

Unit Introduction

Data Distributions

Describing Variability and Comparing Groups

Goals of the Unit

- Apply the process of statistical investigation to pose questions, to identify ways data are collected, and to determine strategies for analyzing data in order to answer the questions posed
- Recognize that variability occurs whenever data are collected
- Describe the variability in the distribution of a given data set
- Identify sources of variability, including natural variability and variability that results from errors in measurement
- Determine whether to use the mean or median to describe a distribution
- Use the shape of a distribution to estimate the location of the mean and the median
- Use a variety of representations, including tables, bar graphs, and line plots, to display distributions
- Understand and use counts or percents to report frequencies of occurrence of data
- Compare the distributions of data sets using their centers (mean, median, and mode), variability (outliers and range), and shape (clusters and gaps)
- Decide if a difference among data values or summary measures matters
- Develop and use strategies to compare data sets to solve problems

Developing Students' Mathematical Habits

The overall goal of *Connected Mathematics* is to help students develop sound mathematical habits. Through their work in this and other data units, students learn important questions to ask themselves about any situation involving data analysis, such as:

- *Is there anything surprising about the data and their distribution?*

- *Where do the data cluster in the distribution?*
- *How can I use the mean or median and range to help me understand and describe a data distribution?*
- *What strategies can I use to compare two different data sets?*

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 ≤ 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

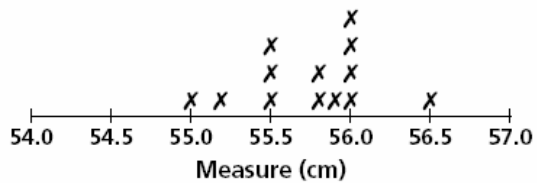
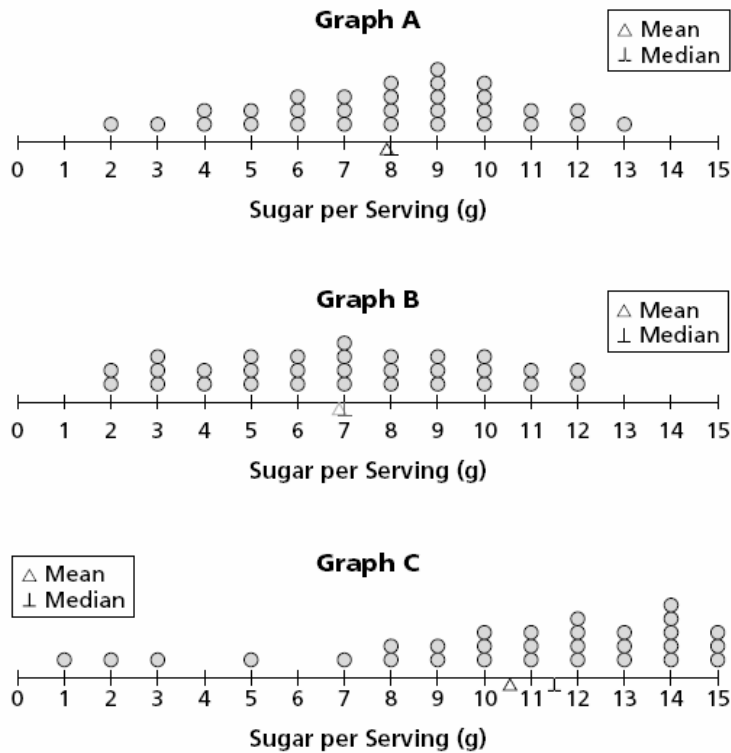
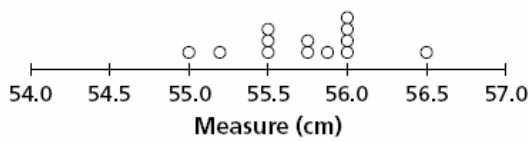


Figure 2

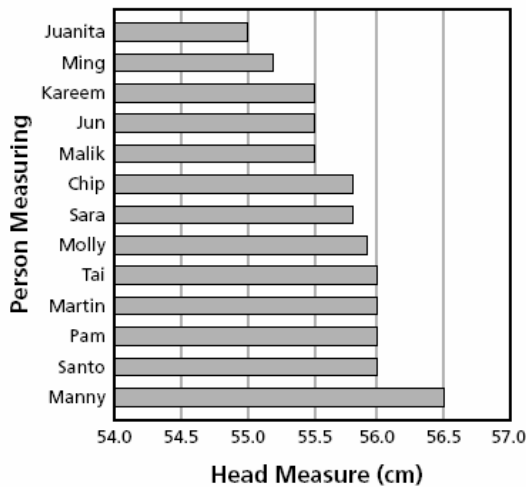


Line Plot With Dots: Measures of Jasmine's Head

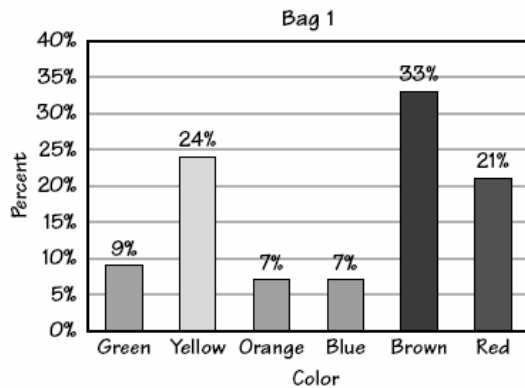


Value bar graph Each case is represented by a separate bar whose relative length corresponds to the magnitude or value of that case.

Ordered Value Bar Graph: Measures of Jasmine's Head

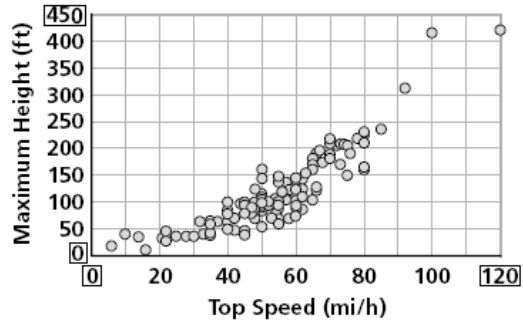


Frequency bar graph A bar's height is not the value of an individual case but rather the number (frequency) of cases that all have that value.



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?

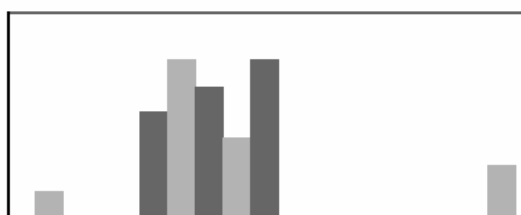
Using Measures of Variability

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.