Chapter 20

Problem Solving and Data Analysis

The Problem Solving and Data Analysis section of the SAT Math Test assesses your ability to use your math understanding and skills to solve problems set in the real world. Problem Solving and Data Analysis questions test your ability 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. Special focus in this domain will be given to mathematical models. You may be asked to create and use a model and to understand the distinction between the model predictions and data collected. Models are a representation of real life. They help us to explain or interpret the behavior of certain components of a system and to predict future results that are as yet unobserved or unmeasured.

The questions involve quantitative reasoning about ratios, rates, and proportional relationships and may require understanding and applying unit rates. Many of the 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. The 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. An important example is understanding the difference between simple interest and compound interest.

Problem Solving and Data Analysis also includes questions that assess your understanding of essential concepts in statistics. You may be asked to analyze univariate data presented in bar graphs, histograms, line graphs, and box-and-whisker plots, or bivariate data presented in scatterplots and two-way tables. This includes computing and interpreting measures of center, interpreting measures of spread, describing overall patterns, and recognizing

the effects of outliers on measures of center. 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 a simple or compound 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 or other contexts.

Some questions require the ability to draw conclusions about an entire population from a random sample of that population and how variability affects those conclusions. The questions may test your understanding of randomization-based inference and the conceptual meaning of the margin of error (although you will not be asked to calculate a margin of error) when the mean or the proportion of a population is estimated using sample data. You may be presented with a description of a study and asked to explain what types of conclusions can be drawn with regard to relationships between variables involved and to what population can the study findings be appropriately generalized.

REMEMBER

Problem Solving and Data Analysis comprise 17 of the 58 questions (29%) on the Math Test.

The questions in Problem Solving and Data Analysis 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. 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.

PRACTICE AT
khanacademy.org/sat

A ratio represents the proportional relationship between quantities, not the actual quantities themselves. Fractions are an especially effective way to represent and work with ratios.

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 is 4 feet high casts a shadow that is 6 feet long. This ratio of the height of the object to the length of the shadow, 4 to 6 or $\frac{2}{3}$, remains the same for any object at the same location and time. So, for example, a person who is 6 feet tall would cast a shadow that is $\frac{3}{2} \times 6 = 9$ 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 proportion, $\frac{L}{H} = \frac{3}{2}$, 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 A, then $A = 36S$.

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. What was the average speed of the Pacific Plate during that time period, in centimeters per year?

A) 1.03
B) 10.3
C) 103
D) 1,030

PRACTICE AT
khanacademy.org/sat

Pay close attention to units, and convert units if required by the question. Writing out the unit conversion as a series of multiplication steps, as seen here, will help ensure accuracy. Intermediate units should cancel (as do the kilometers and meters in Example 3), leaving you with the desired unit (centimeters per year).

Since 1 kilometer = 1,000 meters and 1 meter = 100 centimeters, you get $\frac{1{,}060 \text{ kilometers}}{10{,}300{,}000 \text{ years}} \times \frac{1{,}000 \text{ meters}}{1 \text{ kilometer}} \times \frac{100 \text{ centimeters}}{1 \text{ meter}} = 10.3$ centimeters per year.

Therefore, the correct answer is choice B.

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?

REMEMBER

13 of the 58 questions on the Math Test, or 22%, are student-produced response questions in which you will grid your answers in the spaces provided on the answer sheet.

(Note that this example has no choices. It is a student-produced response question. On an SAT, you would 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 $30 \text{ square miles} \times 370 \frac{\text{people}}{\text{square mile}} = 11{,}100$ 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 $50 \text{ square miles} \times 290 \frac{\text{people}}{\text{square mile}} = 14{,}500$ 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.

PRACTICE AT
khanacademy.org/sat

Percent is a type of proportion that means "per 100." 20%, for instance, means 20 out of (or per) 100. Percent increase or decrease is calculated by finding the difference between two quantities, then dividing the difference by the original quantity and multiplying by 100.

EXAMPLE 5

A furniture store buys its furniture from a wholesaler. For a particular table, the store usually charges its cost from the wholesaler plus 75%. During a sale, the store charged the wholesale cost plus 15%. If the sale price of the table was $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(wholesale cost), and the cost from the wholesaler is $\frac{\$299}{1.15}$ = $260. Therefore, the usual price the store charges for the table is 1.75 × $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.

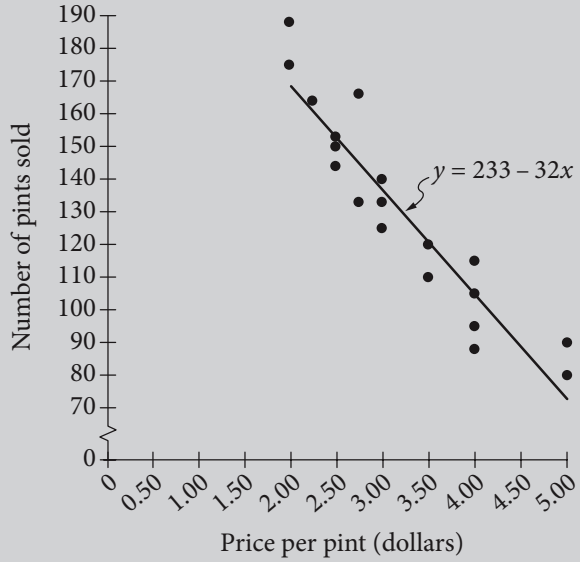
The relationship between two variables may be modeled by a function or equation. The function or equation may be found by examining ordered pairs of data values and by analyzing how the variables are related to one another in the real world. The model may allow very accurate predictions, as for example models used in physical sciences, or may only describe a 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 the 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 good, acceptable, or inappropriate. The questions in Problem Solving and Data Analysis focus on linear, quadratic, and exponential relationships.

REMEMBER

The ability to interpret and synthesize data from charts, graphs, and tables is a widely applicable skill in college and in many careers and thus is tested on the SAT Math Test.

EXAMPLE 6



A grocery store sells pints of raspberries and sets the price per pint each week. The scatterplot above shows the price and the number of pints of raspberries sold for 19 weeks, along with the line of best fit and the equation for the line of best fit.

PRACTICE AT
khanacademy.org/sat

A line of best fit is a straight line that best represents the data on a scatterplot. It is written in $y = mx + b$ form.

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 expect 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 number of pints of raspberries sold, the number of pints the store would be expected 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 amount predicted by the line of best fit?

For a given week, the number of pints of raspberries sold is greater than the amount predicted by the line of best fit if and only if the point representing that week lies above 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 meaning of the slope of the line of best fit?

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 expects to sell 32 fewer pints of raspberries.

D. What is the best interpretation of the meaning of the y-intercept of the line of best fit?

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 terms of 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 — 173 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.

The fact that the y-intercept indicates that 173 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.

(**Note:** Questions interpreting the slope and intercepts of a line of best fit, such as in **C** and **D**, may be classified as part of the Heart of Algebra section and contribute to the Heart of Algebra subscore.)

EXAMPLE 7

Time (hours)	Number of bacteria
0	1×10^3
1	4×10^3
2	1.6×10^4
3	6.4×10^4

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 over 3 hours. Which of the following functions 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) $N(t) = 1{,}000(4^{t})$

PRACTICE AT
khanacademy.org/sat

To determine if a model is linear or exponential, examine the change in the quantity between successive time periods. If the difference in quantity is constant, the model is linear. If the ratio in the quantity is constant (for instance, 4 times greater than the preceding time period), then the model is exponential.

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: $\frac{4 \times 10^3}{1 \times 10^3} = \frac{1.6 \times 10^4}{4 \times 10^3} = \frac{6.4 \times 10^4}{1.6 \times 10^4} = 4$. 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 103 bacteria. This is modeled by the exponential function $N(t) = 1{,}000(4^t)$, which has 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 have questions on simple and compound interest, which are important examples of linear and exponential growth, respectively.

EXAMPLE 8

A bank has opened a new branch and, as part of a promotion, the bank branch is offering $1,000 certificates of deposit at simple interest of 4% per year. The bank is selling certificates with terms of 1, 2, 3, or 4 years. Which of the following functions gives the total amount, A, in dollars, a customer will receive when a certificate with a term of k years is finally paid?

A) $A = 1{,}000(1.04k)$
B) $A = 1{,}000(1 + 0.04k)$
C) $A = 1{,}000(1.04)^k$
D) $A = 1{,}000(1 + 0.04^k)$

For 4% simple interest, 4% of the original deposit is added to the original deposit for each year the deposit was held. That is, if the certificate has a term of k years, $4k$% is added to the original deposit to get the final amount. Because $4k$% is $0.04k$, the final amount paid to the customer is $A = 1{,}000 + 1{,}000(0.04k) = 1{,}000(1 + 0.04k)$. Choice B is the correct answer.

The general formula for simple interest is $A = P(1 + rt)$, where P is the original deposit, called the principal; r is the annual interest rate expressed as a decimal; and t is the length the deposit is held. In Example 8, $P = \$1{,}000$, $r = 0.04$, and $t = k$ years; so A, in dollars, is given by $A = 1{,}000[1 + (0.04)k]$.

In contrast, compound interest is an example of exponential growth.

EXAMPLE 9

A bank has opened a new branch and, as part of a promotion, the bank branch is offering $1,000 certificates of deposit at an interest rate of 4% per year, compounded semiannually. The bank is selling certificates with terms of 1, 2, 3, or 4 years. Which of the following functions gives the total amount, A, in dollars, a customer will receive when a certificate with a term of k years is finally paid?

A) $A = 1{,}000(1 + 0.04k)$
B) $A = 1{,}000(1 + 0.08k)$
C) $A = 1{,}000(1.04)^k$
D) $A = 1{,}000(1.02)^{2k}$

The interest is compounded semiannually, that is, twice a year. At the end of the first half year, 2% of the original deposit is added to the value of the certificate (4% annual interest multiplied by the time period, which is $\frac{1}{2}$ year, gives 2% interest). When the interest is added, the value, in dollars, of the certificate is now $1{,}000 + 1{,}000(0.02) = 1{,}000(1.02)$. Since the interest is reinvested (compounded), the new principal at the beginning of the second half year is $1{,}000(1.02)$. At the end of the second half year, 2% of $1{,}000(1.02)$ is added to the value of the certificate; the value, in dollars, of the certificate is now $1{,}000(1.02) + 1{,}000(1.02)(0.02)$, which is equal to $1{,}000(1.02)(1.02) = 1{,}000(1.02)^2$. In general, after n compounding periods, the amount, A, in dollars, is $A = 1{,}000(1.02)^n$.

When the certificate is paid after k years, the value of the certificate will have been multiplied by the factor (1.02) a total of $2k$ times. Therefore, the total amount, A, in dollars, a customer will receive when a certificate with a term of k years is finally paid is $A = 1{,}000(1.02^{2k})$. Choice D is the correct answer.

The general formula for compound interest is $A = P\left(1 + \frac{r}{n}\right)^{nt}$, where P is the principal, r is the annual interest rate expressed as a decimal, t is the number of years the deposit is held, and n is the number of times the interest is compounded per year. In Example 9, $P = \$1{,}000$, $r = 0.04$, $t = k$, and $n = 2$; so A, in dollars, is given by $A = 1{,}000\left(1 + \frac{0.04}{2}\right)^{2k} = 1{,}000(1.02)^{2k}$.

Note: Although the stated interest rate is 4% per year in Example 9, the value of the account increases by more than 4% in a year, namely 4.04% per year. (You may have seen banks offer an account in this way, for example, 5.00% annual interest rate, 5.13% effective annual yield.) If you take calculus, you will often see a situation in which a stated rate of change differs from the change over an interval. But on the SAT, other than compound interest,

PRACTICE AT
khanacademy.org/sat

Know the formulas for simple and compound interest.

Simple interest: $A = P(1 + rt)$

Compound interest: $A = P(1 + r/n)^{nt}$

A is the total amount, P is the principal, r is the interest rate expressed as a decimal, t is the time period, and n is the number of times the interest is compounded per year.

the stated rate of change is always equal to the actual rate of change. For example, if a question says that the height of a plant increases by 10% each month, it means that $\frac{\text{height of the plant now}}{\text{height of the plant a month ago}} = 1.1$ (or if a question says that the population of a city is decreasing by 3% per year, it means that $\frac{\text{population of the city now}}{\text{population of the city a year ago}} = 0.97$. Then, if the question asks by what percentage the height of the plant will increase in 2 months, you can write

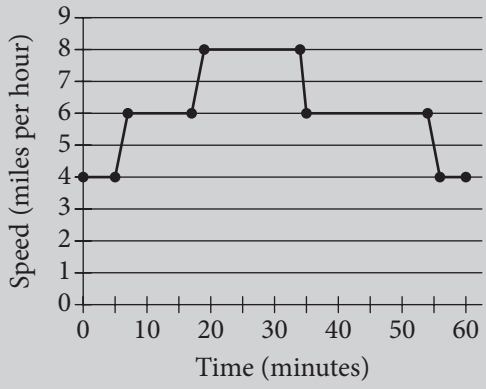
$$\frac{\text{height of the plant in 2 months}}{\text{height of the plant now}} = \frac{\text{height of the plant in 2 months}}{\text{height of the plant in 1 month}} + \frac{\text{height of the plant in 1 month}}{\text{height of the plant now}}$$

$$= 1.1 \times 1.1 = 1.21$$

Therefore, the answer is that the height of the plant increases by 21% in 2 months.

An SAT Math Test question may ask you to interpret a graph that shows the relationship between two variables.

EXAMPLE 10



Each 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, 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, line graph, or other display.

EXAMPLE 11

A store is deciding whether to install a new security system to prevent shoplifting. The security manager of the store estimates that 10,000 customers enter the store each week, 24 of whom will attempt to shoplift. 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) 0.03%
B) 0.35%
C) 0.56%
D) 62.5%

PRACTICE AT
khanacademy.org/sat

Probability is the measure of how likely an event is. When calculating the probability of an event, use the following formula:

probability = number of favorable (or desired) outcomes / total number of possible outcomes

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 $\frac{35}{56} = \frac{5}{8} = 62.5\%$. The correct answer is choice D.

Example 11 is an example of a conditional probability.

You may be asked to answer questions that involve a measure of center for a data set: the average (arithmetic 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:

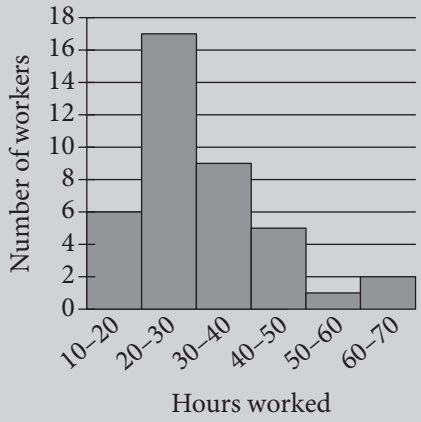
The mean of a set of numerical values is the sum of all the values divided by the number of values in the set.

REMEMBER

Mean, median, and mode are measures of center for a data set, while range and standard deviation are measures of spread.

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 12



The histogram above summarizes the number of hours worked last week 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 number of hours worked 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 up 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 average 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.

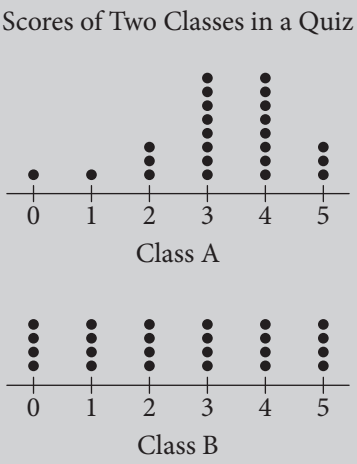
In general, outliers affect the mean but not the median. Therefore, outliers that are larger than the rest of the points in the data set tend to make the mean greater than the median, and outliers that are smaller than the rest of the points in the data set tend to make the mean less than the median. The most evident graphical display used to identify outliers is the box plot.

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 variation, or spread, in the data. One measure of spread is the *standard deviation*, which is a measure of how far away the points in the data set are from the average 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.

REMEMBER

You will not be asked to calculate the exact standard deviation of a set of data on the SAT Math Test, but you will be expected to demonstrate an understanding of what standard deviation measures.

EXAMPLE 13



The dot plots above summarize the scores that two classes, each with 24 students, at Central High School achieved on a current events quiz. Which of the following correctly compares the standard deviation of the scores in each of the classes?

A) The standard deviation of the scores in Class A is smaller.
B) The standard deviation of the scores in Class B is smaller.
C) The standard deviation of the scores in Class A and Class B is the same.
D) The relationship cannot be determined from the information given.

PRACTICE AT
khanacademy.org/sat

When asked to compare the standard deviations of two data sets, first approximate the mean of each data set. Then, ask yourself which data set has values that are more closely clustered around the mean. That data set will have the smaller standard deviation.

In Class A, the large majority of scores are 3 and 4, with only a few scores of 0, 1, 2, and 5; the average score is between 3 and 4. In Class B, the scores are evenly spread out across all possible scores, with many scores not close to the average score, which is 2.5. 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. Or the average income of a household for a city is a parameter describing the population of households in that city. 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 in an election to estimate the outcome of an election. The precision of the estimate depends on the variability of the data and the sample size. For example, 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 (the parameter).

For example, suppose you want to estimate the average amount of time each week that the students at a high school spend on the Internet. Suppose

the high school has 1,200 students. It would be time consuming to ask all 1,200 students, but you can ask a sample of students. Suppose you have time to ask 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. Randomization is essential in protecting against bias and helps to calculate the sampling error reliably. This can be done in different ways. You could 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 a random sample.

If you do not select a random sample, it 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 affect 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 find that the average time is 14 hours. You also find that 6 of the 80 students spend less than 2 hours each week on the Internet. How can these results be used to make a generalization about the entire population of 1,200 students?

Because the sample was selected at random, the average of 14 hours is the most reasonable estimate for average time on the Internet for all 1,200 students. Also, you can use proportional reasoning to estimate the number of students at the school who spend less than 2 hours on the Internet each week. Because 6 of the 80 students in the sample spend less than 2 hours per week on the Internet, the best estimate for the number of students in the entire school who spend less than 2 hours is x students, where $\frac{x}{1{,}200} = \frac{6}{80}$. Solving this equation for x gives $x = 90$. So it is appropriate to estimate that 90 of the 1,200 students at the school spend less than 2 hours per week on the Internet.

But this is not all. 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 likely to be. To describe the precision of an estimate, statisticians use *margins of error* and *confidence intervals*. On the SAT, you will not be expected to compute a margin of error or a confidence interval, but you should understand how different factors affect the margin of error and how to interpret a given margin of error or confidence interval in the context.

If the example above were an SAT question, you might be told that the estimate of an average of 14 hours per week on the Internet from the random sample of 80 students has margin of error 1.2 hours at 95% confidence level.

REMEMBER

You will not need to calculate margins of error or confidence intervals on the SAT Math Test, but you should understand what these concepts mean and be able to interpret them in context.

This means that in random samples of size 80, the actual average will be within 1.2 hours of the true average in 95% of possible samples. In terms of confidence intervals, you can be 95% confident that the interval from 12.8 hours to 15.2 hours includes the true average amount of time on the Internet for all students at the school. (**Note:** In statistics, confidence levels other than 95% can be used, but SAT questions will always use 95% confidence levels.)

There are some key points to note.

1. When the confidence level is kept the same, the size 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. Increasing the size of the random sample provides more information and reduces the margin of error.
2. The margin of error and the confidence interval apply to the estimated value of the parameter for the entire population, *not* for the value of the variable for particular individuals. In the example, we are 95% confident that the interval from 12.8 to 15.2 hours includes *the true average* amount of time on the Internet for all students at the school. It does not imply that 95% of students spend between 12.8 and 15.2 hours on the Internet.

EXAMPLE 14

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 average life of the batteries has a 95% confidence interval of 324 to 360 camera pictures. Which of the following conclusions is the most reasonable based on the confidence interval?

A) 95% of all the batteries produced by the company that day have a life between 324 and 360 pictures.
B) 95% of all the batteries ever produced by the company have a life between 324 and 360 pictures.
C) It is plausible that the true average life of batteries produced by the company that day is between 324 and 360 pictures.
D) It is plausible that the true average life of all the batteries ever produced by the company is between 324 and 360 pictures.

PRACTICE AT
khanacademy.org/sat

When a confidence interval is provided, determine the value for which the interval applies. Confidence intervals concern the average value of a population and do not apply to values of individual objects in the population.

The correct answer is choice C. Choices A and B are incorrect because the confidence interval gives information about the true *average* 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 of the batteries produced by the company on that day. The population of all batteries the company ever produced may have a different average 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 population: the amount of time students spend on the Internet, the 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, while other people who have had knee surgery will be randomly assigned to the usual therapy. The medical results of these patients can be compared. The key questions from a statistical viewpoint are:

- Can the results appropriately be generalized from the sample of patients in the study to the entire population of people who are recovering from knee surgery?
- Do the results allow one to appropriately 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 sampling and random assignment of individuals into groups of different conditions.

- If the sample of all subjects in a study were selected at random from the entire population in question, the results can appropriately 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	Results can be appropriately generalized to the entire population. Conclusions about cause and effect can appropriately be drawn.	Results *cannot* be appropriately generalized to the entire population. Conclusions about cause and effect can appropriately be drawn.
Subjects not randomly assigned to treatments	Results can be appropriately generalized to the entire population. Conclusions about cause and effect *cannot* appropriately be drawn.	Results *cannot* be appropriately generalized to the entire population. Conclusions about cause and effect *cannot* appropriately be drawn.

PRACTICE AT
khanacademy.org/sat

In order for results of a study to be generalized to the entire population, and for a cause-and-effect relationship to be established, both random sampling and random assignment of individuals to treatments is needed.

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

PRACTICE AT
khanacademy.org/sat

Be wary of conclusions that claim a cause-and-effect relationship or that generalize a conclusion to a broader population. Before accepting a conclusion, assess whether or not the subjects were selected at random from the broader population and whether or not subjects were randomly assigned treatments.

EXAMPLE 15

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. Which of the following is an 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 not use the additional audio lessons.