

CHAPTER 17

Problem Solving and Data Analysis

The Problem Solving and Data Analysis questions on the SAT Math Test assess your ability to use your understanding of math and your skills to solve problems set in the real world. The questions ask you to create a representation of a problem, consider the units involved, pay attention to the meaning of quantities, know and use different properties of mathematical operations and representations, and apply key principles of statistics and probability. Special focus in this domain is given to mathematical models. Models are representations of real-life contexts. They help us to explain or interpret the behavior of certain components of a system and to predict results that are as yet unobserved or unmeasured. You may be asked to create and use a model and to understand the distinction between the predictions of a model and the data that has been collected.

Some questions involve quantitative reasoning about ratios, rates, and proportional relationships and may require understanding and applying unit rates. Some problems are set in academic and career settings and draw from science, including the social sciences.

Some questions present information about the relationship between two variables in a graph, scatterplot, table, or another form and ask you to analyze and draw conclusions about the given information. These questions assess your understanding of the key properties of, and the differences between, linear, quadratic, and exponential relationships and how these properties apply to the corresponding real-life contexts.

Problem Solving and Data Analysis also includes questions that assess your understanding of essential concepts in statistics and probability. You may be asked to analyze univariate data (data involving one variable) presented in dot plots, histograms, box plots, and frequency tables, or bivariate data (data involving two variables) presented in scatterplots, line graphs, and two-way tables. This includes computing, comparing, and interpreting measures of center, interpreting measures of spread, describing overall patterns, and recognizing the effects of outliers on measures of center and spread. These questions may test your understanding of the conceptual meaning of standard deviation (although you will not be asked to calculate a standard deviation).

Other questions may ask you to estimate the probability of an event, employing different approaches, rules, or probability models. Special attention is given to the notion of conditional probability, which is tested using two-way tables and in other ways.

Some questions will present you with a description of a study and ask you to decide what conclusion is most appropriate based on the design of the study. Some questions ask about using data from a sample to draw conclusions about an entire population. These questions might also assess conceptual understanding of the margin of error (although you won't be asked to calculate a margin of error) when a population mean or proportion is estimated from sample data. Other questions ask about making conclusions about cause-and-effect relationships between two variables.

Problem Solving and Data Analysis questions include both multiple-choice questions and student-produced response questions. The use of a calculator is allowed for all questions in this domain.

Problem Solving and Data Analysis is one of the three SAT Math Test subscores, reported on a scale of 1 to 15.

Let's explore the content and skills assessed by Problem Solving and Data Analysis questions.

Ratio, Proportion, Units, and Percentage

Ratio and proportion is one of the major ideas in mathematics. Introduced well before high school, ratio and proportion is a theme throughout mathematics, in applications, in careers, in college mathematics courses, and beyond.

Example 1

On Thursday, 240 adults and children attended a show. The ratio of adults to children was 5 to 1. How many children attended the show?

- A) 40
- B) 48
- C) 192
- D) 200

Because the ratio of adults to children was 5 to 1, there were 5 adults for every 1 child. Thus, of every 6 people who attended the show, 5 were adults and 1 was a child. In fractions, $\frac{5}{6}$ of the 240 who attended were adults and $\frac{1}{6}$ were children. Therefore, $\frac{1}{6} \times 240 = 40$ children attended the show, which is choice A.

Ratios on the SAT may be expressed in the form 3 to 1, 3:1, $\frac{3}{1}$, or simply 3.

Example 2

On an architect's drawing of the floor plan for a house, 1 inch represents 3 feet. If a room is represented on the floor plan by a rectangle that has sides of lengths 3.5 inches and 5 inches, what is the actual floor area of the room, in square feet?

- A) 17.5
- B) 51.0
- C) 52.5
- D) 157.5

Because 1 inch represents 3 feet, the actual dimensions of the room are $3 \times 3.5 = 10.5$ feet and $3 \times 5 = 15$ feet. Therefore, the floor area of the room is $10.5 \times 15 = 157.5$ square feet, which is choice D.

Another classic example of ratio is the length of a shadow. At a given location and time of day, it might be true that a fence post that has a height of 4 feet casts a shadow that is 6 feet long. This ratio of the length of the shadow to the height of the object, 6 to 4 or 3 to 2, remains the same for any object at the same location and time. This could be considered a unit rate: the ratio of the length of the shadow to the height of the object would be equivalent to $\frac{3}{2}$ to 1 or the unit rate $\frac{3}{2}$ feet change in length of shadow for every 1 foot change in height of the object. So, for example, a tree that is 12 feet tall would cast a shadow that is $\frac{3}{2} \times 12 = 18$ feet long. In this situation, in which one variable quantity is always a fixed constant times another variable quantity, the two quantities are said to be directly proportional.

Variables x and y are said to be directly proportional if $y = kx$, where k is a nonzero constant. The constant k is called the constant of proportionality.

In the preceding example, you would say the length of an object's shadow is directly proportional to the height of the object, with constant of proportionality $\frac{3}{2}$. So if you let L be the length of the shadow and H be the height of the object, then $L = \frac{3}{2}H$.

Notice that both L and H are lengths, so the constant of proportionality, $\frac{L}{H} = \frac{3}{2}$, has no units. In contrast, let's consider

Example 2 again. On the scale drawing, 1 inch represents 3 feet. The length of an actual measurement is directly proportional to its length on the scale drawing. But to find the constant of proportionality, you need to keep track of units: $\frac{3 \text{ feet}}{1 \text{ inch}} = \frac{36 \text{ inches}}{1 \text{ inch}} = 36$. Hence, if S is a length on the scale drawing that corresponds to an actual length of R , then $R = 36S$, where R and S have the same units.

Many of the questions on the SAT Math Test require you to pay attention to units. Some questions in Problem Solving and Data Analysis require you to convert units either between the English system and the metric system or within those systems.

Example 3

Scientists estimate that the Pacific Plate, one of Earth's tectonic plates, has moved about 1,060 kilometers in the past 10.3 million years. About how far, in miles, has the Pacific Plate moved during this same time period? (Use 1 mile = 1.6 kilometers.)

- A) 165
- B) 398
- C) 663
- D) 1,696

Because 1 mile = 1.6 kilometers, the distance is 1,060 kilometers $\times \frac{1 \text{ mile}}{1.6 \text{ kilometers}} = 662.5$ miles, which is about 663 miles. Therefore, the correct answer is choice C.

Questions may require you to move between unit rates and total amounts.

Example 4

County Y consists of two districts. One district has an area of 30 square miles and a population density of 370 people per square mile, and the other district has an area of 50 square miles and a population density of 290 people per square mile. What is the population density, in people per square mile, for all of County Y?

(Note that this example is a student-produced response question and has no choices. On the SAT, you will grid your answer in the spaces provided on the answer sheet.)

The first district has an area of 30 square miles and a population density of 370 people per square mile, so its total population is $30 \text{ square miles} \times \frac{370 \text{ people}}{\text{square mile}} = 11,100$ people. The other district has an area of 50 square miles and a population density of 290 people per square mile, so its total population is $50 \text{ square miles} \times \frac{290 \text{ people}}{\text{square mile}} = 14,500$ people. Thus, County Y has total population $11,100 + 14,500 = 25,600$ people and total area $30 + 50 = 80$ square miles. Therefore, the population density of County Y is $\frac{25,600}{80} = 320$ people per square mile.

Problem Solving and Data Analysis also includes questions involving percentages, which are a type of proportion. These questions may involve the concepts of percentage increase and percentage decrease.

Example 5

A furniture store buys its furniture from a wholesaler. For a particular style of table, the store usually sells a table for 75% more than the cost of the table from the wholesaler. During a sale, the store sells the table for 15% more than the cost from the wholesaler. If the sale price of the table is \$299, what is the usual price for the table?

- A) \$359
- B) \$455
- C) \$479
- D) \$524

The sale price of the table was \$299. This is equal to the cost from the wholesaler plus 15%. Thus, $\$299 = 1.15(\text{cost from the wholesaler})$, and the cost from the wholesaler is $\frac{\$299}{1.15} = \260 . The usual price is the cost from the wholesaler, \$260, plus 75%. Therefore, the usual price the store charges for the table is $1.75 \times \$260 = \455 , which is choice B.

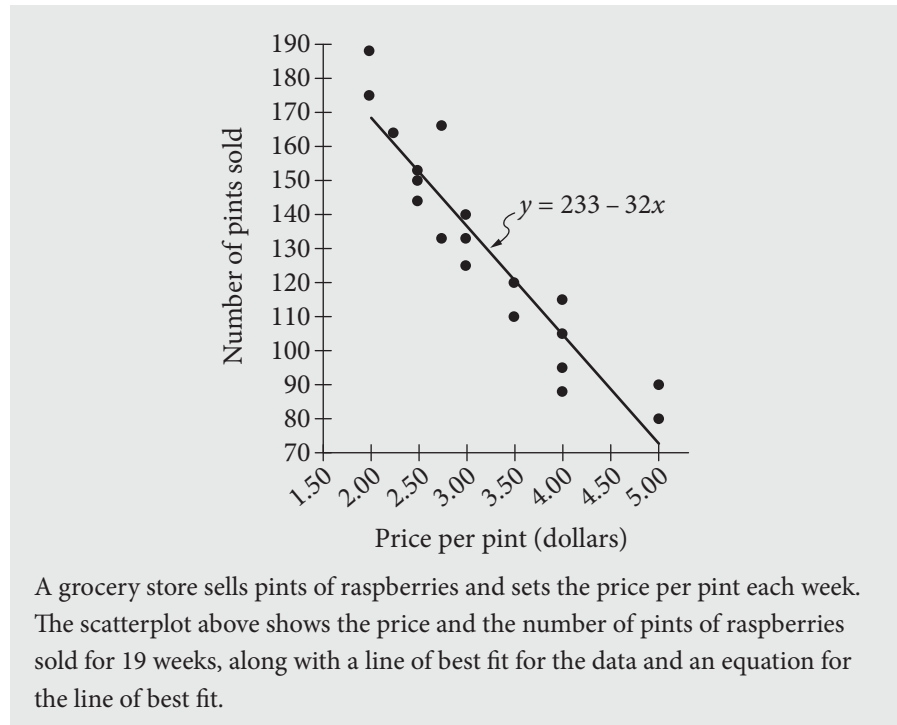
Interpreting Relationships Presented in Scatterplots, Graphs, Tables, and Equations

The behavior of a variable and the relationship between two variables in a real-world context may be explored by considering data presented in tables and graphs.

The relationship between two quantitative variables may be modeled by a function or an equation. The model may allow very accurate predictions, as for example models used in physical sciences, or may only describe a general trend, with considerable variability between the actual and predicted values, as for example models used in behavioral and social sciences.

Questions on the SAT Math Test assess your ability to understand and analyze relationships between two variables, the properties of the functions used to model these relationships, and the conditions under which a model is considered to be an appropriate representation of the data. Problem Solving and Data Analysis questions focus on linear, quadratic, and exponential relationships.

Example 6



There are several different questions that could be asked about this context.

A. According to the line of best fit, how many pints of raspberries would the grocery store be predicted to sell in a week when the price of raspberries is \$4.50 per pint?

Because the line of best fit has equation $y = 233 - 32x$, where x is the price, in dollars, for a pint of raspberries and y is the predicted number of pints of raspberries sold, the number of pints the store would be predicted to sell in a week where the price of raspberries is \$4.50 per pint is $233 - 32(4.50) = 89$ pints.

B. For how many of the 19 weeks shown was the number of pints of raspberries sold greater than the number predicted by the line of best fit?

For a given week, the number of pints of raspberries sold is greater than the number predicted by the line of best fit if and only if the point representing that week lies above the line of best fit. For example, at the price of \$5 per pint, the number sold in two different weeks was approximately 80 and 90, which is more than the 73 predicted by the line of best fit. Of the 19 points, 8 lie above the line of best fit, so there were 8 weeks in which the number of pints sold was greater than what was predicted by the line of best fit.

C. What is the best interpretation of the slope of the line of best fit in this context?

On the SAT, this question would be followed by multiple-choice answer options. The slope of the line of best fit is -32 . This means that the correct answer would state that for each dollar that the price of a pint of raspberries increases, the store is predicted to sell 32 fewer pints of raspberries.

D. What is the best interpretation of the y -intercept of the line of best fit in this context?

On the SAT, this question would be followed by multiple-choice answer options.

In this context, the y -intercept does not represent a likely scenario, so it cannot be accurately interpreted in this context. According to the model, the y -intercept means that if the store sold raspberries for \$0 per pint—that is, if the store gave raspberries away—233 people would be expected to accept the free raspberries. However, it is not realistic that the store would give away raspberries, and if they did, it is likely that far more people would accept the free raspberries. Also notice that in this case, the left-most line on the graph is not the y -axis. The lower-left corner shows the x - and y -coordinates of $(1.5, 70)$, not $(0, 0)$.

The fact that the y -intercept indicates that 233 people would accept free raspberries is one limitation of the model. Another limitation is that for a price of \$7.50 per pint or above, the model predicts that a negative number of people would buy raspberries, which is impossible. In general, you should be cautious about applying a model for values outside of the given data. In this example, you should only be confident in the prediction of sales for prices between \$2 and \$5.

Giving a line of best fit, as in this example, assumes that the relationship between the variables is best modeled by a linear function, but that is not always true. On the SAT, you may see data that are best modeled by a linear, quadratic, or exponential model.

Example 7

Time (hours)	Number of bacteria
0	1,000
1	4,000
2	16,000
3	64,000

The table above gives the initial number (at time $t = 0$) of bacteria placed in a growth medium and the number of bacteria in the growth medium each hour for 3 hours. Which of the following functions best models the number of bacteria, $N(t)$, after t hours?

- A) $N(t) = 4,000t$
- B) $N(t) = 1,000 + 3,000t$
- C) $N(t) = 1,000(4^{-t})$
- D) $N(t) = 1,000(4^t)$

The given choices are linear and exponential models. If a quantity is increasing linearly with time, then the *difference* in the quantity between successive time periods is constant. If a quantity is increasing exponentially with time, then the *ratio* in the quantity between successive time periods is constant. According to the table, after each hour, the number of bacteria in the culture is 4 times as great as it was the preceding hour: $\frac{4,000}{1,000} = \frac{16,000}{4,000} = \frac{64,000}{16,000} = 4$. That is, for each increase of 1 in t , the value of $N(t)$ is multiplied by 4. At $t = 0$, which corresponds to the time when the culture was placed in the medium, there were 1,000 bacteria. This is modeled by the exponential function $N(t) = 1,000(4^t)$, which has the value 1,000 at $t = 0$ and increases by a factor of 4 for each increase of 1 in the value of t . Choice D is the correct answer.

The SAT Math Test may include questions that require you to know the difference between linear and exponential growth.

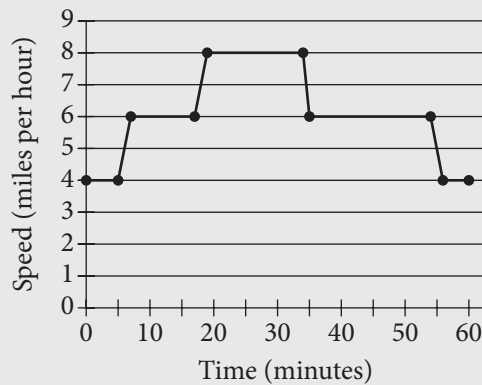
Example 8

Every month Jamal adds two new books to his library. Which of the following types of functions best models the number of books in Jamal's library as a function of time?

- A) Increasing linear
- B) Decreasing linear
- C) Increasing exponential
- D) Decreasing exponential

Over equal intervals, linear functions increase or decrease by a constant amount, while exponential functions increase or decrease by a constant factor. Since the number of books is increasing by a constant amount (2 books) over equal intervals (each month), the function is linear. Also, since the number of books is increasing as time increases, the function is increasing, and therefore choice A is correct.

Example 9



One evening, Maria walks, jogs, and runs for a total of 60 minutes. The graph above shows Maria's speed during the 60 minutes. Which segment of the graph represents the times when Maria's speed is the greatest?

- A) The segment from (17, 6) to (19, 8)
- B) The segment from (19, 8) to (34, 8)
- C) The segment from (34, 8) to (35, 6)
- D) The segment from (35, 6) to (54, 6)

The correct answer is choice B. Because the vertical coordinate represents Maria's speed, the part of the graph with the greatest vertical coordinate represents the times when Maria's speed is the greatest. This is the highest part of the graph, the segment from (19, 8) to (34, 8), when Maria runs at 8 miles per hour (mph). Choice A represents the time during which Maria's speed is increasing from 6 to 8 mph; choice C represents the time during which Maria's speed is decreasing from 8 to 6 mph; and choice D represents the longest period of Maria moving at the same speed (6 mph), not the times when Maria's speed is the greatest.

More Data and Statistics

Some questions on the SAT Math Test will assess your ability to understand and analyze data presented in a table, bar graph, histogram, dot plot, box plot, line graph, or other display.

Example 10

A store is deciding whether to install a new security system to prevent shoplifting. Based on store records, the security manager of the store estimates that 10,000 customers enter the store each week, 24 of whom will attempt to shoplift. Based on data provided from other users of the security system, the manager estimates the results of the new security system in detecting shoplifters would be as shown in the table below.

	Alarm sounds	Alarm does not sound	Total
Customer attempts to shoplift	21	3	24
Customer does not attempt to shoplift	35	9,941	9,976
Total	56	9,944	10,000

According to the manager's estimates, if the alarm sounds for a customer, what is the probability that the customer did *not* attempt to shoplift?

- A) 0.0003
- B) 0.0035
- C) 0.0056
- D) 0.625

According to the manager's estimates, the alarm will sound for 56 customers. Of these 56 customers, 35 did *not* attempt to shoplift. Therefore, if the alarm sounds, the probability that the customer did *not*

attempt to shoplift is $\frac{35}{56} = \frac{5}{8} = 0.625$. The correct answer is choice D.

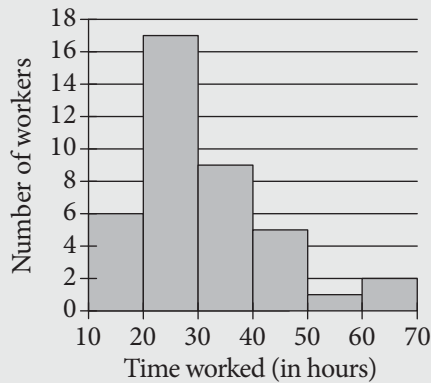
Example 10 is an example of a conditional probability, the probability of an event given that another event is known to have occurred. The question asks for the probability that a customer did not attempt to shoplift given that the alarm sounded.

You may be asked to answer questions that involve a measure of center for a data set: the mean or the median. A question may ask you to draw conclusions about one or more of these measures of center even if the exact values cannot be calculated. To recall briefly:

The mean of a set of numerical values is the sum of all the values divided by the number of values in the set.

The median of a set of numerical values is the middle value when the values are listed in increasing (or decreasing) order. If the set has an even number of values, then the median is the average of the two middle values.

Example 11



The histogram shown summarizes the distribution of time worked last week, in hours, by the 40 employees of a landscaping company. In the histogram, the first bar represents all workers who worked at least 10 hours but less than 20 hours; the second represents all workers who worked at least 20 hours but less than 30 hours; and so on. Which of the following could be the median and mean amount of time worked, in hours, for the 40 employees?

- A) Median = 22, Mean = 23
- B) Median = 24, Mean = 22
- C) Median = 26, Mean = 32
- D) Median = 32, Mean = 30

(Note: On the SAT, all histograms have the same type of boundary condition. That is, the values represented by a bar include the left endpoint but do not include the right endpoint.)

If the number of hours the 40 employees worked is listed in increasing order, the median will be the average of the 20th and the 21st numbers on the list. The first 6 numbers on the list will be workers represented by the first bar; hence, each of the first 6 numbers will be at least 10 but less than 20. The next 17 numbers, that is, the 7th through the 23rd numbers on the list, will be workers represented by the second bar; hence, each of the next 17 numbers will be at least 20 but less than 30. Thus, the 20th and the 21st numbers on the list will be at least 20 but less than 30. Therefore, any of the median values in choices A, B, or C are possible, but the median value in choice D is not.

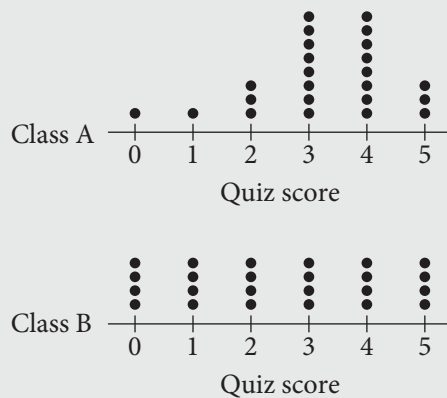
Now let's find the possible values of the mean. Each of the 6 employees represented by the first bar worked at least 10 hours but less than 20 hours. Thus, the total number of hours worked by these 6 employees is at least 60. Similarly, the total number of hours worked by the 17 employees represented by the second bar is at least 340; the total number of hours worked by the 9 employees represented by the third bar is at least 270; the total number of hours worked by the 5 employees represented by the fourth bar is at least 200; the total number of hours worked by the 1 employee represented by the fifth bar is at least 50; and the total number of hours worked by the 2 employees represented by the sixth bar is at least 120. Adding all these hours shows that the total number of hours worked by all 40 employees is at least $60 + 340 + 270 + 200 + 50 + 120 = 1,040$. Therefore, the mean number of hours worked by all 40 employees is at least $\frac{1,040}{40} = 26$. Therefore, only the values of the mean given in choices C and D are possible. Because only choice C has possible values for both the median and the mean, it is the correct answer.

A data set may have a few values that are much larger or smaller than the rest of the values in the set. These values are called *outliers*. An outlier may represent an important piece of data. For example, if a data set consists of rates of a certain illness in various cities, a data point with a very high value could indicate a serious health issue to be investigated.

In general, outliers affect the mean more than the median. Therefore, outliers that are larger than the rest of the points in the data set tend to make the mean greater than the median, and outliers that are smaller than the rest of the points in the data set tend to make the mean less than the median. In Example 11, the mean was larger than the median due to the unusually large amount of time worked by a few employees.

The mean and the median are different ways to describe the center of a data set. Another key characteristic of a data set is the amount of variability, or spread, in the data. One measure of spread is the *range*, which is equal to the maximum value minus the minimum value. Another measure of spread is the *standard deviation*, which is a measure of how far away the points in the data set are from the mean value. On the SAT Math Test, you will *not* be asked to compute the standard deviation of a data set, but you do need to understand that a larger standard deviation corresponds to a data set whose values are more spread out from the mean value.

Example 12



The dot plots show the distributions of scores on a current events quiz for two classes of 24 students each. Which of the following statements about the standard deviations of the two distributions is true?

- A) The standard deviation of quiz scores in Class A is less than that of quiz scores in Class B.
- B) The standard deviation of quiz scores in Class A is greater than that of quiz scores in Class B.
- C) The standard deviation of quiz scores in Class A is equal to that of quiz scores in Class B.
- D) There is not enough information to compare the standard deviations.

In Class A, the mean score is between 3 and 4. The large majority of scores are 3 and 4, with only a few scores of 0, 1, 2, and 5. In Class B, the mean score is 2.5, and scores are evenly distributed across all possible scores, with many scores not close to the mean score. Because the scores in Class A are more closely clustered around the mean, the standard deviation of the scores in Class A is smaller. The correct answer is choice A.

A *population parameter* is a numerical value that describes a characteristic of a population. For example, the percentage of registered voters who would vote for a certain candidate is a parameter describing the population of registered voters in an election. In another example, the average income of a household in a city is a parameter describing the population of households in that city. We often do not know the value of the population parameter; thus, an essential purpose of statistics is to estimate a population parameter based on a sample from the population. A common example is election polling, where researchers will interview a random sample of registered voters to estimate the proportion of all registered voters who plan to vote for a certain candidate. The precision of the estimate depends on the variability of the sample data and the sample size. For instance, if household incomes in a city vary widely or the sample is small, the estimate that comes from a sample may differ considerably from the actual value for the population parameter.

For example, a researcher wants to estimate the mean number of hours each week that the 1,200 students at a high school spend on the internet. Interviewing all 1,200 students would be time consuming, and it would be more efficient to survey a random sample of the students. Suppose the researcher has time to interview 80 students. Which 80 students? In order to have a sample that is representative of the population, students who will participate in the study should be selected *at random*. That is, each student must have the same chance to be selected. Random selection is essential in protecting against bias and increases the reliability of any estimates calculated. The researcher can select students at random in several different ways; for instance, write each student's name on a slip of paper, put all the slips in a bowl, mix up the slips, and then draw 80 names from the bowl. In practice, a computer is often used to select participants at random.

If you do not select a random sample, the sampling method used may introduce bias. For example, if you found 80 students from those attending a game of the school's football team, those people would be more likely to be interested in sports, and in turn, an interest in sports might be related to the average amount of time the students spend on the internet. The result would be that the average time those 80 students spend on the internet might not be an accurate estimate of the average amount of time *all* students at the school spend on the internet.

Suppose you select 80 students at random from the 1,200 students at the high school. You ask them how much time they spend on the internet each week, and you calculate that the mean time is 14 hours. You also find that 6 of the 80 students spend less than 2 hours each week on the internet. Based on these results, what conclusions should be made about the entire population of 1,200 students?

Because the sample was selected at random, the mean of 14 hours is a plausible estimate for the mean time spent on the internet for all 1,200 students. Also, we can use the sample data to estimate how many students spend less than 2 hours on the internet each week. In the sample, the percentage is $\frac{6}{80}$, or 7.5%. Applying this percentage to the entire population of 1,200 students, the best estimate is that 90 students at the school spend less than 2 hours per week on the internet.

However, the estimates of the population parameters need to be interpreted carefully. An essential part of statistics is accounting for the variability of the estimate. The estimates above are reasonable, but they are unlikely to be exactly correct. Statistical analysis can also describe how far from the estimates the actual values are expected to be, at most. To describe the precision of an estimate, statisticians use *margins of error*. On the SAT, you will not be expected to compute a margin of error, but you should understand how sample size affects the margin of error and how to interpret a given margin of error in the context.

If the example above were an SAT question, you might be given survey results indicating that, for a random sample of 80 students, the estimated mean was 14 hours with an associated margin of error of 1.2 hours. An appropriate interpretation of these results is that it is plausible that the mean number of hours for all 1,200 students in the population is greater than 12.8 hours but less than 15.2 hours.

There are some key points to note.

1. The value of the margin of error is affected by two factors: the variability in the data and the sample size. The larger the standard deviation, the larger the margin of error; the smaller the standard deviation, the smaller the margin of error. Furthermore, increasing the size of the random sample provides more information and typically reduces the margin of error.
2. The margin of error applies to the estimated value of the population parameter only; it does not inform the estimated value for an individual. In the example, plausible values for the population mean are in the interval from 12.8 hours to 15.2 hours. The time, in hours, that an individual spends on the internet may or may not fall in this interval.

Example 13

A quality control researcher at an electronics company is testing the life of the company's batteries in a certain camera. The researcher selects 100 batteries at random from the daily output of the batteries and finds that the life of the batteries has a mean of 342 pictures with an associated margin of error of 18 pictures. Which of the following is the most appropriate conclusion based on these data?

- A) All the batteries produced by the company that day have a life between 324 and 360 pictures.
- B) All the batteries ever produced by the company have a life between 324 and 360 pictures.
- C) It is plausible that the mean life of batteries produced by the company that day is between 324 and 360 pictures.
- D) It is plausible that the mean life of all the batteries ever produced by the company is between 324 and 360 pictures.

The correct answer is choice C. Choices A and B are incorrect because the margin of error gives information about the mean life of all batteries produced by the company that day, not about the life of any individual battery. Choice D is incorrect because the sample of batteries was taken from the population of all the batteries produced by the company on that day. The population of all batteries the company ever produced may have a different mean life because of changes in the formulation of the batteries, wear on machinery, improvements in production processes, and many other factors.

The statistics examples discussed so far are largely based on investigations intended to estimate some characteristic of a group: the mean amount of time students spend on the internet, the mean life of a battery, and the percentage of registered voters who plan to vote for a candidate. Another primary focus of statistics is to investigate relationships between variables and to draw conclusions about cause and effect. For example, does a new type of physical therapy help people recover from knee surgery faster? For such a study, some people who have had knee surgery will be randomly assigned to the new therapy or to the usual therapy. The medical results of these patients can be compared. The key questions from a statistical viewpoint are

- Is it appropriate to generalize from the sample of patients in the study to the entire population of people who are recovering from knee surgery?
- Is it appropriate to conclude that the new therapy *caused* any difference in the results for the two groups of patients?

The answers depend on the use of random selection and random assignment.

- If the subjects in the sample of a study were selected at random from the entire population in question, the results can be generalized to the entire population because random sampling ensures that each individual has the same chance to be selected for the sample.
- If the subjects in the sample were randomly assigned to treatments, it may be appropriate to make conclusions about cause and effect because the treatment groups will be roughly equivalent at the beginning of the experiment other than the treatment they receive.

This can be summarized in the following table.

	Subjects Selected at Random	Subjects Not Selected at Random
Subjects randomly assigned to treatments	<ul style="list-style-type: none">• Results can be generalized to the entire population.• Conclusions about cause and effect can appropriately be drawn.	<ul style="list-style-type: none">• Results <i>cannot</i> be generalized to the entire population.• Conclusions about cause and effect can appropriately be drawn.
Subjects not randomly assigned to treatments	<ul style="list-style-type: none">• Results can be generalized to the entire population.• Conclusions about cause and effect <i>should not</i> be drawn.	<ul style="list-style-type: none">• Results <i>cannot</i> be generalized to the entire population.• Conclusions about cause and effect <i>should not</i> be drawn.

The previous example discussed treatments in a medical experiment. The word *treatment* refers to any factor that is deliberately varied in an experiment.

Example 14

A community center offers a Spanish course. This year, all students in the course were offered additional audio lessons they could take at home. The students who took these additional audio lessons did better in the course than students who didn't take the additional audio lessons. Based on these results, which of the following is the most appropriate conclusion?

- A) Taking additional audio lessons will cause an improvement for any student who takes any foreign language course.
- B) Taking additional audio lessons will cause an improvement for any student who takes a Spanish course.
- C) Taking additional audio lessons was the cause of the improvement for the students at the community center who took the Spanish course.
- D) No conclusion about cause and effect can be made regarding students at the community center who took the additional audio lessons at home and their performance in the Spanish course.

The correct answer is choice D. The better results of these students may have been a result of being more motivated, as shown in their willingness to do extra work, and not the additional audio lessons. Choice A is incorrect because no conclusion about cause and effect is possible without random assignment to treatments and because the sample was only students taking a Spanish course, so no conclusion can be appropriately made about students taking all foreign language courses. Choice B is incorrect because no conclusion about cause and effect is possible without random assignment to treatments and because the students taking a Spanish course at the community center is not a random sample of all students who take a Spanish course. Choice C is incorrect because the students taking the Spanish course at the community center were not randomly assigned to use the additional audio lessons or to not use the additional audio lessons.

THIS PAGE IS INTENTIONALLY LEFT BLANK.