26 Statistics II

The goal of statistics is to be able to make predictions and estimates based on limited information. For example, we might want to find the mean weight of all raccoons in the United States. The problem is that it's impossible to survey the entire raccoon population. In fact, by the time that could be accomplished, not only would the data be out of date but there would be new raccoons in the population. The best we can do is to estimate the mean by selecting a certain number of raccoons at random from the population and weighing them. In statistics, selecting a group of members from a population is called **sampling**, and the resulting group of selected members is called a **sample**. A group of members selected at random is called a **random sample**.

In statistics, randomness is very important because 1) it ensures that each member of the population has an equal chance of being included, and 2) it minimizes bias and the risk of favoring certain characteristics or groups, which could distort the results. Using a random sample to estimate something about the entire population is a core tenet of statistics. Here are some examples of how random samples can be used:

- The mean weight of a random sample of raccoons is used to estimate the mean weight of the entire raccoon population.
- A journalist conducts a poll of a random sample of a city's residents to estimate the total number of residents who support a new law.
- A random sample of college students is surveyed to estimate the percentage of all college students who live with roommates.

On the SAT, statistical conclusions are only valid when random samples are used.

EXAMPLE 1: A gym chain has a total of 18,000 members. A manager selected 1,000 members at random and asked how many workout partners each of them had. The results of the survey are shown in the table.

Number of workout partners	0	1	2	3 or more
Number of members	350	300	200	150

Based on the table, what is the best estimate of the total number of members who have 2 workout partners?

Using the sample data, we can estimate the total number of members who have 2 workout partners to be $\frac{200}{1{,}000} \times 18{,}000 = \boxed{3{,}600}$.

K Education & Technology

EXAMPLE 2: In a study, a random sample of turtles was selected from a certain population. Based on the sample, it is estimated that the proportion of the population that has spotted shells is 0.35, with an associated margin of error of 0.06. Based on this estimate and margin of error, which of the following is the most appropriate conclusion about the proportion of the population that has spotted shells?

A) The proportion is exactly 0.35.
B) It is plausible that the proportion is less than 0.29.
C) It is not possible that the proportion is greater than 0.41.
D) It is plausible that the proportion is between 0.29 and 0.41.

As we've learned, samples of a population can be used to estimate characteristics of the whole population. Because an estimate obtained from a sample is not guaranteed to be the actual value for the population, an **associated margin of error** is often given to express how far off the sample estimate might be (i.e. the room for error). Given a margin of error, we can establish a range of values within which the actual value is likely to fall. Note that on the SAT, you will never have to calculate the margin of error. It will always be given.

In this example, the sample proportion is 0.35, with an associated margin of error of 0.06. Given this margin of error, it is plausible that the population proportion is between $0.35 - 0.06 = 0.29$ and $0.35 + 0.06 = 0.41$.

Answer $\boxed{(D)}$. Note that the word *plausible* means *likely* or *reasonable*.

Answer A is incorrect because an estimate obtained from a sample is just an estimate. It's not guaranteed to be the true value for the whole population. Answer B is incorrect because the plausible range is from 0.29 to 0.41. A proportion less than 0.29 is not plausible. Answer C is incorrect because although a proportion greater than 0.41 is not plausible, it's still *possible*. After all, luck plays a role in statistics and the chosen sample may not be representative of the whole population.

EXAMPLE 3: Based on a random sample of a population of dolphins, the mean dorsal fin height of the population is estimated to be 14.6 inches, with an associated margin of error of 1.5 inches. Based on this estimate and margin of error, which of the following is the most appropriate conclusion?

A) The majority of dolphins in the population have a dorsal fin height that is between 13.1 and 16.1 inches.
B) All dolphins in the population have a dorsal fin height of 14.6 inches.
C) It is plausible that the mean dorsal fin height of the population is between 13.1 and 16.1 inches.
D) No dolphins in the population have a dorsal fin height that is less than 13.1 inches or greater than 16.1 inches.

This example question is essentially the same as the previous one, but we have to watch out for answers that try to assign a population characteristic to the members within a population. Let me explain.

Imagine that we estimated the mean weight of your family members to be 145 pounds, with an associated margin of error of 5 pounds. The key word here is *mean*. Our estimate of the mean is not enough for us to conclude that your grandmother weighs 145 pounds or that your father weighs 145 pounds, nor is it enough to conclude that most of your family members weigh between 140 pounds and 150 pounds. After all, if half your family members weighed 50 pounds each and the other half weighed 240 pounds each, the mean weight would still be 145 pounds. The only conclusion we would be able to make is that the *mean* weight of your family members is likely between 140 pounds and 150 pounds. An estimate of the mean applies only to the mean.

In this question, answers A, B, and D may seem valid but are all incorrect because they are conclusions about the actual dorsal fin heights of the dolphins within the population, rather than the *mean* dorsal fin height.

Only answer $\boxed{(C)}$ is appropriate. Based on the given estimate and margin of error, it is plausible that the population mean is between $14.6 - 1.5 = 13.1$ inches and $14.6 + 1.5 = 16.1$ inches.

K Education & Technology

Before we move onto the next example, let's talk about **generalizability**. In statistics, the results of any survey or study can only be applied to the population from which the random sample was drawn. Here are some examples:

- In a study, a random sample of elephants in Thailand was used. The results of the study can only be generalized to all elephants in Thailand.
- For a survey, 100 pilots over the age of 45 were selected at random. The results of the survey can only be generalized to pilots over the age of 45.
- For a poll, 50 members of an alumni association were selected at random. The results of the poll can only be applied to the members of that particular alumni association.

As you've probably noticed, samples must be randomly selected. If a sample is not randomly selected, then any results based on that sample cannot be generalized to the population.

EXAMPLE 4: Malden is a town in the state of Massachusetts. In a study, a real estate agent randomly selected 50 houses in Malden and estimated that the mean house value in Malden is \$750,000, with an associated margin of error of \$80,000. The agent wants to repeat the study to obtain another estimate of the mean house value in Malden, but one that has a smaller margin of error. Which of the following samples would be most appropriate for the agent to use when the study is repeated?

A) 30 randomly selected houses in Malden
B) 80 randomly selected houses in Massachusetts
C) 80 randomly selected houses in Malden
D) 100 randomly selected houses in one of the neighborhoods of Malden

To obtain an estimate specific to Malden, no houses outside of Malden can be included in the sample, so we can eliminate answer B. The estimate also needs to be generalizable to all houses in Malden, not just one of its neighborhoods, so we can eliminate answer D. Only a sample of houses randomly selected from the entire town of Malden can be used.

Now when it comes to margin of error, its value is influenced by two main factors:

- Variability/Spread in the population data set (often measured by standard deviation)
- Sample size

Let's briefly talk about the first factor, even though it's typically not something we can control. Variability refers to how spread out a data set is. The higher the variability in a population data set, the higher the margin of error for any estimate based on a sample of that data set. If all the houses in Malden had a value between \$700,000 and \$710,000, a sample estimate of the mean value would likely have a smaller margin of error than if all the houses in Malden had a value between \$100,000 and \$800,000. When a population has a wide array of values, obtaining a sample that is representative of the whole population is much more difficult. Therefore, most samples result in estimates that are less accurate.

The second factor is sample size. A larger sample size reduces the margin of error. This makes sense because the more data points we include in a sample, the more representative it is of the whole population. Thus, estimates based on larger samples tend to be more accurate.

Given this discussion of sample size, it's clear that the answer to the question at hand is $\boxed{(C)}$. The real estate agent initially used a random sample of 50 houses. To achieve a smaller margin of error, a larger sample size is required. Between answers A and C, only answer C offers a sample size larger than 50.

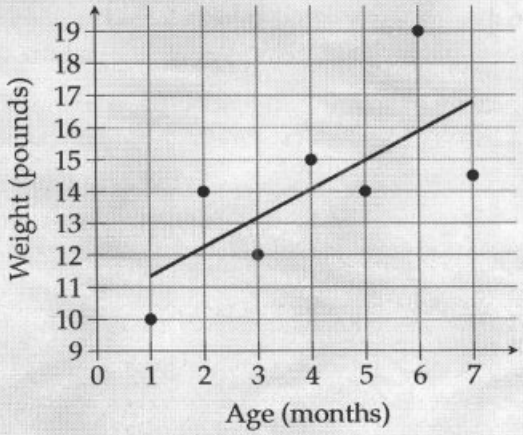
Why don't all studies use large sample sizes? Because gathering more sample data can be too costly and time-consuming.

K Education & Technology

Scatterplots

So far, we've only dealt with univariate data (one variable). In situations where we want to estimate or predict a value based on another value, we need to work with bivariate data (two variables), which can be visualized using a scatterplot.

EXAMPLE 5: The scatterplot shows the relationship between age, in months, and weight, in pounds, for 7 infants. A line of best fit is also shown.



Based on the line of best fit, which of the following is closest to the predicted weight, in pounds, of a 5-month-old infant?

A) 12 B) 14 C) 15 D) 16

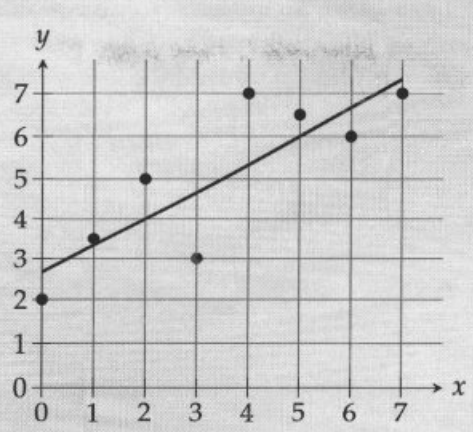
You interpret the points in a scatterplot in the same way you would interpret points in the xy-coordinate plane. For example, the data point at $(1, 10)$ represents an infant who is 1 month old and weighs 10 pounds.

Predictions from scatterplot data can be made using a linear model called the line of best fit, which is the line that best approximates all the points in the data. The line of best fit is found through a process called linear regression, which is not something the SAT will test you on. On the SAT, the line of best fit will either be shown or easily approximated by hand.

For this question, the line of best fit predicts a weight of 15 pounds ($y = 15$) for an infant who is 5 months old ($x = 5$). Answer $\boxed{(C)}$. Note that the answer is not 14 pounds—we want the *predicted* weight given by the line of best fit, not the actual weight of a 5-month-old from the data set.

K Education & Technology

EXAMPLE 6: The scatterplot shows the relationship between two variables, x and y. A line of best fit is also shown.



Which of the following equations best represents the line of best fit shown?

A) $y = 0.67x - 2.7$ B) $y = 0.67x + 2.7$ C) $y = 1.5x - 2.7$ D) $y = 1.5x + 2.7$

The y-intercept of the line of best fit is approximately 2.7. The slope can be estimated from any two points on the line. Using the points $(2, 4)$ and $(5, 6)$, we can calculate the slope to be $\frac{y_2 - y_1}{x_2 - x_1} = \frac{6-4}{5-2} = \frac{2}{3} \approx 0.67$. Given these values, the equation of the line of best fit is best represented by $y = 0.67x + 2.7$. Answer $\boxed{(B)}$.

Experimental design

EXAMPLE 7: Researchers conducted an experiment to determine whether exercise improves student exam scores. They randomly selected 200 students and tracked how much they exercised over the course of a year. At the end of the year, the researchers found that the students who did more exercise performed significantly better on the same exams than the students who did less exercise. Based on the design and results of this study, which of the following is an appropriate conclusion?

A) Exercise likely improves exam scores.
B) Students who do not exercise at all cannot improve their exam scores.
C) Any student who starts exercising more will improve his or her exam scores.
D) There is a positive association between exercise and exam scores.

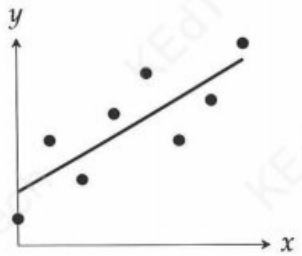
This question deals with a classic case of **association** (also called correlation) vs. **causation** (cause and effect relationship). Just because the students who exercised more got better exam scores doesn't mean that exercise *causes* an improvement in exam scores. All we can conclude is that more exercise is *associated* with better exam scores. Perhaps students who exercise more tend to have more discipline. Perhaps they're more likely to have a consistent routine that allows for dedicated study time. We don't know. Due to the way the experiment was designed, we can't tell if exercise is the cause or if some other underlying factor is the cause.

Therefore, answer A is wrong because it implies causation. Answer B is wrong because it makes an illogical generalization—the experiment wasn't about whether the students were capable of improving their exam scores. Answer C is wrong because it implies causation and makes a strong claim that is not guaranteed to be true. Even if exercise did improve exam scores, not every single student who starts exercising would improve their scores. After all, exercise is not the sole determinant of exam scores.

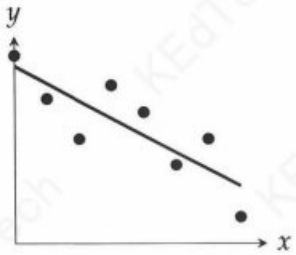
Thus, the answer is $\boxed{(D)}$. There is a **positive association** between exercise and exam scores.

K Education & Technology

Note that a **positive association** refers to a relationship between two variables in which an increase in one variable is associated with an increase in the other variable. A **negative association**, on the other hand, refers to a relationship between two variables in which an increase in one variable is associated with a decrease in the other variable. A positive association is often modeled by a line with a positive slope (as x increases, y also increases), and a negative association is often modeled by a line with a negative slope (as x increases, y decreases).



positive association



negative association

In the experiment, the researchers wanted to determine whether or not exercise *causes* an improvement in exam scores. What should they have done differently to prove causation rather than association? The answer is **random assignment**. Instead of tracking how much each student exercised, they should have randomly assigned each student to one of two groups: the "more exercise" group or the "less exercise" group. Everyone assigned to the "more exercise" group exercises frequently each week, and everyone assigned to the "less exercise" group exercises infrequently each week or not at all. If the "more exercise" group performs significantly better on the same exams than the "less exercise" group, then we can conclude that exercise causes an improvement in exam scores. Of course, conducting this type of experiment can be extremely difficult, which is why proving causation is usually a monumental task.

Why is random assignment needed to prove causation? Imagine that the researchers assigned all the students over the age of 15 to the "more exercise" group and all the students under the age of 15 to the "less exercise" group, and the "more exercise" group still performed significantly better on the same exams. Did they perform better because they exercised more or because they were older? Because age is now a **confounding variable** that could explain the difference in exam scores, we can't conclude that there is a causal link between exercise and exam scores. Random assignment controls for these potential confounding variables by ensuring that participants and their characteristics are equally distributed across experimental and control groups.

K Education & Technology

EXERCISE 1: Answers for this exercise start on page 411.

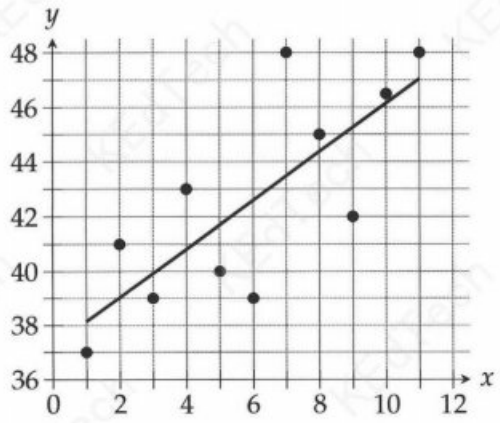
1

An online service currently has 8,400 users. A random sample of 350 users was selected and asked whether they will renew their subscription to the service. Of those surveyed, 12% stated that they will not renew their subscription. Based on this survey, which of the following is the best estimate of the total number of users who will not renew their subscription?

A) 42
B) 200
C) 1,008
D) 7,392

2

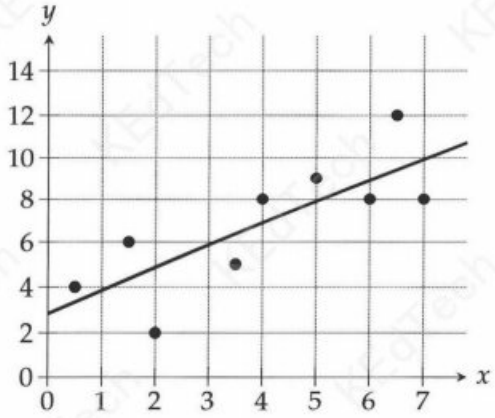
The scatterplot shows the relationship between two variables, x and y. A line of best fit is also shown.



Which of the following is closest to the predicted value of y at $x = 8.5$?

A) 42.1
B) 43.9
C) 44.8
D) 45.3

3



The scatterplot shows the relationship between two variables, x and y. A line of best fit is also shown. For how many of the data points does the line of best fit predict a y-value that is greater than the actual y-value?

4

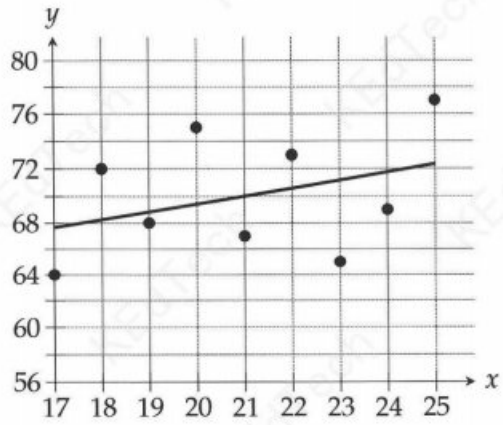
A random sample of 400 people from a town were interviewed about their attitudes toward community wellness programs. Based on this sample, it is estimated that 58% of the town's population supports additional funding for community wellness programs, with an associated margin of error of 6%. Which of the following is the most appropriate conclusion from these data?

A) Exactly 58% of the town's population supports additional funding for community wellness programs.
B) It is plausible that between 52% and 64% of the town's population supports additional funding for community wellness programs.
C) It is plausible that less than 52% of the town's population supports additional funding for community wellness programs.
D) It is plausible that more than 64% of the town's population supports additional funding for community wellness programs.

K Education & Technology

5

The scatterplot shows the relationship between two variables, x and y. A line of best fit for the data is also shown.



At $x = 21$, which of the following is closest to the amount by which the y-value predicted by the line of best fit exceeds the actual y-value?

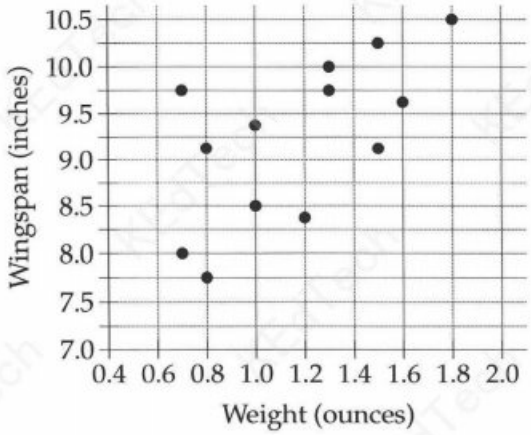
A) 1.5
B) 3
C) 4
D) 6

6

For a poll, an organizer randomly selected 21 professors from a certain liberal arts program at a university. These professors were asked whether they were interested in using a new textbook for their classes, and 14 of them responded "Yes." What is the largest group to which the results of the poll can be generalized?

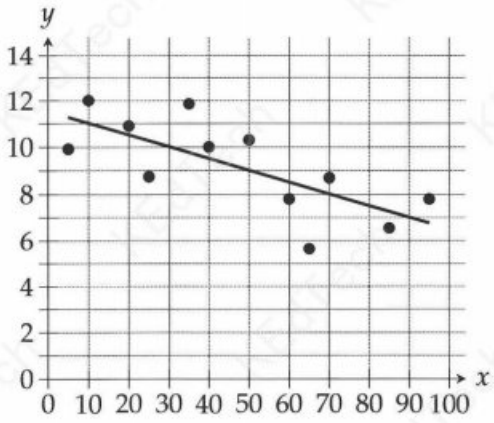
A) The 14 professors who responded "Yes"
B) All liberal arts professors at the university
C) All professors in the program
D) All professors at the university

7



The scatterplot shows the weight, in ounces, and the wingspan, in inches, for 13 sparrows. How many of these sparrows weigh less than 1.4 ounces and have a wingspan greater than 9.25 inches?

8



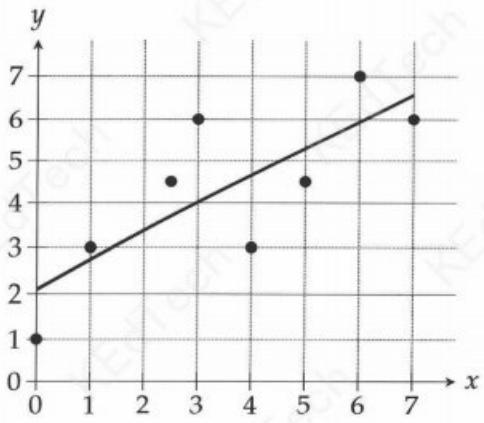
In the scatterplot, a line of best fit for the data is shown. Which of the following equations best represents this line of best fit?

A) $y = 11.5 - 0.05x$
B) $y = 11.5 - 0.5x$
C) $y = -11.5 + 0.05x$
D) $y = -11.5 + 0.5x$

K Education & Technology

9

In the given scatterplot, a line of best fit for the data is shown.



Which of the following is closest to the slope of the line of best fit?

A) -1.56
B) -0.64
C) 0.64
D) 1.56

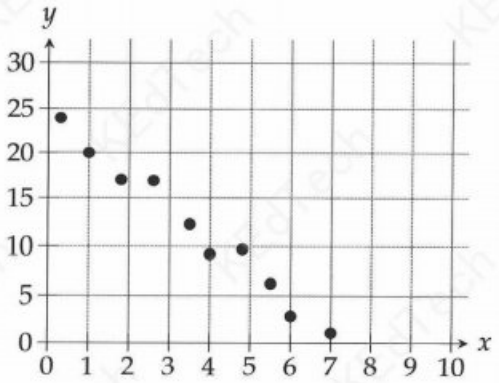
10

An economist conducted research to determine whether there is a relationship between the price of food and population density. He collected data from a random sample of 100 U.S. cities and found significant evidence that the price of food is lower in places with a high population density. Which of the following conclusions is best supported by these results?

A) In U.S. cities, there is a positive association between the price of food and population density.
B) In U.S. cities, there is a negative association between the price of food and population density.
C) In U.S. cities, a decrease in the price of food is caused by an increase in the population density.
D) In U.S. cities, an increase in the population density is caused by a decrease in the price of food.

11

The scatterplot shows the relationship between x and y.



Which of the following equations is the most appropriate linear model for the data shown?

A) $y = -0.7x - 24.2$
B) $y = -0.7x + 24.2$
C) $y = -3.4x - 24.2$
D) $y = -3.4x + 24.2$

12

In a poll, 500 people were chosen at random from a population of 36,000 and asked how many times they used public transportation in the last month. The table shows the results of the poll.

0 times	150
1 time	125
2 times	50
3 times	75
4 or more times	100

Based on these results, which of the following is the best estimate of the total number of people in the population who used public transportation at least two times in the last month?

A) 3,600
B) 9,000
C) 16,200
D) 19,800

K Education & Technology

13

A restaurant manager asked a random sample of 100 customers how long they waited for their order. Based on the responses, the manager estimated that the mean wait time for an order is 7 minutes, with an associated margin of error of 2 minutes. Which of the following is the most appropriate conclusion from this information?

A) It is likely that most customers wait 7 minutes for their order.
B) It is likely that all customers wait between 5 and 9 minutes for their order.
C) It is plausible that the mean wait time for an order is between 5 and 9 minutes.
D) It is not possible that the mean wait time for an order is less than 5 minutes or greater than 9 minutes.

14

A researcher conducted a study to determine if certain memorization strategies could improve the performance of chess players. For the study, the researcher selected a random sample of adults who are competing in a chess tournament. At the end of the study, 80% of those sampled reported that they did not find the memorization strategies to be useful during their chess games. Which of the following is the largest population to which the results of this study can be applied?

A) All chess players who compete in tournaments
B) All adults who compete in chess tournaments
C) All chess players competing in the tournament
D) All adults competing in the tournament

15

For a quality control test, a food manufacturing company selected a random sample of 100 granola bars from a large batch it produced. Based on the results of this test, it was estimated that the mean amount of fiber per bar is 5 grams, with an associated margin of 1.4 grams. The company then repeated the test with a random sample of 200 granola bars from the same batch. If the mean and margin of error were calculated the same way for both tests, which of the following statements is most likely true?

A) The mean from the second test is smaller than the mean from the first test.
B) The mean from the second test is larger than the mean from the first test.
C) The margin of error from the second test is smaller than the margin of error from the first test.
D) The margin of error from the second test is larger than the margin of error from the first test.

16

Researchers must conduct an experiment to see whether a new vaccine is effective in relieving certain allergies. They have selected a random sample of 100 patients suffering from allergies. Some of the patients are assigned to the new vaccine while the rest are assigned to the traditional treatment. Which of the following methods of assigning each patient's treatment is most likely to lead to a reliable conclusion about the effectiveness of the new vaccine?

A) Those who have more than one allergy are assigned to the new vaccine.
B) The median weight of the patients is found. Those who weigh less than the median weight are assigned to the new vaccine.
C) The patients divide themselves evenly into two groups. A coin is tossed to decide which group is assigned to the new vaccine.
D) Each patient is given a random number. Those with an even number are assigned to the new vaccine.

K Education & Technology