

Unit Introduction

Samples and Populations

Data and Statistics

Goals of the Unit

The unit *Samples and Populations* was created to help students:

- Use the process of statistical investigation to explore problems
- Use information from samples to draw conclusions about populations
- Explore the influence of sample size on the variability of the distribution of sample means or medians
- Evaluate sampling plans
- Use probability to select random samples from populations
- Compare sample distributions using measures of center (mean, median), measures of variability (range, minimum and maximum data values, percentiles), and data displays that group data (histograms, box-and-whisker plots)
- Explore relationships between paired values of numerical variables

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 that involves data analysis, such as:

- *What is the population?*
- *What is the sample?*
- *What kinds of comparisons or relationships can I explore using data from the sample?*
- *Can I use my results to make predictions or generalizations about the populations?*

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

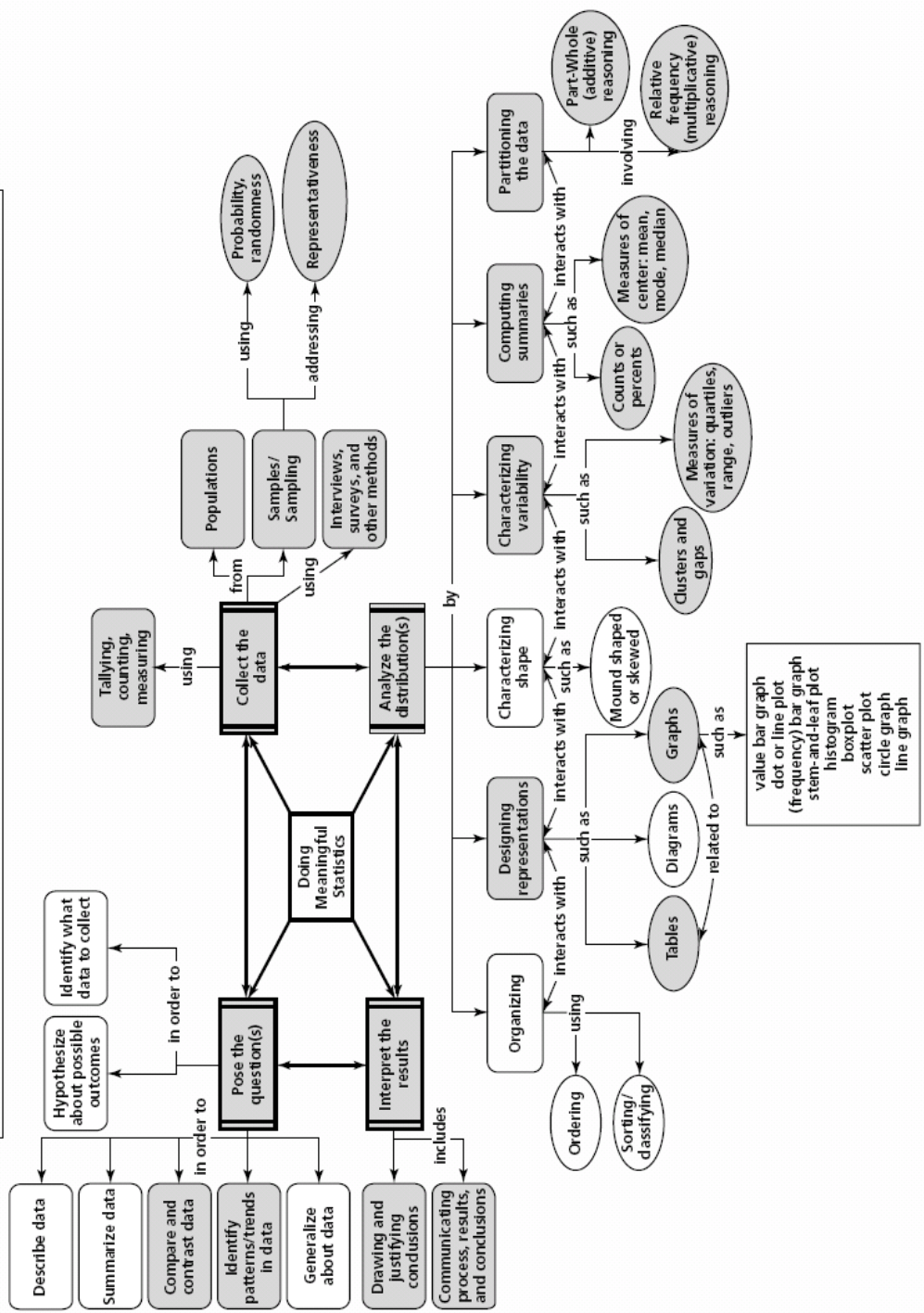
Students apply what they have learned about samples to engaging real-world situations. First, they analyze measurements of Native American arrowheads found at six different archaeological sites. Scientists know the approximate time periods during which four of the sites were settled; the time periods 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.

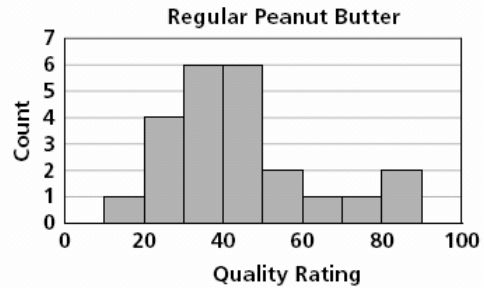


Figure 1

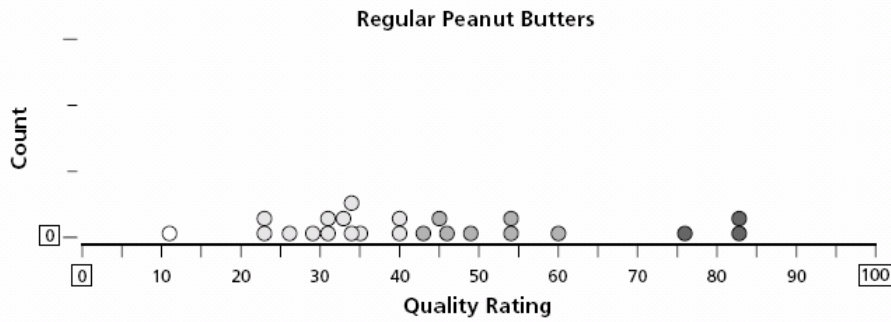
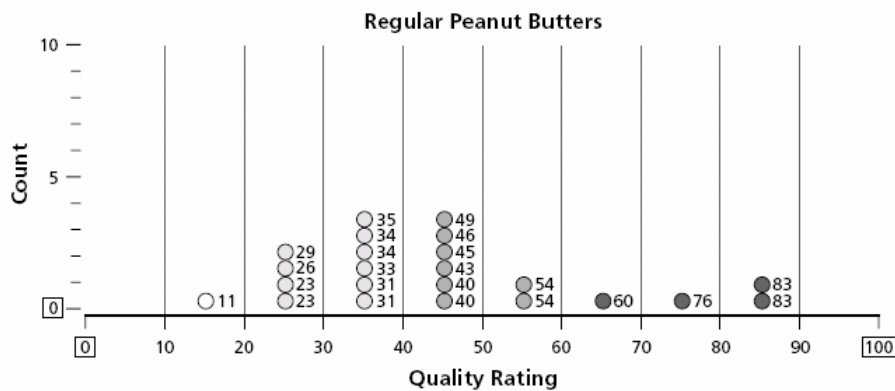
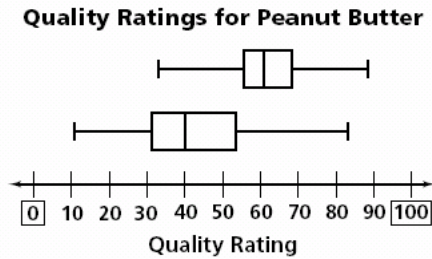


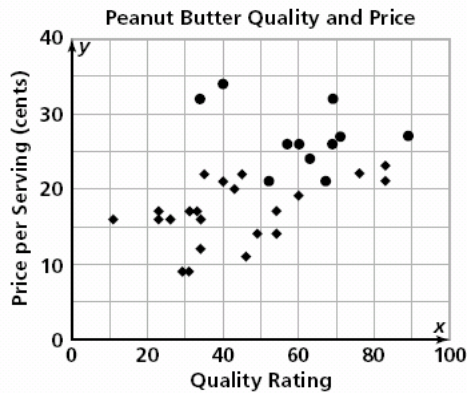
Figure 2



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.



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



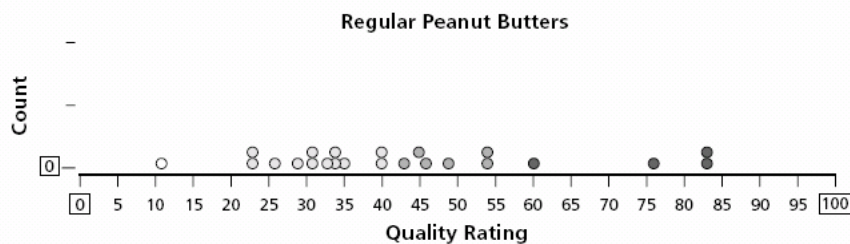
Reading Standard Graphs

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.

Figure 3



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.

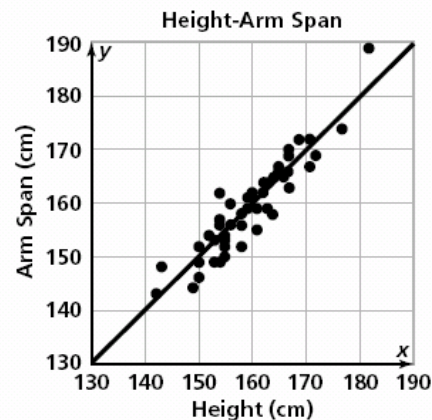
In addition to random sampling, students consider other types of sampling strategies: convenience sampling, voluntary-response sampling, and systematic sampling. It is possible to describe one or more ways in which samples selected using one of these three methods have a greater potential not to be predictive 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.

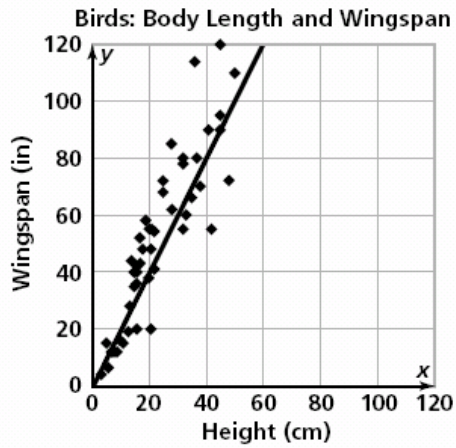
Exploring the concept of covariation or association

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

