



## The Mathematics of the Data and Probability Strand

### How Likely Is It? – Grade 6

#### Unit Overview

#### Summary of Investigations

#### Math Background

- The Meaning of Probability
- Strategies for Finding Outcomes
  - Organized List
  - Tree Diagram
- Experimental vs. Theoretical Probability
  - Comment on the Likelihood of 20 Heads
  - Comment on Random
  - Comment on the Use of Outcomes, Results, and Events
  - Law of Large Numbers
  - Comment on Area and Angles in a Spinner
  - Comment on Experimental and Theoretical Probabilities in Genetics
- Using Probabilities to Make Predictions and Decisions

#### *Content Connections*

### Data About Us – Grade 6

#### Unit Overview

#### Summary of Investigations

#### Math Background

- Different Types of Data
  - Numerical Data
  - Categorical Data
- Distribution
- Data Reduction
  - Standard Graphs
    - Line Plot
    - Frequency Bar Graph
    - Stem-and-leaf plot
    - Coordinate graph
  - Measures of Center
  - Measures of Variation
- Covariation

#### *Content Connections*



## The Mathematics of the Data and Probability Strand

### What Do You Expect? – Grade 7

#### Unit Overview

#### Summary of Investigations

#### Math Background

- Basic Probability Concepts
- Theoretical Probability Models: Lists and Tree Diagrams
- Theoretical Probability Models: Area Models
- Compound Events and Multi-Stage Events
- Expected Value
  - Examples
- More on Independent and Dependent Events
- The Law of Large Numbers
- Binomial Events and Pascal's Triangle

#### *Content Connections*

### Data Distributions – Grade 7

#### Unit Overview

#### Summary of Investigations

#### Math Background

- The Process of Statistical Investigation (Doing Meaningful Statistics)
- Distinguishing Different Types of Data
  - Attributes and Values
  - Categorical or Numerical Values
- Understanding the Concept of Distribution
- Exploring the Concept of Variability
  - What Variability Is and Why It's Important
- Making Sense of a Data Set Using Different Strategies for Data Reduction
  - Using Standard Graphical Representations
    - Line plot
    - Value bar graph
    - Frequency bar graph
    - Scatter plot
  - Reading Standard Graphs
  - Using Measures of Central Tendency
  - Using Measures of Variability
- Comparing Data Sets
- Continuing to Explore the Concept of Covariation

#### *Content Connections*



## The Mathematics of the Data and Probability Strand

### **Samples and Populations – Grade 8**

#### **Unit Overview**

#### **Summary of Investigations**

#### **Math Background**

- The process of statistical investigation (doing meaningful statistics)
- Distinguishing different types of data
  - Attributes and values
  - Categorical or numerical values
  - Understanding the concept of distribution
  - Exploring the concept of variability
  - Making sense of a data set
    - Using standard graphical representations
      - Line plot
      - Histogram
      - Box-and-whisker plot
      - Scatter plot
    - Reading Standard Graphs
    - Using Summary Statistics
  - Comparing data sets
  - Exploring the concept of sampling
  - Exploring the concept of covariation or association

#### ***Content Connections***

## Overview

This is the first unit in the *Connected Mathematics* curriculum that will develop students' abilities to understand and reason about probability. Students will gain an understanding of experimental and theoretical probabilities and the relationship between them. The unit also makes important connections between probability and rational numbers, geometry, statistics, science, and business.

Questions about how likely an event is are asked every day. Such questions ask about the probability of an event happening, and the answers are important to many people. This unit explores different types of probability questions in contexts that are interesting to students, such as games, advertising, contests, and genetics.

## Summary of Investigations

### Investigation 1

#### A First Look at Chance

Investigation 1 introduces students to experimental probabilities and the idea of the chances that some event will occur. Students will have many opportunities to collect data through experimentation using such items as coins and paper cups. Then, they will use the data to assign experimental probabilities to the results. This investigation also introduces students to the notion of equally likely outcomes and that the range of probabilities for a situation is 0 to 1.

### Investigation 2

#### Experimental and Theoretical Probability

This investigation continues students' work with finding experimental probabilities and formally introduces the term *theoretical probability*. The colors of identically-shaped objects chosen from a bag are analyzed both theoretically and experimentally. Students also consider the difference between a particular outcome being *possible* and being *likely* (or probable) and determine if a game or probabilistic situation is fair. Making an organized list or tree diagram are two strategies for finding all the theoretical outcomes.

### Investigation 3

#### Making Decisions With Probability

Investigation 3 introduces spinners as a new context for thinking about probabilities. The crucial difference between spinners and the other objects studied so far is that a spinner has a continuous range of possibilities by subdividing the  $360^\circ$  angle at the end of a spinner into any number of angles. Students also analyze various methods for making fair decisions and devise a simulation to find experimental probabilities.

### Investigation 4

#### Probability, Genetics, and Games

Investigation 4 gives students opportunities to apply and further develop their knowledge about probability in a variety of interesting situations, including applications of probability in genetics and games. Genetics and games can be linked by using a table to find theoretical probabilities of outcomes.

## Mathematics Background

### The Meaning of Probability

The terms *chance* and *probability* apply to situations that have uncertain outcomes on individual trials but a regular pattern of outcomes over many trials. For example, when you toss a coin, you are uncertain whether it will come up heads or tails. But you do know that over the long run, if it is a fair coin, you will get about half heads and half tails. This does not mean you won't get several heads in a row, or that, if you get heads now, you are more likely to get tails on the next toss. Uncertainty of an individual outcome but predictable regularity in the long run is a difficult concept for students to grasp. It often takes a significant amount of time and a variety of experiences that challenge prior conceptions before students understand this basic concept of probability.

If you toss a tack into the air, you know that the tack will land either on its head or its side. If you toss the tack many times, you can use the ratio of the number of times the tack lands on its side to

the total number of tosses to estimate the likelihood that the tack will land on its side. Since you find this ratio through experimentation, the ratio is called an *experimental probability*.

The experimental probability that the tack will land on its side can be expressed as:

$$P(\text{side}) = \frac{\text{number of times the tack lands on its side}}{\text{total number of tosses}}$$

Many uses of probability in daily life are based on experimental probabilities. You collect data for a large number of trials and observe the frequency of a particular result. This is the *relative-frequency interpretation* of probability. The probability that it will rain or that Shaquille O’Neal will make a free throw are two uses of experimental probabilities based on relative frequencies.

Another way to determine probability is to find the *theoretical probability* of a situation. For example, you can examine the theoretical probability of a fair coin landing heads or tails by analyzing the situation. If you toss a fair coin, you know that it will land either heads up or tails up and that each outcome is *equally likely*. Since there are two possible equally likely outcomes, the probability of a fair coin landing heads up is 1 of 2, or  $\frac{1}{2}$ .

You can write this statement as  $P(\text{heads}) = \frac{1}{2}$ . In general, the theoretical probability that a coin will land heads up can be expressed as:

$$\begin{aligned} P(\text{heads}) &= \frac{\text{number of favorable outcomes}}{\text{number of possible outcomes}} \\ &= \frac{1 (\text{heads})}{2 (\text{number of outcomes})} \end{aligned}$$

Another example of a theoretical probability that occurs in this unit involves the roll of a number cube. When you roll a number cube, there are six possible outcomes: 1, 2, 3, 4, 5, and 6.

Each outcome is equally likely on any roll of a number cube. Thus,  $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$ . If you roll a number cube 36 times, you can expect each number to occur about 6 times. You can use this theoretical probability to make an estimate: If a number cube is rolled many times, you can expect each number to occur about  $\frac{1}{6}$  of the time. You can also compute the probability of events that include more than one equally likely outcome. Take the following question, for example:

- What is the theoretical probability of rolling a multiple of 3 on a number cube?

Since two of the six equally likely outcomes, 3 and 6, are multiples of 3, the probability of a multiple of 3 occurring is  $\frac{2}{6}$ , or  $\frac{1}{3}$ .

Some important aspects of the concept of probability are illustrated below using the action of rolling a number cube.

- A probability is a number that is less than or equal to 1 and greater than or equal to 0.
- The sum of the probabilities of all outcomes is equal to 1. In the case of rolling a number cube there are six outcomes, each with a probability of  $\frac{1}{6}$ . The sum is:

$$P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1.$$

- Sometimes problems involve a probability that an outcome A *or* B will occur. For example, to find the probability that the number occurring is either 1 *or* 6, you consider 1 and 6 as favorable outcomes, so

$$P(1 \text{ or } 6) = P(1) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}, \text{ or } \frac{1}{3}.$$

- Sometimes problems may involve a probability that outcomes A *and* B occur. For example, the probability that a number is greater than 3 *and* is even, means that the number must be both. The probability can be thought of as the intersection of the set of numbers that are greater than 3 with the set of even numbers up to 6. Thus,

$$P(> 3 \text{ and even}) = P(4) + P(6) = \frac{2}{6}, \text{ or } \frac{1}{3}.$$

- The probability of rolling the number 7 is  $\frac{0}{6}$ , or 0. It is an impossible outcome.
- The probability of rolling a 1, 2, 3, 4, 5, or 6 is  $\frac{6}{6}$ , or 1. This outcome is certain.
- The probability of rolling a number that is *not* 6 is  $\frac{5}{6}$ , or  $P(\text{not } 6) = P(1, 2, 3, 4, 5, \text{ or } 6) - P(6) = 1 - \frac{1}{6} = \frac{5}{6}$ .

### Strategies for Finding Outcomes

In many situations, making an organized list can help you determine all the possible outcomes. In the situation of rolling a number cube, there are only six outcomes to list. Some situations involve more than one action. For example, suppose you toss a fair coin twice. How many possible outcomes are there? You can list the outcomes as they come to mind, but it is often more efficient to generate the outcomes in a systematic way. This helps to

ensure that you find all the possible outcomes. In the situation of tossing a coin twice, you can list the possibilities for the first toss, namely heads (H) or tails (T). Suppose the first toss resulted in heads (H). What can happen next?

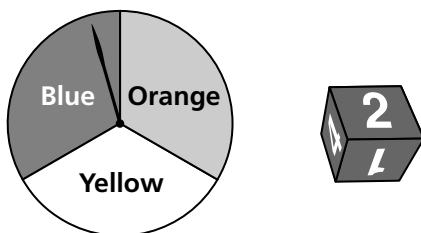
First Toss	Second Toss
H	H
H	T
T	H
T	T

Since the two tosses of the coin are *independent* (the results of one do not affect or depend on the other), you have two possible outcomes (H or T) for the second toss. Thus, you can have either HH or HT. Now suppose the first toss resulted in tails (T). Again there are two possible outcomes for the second toss, H or T. In this case, you can have either TH or TT. Thus, there are four possible outcomes when you toss a coin twice. Since you have considered all the possibilities in a systematic way, you can feel confident that you have found all the possible outcomes.

NOTE: When students toss two coins at once, they may perceive only one way to get a no-match: one head and one tail. Because the two coins are the same, students may not see heads-tails as being different from tails-heads. One way to address this is to toss one coin twice, paying attention to the *order* that the heads and tails come up. Not all students will see this as being the same as tossing two coins at the same time. Another way to investigate the question is to have the coins have different years stamped on them. A 1975 head and 1982 tail are different from a 1975 tail and 1982 head.

**Tree Diagrams and Organized Lists**

Tree diagrams, introduced in Investigation 2, offer students another way to systematically determine all the possible outcomes in a situation. For example, suppose you spin the pointer of a spinner with three sections (made by three angles with the same measure) and you roll a number cube.

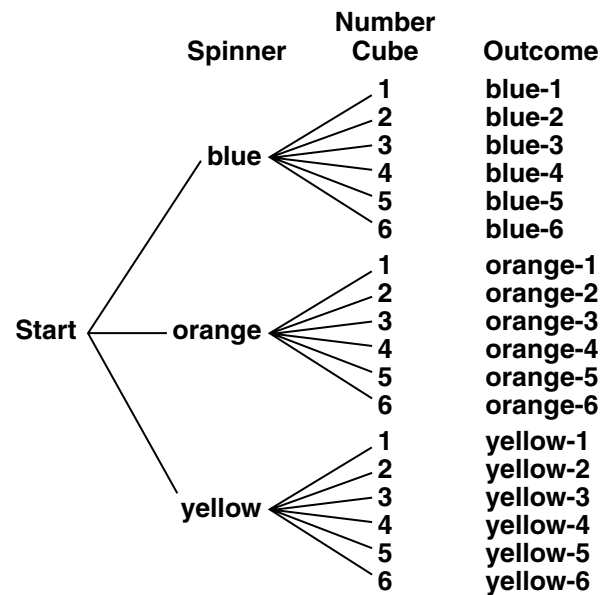


An organized list or a tree diagram can be used to determine all the possible outcomes.

**Organized List**

Color	Number Cube	Outcome
blue	1	blue-1
blue	2	blue-2
blue	3	blue-3
blue	4	blue-4
blue	5	blue-5
blue	6	blue-6
orange	1	orange-1
orange	2	orange-2
orange	3	orange-3
orange	4	orange-4
orange	5	orange-5
orange	6	orange-6
yellow	1	yellow-1
yellow	2	yellow-2
yellow	3	yellow-3
yellow	4	yellow-4
yellow	5	yellow-5
yellow	6	yellow-6

**Tree Diagram**



There are 18 equally likely outcomes, so the probability of each outcome is  $\frac{1}{18}$ .

In this unit, students use tree diagrams to find the number of equally likely outcomes in situations with a large number of possible outcomes. Tree diagrams are particularly useful for listing outcomes in situations involving a series of actions, such as rolling a number cube twice, tossing a coin four times, or choosing several items from a menu, such as a sandwich, drink, and dessert. Tree diagrams can be used as a basis for understanding the multiplication of probabilities, though they are not intended to be used that way in this unit. Students do not yet understand enough about probability to know when and why it is appropriate to multiply probabilities. For example, in the preceding example, the probability of spinning a blue is  $\frac{1}{3}$  and the probability of rolling a 1 is  $\frac{1}{6}$ . The probability of spinning a blue *and* rolling a 1 is  $\frac{1}{3} \times \frac{1}{6}$ , or  $\frac{1}{18}$ .

### Experimental vs. Theoretical Probability

In some situations, it is easier to find theoretical probabilities. In others, it is easier to find experimental probabilities. For example, in this unit students will find experimental probabilities for a tossed paper cup landing on its side or on one of its ends. They will not be able to determine the theoretical probabilities. Although “end” and “side” are the possible outcomes, they are not necessarily equally likely.

Probabilities are useful for predicting what will happen over the long run, yet a theoretical or experimental probability does not tell us exactly what will happen. For example, if you toss a coin 40 times, you may not get exactly 20 heads. However, if you toss a coin 1,000 times, the fraction of heads will be fairly close to  $\frac{1}{2}$ .

### Comment on the Likelihood of 20 Heads

The string HHHHHHHHHHHHHHHHHHHHHHHHH is just as likely as any other string of 20 tosses, such as HTTHHHHTTHTHTHTHTHTT. Since there are two choices for the first position, two for the second position, and so on for all 20 positions, there are  $2 \times 2 \times 2 \times 2 \dots$  (a product of 20 factors of 2), or  $2^{20}$ , different strings that can occur.

A string of 20 heads is one of these 1,048,576 possible strings. Notice that there is only one way to have 20 heads in a row, but there are over a million ways to have a mixture of heads and tails. For example, there are 184,756 ways to get

10 heads and 10 tails. Hence, having some mixture of heads and tails is much more likely than having 20 heads, because there are more ways to arrange them. Still, any one *specific* arrangement of 10 heads and 10 tails is just as likely (or unlikely) as a string of 20 heads. The probability of each specific string is  $(0.5)^{20}$ , or  $\frac{1}{1,048,576}$  (roughly one in a million).

### Comment on Random

In mathematics, *random* has a meaning somewhat different from its everyday usage. In everyday English, random is often used to mean *haphazard* and completely unpredictable, in either the long or short term. In mathematics, random means that any particular outcome is unpredictable, but the long-term behavior is quite stable. When you toss a coin, it is random because you never know whether the next toss will be heads or tails, but you do know that in the long run you will have close to 50% heads.

### Comment on the Use of Outcomes, Results, and Events

Mathematically, an *outcome* is one of the possible results of an experiment or event. Mathematicians also use the term *event* to mean outcome. For example, consider the probability of heads occurring when a coin is tossed. In this situation, heads occurring is the event or outcome. In this unit, we have chosen the language of *outcome* because it seems to be more intuitive for both teachers and students. On occasion, *result* is used if it is more natural. For most purposes, outcome and result are interchangeable terms.

### Law of Large Numbers

Experimental data gathered over many trials should produce probabilities that are close to the theoretical probabilities. This idea is sometimes called the *Law of Large Numbers*. If you can calculate a theoretical probability, you can use it to predict what will happen in the long run rather than having to rely on experimentation. The Law of Large Numbers applies to mathematically random outcomes.

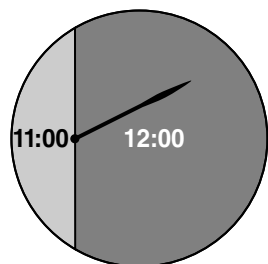
It is important to understand what the Law of Large Numbers says, as well as what it does not say. It does *not* say that you should expect exactly 50% heads in any given large number of trials. Instead, it says that as the number of trials gets

larger, you can expect the percent of heads to be around 50%. For 1 million tosses, exactly 50% (500,000) heads is improbable. But for 1 million tosses, it would be *extremely* unlikely for the percent of heads to be less than 49% or more than 51%.

### Comment on Area and Angles in a Spinner

The probabilities in a spinner are determined by the relative measures of the *angles* in each section rather than by their *areas*. The two are interchangeable in the spinners of Investigation 3 because the center of the spinner is in the center of a circle. The areas of the sections vary in proportion to their angles.

In the spinner below, however, the two outcomes are still equally likely although their areas are not the same. The angles taken up by each section equal  $180^\circ$ .



### Comment on Experimental and Theoretical Probabilities in Genetics

In genetics, the difference between experimental and theoretical probabilities is less clear-cut. Because the basic experiments in genetics deal with reproduction, these experiments can involve causative factors that cannot always be anticipated. You can, of course, study the results of many experiments by studying populations. You can think of the experiment as choosing someone at random. Then, if you know the characteristics of the population as a whole, you have a theoretical probability (just as you did when you knew how many red and blue blocks were in a bag).

We have deliberately chosen not to use the language of *theoretical* and *experimental* with genetics in Investigation 4 in order to avoid confusion. If students raise the question, you might discuss what the experiment is in each case, and the basis for the theoretical probability.

### Using Probabilities to Make Predictions and Decisions

Once you have a probability (theoretical or experimental), you can use it to make predictions. For example, if a coin is tossed 1,000 times, you can predict that heads will occur about 500 times. If you roll a number cube 1,000 times, you can predict that a 3 occurs about  $\frac{1}{6}$  of the time, or about  $\frac{1}{6} \times 1,000 \approx 167$  times. Another way to think of this is as equivalent fractions:  $\frac{1}{6} \approx \frac{167}{1,000}$ . Note that  $\frac{167}{1,000}$  is approximately equal to  $\frac{1}{6}$ .

Students often seek ways to make decisions that are fair. For example, how can you select students for a field trip that can accommodate only ten students? To be fair, the method chosen should give each student an equal (or the same) chance of being chosen. Also, students sometimes find themselves in situations where they would like to know the probabilities of a favorable outcome, such as rubbing two spots on a card containing five spots. Some of the spots, if rubbed, lead to prizes. Such situations can be simulated by an experiment such as choosing colored marbles from a bag. Knowing the probability in these situations can help make decisions about whether to play the game or predict genetic traits.

# Content Connections to Other Units

Big Idea	Prior Work	Future Work
Developing understanding of probability	Performing operations with whole numbers; finding factors and multiples ( <i>Prime Time</i> ); developing understanding of ratio in fraction, percent, or decimal form ( <i>Bits and Pieces I</i> )	Applying rational numbers ( <i>Comparing and Scaling</i> ); finding probabilities and expected values for more complex games and situations ( <i>What Do You Expect?</i> ); using probabilities to make inferences and predictions ( <i>Samples and Populations</i> ); developing counting strategies to help determine probabilities ( <i>Clever Counting</i> © 2004)
Determining experimental probabilities	Working with ratio and proportion ( <i>Bits and Pieces I</i> )	Collecting and organizing data ( <i>Data About Us</i> ); collecting and organizing data from complex games and situations to determine experimental probabilities ( <i>What Do You Expect?</i> ); using experimental probabilities to make inferences and predictions ( <i>Samples and Populations, Clever Counting</i> © 2004)
Determining theoretical probabilities	Analyzing games or situations ( <i>Prime Time</i> ); looking for patterns ( <i>Covering and Surrounding</i> ); working with ratio and proportion ( <i>Bits and Pieces I</i> )	Devising strategies for finding and applying theoretical probabilities ( <i>What Do You Expect?</i> ); making inferences and predictions using theoretical probabilities ( <i>Samples and Populations</i> ); developing counting strategies to determine theoretical probabilities ( <i>Clever Counting</i> © 2004)
Developing understanding of randomness and the Law of Large Numbers	Looking for patterns ( <i>Covering and Surrounding</i> ); working with ratio and proportion ( <i>Bits and Pieces I</i> )	Analyzing and comparing games and situations in which outcomes are random or biased ( <i>What Do You Expect?</i> ); choosing and analyzing random samples to make inferences and predictions about a larger population ( <i>Samples and Populations</i> )
Using probabilities to make predictions	Working with fractions and ratios ( <i>Bits and Pieces I</i> )	Analyzing and comparing games ( <i>What Do You Expect?</i> , <i>Samples and Populations, Clever Counting</i> © 2004)

## Overview

Exploring statistics as a process of data investigation involves a set of four interrelated components (Graham, 1987):

- Posing the question: formulating the key question(s) to explore and deciding what data to collect to address the question(s)
- Collecting the data: deciding how to collect the data as well as actually collecting it
- Analyzing the data: organizing, representing, summarizing, and describing the data and looking for patterns in the data
- Interpreting the results: predicting, comparing, and identifying relationships and using the results from the analyses to make decisions about the original question(s)

This dynamic process often involves moving back and forth among the four interconnected components. For example, collecting the data and, after some analysis, deciding to refine the question and gather additional data. It may involve spending time working within a single component. For example, creating several different representations of the data, some in earlier stages of the process and others at a later time, before selecting the representation(s) to be used for final presentation of the data.

In many of the problems, data are provided. We assume students have had experience collecting data as part of statistical investigations. If they have not, we encourage you to have your class collect their own data for some of the problems. The problems can be applied either to the data provided or to data collected by students.

Even if your students have already had experience collecting data, they may be interested in investigating data about their class. Students will feel empowered if they have the opportunity to use the process of data investigation to explore questions that are of interest to them. Keep in mind that collecting data is time-consuming, so carefully choose the problems for which you will have students generate data.

## Summary of Investigations

### Investigation 1

#### Looking at Data

This first investigation develops some introductory statistical techniques that will be used throughout *Data About Us*. It focuses on describing, interpreting, and comparing distributions. A discussion about the origin of names is used, providing an opportunity to integrate social studies. In addition, students consider lengths of names, and compare distributions of lengths of names from two data sets that are provided and their class's data. Students are introduced to or review the use of tables, line plots, and bar graphs to represent data; ways to describe the shape of a distribution; and the use of measures of center (the mode and median), spread, and range to characterize a distribution.

Students are also introduced to types of data, with a focus on categorical and numerical data. They consider two tables and graphs of data that relate to two questions, one that involves numerical data and one that involves categorical data. Finally, they experiment with using and making horizontal and vertical bar graphs.

### Investigation 2

#### Using Graphs to Explore Data

This investigation first focuses on developing strategies for grouping and displaying data in intervals using stem-and-leaf plots. Data that are collected are often quite spread out or have a great deal of variability. A line plot or bar graph may not be very useful for displaying such data in order to see patterns in the distributions (e.g., clusters, gaps). Students need strategies for grouping and displaying data in equal intervals. The stem-and-leaf plot (or stem plot) is a useful tool for grouping data in intervals of 10, and it helps students see patterns in the data. Students use a stem-and-leaf plot to examine two given data sets. The first data set is about time and distance required for students in a particular class to travel to school. The second data set is about how many times each student in two different classes jumped rope without stopping.

Students then use coordinate graphs to display pairs of data. They begin by collecting data about the lengths of their arm spans and their heights. Using these data, they make a coordinate graph and sketch the  $y = x$  line so they can discuss people who are above, on, or below the line and what this means in terms of the relationship between arm span and height (that is, are most people's arm span and height similar?). They return to the travel time and distance data set and look at a coordinate graph that shows a student's travel time paired with distance traveled in order to discuss whether there is a relationship between travel time and distance traveled (that is, does it take someone who travels farther more time to get to school?).

### Investigation

#### What Do We Mean by *Mean*?

This investigation focuses on developing the concept of mean. The “average” number of people in the families of students in a class provides the setting. The notion of “evening out” or “balancing” the distribution at a point (the mean) located on the horizontal axis is modeled by using cubes and stick-on notes. These models support development of the algorithm for finding the mean: adding up all the numbers and dividing by the number of items.

#### Mathematics Background

In *Data About Us*, several big ideas about statistics are explored. On the next page is a concept map that provides some insights into the overall relationships among these and other important concepts. The shaded portions of the diagram and highlighted graph names are central statistical ideas that are emphasized in *Data About Us*.

#### Different Types of Data

Questions in real life often result in answers that involve one of two general kinds of data: categorical data or numerical data. Knowing the type of data helps us to determine the most appropriate measures of center and displays to use for the data.

#### Numerical Data

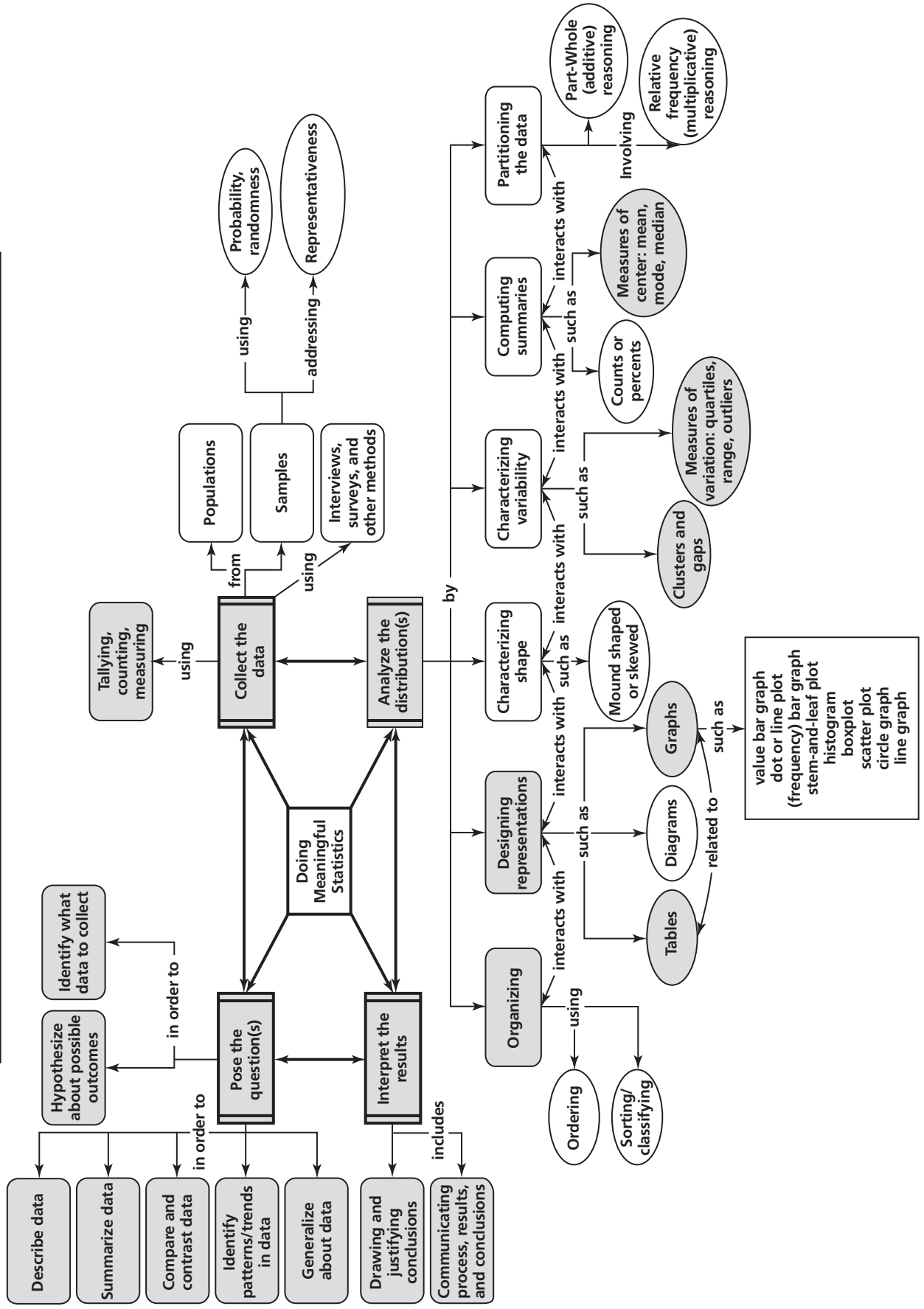
- We can collect data about family size and organize them by using frequencies of how many families have zero children, one child, two children, and so on.
- We can collect data about pulse rates and organize them using intervals by using frequencies of how many people have pulse rates in the intervals of 60 to 69 beats, 70 to 79 beats, and so on.
- We can collect data about height and organize them into intervals by using frequencies of how many people are from 40 to 44 inches tall, 45 to 49 inches tall, and so on.
- We can collect data about time spent sleeping in one day and organize them by frequencies of how many people slept 7 hours,  $7\frac{1}{2}$  hours, 8 hours, and so on.
- We can collect data about responses to a question such as, “On a scale of 1 to 5 with 1 as ‘low interest,’ rate your interest in participating in the school’s field day” and organize them by using frequencies of how many people indicated each of the ratings 1, 2, 3, 4, or 5.
- We can use the mean, median, mode, and range as summary statistics on any numerical data.

#### Categorical Data

- We can collect data about birth years and organize them by using frequencies of how many people were born in 1980, 1981, 1982, and so on.
- We can collect data about favorite type of book to read and organize them by using frequencies of how many people like mysteries, adventure stories, science fiction, and so on.
- We can collect data about hobbies and organize them by using frequencies of how many people collect stamps, build models, knit, and so on.
- Mode is the only summary statistic we can use on categorical data.

At times, categorical data seem to be organized like numerical data. A bar graph of birth months may employ numbers to represent months. For example, 1 is used for January, 2 is used for February, and 3 is used for March. However, we cannot perform numerical operations using months of the year, because months represented numerically are actually categories with a number label representing the category.

Doing Meaningful Statistics – Central Statistical Ideas for Data About Us



## Distribution

The distribution of data refers to the way data occur in a data set. We often use graphs to help us see how data are distributed. A distribution (data as a whole versus individual data values) has characteristics that can be described using statistics such as measures of center or range.

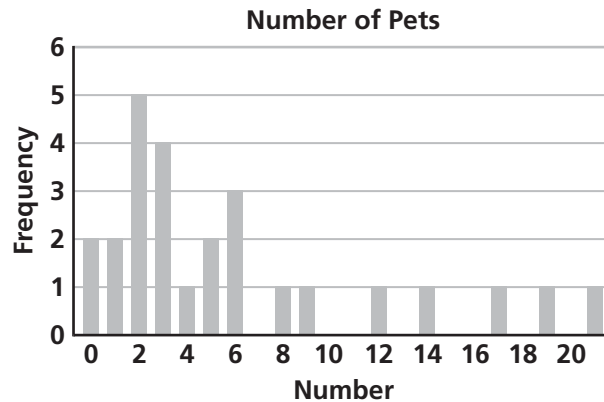
When students work with data, they are often interested in the individual cases, particularly if the data are about themselves. However, statisticians like to look at the overall distribution of a data set and are not interested in individual cases.

We use graphs to help clarify a distribution of data. Distributions (unlike individual cases) have properties such as measures of central tendency (i.e., mean, median, mode) or variation (e.g., outliers, range) and shape (e.g., clusters, gaps).

There appear to be several general ways students think about data:

- Students focus on each data value. For example, they may focus on individual name lengths. They may not see that a group of cases may be related (e.g., several name lengths cluster around lengths of 8 to 10 letters). This kind of thinking is more characteristic of young children. However, when looking at outliers, a focus on individual data values is necessary. How might we explain a name length of 1,019 letters if this data value was part of the data set?
- Students focus on subsets of data values that may be the same or similar like a category or a cluster. This is easier for students when using categorical data (e.g., more students chose dogs as their favorite kind of pet). If students are using numerical data, they might notice clusters (e.g., the number of pets students have at home in the interval of 2 to 3).
- Students view all the data values as an “object” or distribution (see graph of number of pets below). Students look for features of the distribution that are not features of any of the

individual data values (e.g., shape, range, clusters). In looking at the distribution of the number of pets students have, we can see that data are clustered at one end with a kind of tail going off to the right that accounts for several cases in which students have more than six pets.



## Data Reduction

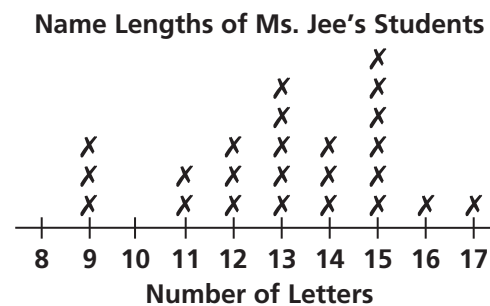
Statisticians use the term “data reduction” to describe what they do when they use representations or statistics during the analysis part of the process of statistical investigation.

## Standard Graphs

Representations in the K–12 curriculum that are addressed in *Data About Us* include the following:

### Line Plot

Each case is represented as an “X” positioned over a labeled number line.



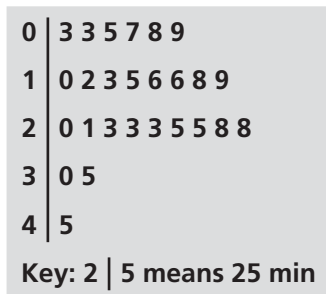
## Frequency Bar Graph

A bar’s height is not the value of an individual case but rather the number (frequency) of cases that all have that value. (See Number of Pets graph above.)

**Stem-and-leaf plot**

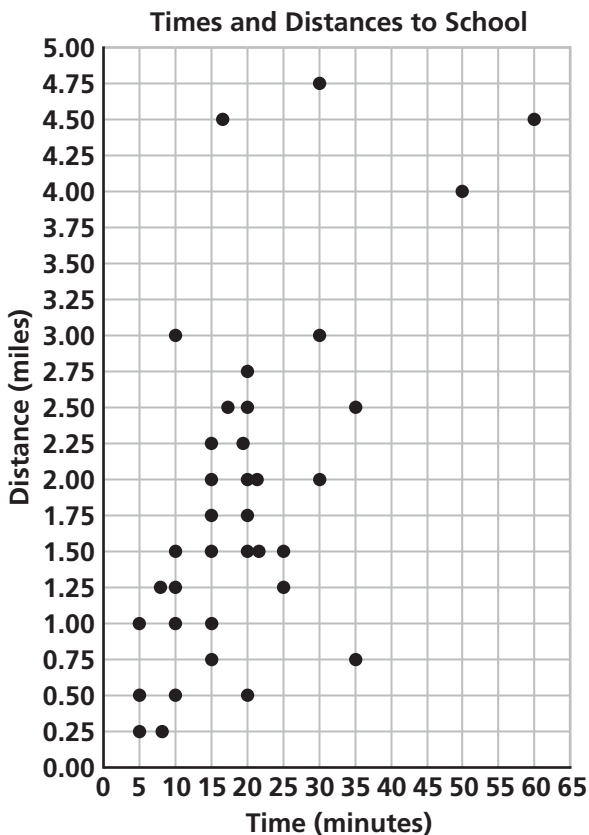
A plot that permits students to group data in intervals (usually by 10's). Stem plots are often introduced as a way to group data that have few repeated values and are spread out. In such situations, the use of line plots provides little information.

**Student Travel Times to School**



**Coordinate graph**

The relationship between two different variables is explored by plotting data values on a Cartesian coordinate system.



As a central component of data analysis, graphs deserve special attention. Three components to graph comprehension that are useful are:

- *Reading the data* involves locating information from a graph to answer explicit questions. For example, “How many students have 12 letters in their names?”
- *Reading between the data* includes using clusters of information presented in a graph. For example, “How many students have more than 12 letters in their names?”
- *Reading beyond the data* involves extending, predicting, or inferring from data to answer questions. For example, “What is the typical number of letters in these students’ names? If a new student joined our class, how many letters would you predict that student would have in his or her name?”

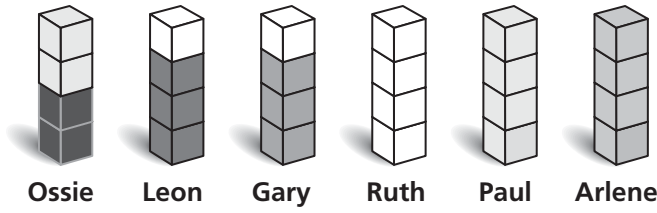
Once students create their graphs, they use them in the interpretation phase of the data-investigation process. This is when they (and you) need to ask questions about the graphs. The first two categories of questions, reading the data and reading between the data, are basic to understanding graphs. However, it is reading beyond the data that helps students to develop higher-level thinking skills such as inference and justification.

**Measures of Center**

We assume mode and median have been addressed during experiences with data analysis in the elementary grades. We also assume that mean may be a relatively new concept. We emphasize the fair share (or evening out) interpretation of mean (average) in *Data About Us*. For example, Ossie has two people in his family. Leon and Gary each have three people in their families. Ruth has four people in her family. Paul and Arlene each have six people. What is the average (mean) number people in these six households?

**Before:**

Ossie	2 people
Leon	3 people
Gary	3 people
Ruth	4 people
Paul	6 people
Arlene	6 people
Total	24 people



**After:**

Ossie	4 people
Leon	4 people
Gary	4 people
Ruth	4 people
Paul	4 people
Arlene	4 people
<hr/>	
Total	24 people

*Mode* is the value that occurs with greatest frequency in a set of data.

*Median* is the value that marks the location that separates an ordered set of data into two equal-sized groups, with the same number of values before the median and after the median.

Although there is one median in a set of data, there may be more than one mode.

In a data set with an even number of values, where the two middle values differ by more than one, the median is the midpoint between these values. For example, the median for the data set 3, 4, 4, 7, 8, 9 is  $5\frac{1}{2}$ , the number that is the midpoint between 4 and 7. If the two middle values are the same number, the median is the value of that number. For example, for the data set 3, 4, 5, 5, 7, 8, the median falls between the two 5's, so the median is 5.

**Measures of Variation**

Measures of variation establish the degree of variability or scatter of the individual data values and their deviations from (or differences from) the measures of center. In *Data About Us*, students use *range* as one measure of variation. Range is the difference between the least and the greatest data values. In addition, students are encouraged to talk about where data cluster and where there are “holes” in the data as further ways to comment about variation.

**Covariation**

Covariation is a way of characterizing a kind of relationship between two variables. It means that information about values from one variable helps us to understand and explain or predict values of the other variable. In *Data About Us*, students are asked to think about whether changes in one variable (e.g., time traveled to school) might be related to changes in another variable (e.g., distance traveled to school). The primary goal for this unit is to review using a coordinate graph to represent data. More formal work with covariation continues in the other data units.

# Content Connections to Other Units

Big Idea	Prior Work	Future Work
Collecting and organizing categorical and numerical data	Analyzing and classifying counting numbers ( <i>Prime Time, Bits and Piece II, Covering and Surrounding, Bits and Pieces III, How Likely Is It?</i> )	Gathering and organizing data collected from conducting experiments or trials of games ( <i>What Do You Expect?, Data Distributions, Samples and Populations</i> )
Representing data with line plots, bar graphs, coordinate graphs, and stem-and-leaf plots	Representing the number of proper factors of a counting number ( <i>Prime Time</i> ), graphing rectangle lengths and widths with constant perimeter or constant area ( <i>Covering and Surrounding</i> )	Representing data to aid with statistical analysis ( <i>Data Distributions, Samples and Populations</i> ); expanding the use of coordinate grids to include negative coordinates ( <i>Accentuate the Negative; Moving Straight Ahead; Thinking With Mathematical Models; Frogs, Fleas, and Painted Cubes; Kaleidoscopes, Hubcaps, and Mirrors; Say It With Symbols; The Shapes of Algebra</i> )
Finding measures of center	Ordering numbers from least to greatest, counting (elementary school, <i>Bits and Pieces I, Bits and Pieces III</i> )	Using measures of center to make inferences and predictions about events or populations ( <i>Data Distributions, Samples and Populations</i> )
Finding measures of variation of a set of data	Comparing, counting, and ordering numbers (elementary school), spread of measures ( <i>Bits and Pieces III</i> )	Using the “variation” or “shape” of a data set to make judgments about the accuracy and reliability of the data and to make inferences and predictions about the group to which the data pertains ( <i>Data Distributions, Samples and Populations</i> )
Calculating the mean	Using arithmetic operations (especially addition and division); learning the meaning of rational numbers (elementary school, <i>Bits and Pieces III</i> )	Developing further understanding about what the mean does and does not measure about a data set; using the mean together with other measures to make predictions and inferences from data ( <i>Data Distributions, Samples and Populations</i> )

## Overview

*What Do You Expect?* is the second probability unit in the *Connected Mathematics* curriculum. The work in this unit assumes that students are familiar with the basic ideas of probability that are presented in the grade 6 unit, *How Likely Is It?* If some or all of your students have not explored the concepts covered in that unit, you will need to prepare them for the mathematics they will encounter in *What Do You Expect?* or consider teaching *How Likely Is It?* If your students have studied *How Likely Is It?*, Investigation 1 of this unit should be a sufficient review, as well as an extension, of the ideas with which they are already acquainted. Through their work in this unit, students will deepen and expand their understanding of basic probability concepts.

## Summary of Investigations

### Investigation 1

#### Evaluating Games of Chance

Investigation 1 uses a variety of situations that provide students a chance to review both experimental and theoretical probabilities, equally likely events, fair/unfair games, and strategies for determining theoretical probabilities. Spinners, choosing marbles from two buckets, and rolling two number cubes provide the settings. These situations also introduce two-stage events. For example, students spin a spinner twice and then look at the outcomes of a match/no-match.

### Investigation 2

#### Analyzing Situations Using an Area Model

Investigation 2 uses the area model as a way to analyze the theoretical probability of two-stage events. The two-stage events used are spinning two spinners, choosing paths in a game, and choosing a marble at random from a container chosen at random.

### Investigation 3

#### Expected Value

In Investigation 3, the two-stage event is a one-and-one free-throw situation. A player with a 60% free-throw shooting average goes for a one-and-one. That is, the player shoots the first free throw and then either takes a second free throw (if the first one was made) or does not get a second chance (if the first free throw was missed). After determining experimental probabilities that the player will get a score of 0, 1, or 2, students find the theoretical probability by using an area model. Students determine the long-term average (expected value) for the situation and explore expected value in a variety of different probability settings.

### Investigation 4

#### Binomial Outcomes

Students are introduced to binomial situations by taking a four-item true-false quiz where each answer is determined by tossing a coin. Students then find the expected value (or average score) for guessing. Students also use lists or trees to determine outcomes. The situations lead naturally to Pascal's Triangle, which is explored in the ACE.

## Mathematics Background

The following is a summary of the basic ideas that are covered in the grade 6 probability unit, *How Likely Is It?*, and descriptions of the new mathematical ideas students will encounter in *What Do You Expect?*

### Basic Probability Concepts

The term *probability* is applied to situations that have uncertain outcomes on individual trials but a predictable pattern of outcomes over many trials. For example, when we toss a fair coin, we are uncertain whether it will come up heads or tails; but we do know that, over the long run, we will get heads about half of the time and tails about half of the time. This does not mean that we can't get several heads in a row. Nor does it mean that if we

get a head on one toss, we are more likely to get tails on the next. This concept of uncertainty on an individual outcome but predictable regularity in the long run is often difficult for students. Students need a variety of experiences that challenge their prior conceptions before they grasp this basic concept of probability.

If we toss a tack into the air, we know that it will land either on its head or its side. If we toss a tack many times, we can use the ratio of the number of times it lands on its side to the total number of tosses to estimate the likelihood that the tack will land on its side. Since this ratio is found by experimentation, it is called an *experimental probability*. Many uses of probability in daily life, such as weather forecasts and sports predictions, are based on experimental probabilities.

This unit offers many opportunities for students to collect data through experimentation and to use their data to assign experimental probabilities to the possible outcomes. It is important for students to realize that comparison of samples with small numbers of trials may show wide variation among the samples, and that only through experimentation over many trials can good estimates be made about what will happen in the long run. In other words, experimental probabilities must be based on a great number of trials relative to the number of possible outcomes in order to have reasonable predictability. In some situations, such as tossing a fair coin, we can also find a *theoretical probability*. We know that a fair coin will land either heads up or tails up and that each outcome is *equally likely*. Since each of the two outcomes is equally likely, the probability that a fair coin will land heads up is 1 out of 2, or  $\frac{1}{2}$ . In a situation where all events are equally likely, the theoretical probability can be expressed as:

$$P(\text{outcome}) = \frac{\text{number of possible favorable outcomes}}{\text{total number of possible outcomes}}$$

The theoretical probability of getting a head on one toss of a fair coin is:

$$P(\text{head}) = \frac{\text{number of possible favorable outcomes}}{\text{total number of possible outcomes}} = \frac{1}{2}$$

Another example of a situation for which we can find a theoretical probability is the rolling of a number cube. The six possible outcomes are 1, 2, 3, 4, 5, and 6 and each are equally likely to occur on any single roll. Thus,

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}.$$

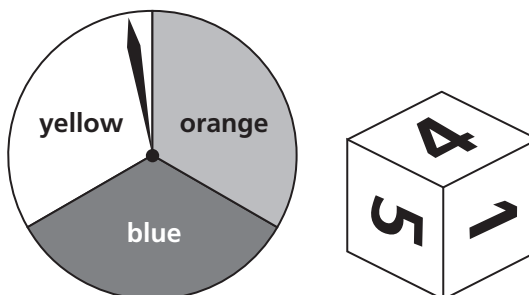
We can use this theoretical probability to estimate that if a number cube is rolled many times, we could expect each number to be rolled about  $\frac{1}{6}$  of the time.

Probabilities, whether obtained through theoretical analysis or experimentation, are useful for predicting what should happen over the long run. Yet, a probability does not tell us exactly what will happen. If we toss a coin 40 times, we may not get exactly 20 heads; but if we toss a coin 1,000 times, the ratio of heads to the number of tosses is likely to be fairly close to  $\frac{1}{2}$ . Experimental data gathered over many trials should produce probabilities that are close to the theoretical probabilities; this idea is sometimes called the *Law of Large Numbers* (see discussion of this on page 9). If we can calculate a theoretical probability, we can use it to predict what will happen in the long run rather than having to rely on experimentation alone.

### Theoretical Probability Models: Lists and Tree Diagrams

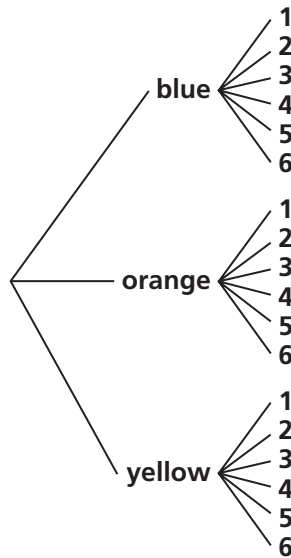
Students who have studied the grade 6 probability unit, *How Likely Is It?*, have already learned quite a bit about conducting simulations to find experimental probabilities and making organized lists of possible outcomes or tree diagrams to find theoretical probabilities. In this unit, they will continue to work with these familiar strategies, while learning a new strategy for finding theoretical probabilities for two-stage events—constructing area models to represent the possible outcomes.

Tree diagrams can be used throughout the unit. They offer students a way to determine all the possible outcomes in a situation systematically, particularly those that are two-stage situations. For example, suppose a spinner divided into three equal sections is spun (stage 1) and a six-sided number cube is rolled (stage 2).



The possible outcomes can be shown in a list and a tree diagram.

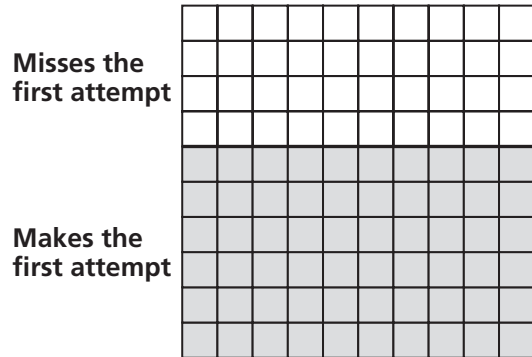
Spinner	Number Cube
blue	1
blue	2
blue	3
blue	4
blue	5
blue	6
orange	1
orange	2
orange	3
orange	4
orange	5
orange	6
yellow	1
yellow	2
yellow	3
yellow	4
yellow	5
yellow	6



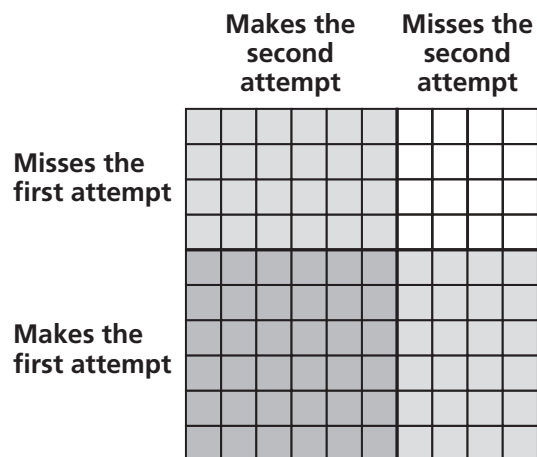
successive events, such as a basketball player who is allowed to attempt a second free throw only if the first succeeds. Unlike tree diagrams, an area model is particularly powerful in situations in which the possible outcomes are not equally likely.

The following steps demonstrate how to create an area model to show the probability that Nishi, a player with a 60% free-throw average, will score 0, 1, or 2 points in a two-try free-throw situation in basketball. In a two-try situation, the player will get to attempt a second free throw whether or not the first free throw succeeds.

The first try has two possible outcomes, making or missing the basket. The probability of missing the basket is 40% or 0.4 or  $\frac{40}{100}$ . The probability of making the basket is 60% or 0.6 or  $\frac{60}{100}$ . The grid below is shaded to indicate this.



The second try has the same two possible outcomes. These are marked vertically on the grid.



The probability that Nishi will make her second try is 60% of the time that she has already made her first try, or 36% of the time. The probability that Nishi will miss her second try is 40% of the time that she makes her first try, or 24% of the time.

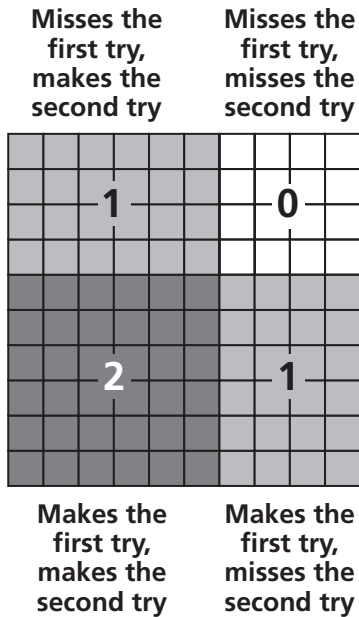
In this unit, students use tree diagrams to find the number of equally likely outcomes in situations with a great number of possible outcomes. Tree diagrams are particularly useful for listing outcomes in situations involving a series of actions in which each outcome of a particular action is equally likely. Such situations include rolling a number cube twice, rolling two number cubes, tossing a coin four times, tossing four coins; or choosing several items from a menu, such as a sandwich, a drink, and a dessert. However, when there are many possibilities at a particular stage, tree diagrams can become unwieldy.

Tree diagrams can be used as a basis for understanding the multiplication of probabilities. Multiplication occasions do arise, but building facility with determining such situations is beyond the scope of this unit. Students do not yet understand enough about probability to know when and why it is appropriate to multiply probabilities.

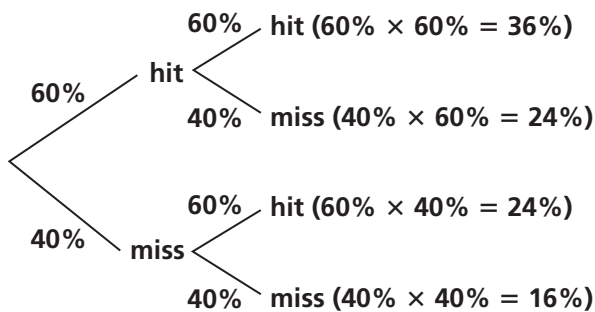
**Theoretical Probability Models: Area Models**

Area models, like tree diagrams, are useful for finding probabilities in situations involving

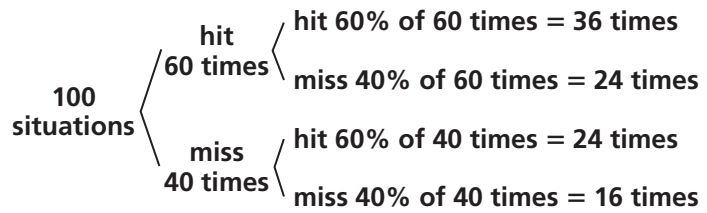
If she missed the first try, she still hits the second one 60% of the time. So the probability of getting a score of 0 is  $\frac{16}{100}$ , getting a score of 1 is  $\frac{48}{100}$ , and getting a score of 2 is  $\frac{36}{100}$ . The grid below indicates this, and each region is labeled with the number of points it represents.



To use a tree-diagram approach in a situation where outcomes are not equally likely, each branch of the tree must be weighted by the probability that it will be chosen. This idea is quite difficult for students at this stage to understand; they have used tree diagrams only in situations involving equally likely outcomes. An area analysis makes the weighting more obvious. It is not recommended that you introduce this idea to your students now, but shown here is a tree diagram that works.

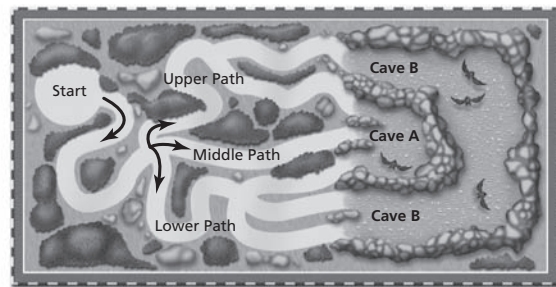


Students, however, will sometimes make a modified version of a weighted tree diagram, pictured below. Such a student might choose a large number of situations (here: 100), then indicate how many of these he would expect to occur on each first branch (here: 60 and 40, corresponding to Nishi's percent of free-throw success). Then each of these numbers is broken down proportionately for the next stage. In effect, this is the same idea as above, but is more accessible to students at this stage of their study of probability.

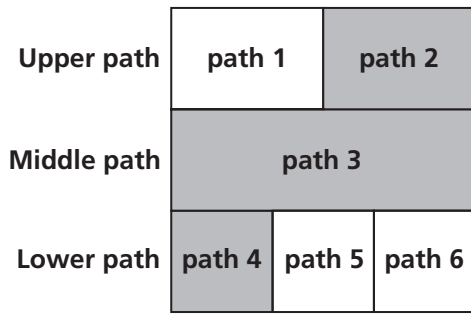


In Problem 3.1, students explore the probabilities of getting a score of 0, 1, or 2 for a person with a 60% free-throw average in a one-and-one situation.

Consider one more example of these ideas. In Investigation 2, students consider a path game (below) in which a player chooses a path at random at each intersection. Students are to figure out the probability of landing in either Cave A or Cave B. Note: The diagram and the analysis here show a different way of labeling the analysis than in the Investigation. This gives you an alternative strategy in case your students are struggling.

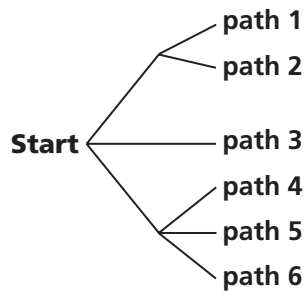


The area model for this game is first split into thirds to indicate the three equally likely paths at the first intersection: the upper path, the middle path, and the lower path. Then each of these thirds is split according to the later intersections (if any), resulting in the model on the following page.

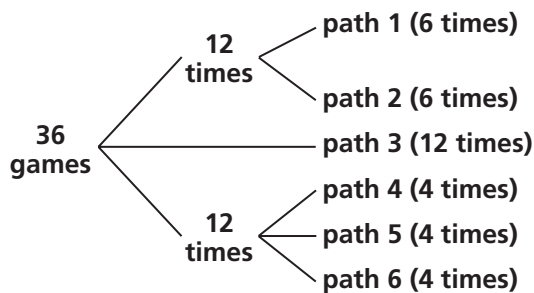


From the area model, it is clear that the 6 paths are not equally likely. Path 3, for instance, has a probability of  $\frac{1}{3}$ , while path 4 has probability of  $\frac{1}{9}$ .

A simple tree diagram would not show this:



But the modified tree diagram described below would in fact represent the differences in the probabilities for each path. Path 3 occurs 12 out of the 36 games, more than any of the other paths.



### Compound Events and Multi-Stage Events

If you are interested in the probability of an event, A, happening, and there are several ways that A can happen, then A is a *compound event*. The probability of A happening is the sum of the probabilities of each possible way that A can happen. For example, if you toss two coins and are interested in finding the probability that you will get a match, there are two ways that A can happen. You can get two heads or two tails. The probability of A,  $P(A)$ , is the sum of the probabilities

of each outcome where two coins match.

$$P(A) = P(t, t) + P(h, h) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

An event is a *multi-stage event* if it takes more than one action to create an outcome. In the example above, event A is a two-stage event, since it takes the toss of two coins (or one coin, twice) to get an outcome. The possibilities are  $(t, t)$ ,  $(t, h)$ ,  $(h, t)$ , and  $(h, h)$ . The question is, what is the probability of each of these outcomes? If the coin is fair, then each coin toss has a probability of landing tails or heads. The coin tosses are *independent* of each other. How the coin lands on a given toss is not affected by any previous toss. Here  $P(t, t) = P(t) \times P(t)$  or  $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ . The same is true for each of the four possible outcomes.

For a player with a 60% free-throw average in a one-and-one situation, whether or not the player gets to take a second try depends on the result of the first try. Here the second try is *dependent* on the result of the first attempt. Thus,  $P(0 \text{ points})$  can only be achieved in one way and that is to miss the first try. The probability of a miss on the first try is 0.4. Thus,  $P(0 \text{ points}) = 0.4$ . There are two possible outcomes resulting from a hit on the first try; the player can hit or miss the second try. So we have  $P(1 \text{ point}) = P(h, h) = 0.6 \times 0.6 = 0.36$  and  $P(h, m) = P(1 \text{ point}) = 0.6 \times 0.4 = 0.24$ . In a one-and-one free-throw situation, we have three possible outcomes, one of which is a one-stage event (0 points) and two of which are two-stage events (1 point or 2 points). The deciding factor is that the second action is dependent on the result of the first action. The sum of all possible outcomes is  $P(0 \text{ points}) + P(1 \text{ point}) + P(2 \text{ points}) = 0.4 + 0.24 + 0.36 = 1$ .

### Expected Value

The “in the long run” perspective of probability is key to understanding probability. Rather than *guarantee* what will happen on a particular trial or even in the short run, probability models *predict* what will happen in the long run over many trials. Often, this is the most valuable information we can gain about a probability situation: a prediction of the expected value of the situation. The *expected value* is the average of the payoff of each outcome weighted by its probability. It predicts long-run expectations.

In this unit, students are introduced to expected value in an informal yet concrete way. We do not

expect them to develop a formal definition of expected value or to use a formula for finding it. In fact, students might never use the term expected value in their work in this unit, instead thinking of the concept as “what is expected in the long run.” However, expected value is vocabulary that the student text uses frequently once it is introduced.

Expected value goes beyond basic probabilities. It uses value, such as points earned in a game or money won in a contest, to weight each possible outcome by then computing the average points or dollars we can expect per game or contest in the long run. You can think of expected value as a “weighted average.”

### Example 1

Consider the long-term average or expected value for a player with a 60% free-throw average in two-try free-throw situations.

If the player goes to the line 100 times, then he/she expects:

A score of 0 to occur 16 times for a total of 0 points and a score of 1 to occur 48 times for a total of 48 points and a score of 2 to occur 36 times for a total of 72 points.

The total number of points expected in 100 situations is  $0 + 48 + 72 = 120$  points. The average number of points expected in 100 trials is  $120 \text{ points} \div 100 \text{ trials} = 1.2$  points per trial.

### Example 2

We could also arrive at this result by the computation

$\frac{16}{100}(0) + \frac{48}{100}(1) + \frac{36}{100}(2) = \frac{0}{100} + \frac{48}{100} + \frac{72}{100} = 1.2$ , which shows each payoff weighted by the probability that it will occur.

The second example is closer to the mathematical definition of expected value but more conceptually difficult for students and is not directly addressed in this unit. Rather, students compute the expected value in steps, as shown in example 1.

### Example 3

The expected value for a player with a 60% free-throw average in a one-and-one situation is:

$$\frac{40}{100}(0) + \frac{24}{100}(1) + \frac{36}{100}(2) = \frac{96}{100} = 0.96$$

A natural question is what free-throw average for a player gives an expected value of exactly 1 point. This question has a surprising answer. The following is a mathematical analysis that shows

how problems such as this one may be revisited in high school when students are ready to solve quadratic equations.

Let  $p$  represent the probability that a player will make the free throw. Then,  $(1 - p)$  represents the probability that the player will miss the free throw. Thus, the probability of

- making 0 points is  $(1 - p)$
- making 1 point is  $p(1 - p)$
- making 2 points is  $p \times p$ , or  $p^2$

The expected value is thus:

$$P(2 \text{ points}) \times 2 + P(1 \text{ point}) \times 1 + P(0 \text{ points}) \times 0$$

Symbolically,

$$p^2 \times 2 + p(1 - p) \times 1 + (1 - p) \times 0$$

Setting the expected value equal to 1, we can solve for  $p$ :

$$p^2 \times 2 + p(1 - p) \times 1 + (1 - p) \times 0 = 1$$

$$2p^2 + p - p^2 + 0 = 1$$

$$p^2 + p = 1$$

Using the quadratic formula to solve the resulting quadratic equation,  $p^2 + p - 1 = 0$ , yields:

$$p = \frac{-1 + \sqrt{1 + 4}}{2} = \frac{-1 + \sqrt{5}}{2}$$

This is the golden ratio, which is approximately 0.6180339887. The golden ratio is the proportion of length to width of a rectangle that many people consider to be the most beautiful rectangle. Many ancient Greek buildings were built with facades that incorporate this ratio.

### More on Independent and Dependent Events

Please note the terms *independent* and *dependent events* are not mentioned in this unit. Naming these ideas can wait until a later course in probability. In this unit, students need only to make sense of each situation and apply the appropriate probability at each stage.

The idea of *independent* and *dependent events* is introduced informally. A more formal approach is often a major focus of probability study in high school and college courses. Yet, we feel it is important to introduce this concept because many students working through a basic probability unit such as this one develop the belief that *all* events are independent.

Suppose you twice choose a marble from a bag containing two red marbles and two blue marbles. If you replace the chosen marble after the first

choice, the two choices will be independent of each other, because what you choose the first time will not affect what you choose the second time. If you do not replace the chosen marble, the second choice will be dependent on the first choice, because the probability of choosing each color the second time depends on the color chosen on the first choice. For example, if you choose a red marble the first time and do not replace it, the probability of choosing a red marble the second time is  $\frac{1}{3}$  rather than  $\frac{1}{2}$ . Yet if you had chosen a blue marble the first time, the probability of choosing red the second time would be  $\frac{2}{3}$ . It is in this sense that the probability of choosing a red on the second choice is a dependent probability.

In this unit, students analyze dependent events by using the situation to help make sense of the sequence of actions. They look at the context and determine the sequence of actions and the possibilities at each step in the sequence. The steps in the sequence guide the apportioning of the total area in an area model, or the designing of a tree diagram representing all possible outcomes. Then, each portion of area in an area model, or each path on a tree diagram, is compared to the total area or the total number of possible outcomes to form probability statements.

Consider an area model for the marbles without replacement:

		Second Choice (with red removed)		
		B	B	R
First Choice	B	BB	BB	BR
	B	BB	BB	BR
	R	RB	RB	RR
	R	RB	RB	RR

The probability of choosing two reds is  $\frac{1}{6}$ . Note that the probability of choosing a red on the second choice is greater if blue was chosen on the first choice.

As students use an area model to make sense of two-stage probability situations, take any opportunity to help those who seem ready to see the connection to multiplying probabilities. For example, in the preceding 60% two free-throw situation,

$$P(\text{score of } 0) = \frac{40}{100} \times \frac{40}{100} = \frac{16}{100}$$

$$P(\text{score of } 1) = \frac{60}{100} \times \frac{40}{100} + \frac{40}{100} \times \frac{60}{100} = \frac{48}{100}$$

$$P(\text{score of } 2) = \frac{60}{100} \times \frac{60}{100} = \frac{36}{100}$$

As an area model is also used to develop an understanding of the multiplication of fractions, many students will see this connection naturally.

### The Law of Large Numbers

The Law of Large Numbers tells us that as we conduct more and more trials, the probabilities drawn from the experimental data should grow closer to the actual probabilities. This idea is difficult for students to grasp; they need time to experiment to develop an understanding of this concept. As you work with the class, talk about the need for many trials in conducting an experiment to find experimental probabilities.

### Binomial Events and Pascal's Triangle

Many interesting probability situations are of the type where there are exactly two equally likely possible outcomes: yes or no, boy or girl, true or false, heads or tails, etc. These are called binomial events. If students guess at every answer for a five-item true/false quiz, there are 32 ways to answer the quiz, but only one of them has all five answers correct. The probability of getting all five answers correct is  $\frac{1}{32}$ . A similar situation involves the families in the town of Ortonville. Each family has exactly five children and they all agree to name their children the same names. There are 32 ways to arrange five children according to numbers of boys and girls (BBGGG, BGBGG, GGGGG, etc.) The probability of a family having exactly five girls is  $\frac{1}{32}$ .

The probability of having two boys and three girls is  $\frac{10}{32}$ . Once one binomial situation has been analyzed it is easy to analyze another binomial situation.

Pascal's Triangle is used to analyze binomial probabilities. The triangle of numbers is named after the seventeenth century mathematician Blaise Pascal. However, the array was in existence long before this. The first five rows are below:

Pascal's Triangle					
1	1				
1	2	1			
1	3	3	1		
1	4	6	4	1	
1	5	10	10	5	1

**Pascal's Triangle and a Coin Toss**

Row	Number of Outcomes
1	Tossing 1 coin
2	Tossing 2 coins
3	Tossing 3 coins
4	Tossing 4 coins
5	Tossing 5 coins

**Pascal's Triangle and a True/False Test**

Row	Number of Outcomes
1	True/false test with 1 question
2	True/false test with 2 questions
3	True/false test with 3 questions
4	True/false test with 4 questions
5	True/false test with 5 questions

The first row states that there are two possible outcomes for tossing a coin, a head or a tail, and there are two possible outcomes for answering a true/false question, true or false. The fifth row states that there is 1 way to get five heads, (1 way to answer all questions true), 5 ways to get four heads and one tail (5 ways to answer four questions true and one question false), 10 ways to get three heads and two tails (10 ways to answer three questions true and two question false), 10 ways to get two heads and three tails (10 ways to answer two questions true and three question false), 5 ways to get one head and four tails (5 ways to answer one questions true and four question false), and 1 way to get five tails (1 way to answer all questions false). A similar analysis can be used for any other binomial situation.

Pascal's Triangle is only presented in an ACE, but students recognize the similarity between the binomial situations and can use previous results to analyze a new situation. An example is a problem that involves a Baseball Series between the evenly matched Gazelles (G) and Bobcats (B). The Gazelles have won the first two games. What is the probability that the series will end in four games? Five games? Six games? Seven games? To answer these questions students analyze the possible outcomes of the last five games. Again there are 32 outcomes. The probability of ending in 4, 5, 6, or 7 games equals  $(\frac{1}{4})$ . However, the Gazelles have a greater chance of winning the series.

# Content Connections to Other Units

Big Idea	Prior Work	Future Work
Gathering data	Gathering, analyzing, and displaying data to show trends ( <i>Data About Us</i> )	Understanding and describing data distributions, sampling techniques, and using samples to predict population behaviors ( <i>Data Distributions, Samples and Populations</i> )
Understanding probability	Understanding chance as the likelihood of a particular event occurring; studying equally likely outcomes and randomness ( <i>How Likely Is It?</i> )	Using probabilities to make inferences and predictions about populations based on analysis of population samples ( <i>Data Distributions, Samples and Populations</i> )
Understanding, determining, and reasoning with experimental probabilities	Conducting trials of a game or experiment to determine experimental probabilities ( <i>How Likely Is It?</i> ); organizing data collected from experiments ( <i>Variables and Patterns; Moving Straight Ahead</i> )	Using data collected from samples of populations to determine experimental probabilities; developing techniques for simulating situations in order to collect and organize data ( <i>Data Distributions, Samples and Populations</i> )
Understanding, determining, and reasoning with theoretical probability	Analyzing simple games to determine theoretical probabilities ( <i>How Likely Is It?</i> ); using an area model for understanding addition and multiplication of fractions ( <i>Bits and Pieces II</i> )	Developing strategies for analyzing complex games or situations to determine theoretical probabilities ( <i>Data Distributions, Samples and Populations</i> ); developing counting strategies to calculate theoretical probabilities ( <i>Clever Counting</i> © 2004)
Finding and reasoning with expected value	Studying favorable outcomes, equally likely outcomes, and random outcomes ( <i>How Likely Is It?</i> )	Using expected values of favorable and unfavorable outcomes to make inferences and predictions; using expected values to make recommendations or to develop solutions to real-world problems ( <i>Data Distributions, Samples and Populations; Clever Counting</i> © 2004)

## Overview

This unit is a new unit for CMP2. It has four investigations that focus students' attention on distributions of data, variability, measures of center, and comparing data sets. The big ideas of the unit are addressed in more detail in the Mathematics Background.

Exploring statistics as a process of data investigation involves a set of four interrelated components:

- Posing the question: formulating the key question(s) to explore and deciding what data to collect to address the question(s)
- Collecting the data: deciding how to collect the data as well as actually collecting it
- Analyzing the data: organizing, representing, summarizing, and describing the data and looking for patterns in the data
- Interpreting the results: predicting, comparing, and identifying relationships and using the results from the analyses to make decisions about the original question(s)

This dynamic process often involves moving back and forth among the four interconnected components—for example, collecting the data and, after some analysis, deciding to refine the question and gather additional data.

In many of the problems, data are provided. We assume students have had prior experience collecting data as part of statistical investigations. If they have not, we encourage you to have your class collect their own data for some of the problems. The problems can be applied either to the data provided or to data collected by students.

Even if your students have already had experience collecting data, they may be interested in investigating data about their class. Students' interest is often enhanced if they have the opportunity to use the process of data investigation to explore questions that are of interest to them. Keep in mind that collecting data is time consuming, so carefully choose the problems for which you will have students generate data.

Problems in contexts are used to help students informally reason about the mathematics of the unit. The problems are deliberately sequenced to provide scaffolding for more challenging problems. Contexts, representations, and describing variability help students develop statistical reasoning.

## Summary of Investigations

### Investigation 1

#### Making Sense of Variability

The first investigation engages students in looking at the variability in data distributions using a variety of contexts involving different kinds of data. Students focus on finding ways to describe distributions. They begin by examining the distribution of colors found in M&M™ candies; there is a consistent pattern that was established by the company making the candies. To what extent is this pattern evident when one bag or many bags are opened and colors counted? Next, students look at numbers of immigrants coming to the United States. They look at two ways to report frequencies: as counts and as percents, or relative frequencies. Finally, students consider measurement error in data. They do this in the context of measuring head sizes to discover what size caps to order.

### Investigation 2

#### Making Sense of Measures of Center

This investigation deepens students' understanding of the three measures of center, their use, and their relationships to shapes of distributions. The mean is reviewed and modeled both as an “equal share” and as a “balance point” in a distribution. The occurrence of repeated values in distributions is examined, and its impact on determining the mode and the location of the median is explored. Students consider a variety of contexts, each represented visually with a graph, and make decisions about the best way to respond to questions using measures of center. Finally, students investigate how changing data values in a distribution—and, consequently, the shape of the distribution—impacts the location of the mean or the median.

### Investigation 3

#### Comparing Distributions: Equal Numbers of Data Values

Students compare data sets with equal numbers of data values. This permits comparisons of frequencies reported using counts. Students explore the data from a computer reaction time game used by a middle-grades class. The data for each person are “scores” (time in seconds to respond) in five trials. Students develop ways to compare individuals and then a group of 40 students. Eventually, they are asked to use these data to make recommendations to a video game designer about the time she needs to give students to react to objects that appear on the screen in her video game.

### Investigation 4

#### Comparing Distributions: Unequal Numbers of Data Values

Students explore comparing data sets with unequal numbers of data values. They use relative frequencies expressed as percents rather than counts. The context is a data set of 150 roller coasters—100 steel coasters and 50 wood coasters. The question involves comparing which coasters are faster, steel or wood. Once that is determined, students look at what other attributes might influence speed, and then they do some informal work with covariation and the use of scatter plots.

### Mathematics Background

In *Data Distributions*, several big ideas about statistics are explored. The sections that follow highlight these important ideas. On the next page is a concept map that provides some insights into the overall relationships among these and other important concepts.

#### The Process of Statistical Investigation (Doing Meaningful Statistics)

This process involves four parts: pose a question, collect the data, analyze the data, and interpret the analysis in light of the question. When completed, students need to communicate the results.

Students need to think about the process of statistical investigation whether they are collecting their own data or are using data provided for them. When students are involved in a problem in which

they do their own data collection, following through with the process of statistical investigation is a natural part of the task. When students are analyzing a data set they have not collected, it is important to help them first understand the data. You can do this by having students ask themselves the same kinds of questions they would ask if they were carrying out the data collection process themselves.

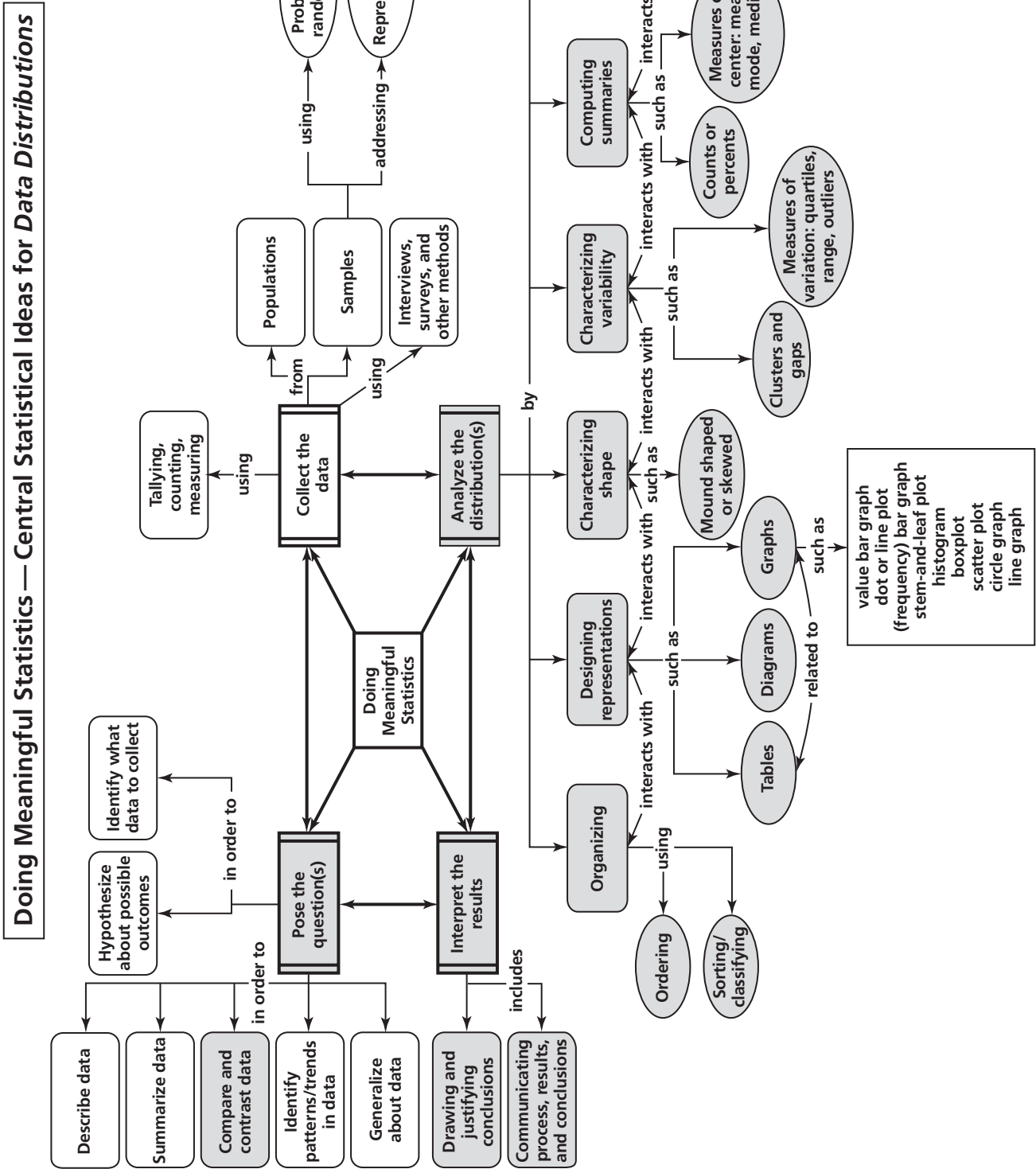
Questions such as these are helpful:

- *What question was asked that resulted in these data being collected?*
- *How do you think the data were collected?*
- *Why are these data represented using this kind of presentation?*
- *What are ways to describe the data distribution?*

In *Data Distributions*, there are several data sets that are provided for your use. The benefit of using provided data is that you know the content that can be developed by using these data sets. However, if you have time, many of the tasks in *Data Distributions* lend themselves to having your students collect their own data, e.g., counting colors of M&M candies in small bags of M&Ms, collecting data about the numbers of grams of sugar in different cereals on different shelves in the local supermarket, and trying out a reaction time game. Students can analyze their own data for some of the problems in this unit in addition to analyzing the data provided.

#### Distinguishing Different Types of Data Attributes and Values

To avoid any confusion with prior algebra work, in *Data Distributions* we refer to attributes (rather than variables) and the values associated with those attributes. An *attribute* is a name for a particular characteristic of a person, place, or thing about which data is being collected. For example, we can have the attribute of “red” to characterize a color of some M&M candies or the attribute of “Fastest Time” to characterize the fastest time taken in five trials reported from a computer reaction time game. *Values* are the data that occur for each individual case of an attribute—that is, the number of red candies recorded for the attribute “red” from one bag of M&M candies or the time in seconds recorded for the attribute “Fastest Time” for one student who played the computer reaction time game.



The data card below shows data about one student, Diana. There are a number of different attribute names on the left that are related to the times reported in playing a computer reaction time game five times. On the right, there is a value for each of these attributes. Diana is one case in a data set of 40 cases.

Attribute	Value	Unit
Name	Diana	
Gender	F	
Age	twelve	
Fastest Time	0.59	sec
Slowest Time	1.08	sec
Trial 1	1.02	sec
Trial 2	0.83	sec
Trial 3	0.73	sec

A second data card for a student named Andrew is also shown.

Attribute	Value	Unit
Name	Andrew	
Gender	M	
Age	eleven	
Fastest Time	0.76	sec
Slowest Time	1.12	sec
Trial 1	1.01	sec
Trial 2	0.8	sec
Trial 3	1.12	sec

Each case has the same attributes; the values for the attributes will be different because each case shown in a data card is about a different student.

### Categorical or Numerical Values

Questions in real life often result in answers that involve one of two general kinds of data values: categorical or numerical. Knowing the type of data values that an attribute has helps us to determine the most appropriate measures of center and displays to use. Students learned to distinguish between categorical and numerical data in *Data About Us*. This unit provides a finer distinction for numerical data, having students focus on both counting and measuring as ways to collect data.

Counted data also are called discrete data. When we use counted data (discrete data values), there are no values possible between consecutive counts; for example:

- We can collect data about family size and organize them by using frequencies of how many families have zero children, one child, two children, and so on, but 1.5 children do not exist in reality.
- We can collect data about responses to a question such as, “On a scale of 1 to 5 with 1 as ‘low interest,’ rate your interest in participating in the school’s field day” and organize them by using frequencies of how many people indicated each of the ratings 1, 2, 3, 4, or 5. In this case, responses between 4 or 5 are not possible because of the stipulation on what choices can be made.
- We can collect data about pulse rates and organize them using frequencies of how many people have pulse rates in the intervals of 60–69 beats, 70–79 beats, and so on. A pulse rate of 65.5 beats is not an option.

With counted data, the mean or median may be decimal numbers but the actual data are reported as whole numbers.

Measurement data also are called continuous data. When we use measurements (or continuous data values), it is possible to measure “between” any two measurements we may have. Of course, the measurement tools we use determine the reality of doing this. Examples include:

- We can collect data about height and organize them into intervals by using frequencies of how many people are between 40–44 inches tall, 45–49 inches tall, and so on. We can measure more exactly to the nearest half-inch, quarter-inch, and so on.
- We can collect data about time spent sleeping in one day and organize them by frequencies of how many people slept 7 hours,  $7\frac{1}{2}$  hours, 8 hours, and so on. We can measure more exactly to the nearest minute or second.

### Understanding the Concept of Distribution

When students work with data, they are often interested in the individual cases, particularly if the data are about themselves. However, statisticians like to look at the overall distribution of a data set. We use graphs to help clarify a distribution of data. Distributions (unlike

individual cases) have properties such as measures of central tendency (i.e., mean, median, mode), or variability (e.g., outliers, range), or shape (e.g., clumps, gaps).

There appear to be several general ways students think about data:

- At the beginning level, students often may focus only on each data value (e.g., each student's own fastest reaction time). They may not see that a group of cases may be related (e.g., several fastest reaction times cluster around 0.7–0.9 seconds). However, when looking at outliers, a focus on individual data values is necessary. For example, how might we interpret a single reaction time of 2.4 seconds if median times in five trials for each of 40 students are  $\leq 1.4$  seconds?
- A next level is to pay attention to subsets of data values that may be the same or similar (i.e., a category or a cluster). For example, if students are using numerical data, they might notice a cluster in the interval of 0.85 and 0.9 seconds for fastest reaction times.
- A final level involves viewing all the data values as an “object” or distribution (Figure 1). Students look for features of the distribution that are not features of any of the individual data values (e.g., shape or clusters). In looking at the distribution of the fastest reaction times, we can see that much of the data are less than 1 second. The distribution is somewhat flat in shape, with data that vary from a little less than 0.6 second to almost 1.2 seconds.

## Exploring the Concept of Variability

### What Variability Is and Why It's Important

When we look at distributions, we often are interested in the measures of center—what's typical (i.e., mean, mode, median). However, any measure of center alone can be misleading. We need to consider the variability of the distribution. Generally, students' earlier work with data analysis has

emphasized describing what is typical about a distribution of data. During the middle grades, there is a shift toward consideration of variability; students are better prepared mathematically and developmentally to consider this concept. Describing variability includes looking at measures of center, range, at where data cluster or where there are gaps in a distribution, at the presence of outliers, and at the shape of the distribution.

Variability refers to the similarities and differences we find among data values in a distribution. There are various causes for variability. In *Data Distributions* students encounter both variability that comes from measurement errors and the natural variability that occurs when studying individual cases in a sample or population. Using statistics and data analysis is all about describing areas of stability (or consistency) in the natural variability that occurs in a distribution. One way to think about variability and stability is to consider addressing the following questions about any set of data with which students are working:

Suppose we are analyzing the distribution of the fastest reaction times for 40 students when they use their dominant hands. (Figure 1)

- If data from a different group of 40 seventh-grade students (who had not played the reaction time game before) were collected, would we expect the distribution of these new data to be the same as or different from the distribution of data for the original 40 students?
- If we expect the distribution to be different, how different and in what ways would it be different? (This question addresses differences among data values, shapes of distributions, locations of data values, and so on.)
- What might we expect to be the same about the two distributions? (This question addresses the use of measures of center, variability, descriptions of shapes of distributions, and so on.)

Figure 1 Fastest Reaction Times for 40 Students (Dominant Hand)



- Several questions highlight interesting aspects of variability. What does a distribution look like? How much do the data points vary from each other? How consistent are the data? What are possible reasons why there is variability in the data?

A distribution's shape is most obvious when we look at a graph—line plot, bar graph, or histogram—of the data. There is a relationship between the shape of a distribution and the locations of the mean and the median. At a gross level, there are distributions in which the mean and median are located close together and there are distributions in which the mean and median are located farther apart. Three different examples of data about the amount of sugar per serving in different cereals are shown below (Figure 2). The shape of the data influences these locations. For graphs A and B, the data are either clustered together or evenly distributed without obvious peaks or clusters. For graph C, the “skewness” of the distribution (a cluster at one end with data values spread out on the other) affects the computation of the mean so that both statistics are not in similar locations.

## Making Sense of a Data Set Using Different Strategies for Data Reduction

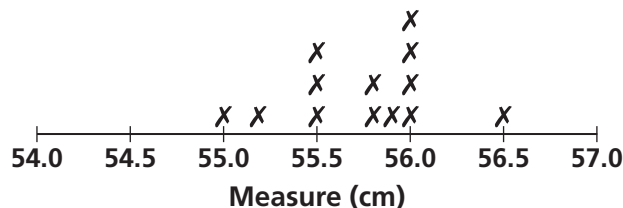
Statisticians use the term *data reduction* to describe what they do when they use representations or statistics during the analysis part of the process of statistical investigation.

### Using Standard Graphical Representations

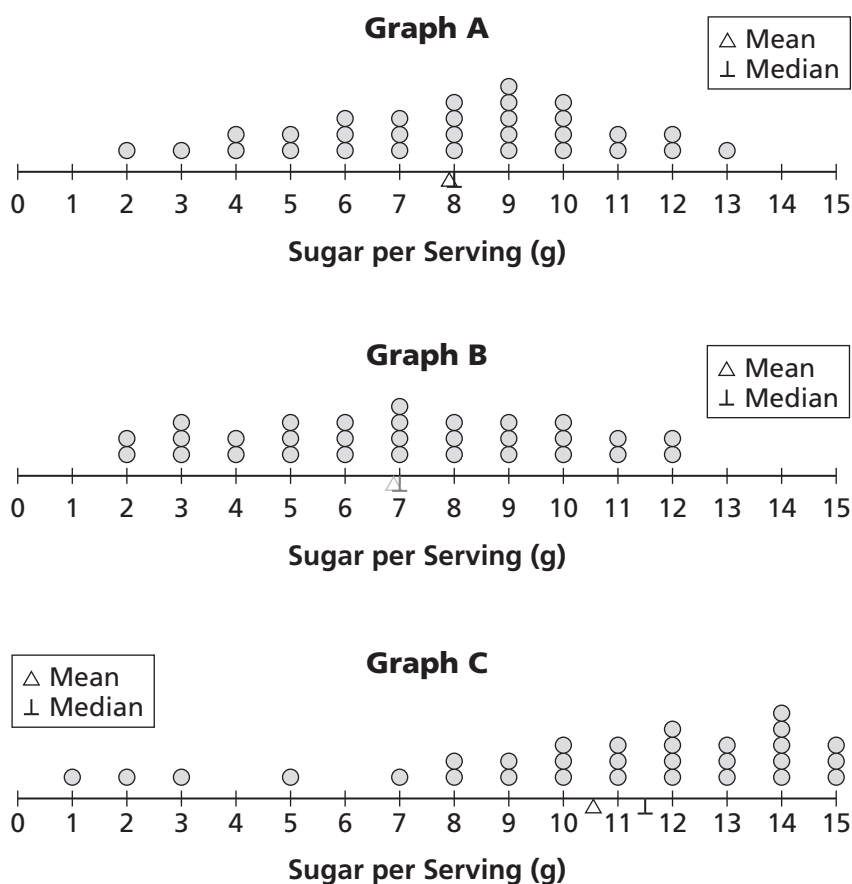
Some often-used representations in the K–12 curriculum that are addressed in *Data Distributions* are shown on the next page.

**Line plot** Each case is represented as an “X” (or a dot) positioned over a labeled number line.

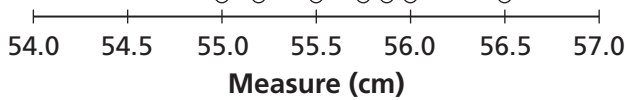
**Line Plot With X's: Measures of Jasmine's Head**



**Figure 2**

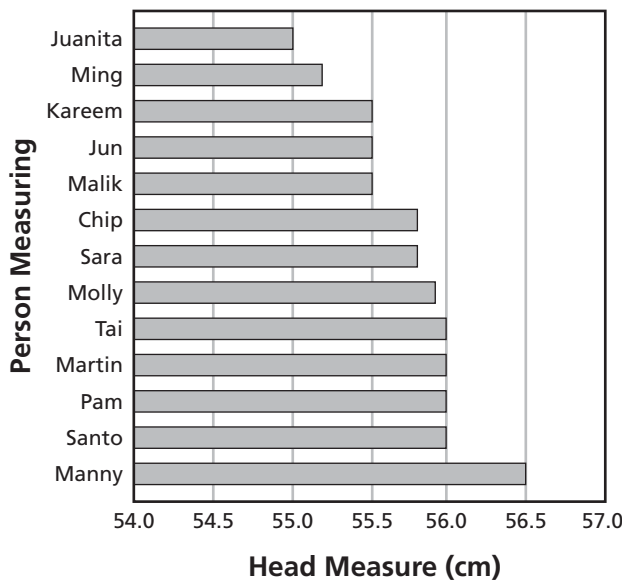


### Line Plot With Dots: Measures of Jasmine's Head

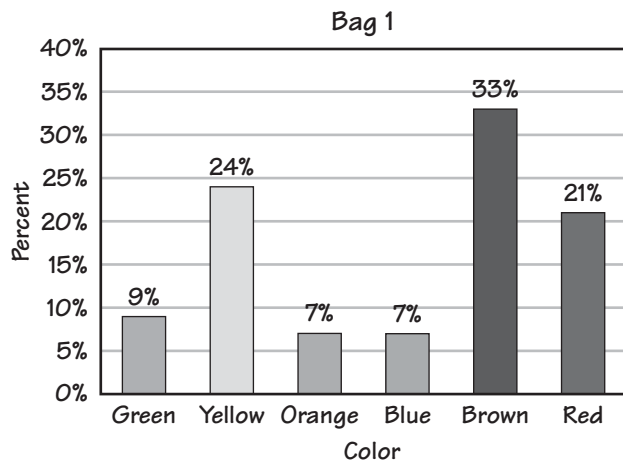


**Value bar graph** Each case is represented by a separate bar whose relative length corresponds to the magnitude or value of that case.

### Ordered Value Bar Graph: Measures of Jasmine's Head

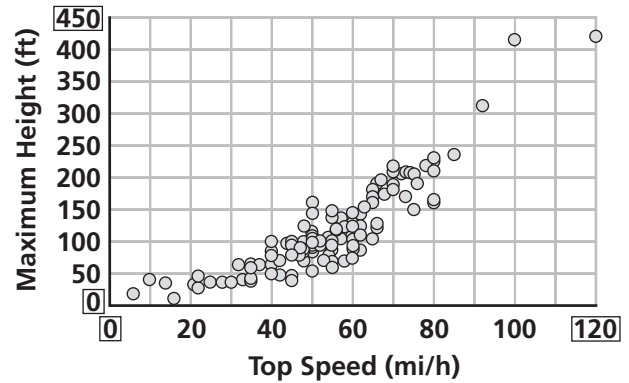


**Frequency bar graph** A bar's height is not the value of an individual case but rather the number (frequency) of cases that all have that value.



**Scatter plot** The relationship between two different attributes is explored by plotting values of two numeric attributes on a Cartesian coordinate system.

### Relationship Between Maximum Height and Top Speed for 150 Roller Coasters



### Reading Standard Graphs

As a central component of data analysis, graphs deserve special attention. In a study of graph comprehension to assess the understanding of students in grades 4 and 7 of four traditional graphs (pictographs, bar graphs, circle or pie graphs, and line graphs), three components to graph comprehension were identified that are useful here.

- *Reading the data* involves “lifting” information from a graph to answer explicit questions. For example, using the data at the left, how many students measured Jasmine’s head size as 56 cm?
- *Reading between the data* includes the interpretation and integration of information presented in a graph. For example, what percent of students’ measures for Jasmine’s head were greater than 55.5 cm?
- *Reading beyond the data* involves extending, predicting, or inferring from data to answer implicit questions. For example, what is the head size you would recommend be used for Jasmine’s head when ordering her cap?

Once students create their graphs, they use them in the interpretation phase of the data-investigation process. This is when they (and you) need to ask questions about the graphs. The first two categories of questions—reading the data and reading between the data—are basic to understanding graphs. However, it is reading beyond the data that helps students to develop higher-level thinking skills such as inference and justification.

### Using Measures of Central Tendency

The three measures of central tendency have been addressed in *Data About Us*. In *Data Distributions*, the intent is to deepen understanding and to explore relationships among the three measures and shapes of distributions.

*Mode* is the data value or category occurring with the greatest frequency. It is ill-defined and sometimes has more than one value. It is unstable because a change in one or a few data values can lead to a change in the mode. It is not usually used for summarizing numerical data. A distribution may be unimodal, bimodal, or multimodal.

*Median* is the numerical value that is the midpoint of an ordered distribution. It is not influenced by extreme data values, so it is a good measure to use when working with distributions that are skewed.

*Mean* is the numerical value that marks the balance point of a distribution; it is influenced by all values of the distribution, including extremes and outliers. It is a good measure to use when working with distributions that are roughly symmetric.

The mean is the same thing as what is usually called the average. There are two interpretations of mean (or average) used in *Data Distributions*:

*Equal share*: If everyone received the same amount, what would that amount be?

*Balance model*: Differences from the mean “balance out” so that the sum of differences for data values below and above the mean equal 0.

Sometimes the mean or median is used to answer the question: What is a typical value that could be used to characterize these data?

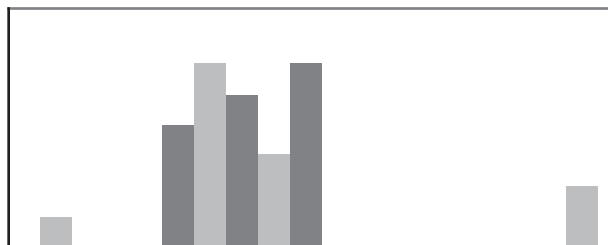
### Using Measures of Variability

Measures of variability establish the degree of variability of the individual data values and their deviations (or differences) from the measures of center. In *Data Distributions*, students use the range as one measure of variability; range

depends on only the minimum and maximum values. Students are encouraged to talk about where data cluster and where “gaps” appear in the data as further ways to comment on variability.

In CMP2 data units, we use minimum and maximum values as terms to specify the least and the greatest values (e.g., the minimum and maximum values are 55 cm and 56.5 cm). The range is a number found by subtracting the minimum value from the maximum value (e.g., the range of the data is 1.5 cm).

In some cases, the range may give you an idea about consistency. At other times, the data can be very consistent, but have outliers that affect the range, as in the following graph.



### Comparing Data Sets

Statistics are useful when comparing two or more data sets. Students must sort out what it means to compare data sets with equal numbers of data values (counts can be used as frequencies) and data sets with unequal numbers of data values (percents must be used as frequencies). It appears that starting with data sets with equal numbers of data values (Investigation 3) and then moving to data sets with unequal numbers of data values (Investigation 4) more readily motivates students to move from counts to percentages to report frequencies.

### Continuing to Explore the Concept of Covariation

Covariation is a way of characterizing a relationship between two (most often) numerical attributes. It means that information about values from one attribute helps us understand, explain, or predict values of the other attribute. In *Data Distributions*, students are asked to consider whether one attribute might help understand the variability in another attribute; for instance, is speed of a roller coaster related to its maximum height? Work with covariation continues to be informal and very concrete.

Big Idea	Prior Work	Future Work
Applying the process of statistical investigation to pose questions, identify ways data are collected, determine strategies for analyzing data and interpreting the analysis to answer the question posed	Collecting and organizing data in different contexts ( <i>How Likely Is It?</i> , <i>Data About Us</i> , <i>What Do You Expect?</i> )	Gathering and organizing data collected from conducting experiments or playing trials of games ( <i>Samples and Populations</i> )
Explaining variability in categorical and numerical data	Finding the range of a set of data ( <i>Data About Us</i> )	Using endpoints, range, and shape of distribution of data to make judgments about the usefulness of the data in helping make inferences and predictions about the group to which the data pertain ( <i>Samples and Populations</i> )
Explaining the difference between collecting numerical data by counting and collecting numerical data by measuring	Collecting and organizing categorical and numerical data ( <i>Data About Us</i> )	Understanding that a measurement has two components, a unit of measure and a count ( <i>Data Around Us</i> ©2004)
Making effective use of a variety of representations to display distributions, including tables, value bar graphs, line plots, and (frequency) bar graphs	Representing the number of proper factors of a counting number ( <i>Prime Time</i> ); representing data with line plots, bar graphs, coordinate graphs, and stem-and-leaf plots ( <i>Data About Us</i> , <i>Accentuate the Negative</i> , <i>Moving Straight Ahead</i> )	Representing data to aid with statistical analysis ( <i>Samples and Populations</i> ); expanding the use of coordinate grids to include negative coordinates ( <i>Thinking With Mathematical Models</i> ; <i>Frogs, Fleas, and Painted Cubes</i> ; <i>Say It With Symbols</i> ; <i>The Shapes of Algebra</i> ; <i>Kaleidoscopes, Hubcaps, and Mirrors</i> )
Understanding and deciding when to use the mean and median to describe a distribution	Finding measures of center ( <i>Data About Us</i> )	Using measures of center to make inferences and predictions about events or populations ( <i>Samples and Populations</i> )
Understanding and using counts or percents to report frequencies of occurrence of data	Percent defined as a ratio of "out of 100" with connections to fractions and decimals ( <i>Bits and Pieces I, II, and III</i> ); using counts to report frequencies ( <i>Data About Us</i> )	Using percentiles to compare samples ( <i>Samples and Populations</i> )
Developing and using strategies for comparing equal-sized and unequal-sized data sets to solve problems	Comparing data sets ( <i>Data About Us</i> )	Making comparisons between groups of different size data sets ( <i>Data Around Us</i> ©2004, <i>Samples and Populations</i> )
Describing how you can use fractions, percents, and ratios as ways to compare sets	Comparing quantities using ratios, proportions, rates, or percents ( <i>Comparing and Scaling</i> , <i>What Do You Expect?</i> )	Comparing samples ( <i>Samples and Populations</i> ), comparing data sets ( <i>Data Around Us</i> ©2004)

## Overview

Statistics is a tool for representing and analyzing data that may then be used to describe a population. Probability is a tool for understanding sampling issues in statistics. The problems in *Samples and Populations* help students make connections between probability and statistics.

This unit applies statistics concepts introduced in grade 6 and reinforced in grade 7. Students begin with an introduction to histograms and box-and-whisker plots as tools for grouping data and comparing distributions. In Investigations 2 and 3, students explore what samples are, how they are related to populations, and ways to select samples, including random samples. In Investigation 4, students look at relationships between two variables and explore how values from one variable can be used to understand, explain or predict values of another variable.

Statistics is the science that relies on data to answer questions. A statistical investigation typically encompasses four interrelated components:

- **Pose the question:** Key questions are formulated and are used to explore and identify what data to collect
- **Collect the data:** Decisions about how to collect the data are made and data are collected
- **Analyze the data:** Data are organized, represented, summarized, and described and patterns in the variability of the distribution are investigated
- **Interpret the results:** The results are used to identify and/or compare relationships, and to make decisions or predictions about answers to the original questions

This dynamic process often involves moving back and forth among the four interconnected components. For example, after collecting some data and doing some analysis of the data, we may decide to refine the question and gather additional data.

In many of the problems in this unit, data are provided. We assume students have had prior experience collecting data as part of statistical investigations. If they have not, we encourage you to have your class collect their own data for some of the problems. The problems can be explored using either the data provided or data collected by students.

## Summary of Investigations

### Investigation 1

#### Comparing Data Sets

Students analyze data from a study on the quality, price, and sodium content of a variety of peanut butters, which are classified by four attributes: natural or regular, creamy or chunky, salted or unsalted, and name brand or store brand. Students review the use of measures of center and are introduced to histograms and box-and-whisker plots as tools for comparing data.

### Investigation 2

#### Choosing a Sample From a Population

Students consider samples and populations, and also use results of analyses of data from samples to make estimates about population characteristics or behaviors. First, students consider the implications of making estimates about the entire U.S. population based on a computer Internet survey involving a few thousand people. The survey raises issues about projecting to an entire population the results from analysis of a sample.

Next, students consider the differences among convenience samples, voluntary-response samples, and random samples. They explore techniques for choosing samples randomly from a population—such as using spinners, number cubes, and random-number generators on calculators—and think about why random samples are often preferable. They then investigate the idea that sample size affects the accuracy of population estimates. Through sampling and determining mean and median statistics for each sample, students learn that the statistics of larger samples are more reliably predictive of the population than statistics from smaller samples.

### Investigation 3

#### Solving Real-World Problems

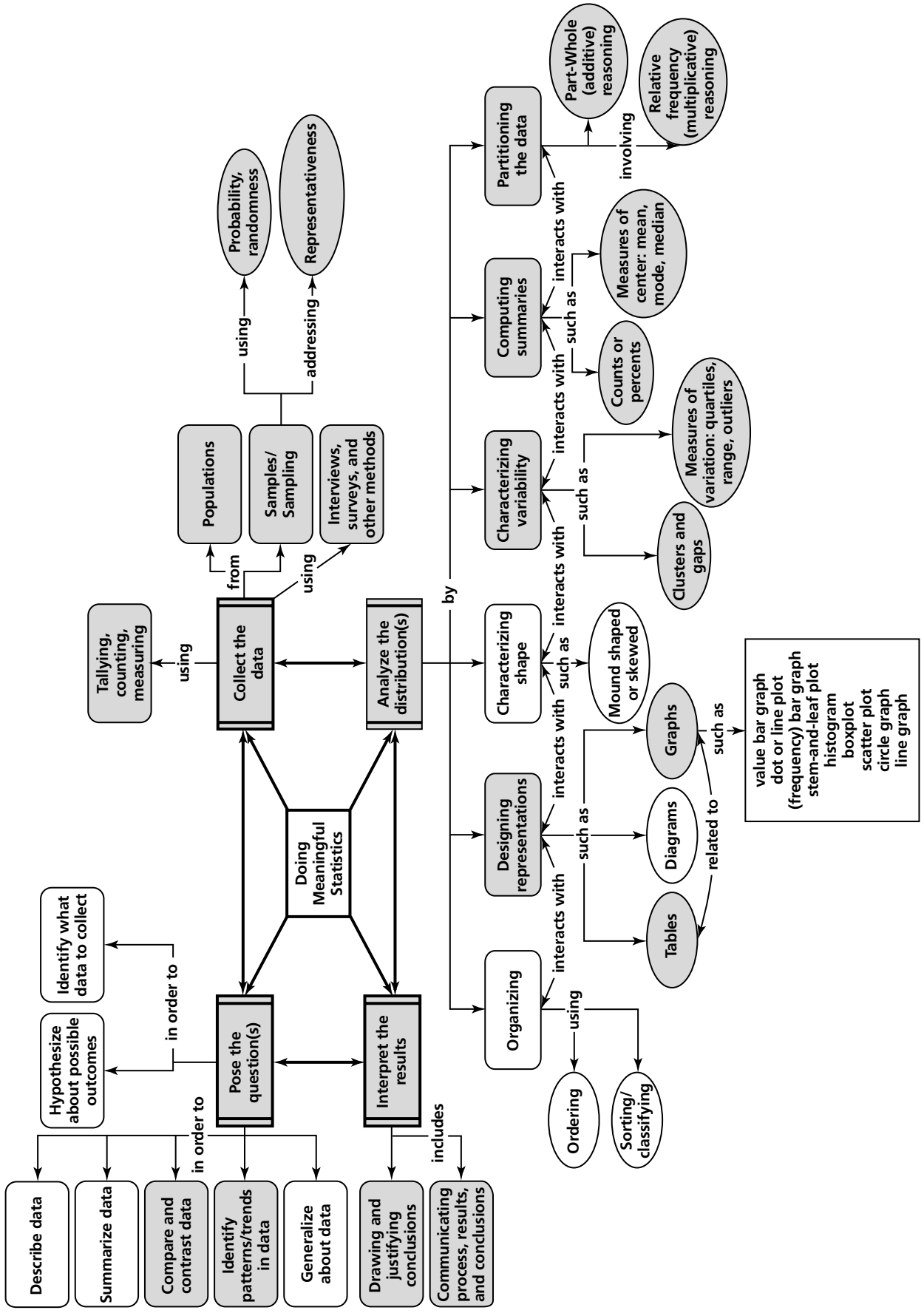
Students apply what they have learned about samples to engaging real-world situations. First, they analyze measurements of Native American arrowheads found at six different archaeological sites. Scientists know the approximate time periods during which four of the sites were settled; the time periods for two newer sites are unknown. Students explore how data from the known sites may be used to make conjectures about the newer sites. Next, they employ a sampling procedure to investigate how many chocolate chips must be added to a batch of cookie dough to ensure that each cookie in the batch will contain at least five chips.

### Investigation 4

#### Relating Two Variables

Students explore how pairs of variables in a given data set may or may not be related, i.e., how one variable varies in relation to the other variable. Students analyze scatter plots of paired values for two different variables in a data set. They consider the question, “If you know a value for one variable, can you make an estimate about the value for a second variable?” To do this, they must consider how strong the relationship is, i.e., how spread out or clustered the paired data values appear to be. They write equations for linear relationships. One kind of linear relationship involves proportions—height and arm span, body length and wingspan, height and foot length. Equations for proportional relationships have a  $y$ -intercept at 0; the scatter plots provided are scaled to facilitate students visualizing a relationship such as  $y = x$  or  $y = 2x$ . The other kind of situation they encounter is one in which the relationship is a constant, e.g., when looking at the scatter plot of quality ratings and sodium content (ACE 1), one could draw a line at  $y = 225$  and then observe how several of the points vary around this line.

**Doing Meaningful Statistics — Central Statistical Ideas for Samples and Populations**



## Mathematics Background

In *Samples and Populations*, several big ideas about statistics are explored. The sections that follow highlight these important ideas. On the preceding page is a concept map that provides some insights into the overall relationships among these and other important concepts.

### The process of statistical investigation (doing meaningful statistics)

This process involves four parts: pose a question, collect the data, analyze the distribution, and interpret the analysis in light of the question. When completed, students need to communicate the results.

Students need to consider the process of statistical investigation whether they are collecting their own data or are using data provided for them. When students collect their own data, following through with the process of statistical investigation is a natural part of the task. When students are analyzing a data set they have not collected, it is important to help them first understand the data. You can do this by having students ask themselves the same kinds of questions they would ask if they were carrying out the data collection process themselves. Questions such as these are helpful: *What question was asked that resulted in these data being collected? How do you think the data were collected? Why are these data represented using this kind of presentation? What are ways to describe the data distribution?*

### Distinguishing different types of data

#### Attributes and values

In order to avoid any confusion with prior algebra work, in *Samples and Populations* we refer to *attributes* (rather than variables) and the *values* associated with those attributes. An *attribute* is a name for a particular characteristic of a person, place, or thing about which data are being collected. For example, we can have the attribute of *kind of peanut butter* to characterize whether a peanut butter is natural or regular or the attribute of *quality rating* to characterize the quality (using a number on a scale) of a given type of peanut butter. Values are the data that occur for each individual *case* of an attribute—that is, for Jif peanut butter, the value for kind of peanut butter is *regular*, the value for consistency is *creamy* and the value for quality rating is 76.

### Categorical or numerical values

Questions in real life often result in answers that involve one of two general kinds of data values: categorical or numerical. Examples of categorical values are *regular* and *natural* for the kind of peanut butter. Examples of numerical values are the numbers used in the quality ratings for peanut butter. Students use both categorical and numerical data in *Samples and Populations*. In *Data Distributions*, students were introduced to discrete and continuous data. Counts are called discrete data and measurements are called continuous data.

### Understanding the concept of distribution

When students work with data, they are often interested in the individual cases, particularly if the data are about themselves. However, looking at the overall distribution of a data set rather than individual cases can reveal important information. We use graphs to help provide a picture of a distribution of data. Distributions (unlike individual cases) have properties that include statistics such as measures of central tendency (i.e., mean, median, mode) or variability (e.g., outliers, range) and characteristics such as shape (e.g., clumps, gaps, skewed distributions). This concept of distribution was addressed in depth in the unit, *Data Distributions*.

### Exploring the concept of variability

In *Samples and Populations*, students encounter situations in which different samples can produce different data and different characteristics of the data. Natural variability is inherent both within and between different samples taken from the same population.

Questions may be used to highlight differences and similarities of the distributions of data for different samples. *What is the shape of a distribution? How much do the data points vary from the mean or median? What are possible reasons why the distributions of data from different samples are different?*

**Making sense of a data set**

Students can use summary statistics, graphical representations, or both during the analysis part of the statistical investigation process.

**Using standard graphical representations**

Often-used representations for the K-12 curriculum addressed in *Samples and Populations* are:

*Line plot:* In a line plot, each case can be represented as a dot (or another mark such as an X) positioned over a labeled number line. A grouped line plot shows data grouped in equal-width intervals. (Figures 1 and 2 below)

*Histogram:* The height of the bar over that interval shows the frequency data values in each interval along the range of data values; frequencies may be displayed as counts or as percentages.

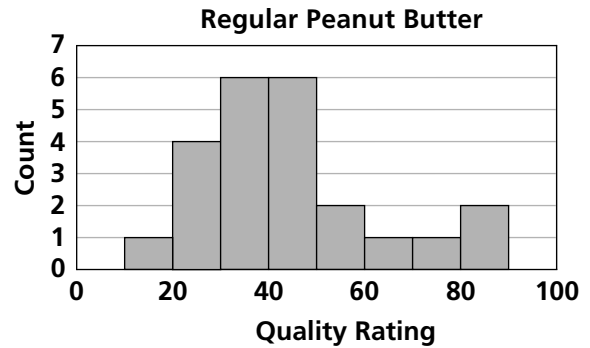


Figure 1

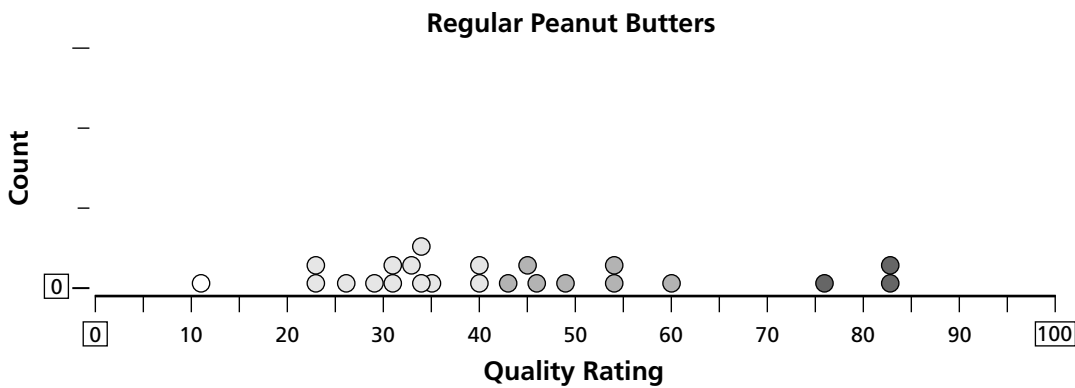
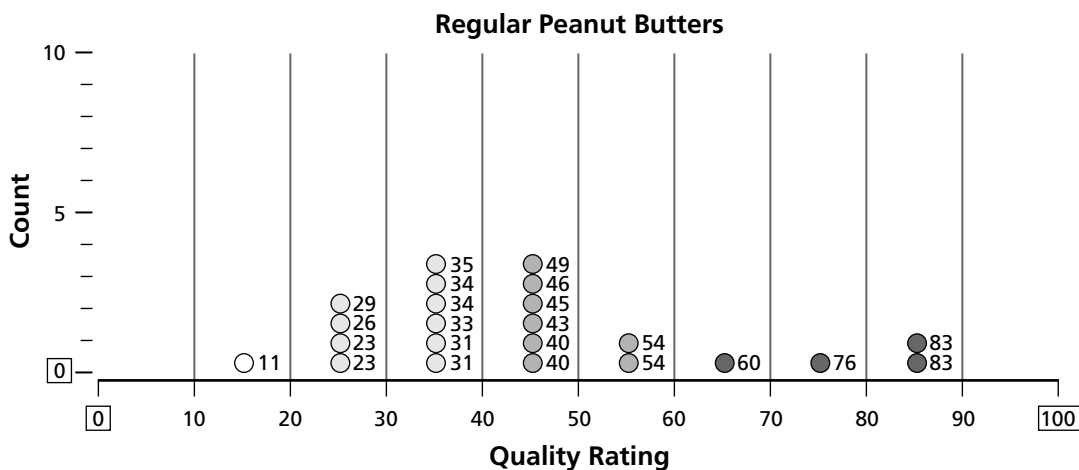
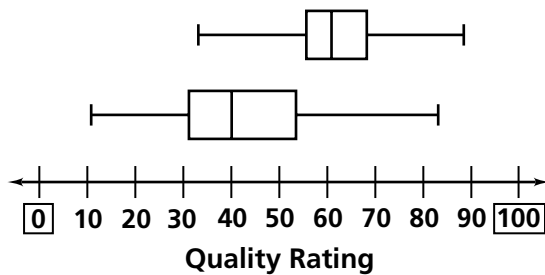


Figure 2



*Box-and-whisker plot:* The box plot is divided into quartiles and displays the properties of distributions, such as symmetry or skewness. This plot was developed largely because comparing data using frequency bar graphs can often be confusing, especially if one is comparing more than two bar graphs.

**Quality Ratings for Peanut Butter**



*Scatter plot:* The relationship between two different attributes is explored by plotting values of the two numeric attributes on a Cartesian coordinate system.

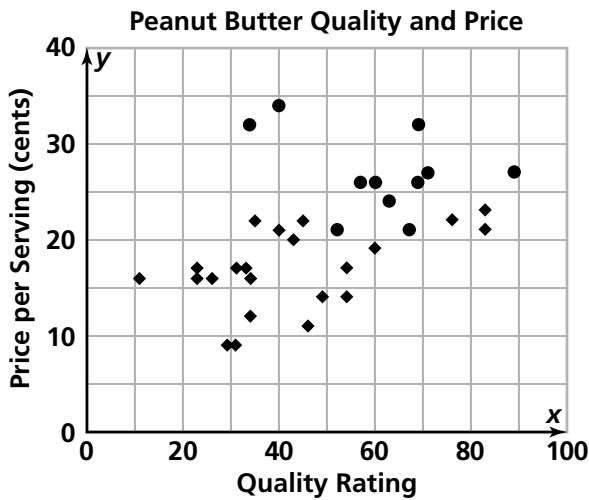
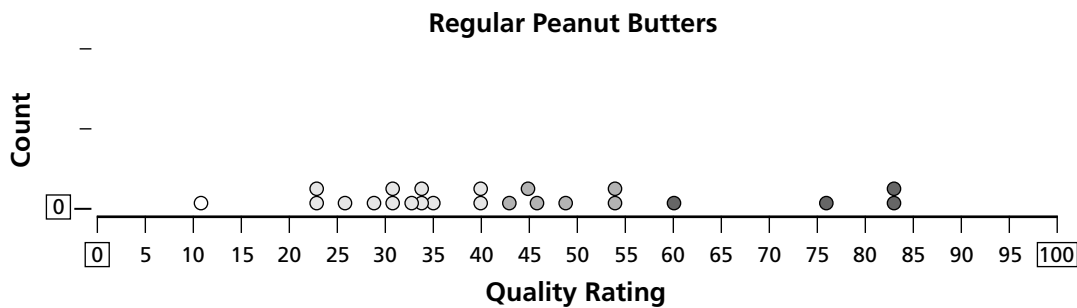


Figure 3



**Reading Standard Graphs**

As a central component of data analysis, graphs deserve special attention. Curcio (1989) identified three components to graph comprehension that are useful here: (Figure 3)

- *Reading the data* involves lifting information from a graph to answer explicit questions. For example, how many regular peanut butters received a quality rating of 40?
- *Reading between the data* includes the interpretation and integration of information presented in a graph. For example, what percent of quality ratings for regular peanut butter are greater than 50?
- *Reading beyond the data* involves extending, predicting, or inferring from data to answer implicit questions. For example, what is the typical quality rating for regular peanut butters?

Once students create graphs, they can use them in the interpretation phase of the statistical investigation process. This is when they (and you) need to ask questions about the graphs. The first two categories of questions—reading the data and reading between the data—are basic to understanding graphs. However, it is reading beyond the data that helps students to develop higher-level thinking skills such as inference and justification.

### Using Summary Statistics

Students use measures of center or variability to summarize a data set.

*Using measures of central tendency or location:*

The three measures of central tendency – mode, mean, median, – have been addressed in *Data About Us*. In *Data Distributions*, students deepened their understanding and explored relationships among the mean and median and shapes of distributions. In *Samples and Populations*, understanding and fluency in the use of the measures is assumed. Students use their knowledge of median to help them understand the basic structure of box plots. They now explore distributions of medians and means taken from several samples as part of the work they do when they explore sample sizes.

*Using measures to describe variability:* Measures of the individual data values and their deviations from (or differences from) measures of center can describe variability. In *Data About Us* and *Data Distributions*, students used both the range and their observations about how the data vary from least and greatest data values as two ways to describe variability. In *Samples and Populations*, students extend their understanding of the central idea of variability by considering quartiles (via box plots) and by developing a method to identify outliers as part of what they do when making box plots. In addition, students continue to be encouraged to talk about where data cluster and where there are *holes* in the data.

### Comparing data sets

Statistics – as attributes of any distribution – serve as useful tools when comparing two or more data sets. The ideas associated with comparing data sets were developed in *Data Distributions*. Students must sort out what it means to compare data sets with equal numbers of data values (counts can be used as frequencies) and data sets with unequal numbers of data values (relative frequencies/percentiles need to be used). Starting with data sets with equal numbers of data values and then moving to data sets with unequal numbers of data values more readily motivates students to move from counts to percentiles to label frequencies. In *Samples and Populations*, most comparison work involves same-size

samples; there are a few cases where students compare unequal-size data sets primarily using box plots, a representation that already is organized using percentiles. However, when students work with histograms, they encounter both counts and percents to report frequencies.

### Exploring the concept of sampling

The essential idea behind sampling is to gain information about the whole by analyzing only a part of it. A census is a sample that consists of the entire population; generally, conducting a census is not possible or reasonable because of such factors as cost and the size of the population. Thus, a primary issue in sampling is choosing a sample. This includes identifying a selection method that avoids bias in the sampling process.

A central issue in sampling is the need for choosing unbiased samples. Students often have intuitive notions about what makes a good sample. They can discuss ways in which certain samples may or may not be fair. Fair means that all samples of the chosen size have the same likelihood of being selected.

To ensure fairness in selecting samples, we try to choose random samples. The concept of randomness is not an easy one for many students to grasp. Every possible sample of the desired size should have an equally likely chance of being selected. The situations involving randomly choosing a sample that are encountered in this unit may all be likened to the idea of “writing each data value on an identical slip of paper, putting each piece of paper in a hat and mixing thoroughly, and then drawing out one or more slips of paper to constitute a sample.”

A number of strategies for selecting random samples are mentioned in this unit, such as spinning spinners, tossing number cubes, and generating lists of values using a calculator. These strategies rely on prior knowledge of probability that students bring to the unit from earlier probability units, i.e., that there is an equally likely chance for any number to be generated by any spin, toss, or key press, and that this number may be used to select a member of a population as part of a sample.

If you use a calculator to generate random numbers, you will need to think about how random digits are generated on the calculators students are using. Most graphing calculators and

many non-graphing calculators have a function for generating decimal numbers; the number of digits in each decimal may be specified (for example, .42 is a two-digit decimal). Students can treat the decimal numbers .00 to .99 as whole numbers for selecting students from the database, with .00 representing student 100 and .01 representing student 1 and so on. Some calculators have a random-integer generator, which takes an argument; that is, one or more numbers are entered as part of the command. The argument consists of the lower and upper bounds of the range within which you are working. For example, on some graphing calculators, `RANDINT(1, 100)` designates the range of whole numbers from 1 to 100.

It is also important to check whether students' calculators generate the same ordered set of random numbers each time the calculator is turned on. If so, the calculator uses a *seed value* that causes it to begin generating random numbers in a specific way. Consult the manual for each calculator to learn how to change the seed value so that each student can generate a different list of random numbers.

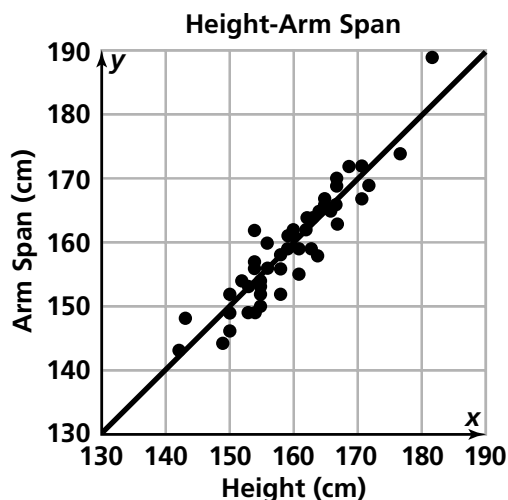
In addition to random sampling, students consider other types of sampling strategies: convenience sampling, voluntary-response sampling, and systematic sampling. It is possible to describe one or more ways in which samples selected using one of these three methods have a greater potential not to be predictive of the population from which they are drawn. Each of these strategies is influenced by factors other than randomness, which means that probability tools cannot be applied.

We want students to develop a general sense about what makes a good sample size. Even with a good sampling strategy, descriptive statistics such as means and medians of the samples will vary in value. However, the accuracy of a sample statistic (i.e., as a predictor of the population statistic) improves with the size of the sample. In Investigation 2, students demonstrate that distributions of means or medians of samples of size 30 generally clustered fairly closely around the actual population mean or median. As a rule of thumb, sample sizes of 25 to 30 are appropriate for most of the settings that students encounter in this unit.

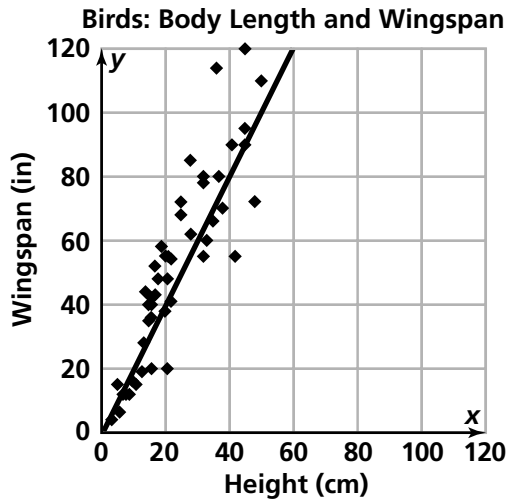
### Exploring the concept of covariation or association

*Covariation*—how two attributes vary in relation to each other—is a way of characterizing a kind of association that is found in a relationship between two numerical attributes and involves analyzing a bivariate distribution displayed using a scatter plot. When the behavior of the values of two different attributes is related in a meaningful way, then information about values from one attribute can help us understand, explain or predict values of the other attribute. Ideas such as fitting a line to and characterizing the strength of a relationship between paired data values for two attributes emerge as ways of describing how the data are distributed. Students develop an awareness that attributes of a data situation can co-vary in some way and that the way they co-vary can be read from a scatter plot.

Fitting a line may be explored informally using a basic understanding of linearity. Many of the relationships explored in Investigation 4 in this unit involve proportional relationships (e.g., height and arm span for people, body length and wingspan for birds); equations for lines characterizing these relationships have a  $y$ -intercept of 0.



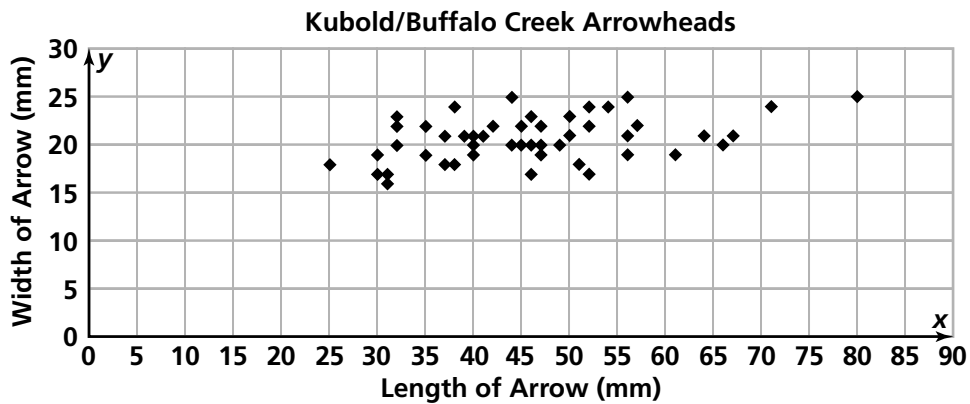
Line for *Arm span* = *height*



Line for  $Wingspan = 2 * body\ length$

In a second kind of situation, the values for one attribute remain relatively constant as the values for the other attribute change. In this case, the fitted line is one in which  $y = a\ constant\ value$ . (Figure 4)

Figure 4



Big Idea	Prior Work	Future Work
Applying the process of statistical investigation to pose questions, identify ways data are collected, determine strategies for analyzing data and interpreting the analysis in order to answer the question posed	Collecting and organizing data into different contexts <i>(Data About Us, Data Distributions, How Likely Is It?, and What Do You Expect?)</i>	Continuing to frame exploration of statistical concepts within the process of statistical investigation ( <i>high school</i> )
Explaining variability in categorical and numerical data	For numerical data: Finding the range of distribution ( <i>Data About Us</i> ) Using range and shape of distribution of data to make inferences and predictions ( <i>Data Distributions</i> ) For categorical data: Analyzing frequencies as counts or percents ( <i>Data About Us and Data Distributions</i> )	Working with graphs, particularly extending measures of spread to include standard deviations ( <i>high school</i> )
Explaining the difference between collecting numerical data by counting or by measuring	Collecting and organizing numerical data ( <i>Data About Us and Data Distributions</i> ) Understanding units of measure and counts ( <i>Numbers Around Us ©2004</i> )	
Making effective use of representations to display distributions, including tables, value bar graphs, dot plot or line plots, and (frequency) bar graphs	Representing data with line plots, value or frequency bar graphs, stem-and-leaf plots, and coordinate graphs ( <i>Data About Us and Data Distributions</i> )	
Deciding when to use the mean and median to describe a distribution	Finding measures of center ( <i>Data About Us</i> ) Examining the behavior of mean and median and shapes of distributions ( <i>Data Distributions</i> )	
Using counts or percents to reports frequencies of occurrence of data	Percents ( <i>Bits and Pieces I, II, III; Comparing and Scaling</i> ) Using counts to report frequencies ( <i>Data About Us and Data Distributions</i> )	
Exploring ways to select samples, including using random sampling techniques	Tying together work with statistics and probability ( <i>How Likely Is It? and What Do You Expect?</i> )	Using and selecting samples and ways to address bias ( <i>high school</i> )
Describing how you can use differences, fractions, percents, and ratios as ways to compare equal or unequal-sized sets	Comparing data sets using ratios, proportions, rates, or percents ( <i>Data About Us, Comparing and Scaling, Data Distributions, and What Do You Expect?</i> )	
Developing a linear equation to characterize a display of data on a scatter plot	Exploring relationships between tables, graphs, and equations and investigating linearity ( <i>Variables and Patterns, Moving Straight Ahead</i> )	Creating and using linear regression models ( <i>high school</i> )