# Analyzing Data

## What Is It?

Data analysis is the process of interpreting the meaning of the data we have collected, organized, and displayed in the form of a table, bar chart, line graph, or other representation. The process involves looking for patterns—similarities, disparities, trends, and other relationships—and thinking about what these patterns might mean.
When analyzing data, ask students questions such as:

• What pattern do you see?
• What does this graph tell you?
• Who could use this data? How could they use it?
• Why is this data shown in a line graph?

The process of collecting, organizing, and analyzing data is not always a simple, sequential process; sometimes a preliminary analysis of a data set may prompt us to look at the data in another way, or even to go back and collect additional data to test an emerging hypothesis. For example, students could survey their classmates on how they are transported to school (such as by car, by bus, by foot, or another way), and then display the data in a circle graph.

After analyzing the data in this graph, students might look at the data in a different way. Students might be interested in finding out more about people who are transported to school by car. Why do they ride in a car to school? Are they on a bus route? Do they carpool with other students? Are they close enough to school to walk, but choose to ride? Is the neighborhood between home and school too dangerous to walk through? Do the people who walk sometimes ride in a car, also? They might discover that most students in the "other" category ride their bikes to school, and decide to create an additional category.

In all grades, students look at graphical displays and describe them by identifying aspects such as the greatest value, the least value, and the relationship of one data point to another. Students in the intermediate grades learn how to summarize or characterize a data set in greater depth by determining the range and two measures of center, the mode and median. Students in the upper grades learn to find the third measure of center, the mean, and also to determine quartiles, identify outliers, and, for scatterplots, calculate a line or curve of best fit and describe any resulting correlation. High-school students should be able to design their own investigations that include effective sampling, representative data, and an unbiased interpretation of the results.

At every grade level, you should encourage students to think about the meaning of the data they have collected and displayed. The crucial question is "Why?"

## Why Is It Important?

The ability to make inferences and predictions based on data is a critical skill students need to develop.

In studying data and statistics, students can also learn that solutions to some problems depend on assumptions and have some degree of uncertainty. The kind of reasoning used in probability and statistics is not always intuitive, and so students will not necessarily develop it if it is not included in the curriculum. (NCTM, 2000).

Data analysis is crucial to the development of theories and new ideas. By paying close attention to patterns, the stories behind outliers, relationships between and among data sets, and the external factors that may have affected the data, students may come to have a deeper understanding of the crucial distinction between theory and evidence.

## How can you make it happen?

Students in the primary grades view a graphical display as a collection of individual parts and are chiefly interested in the part that represents information about them. As they progress into the intermediate grades, they begin to consider the whole picture represented by the graphical display, and look for ways to describe and summarize it. Students can make observations about the shape (for example, the arc of a bell curve or the clustering of scatter plots) and range of the data, and determine the median and mode.

As students continue into the middle grades, their analysis broadens to include comparing two data sets or two characteristics within one data set. At this stage, it is important to help students develop an understanding of a data set as a subset, or sample, of a larger data set or population. Help them learn to make inferences, extrapolations, and conjectures on the relationship of a particular data set to other sets within that larger population. Data collected from a fourth grade class could be used to make inferences about the entire fourth grade population at a school. Data collected from a local pond could be used to make predictions about all pond habitats in the geographic region.

In a scatterplot, students can begin to analyze the data by asking questions such as:

- Which two values are being compared?
- What is the relationship between the data sets?

- Why do you think there is a relationship?
- Was this a valid sample of the population? Are all second-graders represented?
- Would the data be the same if the sample were first-grade students?
- What can you infer about the data?
- What conclusions can be drawn about the data? Why?

Encourage students to support and justify their answers by referring to the data.

With older students, discuss the margin of error that is possible in surveys, and how sampling can misrepresent a population. For example, surveying a crowd outside a bookstore and asking them about their favorite actors may yield a very different result than surveying a crowd outside an acting school. Students should be able to evaluate the sampling of a population and understand how the samples might be distributed throughout a population to determine if there was any bias in reporting the data.

Your role in helping students' progress through the different skill levels in data analysis is complex. The key is to:

- Provide students with lots of practice looking at graphs and other forms of representation, both those made by the students themselves as well as those from other sources. These should include representations that accurately reflect the data set on which they are based as well as those that do not.
- Teach students the skills they can use to summarize data. You should introduce skills, such as determining the measures of center (mean, median, and mode), identifying range, outliers, quartiles, and so on, as they are appropriate.
- Teach students the correct statistical terms so they have the words necessary to clearly convey their thinking; terms such as axis, range, median, correlation, interval, and so on, should become part of their vocabulary.
- Encourage students to look at different aspects of the data displayed in graphs and charts, not only those aspects that support an author's findings.
- Guide students in evaluating graphs to see if they answer the questions for which they were designed. For example, if students conduct a survey of which soft drink is preferred in the school cafeteria for the purpose of ordering drinks for the vending machines, the data should be displayed in a way that clearly shows the students' preferences. Displaying data in a line graph that shows what times of day students typically buy soft drinks would provide inappropriate data that does not answer the question.
- Have students not only communicate their opinions and arguments, but also back up them up with data.
- Encourage students to think of extensions or follow-up research for their analyses.

When students analyze data, have them ask questions such as:

- Is my data source valid?
- What are the different categories of data?
- What problem am I trying to solve?
- Is there missing information?
- Is there extra, unnecessary data?

- Are there places where the data are concentrated or clumped?
- What trends are visible in the data?
- Are there values for which there are no data?
- Are there data points that have unusual values?
- Are these values consistent with my predictions?
- Do I have enough background information to analyze this data?
- What level of accuracy is required in my analysis?
- What is the range, mean, median, mode, and so on?
- Why might there be an increase/decrease in the data?
- How might this data change if...? (This may be a change in history, market value, audience questioned, etc.)
- What is the most appropriate way to summarize my findings?
- Based on this data, what will happen in the future?
- For what other data sample might I use a similar graph?

## How Can You Stretch Students' Thinking?

Students often have difficulty understanding how the measures of center—the mean, median, and mode—differ from one to another, particularly the mean and median, when describing an average value for a set of data. Have students find all three measures for a set of data, then change one value in the set, and have them determine if and how that change affected each measure. Change a different value, and repeat the process.

Continue, choosing values both near the middle of the set as well as those at either end. Have students try this with different data sets. They should find that, generally, the mean is more sensitive than the median or mode to a single changing value and, as such, is more affected by extreme values in a data set. Help them conclude that the mean is useful to describe an average value for a set of data where there are no outliers; when there are outliers, using the median may be more appropriate; and lastly, if there are many identical values, using the mode may be the better choice.

Show students graphs that aren't labeled and challenge them to describe types of data that would fit the graphs. For example, a line graph that shows a line going up could represent the rising cost of movie tickets; a graph with a line going down could represent the shrinking cost of DVD players; a graph with a line that repeatedly goes up and down could represent the seasonal fluctuations in ice cream sales or bike rentals.

To discuss possible bias in representing data, display a set of data and a graph that misrepresents that data. For example, a bar graph in which the proportionality of the values in the data set are not preserved—a bar is three times the height of another, yet the value it represents is not three times as great as the other value. Have students write about whether they think the graph is accurate or not and why.