$

So the president should take a sample of size 864.4, or actually 865, to estimate the proportion p to within .02 with 95% confidence.

PRACTICE: CHOOSING A SAMPLE SIZE

Suppose you wish to take a survey at your school to learn about the proportion of students p who currently own cars.

1. Suppose you initially have little information about the value of p. You wish to estimate p within a margin of error of .02 with 90% confidence. How many students should you sample?
2. You decide to take a preliminary sample to learn about this proportion. Of 30 students, you find out 18 own their own cars. Using this information, find the sample size so that p is estimated to within .02 with 90% confidence.
3. If you wish to estimate p to within .01 instead of .02, how do you think the required sample size will change? Why?
4. If you wish to estimate p with 95% confidence instead of 90% confidence, how will the required sample size? Explain why this makes sense.

SOME CAUTIONS

Although it is relatively easy to compute a confidence interval, the validity of this procedure rests on several important assumptions about how the sampling was done. Recall our example about the Harris poll about adults' use of seatbelts in their cars. This article has an important disclaimer about their methodology that brings up several important cautions about confidence intervals.

> *The Harris Poll*® was conducted by telephone within the United States between February 8 and 13, 2005, among a nationwide cross section of 1,012 adults aged 18 and over.
>
> In theory, with a probability sample of this size, one can say with 95 percent certainty that the results have a sampling error of plus or minus 3 percentage points of what they would be if the entire U.S. adult population had been polled with complete accuracy.
>
> Unfortunately, there are several other possible sources of error in all polls or surveys that are probably more serious than theoretical calculations of sampling error. They include refusals to be interviewed (non-response), question wording and question order, interviewer bias, weighting by demographic control data and screening (e.g., for likely voters). It is impossible to quantify the errors that may result from these factors.

The first paragraph simply describes how and when the sample was taken. We see that this poll was a telephone sample of 1012 adults aged 28 and over taken in February 2005.

The second paragraph describes the confidence that the sample proportion will be close to the population proportion p. It says that one can say with 95% certainty that the sampling error would be plus or minus 3 percentage points of the results if the entire U.S. adult population were surveyed. In other words, the margin of error of estimating p by $\hat{p}$, with 95 confidence, would be no larger than .03. We can check this calculation by computing the margin of error of our confidence interval. Recall that we observed 870 seatbelt users with a sample of n=1012, so $\hat{p}$=870/1012 = .86. The margin of error of a 95% confidence interval would be

$$\mathrm{ME}=1.96\sqrt{\frac{\hat{\mathrm{p}}(1\text{-}\hat{\mathrm{p}})}{\mathrm{n}}}=.021$$

which is smaller than the 3% stated in the disclaimer, so it seems like a correct statement.

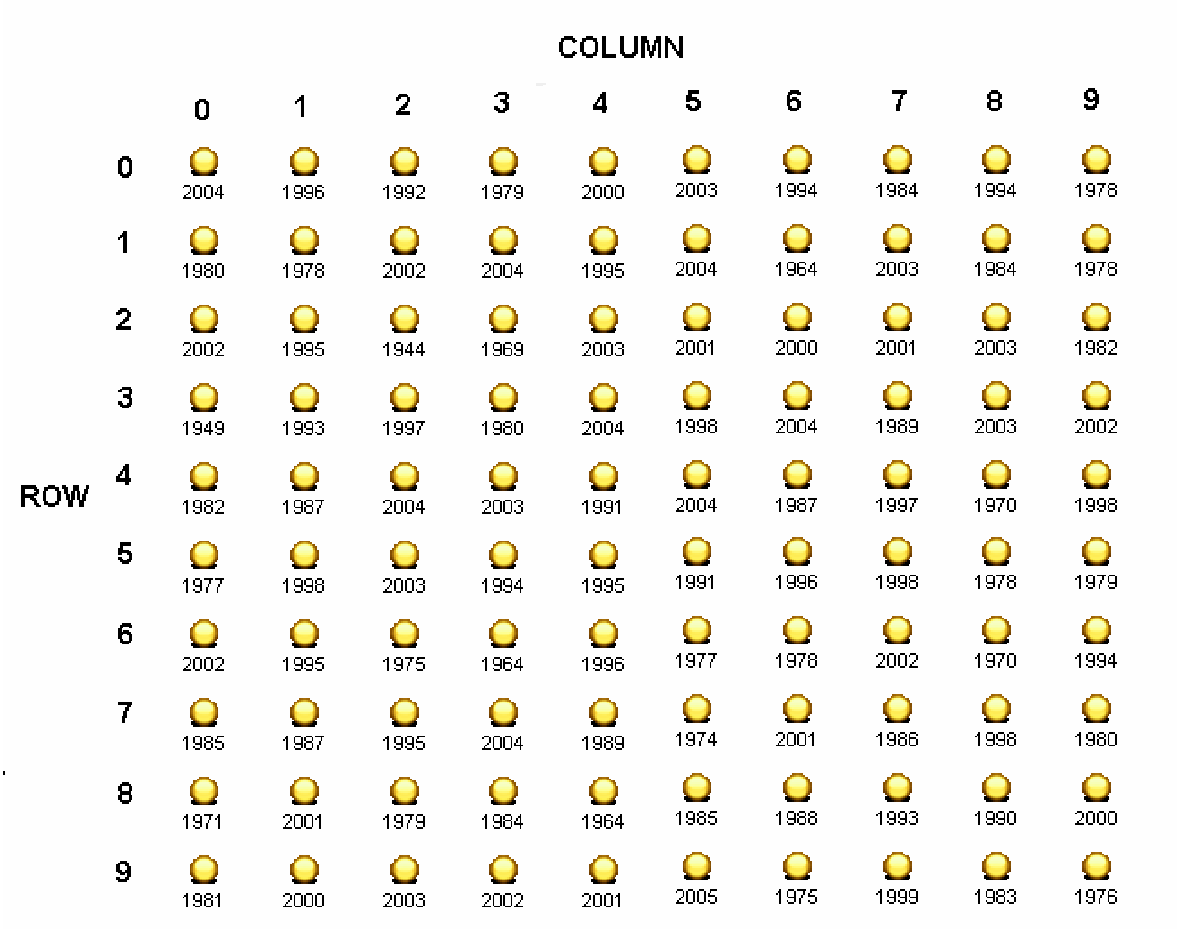
The last paragraph states perhaps the most important aspect of this confidence interval. The above margin of error calculation makes the important assumption that our sampling is taken at random, and the only source of error is sampling error. However, in many cases this may not be true. This disclaimer mentions other sources of error which potentially are much more serious than sampling error. These other error sources include

- people's refusal to be interviewed
- errors due to the wording of the question
- bias due to the manner in which the interviewer gave the questions

In our sampling topic, we described some of these nonsampling errors in detail. Unfortunately, as this disclaimer states, it is difficult to quantify these nonsampling errors when they may exist.

TECHNOLOGY ACTIVITY: PENNY AGES

Suppose there are 100 pennies in my dresser drawer. I'm interested in the ages of these pennies. Specifically, suppose I'm interested in the proportion of pennies p that are dated 1985 or earlier. What can we learn about this population of penny ages by taking a random sample and computing a confidence interval? For convenience of sampling, we have laid the population of pennies out in a 10 by 10 grid. Each gold ball represents a penny with the date shown below.



		COLUMN									
		0	1	2	3	4	5	6	7	8	9
	0	2004	1996	1992	1979	2000	2003	1994	1984	1994	1978
	1	1980	1978	2002	2004	1995	2004	1964	2003	1984	1978
	2	2002	1995	1944	1969	2003	2001	2000	2001	2003	1982
	3	1949	1993	1997	1980	2004	1998	2004	1989	2003	2002
ROW	4	1982	1987	2004	2003	1991	2004	1987	1997	1970	1998
	5	1977	1998	2003	1994	1995	1991	1996	1998	1978	1979
	6	2002	1995	1975	1964	1996	1977	1978	2002	1970	1994
	7	1985	1987	1995	2004	1989	1974	2001	1986	1998	1980
	8	1971	2001	1979	1984	1964	1985	1988	1993	1990	2000
	9	1981	2000	2003	2002	2001	2005	1975	1999	1983	1976

1. Using a random digit table, select 40 digits on any row of your choice. The first two digits correspond to the number of the row and the number of column of the first penny in your sample. The next two digits correspond to the second penny in your sample, and so on. Continue in this way until you have selected 20 pennies. In the table below, write down the random digits and the year of each penny selected.

Penny	1	2	3	4	5	6	7	8	9	10
Random digits										
Year										
Penny	11	12	13	14	15	16	17	18	19	20
Random digits										
Year										

2. Compute $\hat{p}$ = the proportion of pennies of age 1985 or earlier in your sample.

3. Compute a 90% confidence interval for p.

4. Collect the 90% intervals for all students in your class. Graph these intervals on the below graph.

5. What is the value of the population proportion p in this problem? What fraction of intervals contained the value of p? Is it close to what you expected? Explain.

EXERCISES

1. **Gas Prices**

In the fall of 2005, gas prices increased sharply and U.S. car manufacturers were wondering about the impact of the cost of gas on new car purchases. Harris Interactive and Kelley Blue Book Marketing Research took an on-line survey of 2000 adults who planned on buying or leasing a new vehicle in the next twelve months. Fifty-nine percent of these shoppers said that gas prices either changed their minds or strongly influenced their purchase decision.

a. Describe the population of interest in this study.

b. Describe the sample.

c. Define in words the parameter p and the statistic $\hat{p}$ in this study.

d. Suppose that a new sample of 2000 adults was taken a year after this initial study. Would you expect the percentage of shoppers saying that gas prices influenced their purchase decision be different from fifty-nine percent? Why?

2. **Prime Minister of Great Britain**

USA Today reported the results of a Gallup poll of Americans adults in January 2006 to learn about their opinion of Tony Blair, the current Prime Minister of Great

Britain. Of the 506 adults sampled, 67% were favorable, 15% had no opinion, 9% were unfavorable, and 9% never heard of him.

a. Describe the population of interest in this study. How large is this population?

b. Describe the sample.

c. If the focus is on a person's opinion being "favorable", define the population proportion p.

d. Define the sample proportion $\hat{p}$.

e. Suppose USA Today wishes to report American's opinion in July, 2006 about Tony Blair. Would it be reasonable to conclude in July that 67% of Americans are favorable of the Prime Minister?

3. **Sexual Harassment**

A headline in the USA Today states ""Students: Sexual harassment all too common on campus". This conclusion is based on the results of an online survey in May of 2,026 full and part-time undergraduates of ages 18 to 24 enrolled in a two or four-year college last spring. Data were adjusted to be nationally representative. It found that 62% of female students and 61% of male students say they have been sexually harassed while in college.

a. What is the population in this study?

b. Define the population proportion p.

c. Define the sample and the statistic $\hat{p}$.

d. Are you surprised by the results of the survey? Do you believe that the sample accurately represents the population in this situation?

4. **Drinking and Drugs at Teen Parties**

The National Center on Addiction and Substance Abuse at Columbia University take an annual survey of American teenagers and their parents. In a telephone sample of 562 parents taken in spring 2006, 80% said that neither alcohol nor marijuana is available at teen parties. In contrast, 50% of the 1297 teenagers surveyed said that alcohol, drugs or both are available at these parties.

a. There are two populations described in this study. What are they?

b. If we consider only the parents, what is the population proportion of interest? What is the value of this proportion?

c. The article states that the margin of error in estimating the population proportion is plus and minus 3 percentage points for the teens and plus and minus 4 percentage points for the adults. Why is the margin of error smaller for the teen proportion?

5. **Sampling Students**

Suppose you are interested in the proportion of students on your campus that regularly attend football and basketball games on campus. Suppose that the population proportion is equal to p = .1 and you plan on taking a random sample of 15 students to learn about p. Each student will be asked the question "Do you regularly attend football and basketball games?" and the possible answers will be "yes" (Y) or "no" (N). In the table below, the results of twenty samples of size n = 15 have been simulated when the population proportion is equal to p = .1. For the first sample, note that we observed two yes's and 13 no's and so the sample proportion is equal to $\hat{p}$ = 2/15.

SAMPLE	RESPONSES	phat	**SAMPLE**	RESPONSES	phat
1	N N N N N N N N N N Y N Y N N	2/15	**11**	N Y N N N N N N Y N N N N N N	____
2	N N N N N N N N N N N N N N N	____	**12**	N N N N N N N Y N N N N N N Y	____
3	N N N N N N N N Y N N N N N N	____	**13**	N N N N N N N N Y N N Y N N N	____
4	N Y N Y N N N N N N N N N N N	____	**14**	N N N N N N N N N N N Y N N Y	____
5	Y N N N N N N N N N Y N N N N	____	**15**	Y N N N N N N N N N N Y N N N	____
6	N N Y N Y N N N N Y N Y N N N	____	**16**	N N N N N N N N N N N N N N N	____
7	N N Y N N N N N N N N N N Y N	____	**17**	N N N N N N N N N N N N N N N	____
8	N N N N N N N N N N N N N Y N	____	**18**	N N N N N N Y N N N N N N N N	____
9	N N N N N Y Y N N Y N N N N N	____	**19**	N N N N N N N N N N N N N N N	____
10	N N Y N N N N N Y Y N N N N N	____	**20**	N N N N N N N N N N N N N N N	____

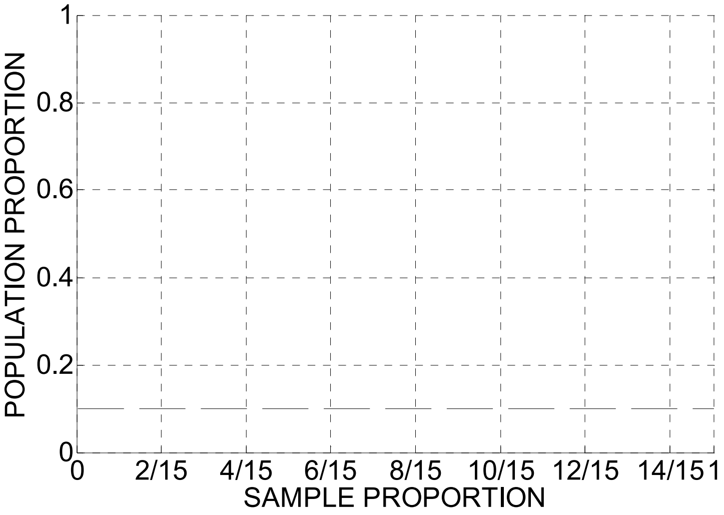
a. For each of the twenty samples, compute the value of the sample proportion $\hat{p}$ in the lines provided.

b. In the table below, write down the count and relative frequency of each sample proportion value. (This is the sampling distribution of $\hat{p}$ when the population proportion p = .1.)

$\hat{p}$	Count	Relative frequency	$\hat{p}$	Count	Relative frequency
0/15			8/15		
1/15			9/15		
2/15			10/15		
3/15			11/15		
4/15			12/15		
5/15			13/15		
6/15			14/15		
7/15			15/15		

c. From the table, find the 2nd smallest and 2nd largest values of $\hat{p}$**.** The probability that $\hat{p}$ falls between these two values is approximately 80%.

d. On the graph below, draw a thick horizontal line at p = .1 between the two values you found in part c. (The line p = .1 has been indicated on the figure by a dashed line.)



6. **Computers Needing Repair**

A computer manufacturer has a one-year warranty on the computers they sell. They are interested in estimating p, the proportion of computers that need repair within the warranty period. They plan on taking a random sample of 50 computers and computing the proportion $\hat{p}$ that need repair. Suppose it is know that the actual value of the population proportion is p = .2. By use of a computer, we simulated taking 40 samples of size 50 from the population and for each sample computing the sample proportion $\hat{p}$. Here are the values of $\hat{p}$ from the 40 samples.

0.16 0.18 0.22 0.30 0.18 0.14 0.20 0.18 0.10 0.18 0.28 0.16 0.12 0.18 0.26
0.18 0.18 0.16 0.16 0.28 0.24 0.10 0.08 0.18 0.20 0.10 0.18 0.10 0.20 0.22
0.14 0.24 0.36 0.20 0.24 0.24 0.22 0.20 0.26 0.18

a. If the sample proportion $\hat{p}$ =.16, how many computers in the sample needed repair?

b. Construct a dotplot of the sample proportions from these 40 samples.

c. What is the shape of the distribution of sample proportions?

d. Find the second smallest and second largest values of $\hat{p}$. The probability that $\hat{p}$ falls between these two values is approximately 90%.

7. **Sampling Students (continued)**

Consider the problem of estimating the proportion p of students at your school that regularly attend football or basketball games. If the population proportion p is equal to .3, twenty samples of size 15 have been simulated in the "p=.3" table below. Similarly, twenty samples of size 15 have been simulated when p = .5, when p = .7, and when p = .9.

**************************** p=.3 **********************************

SAMPLE	RESPONSES	phat	**SAMPLE**	RESPONSES	phat
1	N Y N Y N N N N Y N Y N Y N N	____	**11**	N Y N N N N N N Y N N Y Y N Y	____
2	N Y N Y N N N N Y N N N N N N	____	**12**	Y N N N N N N N N N N N Y Y N	____
3	N Y N N N N N N Y Y N Y Y N Y	____	**13**	N Y N N N Y Y N N Y N N Y N N	____
4	N Y N N N N N N Y Y N Y Y N N	____	**14**	N N N N N N N N N N Y N N N N	____
5	N Y N Y Y N Y N Y Y N Y N N N	____	**15**	Y N N N N N N N N N N N Y Y N	____
6	Y N Y N N N Y N N N N N N N Y	____	**16**	N Y Y N N N N N Y Y Y N N Y Y	____
7	Y Y N Y Y N N Y N N Y N N N Y	____	**17**	N N Y Y N Y N N N N Y Y Y Y Y	____
8	N Y Y N N N Y N Y N N N N N Y	____	**18**	Y Y N N N Y N N Y N Y N Y N N	____
9	N N N N Y Y N N Y Y N N N N Y	____	**19**	N N N N N N N N N N N Y N N Y	____
10	N N N Y N Y Y N N N N Y Y Y N	____	**20**	N N Y Y N N N N N N N Y N Y N	____

**************************** p=.5 **********************************

```
SAMPLE RESPONSES                        phat SAMPLE RESPONSES                        phat
 1   Y Y Y N Y Y N N N Y N N N Y N ____ 11 Y N Y N N N Y Y N N Y Y Y N N ____
 2   N N Y Y Y Y N Y N N Y Y Y Y Y ____ 12 Y Y N N N N Y N N Y Y N N Y Y ____
 3   N N N N Y Y Y N Y Y Y Y N N Y ____ 13 N Y N Y N Y N N Y N Y Y N Y N ____
 4   Y N Y N Y Y Y Y Y Y Y Y Y N N ____ 14 Y Y N N N N Y Y Y Y Y Y Y N Y ____
 5   N Y N Y Y N Y Y Y Y N N N Y N ____ 15 N N N Y Y Y N Y N N Y N N Y Y ____
 6   Y Y N N N Y Y Y N Y Y Y N Y Y ____ 16 Y N N N N N Y Y Y N Y Y N Y Y ____
 7   Y Y N Y N Y Y Y Y N N N Y Y Y ____ 17 Y N N Y Y Y Y N Y Y N N Y N Y ____
 8   Y Y N N Y Y Y Y N N N N N N Y ____ 18 Y Y N Y Y N Y N Y Y N Y Y N Y ____
 9   Y Y Y Y N N N N Y N N Y N Y Y ____ 19 N Y N N N N N N N Y N Y N N N ____
 10  Y Y Y Y N N Y Y Y N Y Y N Y N ____ 20 N Y N Y N Y N N N N Y N Y N Y ____
```

```
**************************** p=.7 **********************************
SAMPLE RESPONSES                        phat SAMPLE RESPONSES                        phat
 1   Y N N Y Y Y Y Y Y Y N Y Y Y Y ____ 11 N Y N Y Y N Y N Y Y N Y Y Y Y ____
 2   N Y Y Y Y Y N Y Y Y Y N N Y Y ____ 12 Y Y Y Y N Y Y N N N Y Y N Y Y ____
 3   N N Y N N N Y Y Y Y Y Y Y N Y ____ 13 N Y Y Y Y Y Y N Y N Y N Y Y Y ____
 4   Y Y N N Y Y Y N N Y Y Y Y Y N ____ 14 N Y Y Y N N Y Y Y N N Y Y Y N ____
 5   N N N Y Y Y N Y Y N Y Y N Y Y ____ 15 Y Y Y Y Y Y Y N Y Y Y N Y Y Y ____
 6   Y Y Y Y N Y Y Y N Y Y Y Y Y N ____ 16 N Y Y Y N Y Y Y Y N Y N N N Y ____
 7   Y Y Y N N N N Y N Y Y Y Y N Y ____ 17 Y Y N Y Y Y Y Y Y Y Y N N N Y ____
 8   Y Y Y N Y N Y Y Y Y Y Y Y N N ____ 18 Y N Y N N Y Y N Y Y Y Y Y Y Y ____
 9   N Y Y Y Y N N N N N Y N N Y Y ____ 19 Y N N Y Y Y Y Y N Y Y Y N Y N ____
 10  Y Y Y N N Y Y N Y Y Y N Y N Y ____ 20 Y Y N Y Y Y Y Y N N Y Y Y Y N ____
```

```
**************************** p=.9 **********************************
SAMPLE RESPONSES                        phat SAMPLE RESPONSES                        phat
 1   N Y Y Y Y Y Y Y Y Y Y Y Y Y N ____ 11 Y Y Y Y Y Y Y Y Y Y Y Y N Y Y ____
 2   Y Y Y N Y Y N Y Y Y Y N Y Y N ____ 12 Y Y Y N Y Y Y Y N Y Y Y Y Y Y ____
 3   Y Y Y Y Y N Y Y Y Y Y Y Y Y Y ____ 13 Y N Y Y Y Y Y Y Y Y Y Y N Y Y ____
 4   Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y ____ 14 Y Y Y Y Y Y Y Y Y Y Y Y N Y Y ____
 5   Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y ____ 15 Y Y N Y Y Y N Y Y Y Y Y N Y Y ____
 6   Y Y Y Y Y Y Y Y Y Y Y Y Y Y Y ____ 16 Y N Y N Y Y Y Y Y Y Y Y Y N Y ____
 7   Y Y Y Y Y Y Y N Y Y Y Y Y Y Y ____ 17 Y Y Y Y Y Y Y Y Y Y Y Y N N Y____
 8   Y Y Y Y Y Y Y Y Y N N N Y Y Y ____ 18 Y Y Y Y Y Y Y Y Y N Y Y Y Y Y ____
 9   Y Y Y Y Y Y Y Y N Y Y N Y Y Y ____ 19 Y Y N Y Y Y Y Y Y Y Y Y Y N Y ____
 10  Y N Y Y Y Y Y N Y Y Y Y Y Y Y ____ 20 Y N Y Y Y Y Y Y Y Y Y Y Y Y Y ____
```

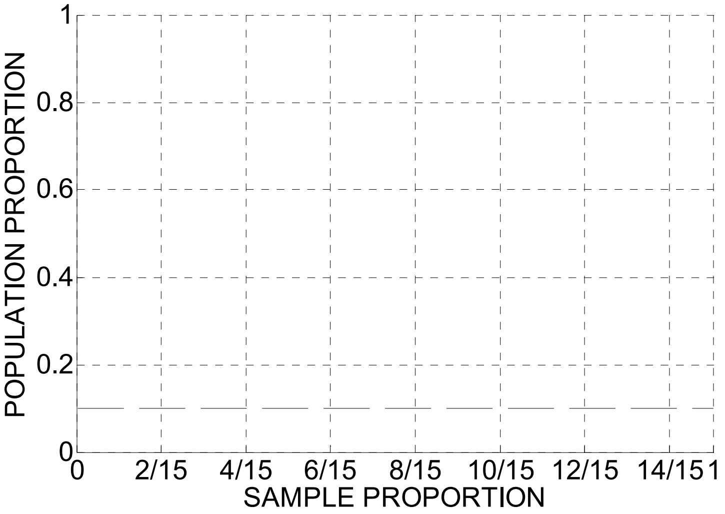
a. For each simulated sample, compute the proportion $\hat{p}$ of yes's. Put your sample proportions in the lines provided.

b. In the table below, find the count and relative frequency of each value of $\hat{p}$ in the cases when the population proportion p is equal to .3, .5, .7 and .9.

c. For the sampling distribution for p = .3, find the second smallest value of $\hat{p}$ and the second largest value of $\hat{p}$. Put these values in the last two rows of the table. Likewise find the second smallest and second largest values of $\hat{p}$ for each of the cases p = .5, p = .7, and p = .9, and put these values in the table.

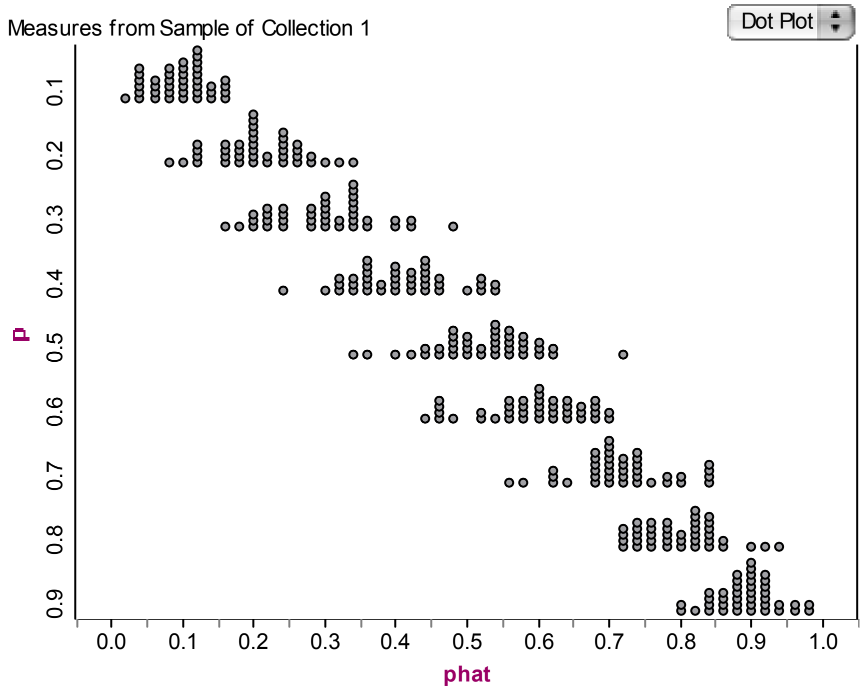
	p = .3		p = .5		p = .7		p = .9	
$\hat{p}$	Count	Rel. freq.	Count	Rel. freq.	Count	Rel. freq.	Count	Rel. freq.
0/15								
1/15								
2/15								
3/15								
4/15								
5/15								
6/15								
7/15								
8/15								
9/15								
10/15								
11/15								
12/15								
13/15								
14/15								
15/15								
2nd smallest value of $\hat{p}$								
2nd largest value of $\hat{p}$								

d. For each of the values p = .3, .5, .7, and .9, graph a horizontal line between the 2nd largest and 2nd largest values of $\hat{p}$. These lines represent 80% probability intervals for the sample proportions $\hat{p}$ for each population proportion value p.



8. **Computers Needing Repair (continued)**

A computer manufacturer is interested in estimating p, the proportion of computers that need repair within the warranty period. In Exercise 6, values of the sample proportion $\hat{p}$ were simulated for 40 samples of size 50 assuming the population proportion was equal to .2. This procedure was repeated for 40 samples of size 50 from populations with p = .1, .3, .4, …, .9. The Fathom output shows dotplots of the sample proportions plotted against the value of the population proportion p.



a. On each of the nine individual dotplots, draw a thick horizontal line that extends from the second smallest value of $\hat{p}$ to the second largest value of $\hat{p}$. This line contains (approximately) the middle 90% of the sampling distribution of $\hat{p}$.

b. Suppose 50 computers are sampled and 20 need repair. Find the value of the sample proportion $\hat{p}$.

c. On the graph, draw a vertical line at the value of the sample proportion you found in part b. Write down the values of p where the vertical line intersects with the horizontal lines you drew in part a.

d. Based on your work, _______ to ________ is a 90% confidence interval for p.

e. Using this same method, find a 90% confidence interval for p if 10 in your sample need repair.

9. Sampling Students (continued)

In Exercise 7, you plotted 80% probability intervals for the sample proportions $\hat{p}$ for the five population proportion values p = .1, .3, .5, .7, and .9.

a. Suppose you take a sample of 15 students and 4 say that they regularly attend football and basketball games. Find the value of the sample proportion $\hat{p}$.

b. On the graph, draw a vertical line at the value of the sample proportion that you found in part a. Write down the values of p where the corresponding 80% probability interval intersects with the vertical line.

c. From your work, you can say that ________ to __________ is a 80% confidence interval for the proportion p.

d. Suppose instead that 8 students in your sample regularly attend football and basketball games. Find the value of $\hat{p}$ and find the 80% confidence interval for p.

10. Computers Needing Repair (continued)

A computer manufacturer is interested in estimating p, the proportion of computers that need repair within the warranty period. In Exercise 8, values of the sample proportion $\hat{p}$ were graphed from repeated samples of size 50 for different values of the population proportion p.

a. Suppose that the manufacturer claims that p = .1 and you observe 20 computers needing repair from a sample of 50. Looking at the figure of dotplots, is this a reasonable result? Why or why not?

b. If the answer to part a is no, what does this say about the manufacturer's claim that the population proportion is p = .1?

c. Suppose you observe 20 computers needing repair in a sample of 50. A large sample 90% confidence interval for p is (.286, .514). Is it correct to say that the probability p falls in the interval (.286, .514) is .9? Why or why not?

11. Holiday Greetings

The Pew Research Center surveyed 1502 adults of age 18 years and old about the holiday greeting they are given at the entrance of stores such as Wal Mart. When asked to choose between "Merry Christmas" and non-religious terms, 901 preferred that stores and business welcome customers with a greeting of "Merry Christmas".

a. Find a 90% confidence interval for the proportion of all adults that would prefer a greeting of "Merry Christmas".

b. Find the margin of error of estimating p by the sample proportion $\hat{p}$

c. Is it correct to say that the probability that your confidence interval computed in part a contains p is .9? Explain.

12. Literacy of Students

In a study of the literacy of college students by the American Institutes for Research, it was found that twenty percent of students completing 4-year degrees had only basic quantitative literacy skills. This means that these students were unable to estimate if their car would have enough gas to get to the next gas station or calculate the total cost of ordering office supplies. This study was based on a sample of 1000 students from a random selection of 4-year colleges and universities across the United States.

a. Identify the population and the sample in this problem.

b. Let p denote the proportion of all students completing 4-year degrees that have only basic quantitative literacy skills. Find a 95% confidence interval for p.

c. Find the margin of error in estimating p.

13. **Super Bowl**

On the foxsports.com website before Super Bowl XL, people were asked to predict the winner of the Super Bowl. A total of 236,000 people responded to this poll and 60% predicted that the Steelers would defeat the Seahawks.

a. Let p denote the proportion of all football fans who predict the Steelers to win the Super Bowl. Use this data to construct a 90% confidence interval for p.

b. Would it be reasonable to assume that these 236,000 people represent a random sample from the population of all football fans?

c. Based on your answer in part b, how does that affect the "confidence" you have in the interval estimate you found in part a?

14**. Electronic Games at Home**

In December 2005, Ipsos-Public Affairs polled 1,006 adults nationwide to see if anyone in their household has or uses each of a list of devices. Of this sample, 39% said that they had or used an electronic gaming device such as a Playstation or Xbox.

a. Construct a 95% confidence interval for the proportion of all households that have an electronic gaming device.

b. The article describing the survey states that the margin of error in this estimate of the population proportion is 3.1 %. Check if this is indeed the correct value for the margin of error for this problem.
c. Would this confidence interval still be accurate for estimating the proportion of households that have electronic gaming devices in December 2007? Explain.

15. **Global Warming**

A Fox News poll taken in November 2005 found that 77 percent of Americans believe global warming is happening. The article states that these conclusions are based on a telephone survey of 900 registered voters taken in two days in October.
a. Find the margin of error of this estimate of the proportion of all registered voters in American who believe that global warning is happening. You can assume 95% confidence.
b. Suppose that Fox News wishes to take a new poll so that the margin of error in estimating the proportion of Americans who believe in global warning is no larger than .01. How large a sample should they take?

16. **Global Warming (continued)**

Fox News makes the following statement regarding the accuracy of their poll results regarding people's opinion on global warning:
"For a sample of about 900 interviews, the error due to sampling is plus or minus three percentage points. For example, when the survey says "47% of voters..." then chances are very strong that no less than 44% and no more than 50% of all voters would have responded the same way."
a. Is this statement correct? (To check the correctness, write down an interval that you believe will contain 95% of the sample proportions, and compare this statement with the Fox News statement.)
b. In addition, the article states that
"In addition to sampling error, question wording and question order can influence poll results."

Explain how, in this global warning survey, the wording of the question could influence the results.

INTRODUCTION TO HYPOTHESIS TESTING

NCTM Standards

✓In Grades 9-12, all students should use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions.

✓In Grades 9-12, all students should understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference.

INTRODUCTION: A TASTE TEST

Suppose you go to the convenience store and want to purchase a cola. What brand do you buy? You see different brands such as Pepsi, Coke, and RC Cola – perhaps you will choose the brand that is currently selling at the lowest price. But perhaps you will buy one brand, say Pepsi that you believe tastes the best.

That raises the interesting question: can people really distinguish between the two popular brands of cola, Coke and Pepsi?

Let's put this question in our framework of statistical inference. First, we have to decide on what we mean by "people". What group are we are interested in learning about? You note that college students purchase a lot of pop and so you decide that the population will be the group of undergraduate students at your school.

In this population there will be two groups relative to your main question. Some people will be able to correctly distinguish Coke from Pepsi and other students will be unable to tell the two colas apart. We let p be the proportion of students that can correctly make this distinction.

Based on your experience and conversations with your friends, you believe that the value of p is small. Many students don't like cola and haven't tried Pepsi and Coke very often. These students would have a tough time correctly distinguishing the two colas. Also, you believe that there is a subtle difference in the taste in the two colas, so there will be other cola drinkers who you believe will have difficulty telling the colas apart. Based on some thought, you believe that the population proportion is equal to p = 3. That is, you believe that only 30% of college students can correctly distinguish Coke from Pepsi.

Now you have a friend who is an avid Coke drinker. She believes that Coke and Pepsi have very different flavors and she believes that many students would be able to distinguish between the two colas in a taste test. Specifically, she disagrees with your belief that p = .3 – she thinks that the proportion of college students would be larger than 30%.

Who is right – you or your friend? We don't know, since it would be difficult to conduct a taste test with every single student at your college. The value of the population proportion is unknown, so we can't say for sure who is right. But your friend wants to provide evidence that you are wrong. She plans to take a random sample of students, conduct a taste test, and then use the data from this sample to show that you are in error.

STATING THE HYPOTHESES

The first step in this process is to clearly state the two statements about the proportion p. There are two statements about the population population:

- You say the proportion is p = .3.
- Your friend says that the proportion p > .3

We call these statements *hypotheses* about the proportion. The null hypothesis, represented by the symbol H_0, is the hypothesis of "equality" or "no difference". Here since the equality statement is p = .3, the null hypothesis would be your claim: H_0: p = .3.

The alternative hypothesis, denoted by H_a, is the statement about p that the person is interested in showing. If your friend is conducting this test, then she is interested in showing that the proportion of students who can distinguish colas is larger than .3. So the alternative hypothesis, in this case, is equal to H_a: p > .3.

Here we wish to make a decision about the appropriate hypothesis. Your friend will conduct a taste test of a sample of students and count the number who are able to correctly distinguish Pepsi and Coke. On the basis of this data, she would like to conclude that the proportion of all students who can correctly distinguish colas, p, is larger than 30%.

PRACTICE: STATING THE HYPOTHESES

Each of these situations describes a problem where one wishes to make a decision regarding a population proportion. For each situation, describe the population proportion of interest and give the null and alternative hypotheses.

1. A standard test for extrasensory perception (ESP) asks a subject to identify which of four shapes (star, circle, wave, or square) appears on a hidden card. The subject guesses at the shape appearing on 16 cards and the experimenter records the number of correct guesses. One is interested in deciding if the subject is truly guessing or if the subject has some extraordinary ability to detect the shape on the card.

The population proportion p:
The null hypothesis H_0:
The alternative hypothesis H_a:

2. Suppose that a particular brand of toothpaste currently has a market share of 15%. That means that of all the toothpaste purchased, 15% is of this brand. Suppose that the company uses a new advertising campaign that they hope will help will increase sales. Six months later, the company conducts a new market survey to learn about the new market share of their brand of toothpaste.

The population proportion p:
The null hypothesis H_0:
The alternative hypothesis H_a:

3. A company manufactures electronic components and it is known that 5% of the components that come off the assembly line are defective. The company implements

new quality control program that they believe will reduce the number of defectives. They sample 200 components and find that only 5 are defective.

The population proportion p:
The null hypothesis H_0:
The alternative hypothesis H_a:

A STATISTICAL TEST

Suppose your friend obtains a sample of 20 students and does a taste test experiment. Each student is given samples of Pepsi and Coke in plain white cups and asked to identify the brand of each cola. She will compute

X = number of sample who correctly identify the two colas

The random variable X is called a *test statistic* – it is a measurement from our sample that will be used to decide between the two hypotheses p = .3 and p > .3.

How do we use the value of X to make a decision? The method we use is analogous to the procedure used in a trial to convict a person who is accused of a crime. The person is presumed innocent of the crime. If the prosecuting attorney can present sufficient evidence that the person did indeed commit the crime, then the jury can decide that the person is guilty.

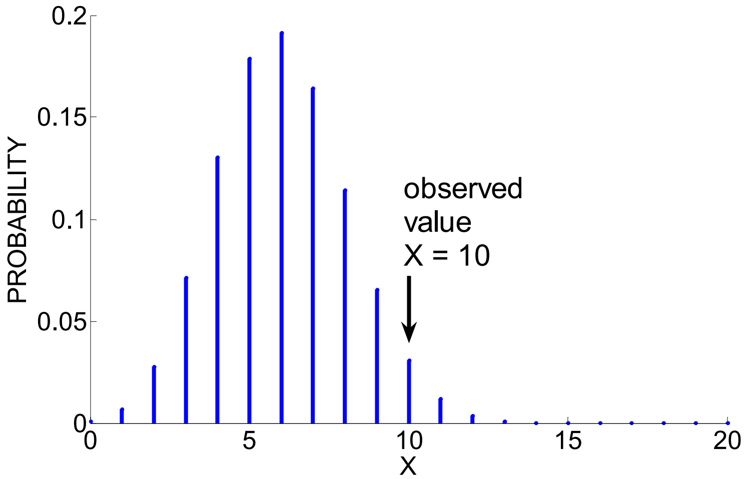
In our situation, we presume that the null hypothesis H_0 that p = .3 is true. If we can produce sufficient evidence from our sample that is contrary to the assumption that p = .3, we reject the null hypothesis and conclude that the alternative hypothesis H_a is true.

In particular, suppose that our friend does the taste test of 20 students and X = 10 students can correctly distinguish Coke from Pepsi. Is this sufficient evidence that the null hypothesis p = .3 is not true? To start to answer this question, note that if really 30% of the 20 students could distinguish the two colas, then we would expect n × p = 20 × .3

= 6 students to distinguish the drinks in our sample. Since 10 students actually did, this is surprising since 10 is larger than 6.

To see how surprising this result is, we need to look at the possible values of X when p = .3. The binomial probability distribution of X when n = 20 and p = .3 is presented in the below table. Following the table is a graph of the binomial probabilities.

X	P(X)
0	.001
1	.007
2	.028
3	.072
4	.130
5	.179
6	.192
7	.164
8	.114
9	.065
10	.031
11	.012
12	.004
13	.001
14	.000
15	.000
16	.000
17	.000
18	.000
19	.000
20	.000



If really 30% of the students from the population were able to distinguish the two colas, how likely is our sample result X = 10? We see from the table that X = 10 is a relatively rare result – in fact, we see from the table that the probability of observing X = 10 or greater is

$$P(X > = 10) = .031 + .012 + .004 + .001 = .048.$$

If this probability is very small, then we have sufficient evidence against the assumption that the population proportion p = .3.

It is common practice to say that a probability is "small" if it is less than the value .05. If we use this convention, then since P(X > = 10) is smaller than .05, we have sufficient evidence to reject the null hypothesis p = .3 and so we conclude that p is larger than .3.

What would we conclude if it turned out that X = 8 students correctly distinguished the colas in the sample? We ask the question "what is the chance of having 8 or more students correctly distinguishing colas if really the proportion of the population who can correctly distinguish colas is p = .2?" Again we consult the table above and find the probability P(X > = 8). We see that this probability is

$$P(X > = 8) = 0.228.$$

This probability of getting our result X = 8 or more extreme is not smaller than .05. This means that observing 8 correct is a relatively common result if p = 2, and so in this case we don't have sufficient evidence to conclude that the null hypothesis is false.

PRACTICE: A STATISTICAL TEST

Consider again the test for extrasensory perception (ESP) that asks a subject to identify which of four shapes (star, circle, wave, or square) appears on a hidden card. The subject guesses at the shape appearing on 16 cards and the experimenter records the number of correct guesses.

1. Suppose that the person does not possess ESP and is guessing at random at the shape on the card. If p is the probability of being correct, find the value of p.

2. State the two hypotheses for this problem.

3. If the person doesn't have ESP and X represents the number of correct guesses, then X has a binomial distribution with n = 16 trials and probability of success p, where you found p in part 1. Find the probabilities for all values of X and place the values in the below table.

X	Prob
0	
1	
2	
3	
4	
5	

X	Prob
6	
2	
8	
9	
10	
11	

X	Prob
12	
13	
14	
15	
16	

3. Suppose the person does not possess ESP. How many cards would you expect the person to guess correctly?

4. Suppose the test is conducted and the person gets exactly 8 correct. Assuming the person doesn't have ESP, find the probability that she gets 8 or more correct.

5. Based on your computation in part 4, what would be your conclusion related to the two hypotheses? Why?

6. Suppose instead suppose that the person got exactly 6 correct. Find the p-value and make a conclusion.

A LARGE SAMPLE TEST FOR A PROPORTION

In our taste test example, a small sample was taken and we made our decision to reject or not reject the null hypothesis by use of a binomial probability table. In the case where a large sample is taken, the distribution of a binomial random variable X is approximately normally distributed and we can base our decision using normal tables.

To illustrate a large sample test, suppose that 20% of the computers manufactured by a company require some repair within the warranty period. The company believes that this repair percentage is too high and a new quality control program is used to hopefully improve the quality of their manufactured computers. In the next year, the company takes a sample of 500 computers and records that only 85 require repair in the warranty period. Does this data provide sufficient evidence that the proportion of computers requiring repair has decreased?

Here the relevant population is the collection of computers manufactured by the company since the new quality control program has been introduced. In particular, we are interested in p, the proportion of these "new" computers requiring repair within the warranty period. The old repair rate was 20% and we are interested in showing that the rate has decreased, so the relevant hypotheses are

$$H_0:\ \mathrm{p} = .2,\ \ H_a:\ \mathrm{p} < .2.$$

The next step is to construct a test statistic that will be used to make a decision between the two hypotheses. A random sample of 500 computers is selected and 85 are found to require repair. The proportion needing repair is

$$\hat{p} = \frac{X}{n} = \frac{85}{500} = .17\ .$$

To decide if this value of $\hat{p}$ is sufficiently small to reject H_0, we need to know something about the sampling distribution of $\hat{p}$ when H_0 is true.

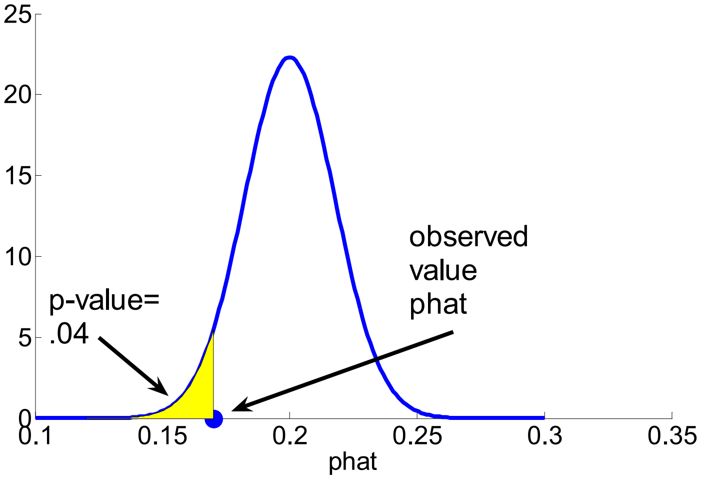
In Topic P9, we showed that the sampling distribution of a proportion $\hat{p}$ is approximately normally distributed for large samples. To state this result in this context, if the population proportion of computers requiring warranty is really p = .2, then the sampling distribution of $\hat{p}$ will be normal with mean p = .2 and standard deviation $\sqrt{p(1-p)/n} = \sqrt{.2(1-.2)/500} = .0179$. We observed a sample proportion

$\hat{p}=.17$. To find the probability of observing a sample proportion at least as extreme as this value, called the *p-value*, we compute the z-score

$$z=\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}=\frac{.17-.2}{\sqrt{.2(1-.2)/500}}=-1.68,$$

and find the probability of observing a z-score at least as extreme as this value. The probability of observing a z-score smaller than -1.68 is given by

P(Z<-1.68) = .046.



Since this probability is smaller than .05, we have sufficient evidence to conclude that the proportion of computers needing repair in the warranty period is smaller than 20%.

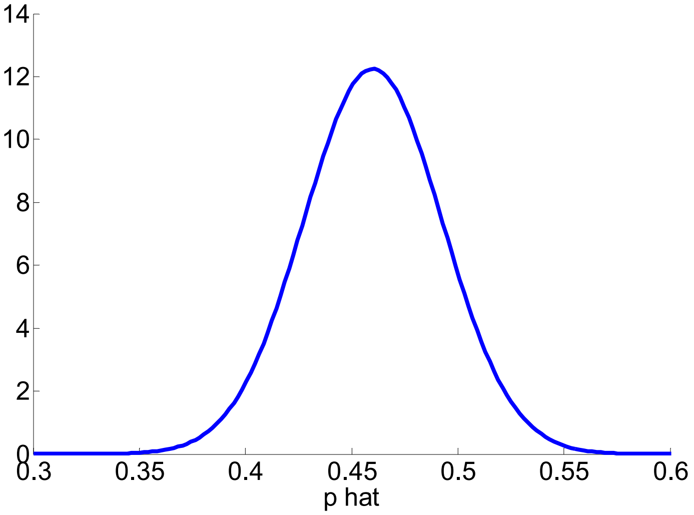
PRACTICE: A LARGE SAMPLE TEST FOR A PROPORTION

During the 2004-2005 basketball season, Shaquille O' Neal only was successful in making 46% of his free throw attempts. At a particular time during the 2005-2006 season, O' Neal was successful in making 112 out of 234 free throw attempts. Is this sufficient evidence that his free throw shooting percentage has increased from the previous season?

1. Define the population proportion p in this problem and state the hypotheses H_0 and H_a.

2. If the null hypothesis H_0 is true, find the mean and standard deviation of the sampling distribution of the proportion $\hat{p}$ of successful shots during the 2005-2006 season.

4. Below the approximate normal distribution of the sample proportion is displayed. Show the observed value of $\hat{p}$ and shade in the p-value.



5. Find the z-score of the observed sample proportion.

6. Compute the p-value.

7. Based on the p-value computation, is there sufficient evidence to conclude that his free-throw shooting percentage has increased from the previous season?

STATISTICAL SIGNIFICANCE

In both of our examples, we said that we had sufficient evidence to reject the null hypothesis if the probability of observing our result or more extreme, given H_0 is true, is smaller than .05. We state how much evidence we need to reject H_0 by specifying a probability α called the *level of significance* and we reject H_0 if the p-value is smaller than α. In our examples, we used a level of significance of α =.05 which is commonly used by practitioners. But we could really use any small value of α depending on the particular application. By stating a value for α, we are indicating how much evidence we require to reject the null hypothesis. If we use a smaller value, say $\alpha = .001$ than the customary value $\alpha = .05$, then we are requiring a greater amount of evidence to conclude the alternative hypothesis is true.

Let us illustrate the use of different significance levels for our computer repair example. We computed a p-value equal to .046 and, using a significance level of α =.05, we rejected H_0 and concluded that less than 20% of the computers require repair within the warranty period. But what if we required initially more evidence to reject and set α =.01? Then our p-value would be larger than the significance level and so we would not reject the null hypothesis. In this case, we would say that we have insufficient evidence to say that the proportion of computers needing repair is smaller than 20%.

What is the right significance level to choose -- α =.05 or α =.01? It really depends on the application and how confident one wishes to be about the conclusion of rejecting H_0. By stating a significance level of α =.05, we can guarantee that, when H_0 is true, we will incorrectly reject it with probability .05. If instead we choose α =.01, we will incorrectly reject H_0 less often – only about 1% of the time. By choosing a value of α, one is guarding the risk of making the wrong decision when the null hypothesis is true.

PRACTICE: STATISTICAL SIGNIFICANCE

Recall the previous practice problem where Shaquille O' Neal was successful in making 46% of his free throw attempts in the 2004-2005 season. At one time during the 2005-2006 season, O' Neal was successful in making 112 out of 234 free throw attempts. Using a significance level of α = .05, we decided that there was insufficient evidence that his probability of making a free-throw shot has increased for the 2005-2006 season.

1. What would have been our conclusion if we used instead a significance level of α = .01? What if we used a level α = .001?

2. Is it possible to choose a significance level for this problem where we would have rejected H_0 and concluded that Shaq's shooting percentage has increased from the previous year?

3. Is the choice of significance level completely arbitrary? What are common choices for α? Would our decision about rejecting or accepting H_0 be the same if we used any of these common choices for α?

TWO-SIDED TESTS AND CONFIDENCE INTERVALS

Let's consider a slight variation of our first example that illustrates a two-sided hypothesis test. Coke and Pepsi are the two major cola drinks – which drink is preferred by students at your school? Imagine that you are able to give a taste test to every single student at your school and each student has to state a preference for one of the two cola drinks. Let p denote the proportion of all students who state a preference for Coke. If p = ½, this means that half of the students prefer Coke and half prefer Pepsi. On the other hand, if there is a general preference between drinks, this means that either p > ½ (Coke is generally preferred) or p < ½ (Pepsi is generally preferred). Since you are interested in showing a preference, the null and alternative hypotheses would be

$$H_0 : p = .5\,,\ H_a : p \neq .5\,.$$

The alternative hypothesis is called a *two-sided hypothesis*, since it includes values of p on *two sides* of .5 –values smaller than .5 *and* values larger than .5. The use of this alternative hypothesis is appropriate when one is showing that p is *different* from some given value like p = .5.

Suppose you obtain a random sample of 50 students and you collect X, the number of students who prefer Coke. In your sample, you find X = 18. Is this sufficient evidence to conclude that the students at your school prefer one type of cola?

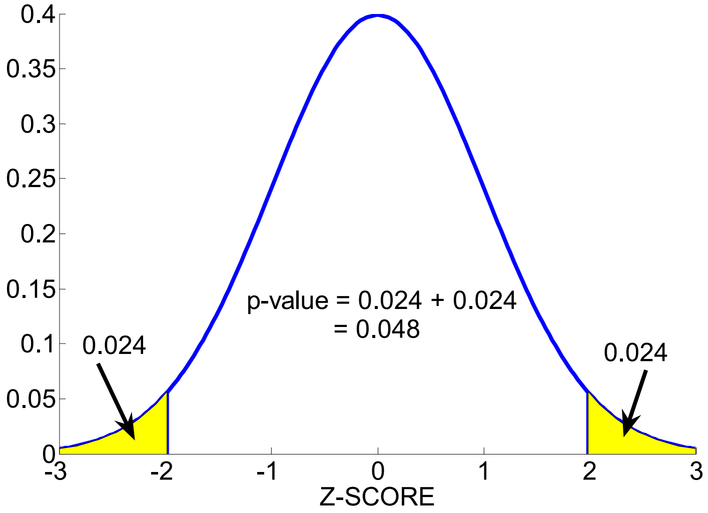
We make a decision by computing a p-value with a slight adjustment to account for a two-sided alternative hypothesis. If the null hypothesis is true (p = .5), we know that the z-score

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} = \frac{\hat{p} - .5}{\sqrt{.5(1-.5)/50}}.$$

will have a standard normal distribution. We observed 18 out of 50 students who preferred Coke – the sample proportion is $\hat{p}$ = 18/50 = .36 and the corresponding z-score is

$$z = \frac{.36 - .5}{\sqrt{.5(1-.5)/50}} = -1.98.$$

Since we have a two-sided alternative hypothesis, we consider the chance that z is smaller than -1.98 or the chance that z is 1.98 or greater. (We do this since we can reject H_0 either for large negative or large positive z-scores.) The shaded areas in the figure below show these tail probabilities. The p-value is the sum of the two areas – we see from the figure that the p-value is equal to 0.024 + 0.024 = 0.048. If we use the standard significance level of α =.05, we reject H_0 and conclude there is significant evidence that students show a preference for one of the two colas.



When we have a two-sided test, there is a relationship between the test result and a confidence interval described in the previous topic. In this setting, suppose we are interested in estimating the proportion of students p that prefer Coke by a 95% confidence interval. Using the large-sample formula, we obtain the interval

$$\begin{aligned}&(\hat{p}-1.96\sqrt{\hat{p}(1-\hat{p})/n},\, \hat{p}+1.96\sqrt{\hat{p}(1-\hat{p})/n})\\&=(.36-1.96\sqrt{.36(1-.36)/50}, .36+1.96\sqrt{.36(1-.36)/50})\\&=(0.227, 0.493).\end{aligned}$$

From this confidence interval, we see that values of the population proportion p between .227 and .493 are all plausible given these data. Specifically, since our null hypothesis value p = .5 is outside of this interval, this value is not plausible and so we can reject this hypothesis.

This is an example of a general rule. We can test the two-sided hypothesis $H_0: p = p_0, H_a: p \neq p_0$ at size α by

- constructing a confidence interval with confidence $100(1-\alpha)$
- rejecting H_0 if the value specified in H_0, p_0, is not inside the confidence interval

In general, it is good practice to compute a confidence interval when you are testing hypotheses about a proportion or about any other parameter. In our situation, if you reject the hypothesis that the proportion is equal to a given value p_0, then really all you can conclude at this point is that the proportion is not equal to p_0. One is typically interested in how far the proportion p is from p_0, and you can learn about plausible values of the proportion by the computation of a confidence interval.

In our example, we did reject the hypothesis that Coke and Pepsi were equally preferable. There is sufficient evidence, at a significance level of α = .05, that the proportion of students preferring Coke is not .5, and the data indicates that Pepsi is preferred. But the next question to ask is whether there is a strong preference for Pepsi. We answer this question by computing a confidence interval. We just did that earlier and found a 95% confidence interval for the proportion of Coke drinkers to be (.227, .493). This is a pretty wide interval for p and includes values close to .5. So although we have rejected the hypothesis that the two drinks were equally preferred, it is possible that there is only a slight preference for Pepsi among the student body.

PRACTICE: TWO-SIDED TESTS

Suppose you have a wooden die that you suspect is not balanced. You plan on rolling it 50 times, count the number of ones rolled, and use this data to make a decision about the fairness of the die.

1. Suppose p is the probability that a one is rolled. If you are interested in showing that the die is not fair, state the hypotheses H_0 and H_a.

2. If you roll the die 50 times and observe 13 ones, find the sample proportion $\hat{p}$.

3. Find the z-score.

4. Find the probability of observing a z-score at least as large as the value you found in part 3.

5. Find the p-value.

6. Using the significance level α =.05, what is your conclusion regarding the hypotheses H_0 and H_a?

7. Construct a 95% confidence interval for the proportion p. Does the value of the proportion in the null hypothesis fall in the confidence interval? Based on your answer, should you reject or accept the null hypothesis H_0? Is this conclusion consistent with your answer in part 6?

DECISIONS, TWO ERRORS, AND CONFIDENCE

When one performs a hypothesis test, there are two possible decisions – either you are going to reject the null hypothesis H_0 or fail to reject H_0. Here we look closer at the consequences of making these two decisions.

Let's return our problem about the proportion p of computers that require some repair within the two-year warranty period. We are testing the hypotheses H_0: p = .2 and H_a: p < .2 and we will make a decision on the basis of the z-statistic

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} = \frac{\hat{p} - .2}{\sqrt{.2(1-.2)/500}}.$$

We will reject H_0 if the p-value, the probability that z is smaller that our computed value, is smaller than the significance level $\alpha = .05$. Since the 5th percentile of a standard normal curve is equal to –1.645, this is equivalent to rejecting when $z < -1.645$; that is,

$$\frac{\hat{p}-.2}{\sqrt{.2(1-.2)/500}} < -1.645,$$

which is equivalent to rejecting H_0 when

$$\hat{p} < .2 - 1.645\sqrt{.2(1-.2)/500} = 0.171.$$

So we will reject the null hypothesis H_0 if the sample proportion $\hat{p}$ is smaller than 0.171 and fail to reject H_0 if $\hat{p}$ is 0.171 or larger.

Whenever we perform a statistical test, there are two possible decisions we can make after taking data – either we will reject the null hypothesis H_0, or fail to reject the hypothesis H_0. (When we fail to reject the null hypothesis, we essentially are accepting the alternative hypothesis H_a.) Also one of the two hypotheses is actually true – either the proportion p is equal to .2 (H_0 is true) or p is smaller than .2 (H_a is true). In this situation, there are two good situations – either we reject the null hypothesis when the alternative hypothesis is true, or we fail to reject H_0 when the null hypothesis is actually true. In each case, we have made the correct decision – we haven't made an error.

But there are two possible mistakes we can make in a statistical test. In one case, we decide to reject the null hypothesis H_0 when H_0 is actually true. This particular error is called a Type I error. In the second case we may fail to reject H_0 when in reality the alternative hypothesis H_a; this error is called a Type II error. We have shown the two possible errors in a statistical test by the below table.

		What is true?	
		H_0 is true (p = .2)	H_a is true (p < .2)

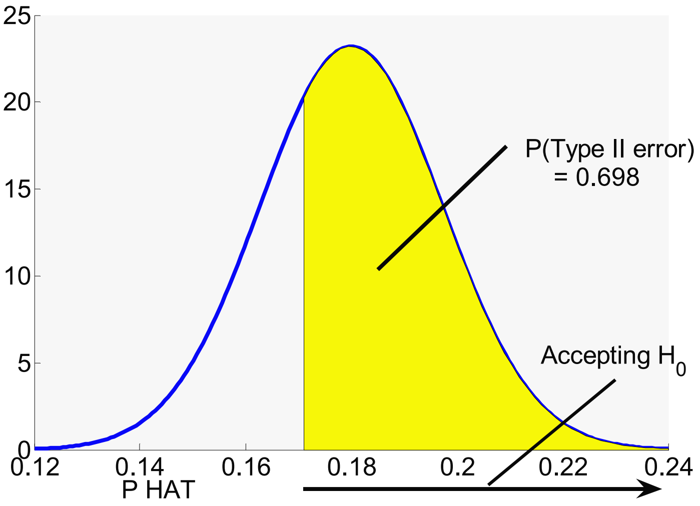
Result of the test	Reject H_0	Type I Error	No Error
	Do not reject H_0	No Error	Type II error

When we perform a statistical test, we don't want to make errors very often. How often do we make Type I and Type II errors? First, how likely is it to make a Type I error? Here we assume that really 20% of the computers require repair in the warranty period (p = .2) and we incorrectly reject H_0 : p = .2. Since we reject when $\hat{p}$ < .171, the probability of a Type I error is the probability of getting a sample proportion smaller than .171 when the proportion p = .2. Using our normal approximation, we find

$$P(\hat{p} < .171\, when\ p = .2) = P(\frac{\hat{p} - .2}{\sqrt{.2(1-.2)/500}} < \frac{.171 - .2}{\sqrt{.2(1-.2)/500}})$$
$$= P(Z < -1.62) = 0.052.$$

So the probability of making a Type I error, that is, incorrectly rejecting the null hypothesis, is .052. Actually, this is a recalculation of the significance level α = .05. A statistical test is always set up so that the chance of a Type I error is equal to α .

But there is a second error we can make in a statistical test. This error, called a Type II error, is incorrectly failing to reject the null hypothesis when really the alternative hypothesis is true. What is the chance of making this error? In this case, a Type II error is failing to reject H_0, that is observing a sample proportion $\hat{p}$ > .171, when really p < .2. By choosing a value of p in the region p < .2, we can compute the probability of a Type II error. Suppose that the proportion of computers needing repair is really p = .18. In this case, the sampling distribution for the proportion $\hat{p}$ will be approximately normal with mean p = .18 and standard deviation $\sqrt{.18(1-.18)/500}$ =.0172. This normal curve for $\hat{p}$ is shown in the figure below. The probability of a Type II error is the chance of observing $\hat{p}$ > .171 that is represented by the shaded area.



We compute the probability of a Type II error by computing this normal probability:

$$P(\hat{p} \geq .171 \; when \; p = .18) = P(\frac{\hat{p} - .18}{.0172} \geq \frac{.171 - .18}{.0172})$$
$$= P(Z \geq -.52) = 0.698.$$

You might be surprised at the large number we just calculated. If the proportion p = .18, which means that the proportion of computers needing repair is only 18%, then the chance of making a Type II error (incorrectly accepting the null hypothesis) is .698. That is pretty large – the chance of making this error is high. Another way of saying this is that the chance of correcting rejecting H_0 when p = .18 is only .302. So if the company has decreased the proportion needing repair from p = .2 to p = .18, the chance our test will correctly detect this is only 30%.

The probability of a Type II error will depend on the value of p < .2 we assume is true under the alternative hypothesis. In an exercise, you will be asked to show that if p = .15, then the probability of a Type II error is equal to 0.0942. This is better – if 15% of the computers really need repair, then the chance of making the wrong decision will only be about 9%. Note that the probability of a Type II will depend on the value of the p you are assuming is true under the alternative hypothesis.

To summarize, when you do a statistical test, there are two errors that you make. By setting the significance level $\alpha = .05$, say, you are controlling the rate of only one of these errors – specifically the rate of incorrectly rejecting the null hypothesis. You have confidence in the conclusion of rejecting H_0 since you are controlling the rate of a Type I error. On the other hand, you don't set the probability of a Type II error in a statistical test. So you don't have confidence in the conclusion "accept H_0" since you don't know the chance of the mistake in making this statement. We can compute the probability of a Type II error as we did above to understand better the properties of our test. But this is typically not done. In practice, we set up our hypotheses so a Type I error is the important error and a significance level controls the rate of making this error.

PRACTICE: DECISIONS, TWO ERRORS, AND CONFIDENCE

We revisit the test for extrasensory perception (ESP) that asks a subject to identify which of four shapes (star, circle, wave, or square) appears on a hidden card. If p denotes the probability the subject correctly identifies the subject on a card, then we are interested in testing the hypotheses H_0: p = .25 (the subject is guessing) against H_a: p > .25 (the subject has some ESP). The student will be presented 16 cards and the number X of correct identifications is recorded. Note that X will have a binomial distribution with n = 16 and probability of success p. Suppose that one decides to reject H_0 if X ≥ 7. Below there are two binomial tables shown for X – one column gives the probabilities for X for n = 16 and p = .25, and the second column shows probabilities for n = 16 and p = .40.

X	p=.25	p=.40	X	p=.25	p=.40
0	0.010	0.000	9	0.006	0.084
1	0.053	0.003	10	0.001	0.039
2	0.134	0.015	11	0.000	0.014
3	0.208	0.047	12	0.000	0.004
4	0.225	0.101	13	0.000	0.001
5	0.180	0.162	14	0.000	0.000
6	0.110	0.198	15	0.000	0.000
7	0.052	0.189	16	0.000	0.000
8	0.020	0.142			

1. Here there are two hypotheses – either the subject is guessing (p = .25) or the subject has ESP (p > .25). Describe Type I and Type II errors in the context of this problem.
2. Recall that one rejects H_0 if X $\geq$ 7. Use the table to find the probability of a Type I error.
3. Suppose that the student does possess a little ESP and his chance of identifying a card correctly is p = .40. Find the probability of rejecting H_0 in this case.
4. Use the answer in 3 to find the probability of a Type II error when p = .4.
5. For this test, are you confident of rejecting H_0 or accepting H_0? Explain.

TECHNOLOGY ACTIVITY: IS THE MACHINE WORKING?

Suppose there is a machine that produces widgets. When it is working, it produces 10% defective parts, which is considered acceptable. Sometimes the machine will malfunction – in this case it will produce 20% defectives. To check the state of the machine, an inspector will periodically test 20 machines from the machine – if 5 or more widgets are found defective, he will stop the machine to get it repaired.

Here our population is all of the widgets produced by this particular machine. We are interested in p, the proportion of all these widgets that are defective. We are testing the null hypothesis that the "machine is working" H_0 : p = .1 against the alternative hypothesis that the "machine is broken" (H_a: p = .2).

The inspector will look at a sample of 20 widgets and observe X, the number that are defective. He will decide the machine is broken if X >4; otherwise he will continue using the machine under the assumption that the machine is not broken.

In this simulation activity on Fathom, we will see how this inspection procedure performs in the long run.

Open up a new Fathom document.

1. Open a new Collection. Define three Attributes – p, X, and decision.
2. With the Collection selected, Select the Collection menu and choose New Cases. Add 1000 cases to your collection.

3. Select the p Attribute and Edit Formula (control-E or apple-E). In the formula box, type

```
randompick(.1,.2)
```

(This command randomly picks a working machine with p = .1 or a broken machine with p = .2.)

4. Select the X Attribute and Edit Formula. In the formula box, type

```
randombinomial(20,p)
```

(This command finds the number of defectives from a random sample of 20 taken from a machine where the probability of defective is p.)

5. Select the decision Attribute and Edit Formula. In the formula box, type

```
if (X > 4) “reject H0”, “accept H0
```

(If the number of defectives in the sample exceeds 4, then we decide to reject H_0 and conclude the machine is broken. Otherwise we accept H_0 and conclude the machine is working.)

6. We are interested in looking at the relationship between the true model (the value of p) and our decision to either reject H_0 or accept H_0. Construct a two-way table of p against decision.

Use this two-way table to answer the following questions.

7. What proportion of the simulations was the machine working (p = .1)?

8. How many times was the machine working and we decided to reject H_0?

9. When the machine was working, approximate the probability of rejecting H_0. Is this probability small?

10. What proportion of the simulations was the machine broken (p = .2)?

11. When the machine was broken, approximate the probability that we incorrectly accept H_0.

12. When we test hypotheses, there are two possible errors we can make: (1) we can incorrectly reject H_0 when the machine is working (p = .1) or (2) we can incorrectly accept H_0 when the machine is broken (p = .2). Which error is less likely to happen in our simulation?

Extensions:

This simulation can be used to find the probabilities of the two types of errors for other situations. Here are some things to try.

1. Change your test procedure. Suppose you reject H_0 when X > 3. How does this change the probabilities of the two errors that you found in parts 9 and 11?

2. Change your test procedure by rejecting H_0 when X > 5. How does this change the probabilities of the two types of errors?

3. Change your size of your random sample. Suppose you take a sample of size 50 (instead of 20) and decide to reject H_0 when X > 8. How does this new sample size and new test procedure affect the probabilities of the two types of errors?

EXERCISES

1. **Typing Errors**

Suppose that you currently mistype 5% of the words that you type. Your supervisor wants you to type more accurately and so you take a special course to help improve your typing skills. Let p denote the proportion of words that you mistype after taking the course.

a. If you are interested in showing that the course improved your typing accuracy, state the hypotheses H_0 and H_a.

b. If you are interested in showing that the course changed your typing accuracy, state the hypotheses H_0 and H_a.

c. Describe a Type I error in the situation of part a.

2. **Specifying Hypotheses**

In each of the following situations, define the population proportion p and give the null and alternative hypothesis H_0 and H_a.

a. Suppose that 30% of patients with migraine headaches get relief using the standard medicine. To test a new painkiller, 200 patients with migraines are given the new drugs and 80 experience relief.

b. One month before the election, suppose that is it known that 60% of the registered voters would vote for a particular Republican candidate for major in a small town. The IRS claims that this candidate has cheated on his income taxes and one is interested in seeing if this news has changed the town's support for this person. A new poll is taken of 200 registered voters and 90 indicate that they would vote for this candidate.

c. Based on previous data, it is known that 30% of the drivers of cars in a local community do not use seat belts. A new advertising campaign is conducted in the newspapers that promote the use of seat belts. To check the success of this campaign, 80 cars are stopped at several checkpoints in the town and 20 drivers are not using their seat belts.

3. **Vacationing this Year?**

Suppose that last year 80% of the families in your community took a vacation out of state each year. Due to the impact of high gasoline prices, you suspect that the proportion of families taking vacations has decreased this year.

a. Define the population proportion and the two hypotheses to be tested.

b. Suppose you sampled 15 families and 10 said that they took a out-of-state vacation this year. The below table displays binomial probabilities for n = 15 and p = .8. Use this table to compute a p-value.

X	P(X)
0	0.000
1	0.000
2	0.000

X	P(X)
8	0.014
9	0.043
10	0.103

3	0.000	11	0.188
4	0.000	12	0.250
5	0.000	13	0.231
6	0.001	14	0.132
7	0.003	15	0.035

c. Based on your computation in part b, what is your conclusion?

4. **Spinning a Quarter**

In an earlier activity in this book, a student conducted an experiment where a quarter was spun (not flipped) 20 times.

a. Let p denote the probability that the quarter lands heads when spun. If one wishes to decide if the coin is fair, give the hypotheses H_0 and H_a.

b. Suppose the student obtains 13 heads in her 20 spins. Compute a p-value using binomial tables to decide if this is sufficient evidence to say that the coin is not fair.

5. **Driving to Work**

You are interested in the proportion of commuters from New York State that travel at least 30 minutes to get to work. A friend of yours claims that this proportion is only 0.1 but you suspect that the proportion is a larger value. You collect a sample of 49 travel times from the 2000 U.S. Census that is summarized in the following tabke.

a. Find the proportion of travel times in the sample that are least 30 minutes.

b. Based on this data, is there sufficient evidence that the proportion of all commuters traveling at least 30 minutes exceeds 10%?

Travel time (minutes)	Count		Travel time (minutes)	Count
2	1		25	2
3	1		30	7
5	3		45	4
10	11		60	6

15	7		65	1
20	5		90	1

6. **Parents' Attitudes about School**

In January 2005, AP-AOL Learning Services conducted a poll of 1085 parents to learn about their attitudes about the schooling of their children. Of this sample, 57% felt that the amount of homework assigned to their children was about right. Is there sufficient evidence from this poll to conclude, as the article suggests, that over half of the American parents believe that the amount of homework assigned is "about right"?

7. **Sleeping Times of Students**

Suppose someone claims that the median sleeping time for college students is 7.2 hours. This means that the proportion p of students getting at least 7.2 hours of sleep is one half. You believe that the proportion p is larger than .5. The following sleep times (in hours) for a sample of 22 students is recorded.

```
7.00    6.50   10.00    6.25    9.50    8.50    6.00    8.00    8.75
9.38    5.75    5.50    8.25    8.50    8.00    8.50    7.80    8.75
5.50    9.00    7.75    6.00
```

a. Write down the hypotheses H_0 and H_a.

b. Find the number of students who get at least 7.2 hours of sleep in the sample.

c. Based on the data in part b, is there sufficient evidence to show that the proportion of students p is larger than .5?

8. **Working Out**

It is known from previous data that one half of American adult women exercise regularly. A telephone poll of 500 American adult men was taken and 270 said that they exercise regularly.

a. Write down the hypotheses H_0 and H_a.

c. Is there sufficient evidence from this data to conclude that American men are more likely to work out than American women?

9. **Mobility of Americans**

In a sample of data collected from the 2000 U. S. Census, one records for each adult (25 years or older), the state where he or she was born (the birth state) and the state in which he or she currently lives. We call an adult "mobile" if the adult is currently living in a different state than the birth state. For a sample of 167 adults, one observes that 82 are mobile.

a. Find a 95% confidence interval for the proportion of all adults who are mobile.

b. Suppose that, based on previous data, a researcher states that the proportion of all adults that are mobile is equal to .4. Based on your confidence interval, do you have sufficient evidence to dispute the researcher's statement?

c. Are you confident of your conclusion in part b? If so, what is the probability you are making an error in your conclusion?

10. **Who Do You Trust?**

A Harris Poll sampled 1002 U.S. adults in July 2006 to learn what occupations were most truthworthy. In particular, they found that 74% of the sample would trust a clergyman or priest.

a. Find a 95% confidence interval for the proportion of all U.S. adults who would trust a clergyman or priest.

b. A similar poll was conducted in 2002 and 64% of the adults would trust a clergyman or priest. Is there sufficient evidence that the proportion of adults trusted religious leaders has increased since 2002?

11. **Rolling a Die**

Suppose a six-sided die is rolled. Either the die is the “fair” die showing 1, 2, 3, 4, 5, 6 on the six dies, or it is a “special” die that shows 2 on two sides, 4 on two sides, and 6 on two sides. Let p denote the probability that the die lands six when rolled.

a. If H_0 is the hypothesis that the die is fair and H_a is the hypothesis that the die is special, write the two hypotheses in terms of the proportion p.

b. Suppose you plan to roll the die 15 times and count Y, the number of sixes. You will reject H_0 if Y ≥ 6. Using binomial calculations, find the probability of a Type I error.

c. Find the probability of a Type II error.

d. Suppose you do the experiment and observe 5 sixes (Y = 5). What is your conclusion? Are you confident of this conclusion? Why or why not?

12. **Cell Phones**

Last year, the proportion of students at your school that owned cell phones was .6. You believe the proportion of students p that currently use cell phones exceeds .6.

a. State the hypotheses H_0 and H_a.

b. Suppose that you plan on taking a sample of 100 students and you will reject H_0 if the proportion of students in your sample owning cell phones exceeds .67. Using a normal approximation, show that the probability of a Type I error is equal to α = .1.

c. Suppose that really 70% of the students own cell phones. Find the probability of a Type II error of your test.

TOPIC I3: LEARNING ABOUT A POPULATION MEAN

NCTM Standards

✓In Grades 9-12, all students should use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions.

✓In Grades 9-12, all students should understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference.

HOW LONG IS A CELL PHONE CALL?

Currently it seems like all students at colleges own cell phones. When one walks across campus between classes, one sees many students engaged in cell phone conversations. These conversations have a cost. Typically, one is billed monthly for cell phone use, and the size of bill can depend on the total number of minutes that the user talks on his or her phone.

The total number of cell phone minutes appearing on a monthly bill depends on the number of calls the student makes in a month, and the length (in minutes) of an average cell phone call.

This brief discussion motivates the question: How long is a typical cell phone call?

This is a relatively imprecise question, so we should define what we mean by a "typical" length of a phone call. People have different habits on cell phones. Some students may use the phone primarily for short calls; other students may use the cell phone for long conversations with friends who are local or live far away. Since the use of cell phone varies a lot between students, it makes sense to focus on the cell phone use for one particular student that we will call Bob.

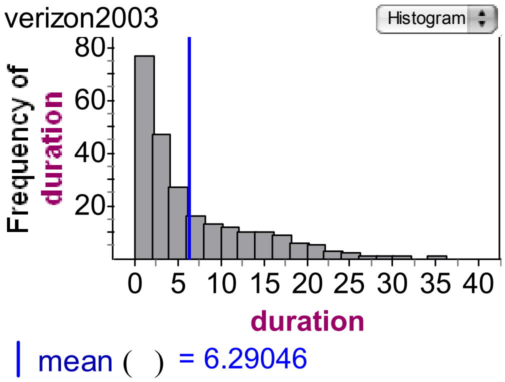
Bob uses his cell phone frequently during a particular semester – on an average day he makes about 10 calls. He is concerned about the size of his cell phone bill, so he would like to learn about the average length of his calls.

In this situation, the variable of interest is TIME, the length of minutes of one of Bob's outgoing phone calls. In the earlier inference topics, we collected categorical data and we wished to learn about the proportion of the population in a particular category. Here TIME is a quantitative variable. The population consists of the lengths of all of Bob's phone calls that he has or will make this semester. Bob is interested in the mean value of this population μ, since it is directly related to the cost of his phone bills. Bob will be unable to collect the durations of all of his phone calls this semester since many of these calls will happen in the future. But he is able to collect the lengths of a random sample of 20 calls that he made this month. He would like to learn about the location of the population mean μ based on the sample mean of the collection of 20 calls.

REVIEW: THE PATTERN OF SAMPLE MEANS

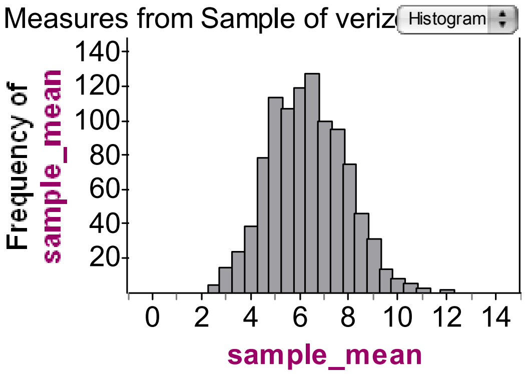
I don't know much about the lengths of cell phone calls by a male college student. But I pay the phone bills for my family and so I am familiar with the durations of cell phone calls by my family. I will use my cell phone data to illustrate what happens when one takes repeated sample means from a known population. These results were previously described in the normal distribution topic. They are especially important here since they will give us methods for estimating and testing an unknown population mean.

I collected the durations (in minutes) for a large number of cell phone calls that my family made during one spring season. Since the number of measurements is large, we can think of this as a known population. A histogram of the measurements is shown in the below figure. Note that the population distribution is strongly right-skewed -- there were a large number of short calls corresponding to "no answers" or quick messages and a relatively small number of long calls. We are interested in the mean value μ of this population. This value μ = 6.29 minutes is shown in the Figure as a vertical line. The spread of this population can be measured by the standard deviation given by σ = 6.73 minutes.



Suppose I take a random sample of n = 20 calls from my population and I compute the sample mean $\bar{x}$ -- say for my sample, it turns out that $\bar{x} = 7.5$ minutes. What does that tell us about the location of the population mean μ? Actually, not much. The value of $\bar{x}$ from a single sample says little about the location of μ . But if we take repeated samples of the same size n from the population, then the Central Limit Theorem tells us something about the pattern of values of $\bar{x}$ from the many samples.

This pattern of sample means from repeated samples can be demonstrated by software such as Fathom. One starts with the collection of durations from the large group of phone calls. One takes a random sample of size n = 20 from this population and computes the sample mean $\bar{x}$. One repeats the process of sampling and computing the sample mean many times. Below the figure displays a histogram of the collection of sample means that are collected – this is the sampling distribution of the mean $\bar{x}$.



Looking at this sample distribution, we notice several things. First, the distribution is normal-shaped – the Central Limit Theorem tells us that the distribution of sample means will be approximately normal for large samples taken from any population. Here the population of durations is clearly not normally distributed, but the pattern of means computed from samples of size 20 from this population is bell-shaped.

The second thing to notice is the center of this sampling distribution. The Central Limit theorem tells us that the mean of this sampling distribution is equal to the mean of the population. Here we can compute from Fathom the mean of the $\bar{x}$'s – this mean is 6.28 minutes which is approximately equal to μ = 6.29, the mean of this population of cell phone durations.

Last, look at the spread of the sampling distribution. The distribution of $\bar{x}$'s is pretty tight – most of the sample means fall between 3 and 10 minutes. This distribution has much smaller spread than the population. In fact, the Central Limit Theorem says that the standard deviation of the distribution of $\bar{x}$'s is given by

$$SD(\bar{x}) = \frac{\sigma}{\sqrt{n}},$$

where σ is the standard deviation of the population and n is the sample size. We can check this result by computing the standard deviation of the $\bar{x}$'s – we compute SD($\bar{x}$) = 1.54 which can be compared to $\sigma / \sqrt{n} = 6.73 / \sqrt{20} = 1.50$.

To sum up, if we take repeated samples of a particular size from a population with mean μ, then the distribution of sample means will be approximately normal-shaped. The average of the sample means will be approximately equal to the population mean of interest. The spread of the sample means will have smaller spread than the population spread. In fact, if you look again at the formula for the standard deviation

$$SD(\bar{x}) = \frac{\sigma}{\sqrt{n}},$$

note that the standard deviation of $\bar{x}$ will be smaller for larger sample sizes. This means that if you take a large sample (that is, large *n*), the sample means from any population will tend to be very close to the population mean.

PRACTICE: THE SAMPLING DISTRIBUTION OF A MEAN

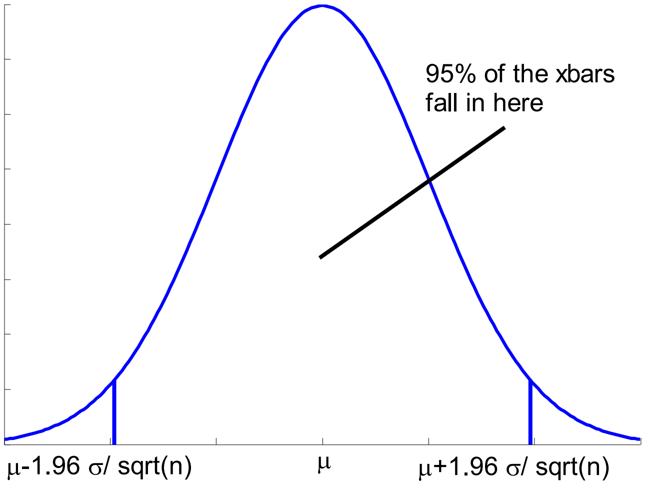
Suppose the admissions office is able to collect the distances from home to school for all of the students at your college. The mean and standard deviation of these distances are given by 30 miles and 20 miles, respectively.

1. Suppose you take a random sample of 25 students. Find the mean and standard deviation of the sample mean $\bar{x}$ of the distances from home for the sample of students.
2. If you take repeated samples of size 25, find an interval that contains the middle 95% of the sample means.
3. If samples of size 100 were taken instead of 25, how would that change the 95% interval of the sample means? Explain.
4. The probability that the sample mean $\bar{x}$ will be within 4 miles of the population mean will increase or decrease if you take a larger sample size? Explain.

A CONFIDENCE INTERVAL FOR A MEAN

We can use the general result about the sampling distribution of means to derive an interval estimate of a population mean.

Suppose we have a population with mean μ and standard deviation σ. If one takes repeated samples of a given size n from this population, then we know that the distribution of sample means $\bar{x}$ will be approximately normally distributed with mean μ and standard deviation $\sigma / \sqrt{n}$. This sampling distribution is displayed below.



Based on the normal shape, we know that 95% of the sample means will fall between $\mu - 1.96\,\sigma/\sqrt{n}$ and $\mu + 1.96\,\sigma/\sqrt{n}$. In probability language, this is equivalent to saying that

$$P(\mu - 1.96\,\sigma/\sqrt{n} < \bar{x} < \mu + 1.96\,\sigma/\sqrt{n}) = .95.$$

This is equivalent to saying that

$$P(\bar{x} - 1.96\,\sigma/\sqrt{n} < \mu < \bar{x} + 1.96\,\sigma/\sqrt{n}) = .95,$$

which means that 95% of the random intervals $(\bar{x} - 1.96\,\sigma/\sqrt{n}, < \bar{x} + 1.96\,\sigma/\sqrt{n})$ will cover the unknown population mean μ. We call this interval a 95% confidence interval for the population mean μ.

Unfortunately, this is not a useful procedure to use in practice. Why? If we don't know the population mean μ, then likely we also don't know the value of the population standard deviation σ. Since the limits of the interval contain σ, we won't be able to compute the endpoints of the interval.

But there is a simple way out of this dilemma. A reasonable estimate at the population standard deviation σ is the sample standard deviation s. If we substitute s for σ in the formula, we obtain the interval estimate

$$(\bar{x} - 1.96\, s/\sqrt{n}, < \bar{x} + 1.96\, s/\sqrt{n}).$$

This is our 95% "large sample" confidence interval for a population mean.

To illustrate the computation of a confidence interval, suppose a student wishes to estimate his or her mean duration of a cell-phone call. She collects a sample of 30 cell phone calls, say from recent phone bills, and finds that the mean and standard deviation of the durations of these 30 calls (in minutes) are given respectively by $\bar{x} = 5.2$, $s = 4.2$. Using the formula, a 95% confidence interval for the mean duration μ of all cell phone calls is

$$(\bar{x} - 1.96\, s/\sqrt{n}, < \bar{x} + 1.96\, s/\sqrt{n}) = (5.2 - 1.96\frac{4.2}{\sqrt{30}}, 5.2 + 1.96\frac{4.2}{\sqrt{30}})$$
$$= (3.70, 6.70).$$

The student is 95% confident that the mean duration of all of his/her cell phone calls is between 3.7 and 6.7 minutes. What is the meaning of "95% confidence"? Is it correct to say that the chance the population mean μ falls in the interval (3.70, 6.70) is .95? No. Actually, the student really doesn't know if any given interval such as (3.70, 6.70) contains the unknown population mean. But the student can be confident about the behavior of the random interval $(\bar{x} - 1.96\, s/\sqrt{n}, < \bar{x} + 1.96\, s/\sqrt{n})$ in repeated sampling. If the student were to compute many 95% confidence intervals, then he or she could expect about 95% of them to contain the population mean μ, and 5% of the intervals would not contain μ. So really "confidence" refers to confidence about the performance of the interval estimate in repeated sampling.

PRACTICE: A CONFIDENCE INTERVAL FOR A MEAN

The number of pairs of shoes was recorded for 23 students with the results

52, 16, 30, 3, 20, 30, 30, 30, 25, 52, 60, 20, 60, 31, 15, 20, 18, 10, 25, 60, 2, 44

1. Assuming this is a random sample from the student body, find a 95% confidence interval for the mean number of pairs of shoes owned by students at your college.

2. Is it correct to say that the probability that your interval contains μ is equal to .95? If not, what is the correct interpretation of confidence?

UNDERSTANDING A CONFIDENCE INTERVAL FOR A MEAN

A confidence interval for a population mean μ has the general form

$$(\bar{x} - z_{\alpha/2}\, \sigma / \sqrt{n}, < \bar{x} + z_{\alpha/2}\, \sigma / \sqrt{n}) .$$

The value $z_{\alpha/2}$ relates to the desired confidence of the interval. In our first example, we were interested in a 95% interval where the confidence $1-\alpha = .95$. The values $(-z_{\alpha/2}, z_{\alpha/2}) = (-1.96, 1.96)$ bracket the middle 95% of the probability under a standard normal curve. In general, the values $(-z_{\alpha/2}, z_{\alpha/2})$ include the middle $1-\alpha$ of the normal curve probability. If we decide that the confidence is equal to $1-\alpha = .90$, the value $z_{\alpha/2}$ = 1.645 would be used, and if the confidence is equal to $1-\alpha = .98$, the value $z_{\alpha/2}$ = 2.33 would be used. In practice, we choose a high confidence level such as .9, .95, or .99. As stated in the earlier section, we don't know if one computed interval will contain the population mean. But by using a high confidence value, we know that it is very likely for the random interval to contain the value μ in repeated sampling.

The length of the confidence interval is given by

$$LENGTH = 2\ ME$$

where $ME =$ $z_{\alpha/2}\, \sigma / \sqrt{n}$ is the margin of error of the interval. We note that the interval length depends on three values – the confidence $1-\alpha$, the standard deviation σ of the population, and the sample size *n*. If we increase the confidence, say from 90% to 95%, then that will increase the value of $z_{\alpha/2}$ and therefore the length of the interval. So we pay for a high confidence with a long interval or a large margin of error. Also, the length of the confidence interval is an increasing function of the population standard deviation σ. If the population of interest has large spread, then the standard deviation σ will be large, and that will result in a long interval for the mean μ. In contrast, the length of the interval decreases as a function of the sample size n. Taking a larger sample means that *n* is large, and one can obtain a shorter interval and a more precise estimate at the population mean μ.

Sometimes one wishes to design an experiment to estimate the mean to a specific accuracy. Recall the margin of error of our interval is given by $ME = z_{\alpha/2}\, \sigma / \sqrt{n}$. Suppose we wish to estimate the mean within a particular margin of error with a given confidence. If we solve this margin of error equation for the sample size, we obtain the *sample size formula*

$$n = \frac{z_{\alpha/2}^2\, \sigma^2}{ME^2}.$$

To illustrate, let's return to our example of estimating the mean duration of our cell phone calls. Suppose we wish to estimate the mean duration of our calls to within an error of .5 minutes with 90% confidence. To use this formula, we need three quantities:

1. The z percentile. Since we ask for 90% confidence, $1-\alpha$ =.9 and $z_{\alpha/2}$ =1.645.

2. The margin of error ME. Since we wish to estimate the mean μ to within an error of .5, we have *ME* = .5.

3. The population variance σ^2. Generally in most applications the variance of the population will be unknown, so we replace σ^2 with a sample estimate from previous data. In the earlier example, we estimated s = 4.2 from a sample of 30 cell phone durations, so we can use this value for σ in the formula.

With these values, we calculate the required sample size to be

$$n = \frac{z_{\alpha/2}^2\, \sigma^2}{ME^2} = \frac{(1.645)^2 (4.2)^2}{(.5)^2} = 190.9$$

Rounding this calculated value to the next whole number, this suggests that we should take a sample of at least 191 cell phone lengths to estimate the mean to within a margin of error of .5 minutes with 90% confidence.

PRACTICE: UNDERSTANDING A CONFIDENCE INTERVAL FOR A MEAN

Consider again the problem of estimating the number of pairs of shoes for students at your school.

1. Find the margin of error for the 95% interval estimate that you found earlier.
2. Suppose you wish to construct instead a 98% confidence interval for the population mean μ. Find the new interval estimate and the margin of error of this estimate.
3. Suppose you wish to estimate the mean number of shoes to within a margin of error of 2 with 95% confidence. How many students do you need to sample?

TESTING ABOUT A MEAN

In the first part of this topic, we focused on learning about the location of a population mean. We learned about the location by constructing a confidence interval. For other situations, the problem is not to learn about the location of μ – rather, one wishes to make a decision about its value. In this case, we wish to test a hypothesis or statement about the mean.

Have you heard of the phrase "the aging of America"? This statement reflects the belief that due to the advances in health care, the population of Americans is increasingly getting older. Suppose you want to investigate this belief. You begin by checking 1900 census records and you find that the mean age of the U.S. population in the year 1900 is equal to 25.8 years. This value is known since this is an average taken over the entire population of Americans in 1900. One hundred years later, you take a sample of 50 residents from the year 2000 census records – from the sample of ages of these residents, you compute the mean is 34.68 years and standard deviation is 22.80 years. Does this data provide sufficient evidence to show that the mean age of all U.S. residents in the year 2000 is larger than the mean age of Americans in 1900?

We first identify the parameter of interest. The relevant population is the ages of all U.S. residents in the year 2000 and we are interested in the mean μ of this population. We would like to make a decision about the value of this mean. Either the mean age of Americans has not changed in the last 100 years or the mean age has increased since

1900. Since we know the mean age of Americans in 1900 was 25.8 years, these two statements are equivalent to the hypotheses

$\mu = 25.8$ (mean of Americans in 2000 hasn't changed since 1900)
$\mu > 25.8$ (mean of Americans in 2000 has increased since 1900)

As in the proportion setting, the null hypothesis is typically the statement of equality or no change, and the alternative hypothesis is the statement about the parameter that we would like to show. Since we want to show that Americans are getting older, we have the hypotheses:

$$H_0 : \mu = 25.8, \quad H_a : \mu > 25.8$$

Our general strategy for making a decision between these two hypotheses can be described as follows:

1. We first assume that the null hypothesis is true – that is, that the mean population of Americans in the year 2000 is really $\mu = 25.8$ years.
2. We observed a sample mean value of $\bar{x} = 34.68$. We ask "what is the probability of getting a sample mean at least as large as 34.68 if the mean value was really 25.8 years?"
3. If the probability we calculate in part 2 is sufficiently small (say smaller than .05), we reject the null hypothesis and conclude that the mean ages of Americans has increased since 1900. If the probability is not small, we say we have insufficient evidence to reject the null hypothesis.

Following this strategy, we conduct our statistical test. We start by assuming that the mean age of the 2000 residents is equal to $\mu = 25.8$ years (the null hypothesis). We are interested in finding the probability

$$P(\bar{X} \geq 34.68 \text{ given that } \mu\text{=}25.8).$$

By the Central Limit Theorem, we know if the population mean is equal to $\mu = 25.8$, then the sample mean $\bar{X}$ will be approximately normally distributed with mean 25.8 and standard deviation $\sigma / \sqrt{n}$, where σ is the standard deviation of the population and n is the sample size. Using this approximation, we write this probability as

$$
\begin{aligned}
&P(\bar{X} \geq 34.68 \text{ given that } \mu{=}25.8) \\
&= \mathrm{P}\left(\frac{\bar{X}-25.8}{\sigma/\sqrt{n}} > \frac{34.68-25.8}{\sigma/\sqrt{n}}\right) \\
&= P\left(Z > \frac{34.68-25.8}{\sigma/\sqrt{n}}\right).
\end{aligned}
$$

This tail probability, called the p-value, is the probability a standard normal variable is greater than the z-score

$$z = \frac{34.68-25.8}{\sigma/\sqrt{n}}.$$

Here we are taking a sample of size 50, so n = 50. We don't know the population standard deviation σ, but we can estimate σ by s = 22.80, the standard deviation from our sample. With these substitutions, the z-score is equal to

$$z = \frac{34.68-25.8}{22.80/\sqrt{50}} = 2.75.$$

The p-value is the probability of obtaining a z-score at least as large 2.75. Using the normal tables, we find

$$P(Z \geq 2.75) = .003.$$

This tail probability is very small and so we conclude we have sufficient evidence to reject H_0. In other words, there is sufficient evidence from our sample to conclude that the mean age of Americans is higher in 2000 than it was a hundred years ago.

PRACTICE: TESTING ABOUT A MEAN

The yearly snowfalls for Rochester, New York for 15 years are given (in inches) as follows:

136, 181, 153, 59, 192, 86, 98, 135, 131, 171, 67, 164, 192, 167, 97

1. You are told that the average yearly snowfall amount for Buffalo, New York is 93.6 inches. If you are interested in showing that Rochester tends to get more snow than Buffalo, define the population mean μ and state the hypotheses H_0 and H_a.
2. Compute the sample mean $\overline{x}$ and the standard deviation s and the z-statistic.
3. Compute the p-value.
3. Using a significance level of α = .05, decide whether to reject or accept H_0.

MORE ABOUT A TEST FOR A MEAN

A test for a population mean follows the same general outline as the test for a population proportion described in topic I2. In a testing problem, you are interested in making a decision about the value of the population mean μ. There are two possible statements about the location of μ that we call the null hypothesis H_0 and the alternative hypothesis H_a. The hypothesis that you are interested in showing, the research hypothesis, is the alternative H_a, and the null hypothesis H_0 is typically the hypothesis that the parameter is equal to some value.

There are three possibilities when you state the hypotheses. There are two possible *one-sided tests* where either you are interested in showing that the mean μ is smaller than some value μ_0

$$H_0 : \mu = \mu_0, \quad H_a : \mu < \mu_0$$

or you are interested in showing that μ is larger than some value μ_0

$$H_0 : \mu = \mu_0, \quad H_a : \mu > \mu_0 .$$

In some cases you may be interested in showing that μ is different from some value μ_0 that would give the *two-sided test*

$$H_0 : \mu = \mu_0, \quad H_a : \mu \neq \mu_0 .$$

After stating the hypothesis, you will make a decision on the basis of the data collected in a random sample. We take a sample size *n* and compute the sample mean $\bar{x}$ and the standard deviation *s*. From the data, we compute the z-statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} .$$

Typically, one doesn't know the value of the population standard deviation σ, so we substitute the sample standard deviation s for σ.

After we compute the z-statistic, we compute a p-value. This p-value is the answer to the question: what is the chance of observing a z score at least as extreme as our observed z if the population mean was really equal to μ_0, the value of the population mean under the null hypothesis? The meaning of "z score as least as extreme as our observed z" depends on the alternative hypothesis H_a.

The figure below shows graphically the p-value as a shaded area in each of the three testing situations. Suppose the alternative hypothesis is the form $H_a : \mu > \mu_0$ and the observed z statistic is equal to z = 1.5. Then as the figure indicates, the p-value is the probability that a standard normal variable z is at least as large as 1.5, that is

$$p - value = P(Z \geq 1.5) .$$

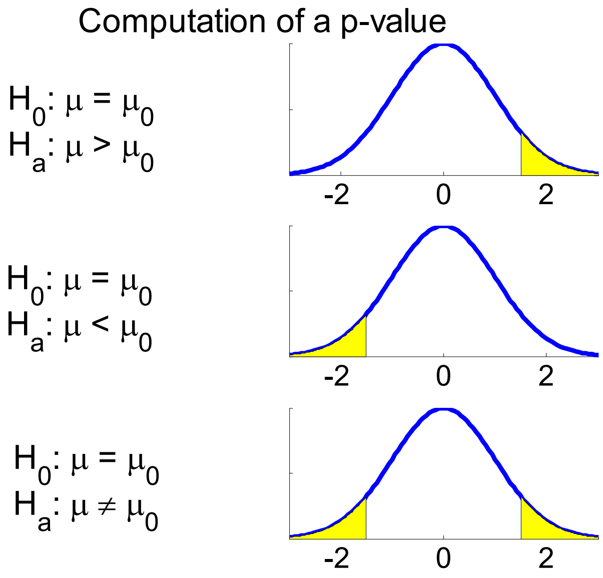
Suppose instead that the alternative hypothesis is the "less-than" form $H_a : \mu < \mu_0$ and we observe z = -1.5. Then the p-value will be the probability of getting a z score less than or equal to -1.5, or

$$p - value = P(Z \leq -1.5) .$$

Last, suppose the alternative hypothesis is of the "two-sided" form $H_a : \mu \neq \mu_0$ and again we observe the z-score from the sample z = 1.5. Now since we can reject H_0 for large positive z-values or large negative z-values, the p-value is defined to be twice the probability of that z-score or more extreme or

$$p - value = 2 \times P(|Z| \geq 1.5).$$

As shown by the figure, this p-value is the area under the normal curve to the right of z=1.5 plus the area under the curve to the left of z=-1.5.



After we compute the p-value, we make a decision. If the p-value is smaller than the stated significance level (that is typically given to be .05), then we reject the null hypothesis H_0 and conclude that we have sufficient evidence to conclude that the alternative hypothesis H_a is true. If the p-value is not smaller than the significance level, we don't reject the null hypothesis. In this case, we really don't know if the null hypothesis is true; we just have insufficient evidence from the data to reject H_0.

To illustrate this procedure, suppose you have read that the average age for a woman in the United States to marry for the first time is 23.7 years. Since this data is based on a 1995 survey, you wonder if the current average age for women to get married has changed since 1995. From courthouse records, you collect the following ages for women who have recently obtained their marriage licenses:

25, 14, 26, 27, 35, 16, 20, 33, 29, 29, 21, 24, 31, 30, 34, 26, 22, 30, 22, 23

Is this sufficient evidence to conclude that the average age for women to get married has changed from 1995?

You begin by stating the hypotheses. Here the population would be the women in the United States this year who are getting married for the first time and μ would be the mean age of these women. There are two statements about the parameter of interest: either μ hasn't changed from the mean marriage age of 23.7 from the 1995 survey, or μ has changed from the 1995 value. Since you are interested in showing that the mean age has changed, you would set

$$H_0 : \mu = 23.7, \quad H_a : \mu \neq 23.7 .$$

Next, you compute the value of the z-statistic from the data. From the sample of 20 ages, you compute $\bar{x}$=25.85 and s = 5.71 and the z-statistic would be equal to

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{25.85 - 23.7}{5.71 / \sqrt{20}} = 1.68.$$

Since this is a two-sided test, the p-value would be twice the probability of getting this z-score or more extreme. So the p-value is equal to

$$p - value = 2 \times P(Z \geq 1.68) = 0.092.$$

Suppose we use the usual significance level α = .05. Since the p-value of .092 is larger than .05, we don't reject H_0. So there is insufficient evidence from these data to say that the mean age at marriage for American women has changed since 1995. It is important to note that this does not mean that the null hypothesis is true – rather the information from this relatively small sample of 20 women is not strong enough to reject H_0. Perhaps we could reject H_0 if we were able to collect a larger sample of marriage ages from courthouse records.

PRACTICE: MORE ABOUT A TEST FOR A MEAN

In the following problems, (a) define the population mean of interest, (b) give the hypotheses H_0 and H_a, (c) compute the z-statistic, (d) compute the p-value, and (e) make a decision about the hypotheses using a significance level of $\alpha = .05$.

1. Suppose you currently type at a rate of 40 words per minute. You need to improve your typing speed and you take a class on the Internet on advanced typing. After the course, you take 15 measurements of your typing speed and your mean number of words per minute is 43 with a standard deviation of 12. Is this sufficient evidence that your typing speed has increased?

2. The mean time to complete the 2005 Grandma's Marathon for all 6888 participants was 262 minutes. A sample of 20 runners between the ages of 30 and 39 were randomly selected and their racing times (in minutes) were

239 215 297 196 234 236 268 259 275 206 232 324 237 258 214 312 285 206 218 185

Is this sufficient evidence that the participants in their 30's had a mean racing time different from 262 minutes?

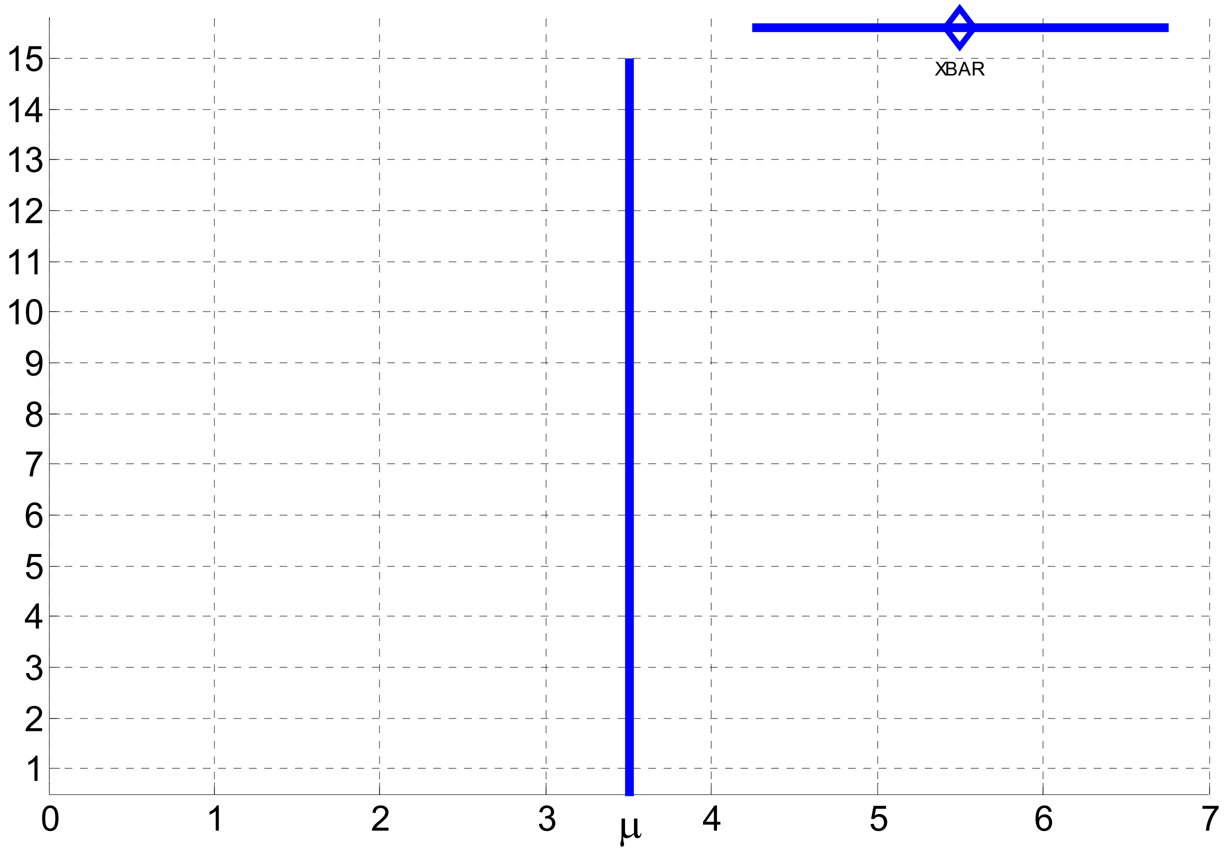
ACTIVITY: ESTIMATING FAMILY SIZE

Suppose you have a population of families and you are interested in estimating the mean family size. For this particular population, 1/6 of the families have a size of 1, 1/6 have size 2, 1/6 have size 3, 1/6 have size 4, 1/6 have size 5, and 1/6 have size 6. Here the population mean size is $\mu = 3.5$ and the standard deviation is $\sigma = 1.71$. We are going to take a sample of five families – compute the mean $\bar{X}$ and find a 90% confidence interval of the form

$$(\bar{X} - 1.645\,\sigma / \sqrt{5}\,, \bar{X} - 1.645\,\sigma / \sqrt{5}) = (\bar{X}\text{ -1.26, }\ \bar{X}\text{ +1.26})$$

What does 90% mean?

1. Roll five dice and find the mean $\bar{X}$.
2. On the graph below draw the 90% confidence interval on the first horizontal line. The thick line below has the right length – just center this line about the value of $\bar{X}$.
3. Continue rolling dice, computing $\bar{X}$, and drawing the confidence interval until you have computed 15 intervals.
4. How many intervals cover the population mean μ = 3.5? What proportion of intervals cover the population mean?
5. Suppose a person computes the interval (2.74, 5.26) and says that the probability that μ is in the interval (2.74, 5.26) is 90%. What is wrong about this statement?


XBAR
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
1
2
3
μ
4
5
6
7

ACTIVITY: ESTIMATING THE TOTAL OF A RESTAURANT BILL

When you are at a restaurant, sometimes you are interested in estimating your total bill. A simple way of doing this is to round each item to the nearest dollar and then add the rounded values to get your estimate. Your estimate may be too low or too high, but you would hope that the low and high estimates would tend to balance out in the long run. If so, then your method would be unbiased. The purpose of this activity is to investigate if this method of estimation is unbiased, and if the method is biased, find an alternative method that could reduce the bias.

1. Order a complete meal (one item from each category) for you and your friend from the below menu. Choose each item at random from a category so that each item is equally likely to be chosen. Write down your order below.

Category	Your order (item and price)	Your friend's order(item and price)
Starter, Soups and Salads		
Main Course		
Desserts and Fountain Specialties		
Drinks		

2. Round each item to the nearest dollar. Find the rounded total, the actual total, and the estimation error ROUNDED TOTAL – ACTUAL TOTAL. (Your estimation error can be negative.)
3. Construct a graph of the estimation errors for all students in your class.
4. Compute the mean error for the dinners in your class. Is this rounding procedure biased? If so, in what direction? Can you suggest a reason for this?
5. Suppose you order one item at random from the 25 items listed on the menu. Complete the table below that lists the ending price, the error when rounded, the frequency, and the probability.

Price ending in	Error when rounded	Frequency	Probability
0.29	-0.29	5	5/25
0.49			
0.69			
0.79			
0.89	0.11	2	2/25
0.95			

6. If X denotes the error when rounding a single item from the menu, find the mean of X. Based on this value, is the method of estimation unbiased?

7. Suppose you ordered 5 items at random from the menu and you estimated the cost of your meal by rounding each item to the nearest dollar and adding the rounded errors. Compute the bias of your estimation procedure.

8. Can you think of a different rounding method that would reduce the size of the bias? Demonstrate that your new method is successful if you purchase a single item from the menu.

Diego and Delilah's Café

Starters, Soups and Salads

Bruschetta	$3.49
Clam Chowder Cup	$2.95
Bowl	$4.29
Spinach, Walnut, and Feta Salad	$4.95
Nachos	$2.29

Main Courses

Bangers and Mash	$7.95
The Six-dollar Burger	$5.95
With Cheese	$6.69
Salmon with Basmati Rice	$12.79
Alamo Baby Back Ribs and Biscuits	$11.79
Roast Chicken with Apricot Sauce	$10.69
Maiale (pork loin with garlic potatoes)	$8.89
Spinach Lasigna	$8.79
Spicy Eggplant with Basil and Bean Sauce	$7.29
Feijoada (Brazilian black beans with orange)	$6.69

Desserts and Fountain Specialties

Shakes	$2.69
Malts	$3.29
Flan	$3.89
Double Chocolate Surprise	$4.95
Fresh Peach Cobbler	$4.95

Drinks

Milk	$0.95
Lemonade	$1.29
Iced Tea	$1.69
Smoothie-of-the-Day	$2.69
Sodas	$0.95

TECHNOLOGY ACTIVITY: IS THE NEIGHBORHOOD EXPANDING?

Suppose you are interested in learning about the mean size of a population of 100 homes in a community. You did a census ten years ago and found that the mean and standard deviation of the home sizes were 4.2 units and 3.6 units, respectively. You wonder if the mean size of a home today is larger than it was ten years ago. If μ denotes the current mean size (in units) of these homes, you test the hypotheses $H_0: \mu = 4.2$, $H_a: \mu > 4.2$. You plan on taking a sample of ten homes, computing the sample mean $\bar{x}$ and rejecting H_0 if $\bar{x} \geq 5.66$ units. In this activity, we use Fathom to investigate the properties of this testing procedure.

1. In Fathom, import the datafile randomrectangles.txt. In the collection, there is an attribute “area” that gives the size (in units) of each of the 100 homes in the population.
2. Define a new slider called “growth”. This slider will represent the growth in house sizes in our community over the last ten years. Give the slider limits from 0 to 5. The current value of the slider should be 0 which means that there is really no growth in the sizes of the homes over the last ten years.
3. Define a new attribute “newarea” that will represent the current area of a home. The formula for newarea is

$$\text{newarea} = \text{area} + \text{growth}$$

4. Take a sample of size 10 with replacement from the population of homes. Define two measures from this sample:

- samplemean = mean(newarea)
- decision which is “reject” if the samplemean ≥ 5.66 and “accept” otherwise

5. Repeat this process of taking a sample and making a decision 1000 times by collecting measures.

6. Use a summary table to find the number of times you reject and accept in your 1000 samples.

7. Now we are set up to answer the following questions:

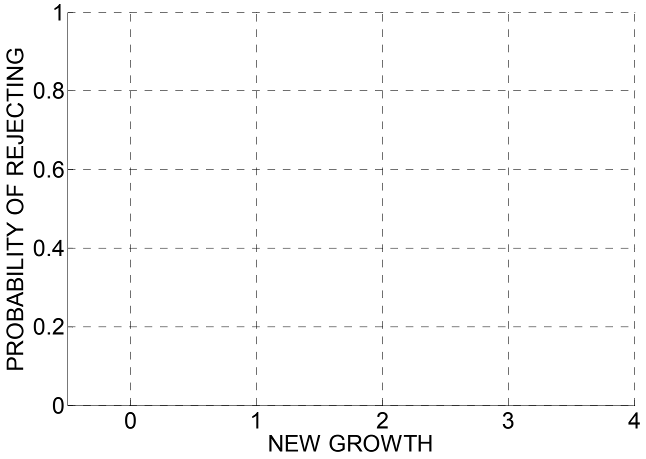
(a) Suppose that there is actually no growth in the community, that is, the value of the slider newgrowth = 0. (You set the slider growth to 0.) What is the probability that you will reject H_0?

(b) Suppose each house has increased in size by one unit. (You set the slider growth to 1.) Find the probability of rejecting.

(c) Suppose each house has increased by two units. Find the probability of rejecting.

(d) Repeat if each house has increased by three units.

(e) A power curve is a graph of the probability of rejecting H_0 plotted as a function of the true growth newgrowth. Draw the power curve of your testing method by plotting the four values above and connecting the points with a smooth curve.



(f) From your work, find the probability of a Type I error. That is, if there is really no growth in the community, find the probability of incorrectly rejecting H_0.

(g) If there is really a 2 unit growth in the homes, find the probability of a Type II error which is the probability of incorrectly accepting H_0 if the true growth is 2 units.

(h) As the true growth increases, does the probability of a Type II error increase or decrease?

EXERCISES

1. **Male Heights**

Heights of American males are approximately normally distributed with mean 70 inches and standard deviation 4 inches.

a. Find the probability a randomly selected male is over 72 inches tall.

b. Suppose you take many random samples of size 25 from the population of American males. Find the mean and standard deviation of the sampling distribution of means.

c. If you take a random sample of 25 men, find the probability the sample mean of heights is larger than 72 inches.

2. **Total Errors in Check Recording**

Suppose you record the amount of a written check to the nearest dollar. Assume that the error between the actual check amount and the written amount is uniformly distributed between -\$0.50 and +\$0.50. The mean and standard deviation of this population are given by $\mu = 0$ and $\sigma = 0.29$, respectively.

a. Suppose you write 30 checks and record the mean error of these checks. What is the shape of the sampling distribution of the mean error?

b. Find the mean and standard deviation of the mean error from writing 30 checks.

c. Find the probability the mean error from writing 30 checks is between -\$0.05 and \$0.05.

3. **Lengths of Phone Calls**

In topic D2, the durations (in minutes) of 58 phone calls of a local company are shown below.

```
2.8    1.8    0.9    0.3    1.2    1.2    2.4   12.2    1.5
5.3    0.9    0.6    4.3    1.0    0.7    0.5    1.2    1.2
6.9    0.6    0.3    4.5    0.6    7.8    0.8    0.9    0.8
5.3    7.5    4.2    2.7    1.5    2.4    6.0    3.6    6.3
1.2    2.7    1.1    0.6    0.4    2.7    1.0    1.4    2.1
0.6    3.1    1.6    1.5    3.0    5.7    0.3    9.9    1.7
3.7    0.6    2.0    3.0
```

The mean and standard deviation of these durations are given by $\bar{x}$ = 2.63 and s = 2.55, respectively.

a. Find a 90% confidence interval for the mean duration of all phone calls μ of this company.

b. Find the margin of error in estimating the mean duration by the sample mean $\bar{x}$.

c. Explain in words what 90% means in the context of this example.

4. **Sleeping of College Students**

The hours of sleep for a sample of 22 college students at a particular school were collected with the following results.

```
7.00    6.50   10.00    6.25    9.50    8.50    6.00    8.00    8.75
9.38    5.75    5.50    8.25    8.50    8.00    8.50    7.80    8.75
5.50    9.00    7.75    6.00
```

a. Identify the population and sample in this problem.

b. Find a 95% confidence interval for the mean number of hours of sleep μ for all college students at this school.

c. Suppose you were interested in estimating the mean μ to with a margin of error of .1 hours with 95% confidence. How large a sample should you take?

5. **Commuting to Work in California**

How long is a typical commuting time in California? A sample was taken of 88 Californians who commuted to work. (The data was collected from the 2000 Census.) The mean commuting time for these workers was 27 minutes with a standard deviation of 22.6 minutes.

a. Define the population of interest and the population mean μ.

b. Construct a 92% confidence interval for the mean commuting time of all Californians who commute.

c. What confidence do you have that μ is actually inside the interval you computed in part b?

6. **Cost of Grocery Shopping**

A family is interested in estimating the mean amount of money they spend at the grocery. For one year, they record the amount of money they spend on 34 visits. The mean purchase amount was $45.74 and the standard deviation was $20.45.

a. Define the population of interest and the population mean μ.

b. Construct a 90% confidence interval for the mean purchase amount for all visits.

c. Find the margin of error of your confidence interval.

d. Suppose you wish to divide this margin of error in half. How many visits should you make to the grocery store to achieve this margin of error?

7. **Worship Attendance**

Last year, the mean Sunday worship attendance at a local church was 368. You believe that the attendance has increased this year. During the first 17 weeks this year, the mean worship was 407 with a standard deviation of 54.

a. Define the population mean m in this problem.

b. State the hypotheses H_0 and H_a.

c. Compute the p-value for the observed sample value $\overline{x}$ =407.

d. Based on your p-value computation, is there sufficient evidence to conclude that the mean worship attendance has increased from last year?

8. **Ages of Women Runners**

The mean age of all of the male runners in the 2005 Grandma's Marathon is equal to 39.6. A sample of 30 women runners was selected and the mean and standard deviation of the ages are given by 32.8 and 10.7, respectively. Is this sufficient evidence to conclude that the mean age of the women runners is different than 39.6?

9. **Cell Phone Lengths**

The lengths of 100 cell phones calls collected from a student's bill are shown below. Suppose this represents the population. The mean and standard deviation of this population are given by $\mu = 6.86$ minutes and $\sigma = 7.32$ minutes, respectively.

COLUMN

ROW	0	1	2	3	4	5	6	7	8	9
0	1	1	2	1	1	1	11	8	3	1
1	35	1	24	3	17	5	23	1	14	1
2	1	1	3	3	10	1	3	1	2	21
3	7	15	16	15	1	12	1	1	3	18
4	2	4	1	14	10	1	28	1	6	24
5	1	2	18	16	18	4	13	1	8	1
6	13	2	15	6	1	4	1	1	7	2
7	1	3	2	5	14	7	3	1	4	1
8	12	4	7	5	6	10	3	5	6	26
9	1	8	10	5	1	4	10	4	2	2

a. Use the random digit table to select a sample of 10 with replacement from this population. Write your sample in the space below.

b. Compute a 80% confidence interval for the population mean using the fact that we know the population standard deviation $\sigma = 7.32$ minutes.

c. Does your interval contain the actual value of the population mean?

d. Repeat the procedure of selecting a sample of 10 and computing a 80% confidence interval nine more times. In the table below, record for each interval if the interval contained the population mean μ.

Interval	1	2	3	4	5	6	7	8	9	10
Cover μ?										

e. By use of the work in this exercise, give an interpretation of 80% confidence.

10. **Cell Phone Lengths (continued)**

Consider again the population of lengths of 100 cell phones calls collected from a student's bill that is displayed in Exercise 9. Recall that the mean of this population is μ = 6.86. Suppose we wish to take a sample of 10 calls to test the hypotheses $H_0 : \mu = 6.86, H_a : \mu > 6.86$. We will decide to reject H_0 if $\bar{x} > 13$ minutes.

a. Use the random digit table to select a sample of 10 with replacement from this population. Write your sample in the space below.

b. Compute the value of $\bar{x}$ and decide whether to reject or accept the null hypothesis H_0.

c. Repeat the procedure of selecting a sample of 10 and making a decision nine more times. In the table below, record for each sample if you decided to reject or accept H0.

Interval	1	2	3	4	5	6	7	8	9	10
Reject H_0?										

d. Based on your work, estimate the probability of making a Type I error.

11. **States Visited by Students**

In topic D2, undergraduate students at a particular college in Ohio were asked to give the number of states in the U.S. that they have actually visited. Here are the answers:

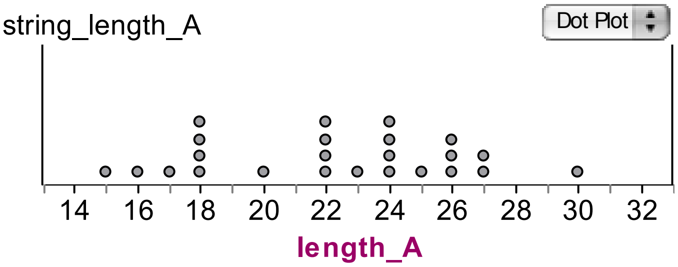
```
46      10      15      14       8      10      13      11      18      13
12      10      11       5      30      32      11      25      14      12
13      11
```

a. Suppose you are interested in estimating the mean number of states μ visited by all undergraduate students at this Ohio college. Find a 90% confidence interval for μ.

b. Would the interval computed in part (a) be an appropriate estimate of the mean number of states visited by undergraduate students from a college in Maryland? Explain.

12. **Measurement Bias**

In the measurement bias activity from topic D3, students were asked to guess at the length of a string with an actual length of 26 inches. A dotplot of the guesses (in inches) from 23 students is shown below.



Let μ denote the mean guess (in inches) of all statistics students who might participate in this activity. Assuming that these measurements are a random sample from the population, is there sufficient evidence from these data to conclude that μ is shorter than the actual string length (that is, $\mu < 26$ inches)?

13. **Words Used in a Sentence**

Suppose you are interested in learning about the mean number of words per sentence for your favorite author. (The mean number of words per sentence is a measure of the reading level of a book.) From one of the books of your author, choose a random sample of 20 sentences and count the number of words in each sentence. (A convenient way of taking this sample is to use the table of random digits to select 20 pages at random and then select the first full sentence on each page.)

a. Record the number of words for your 20 sentences in the space below.

b. Using your collected data, find a 95% confidence interval for the mean number of words for all sentences of all books written by your author.

c. Would there be any possible bias in your sampling procedure since you only sampled sentences from a single book of your author? Explain.

14. **Candy Colors**

The M&M Candy website states that the proportion of orange candies in their bags of M&M's is equal to 20%. You are interested in checking if this statement is correct. You purchase 20 bags of M&M candies where each bag contains 50 candies. If the manufacturer's claim is correct, you would expect to find ten orange candies in each bag. In your twenty bags, you observe the following number of orange candies:

5 14 10 3 6 11 8 7 6 5 10 7 12 8 5 10 4 7 10 7

a. Does these data provide sufficient evidence to reject the manufacturer's claim that 20% of the candies are orange?

b. If the answer to part a is yes, find a confidence interval for the mean number of orange candies in a bag.