25 Statistics I

A data set is simply a collection of values. Let data set A consist of the following 6 values:

$$5, 6, 2, 2, 2, 7$$

The **mean** of data set A is the average of its values:

$$\frac{5+6+2+2+2+7}{6} = \boxed{4}$$

The **median** of a data set is the number in the middle when the values in the data set are ordered from least to greatest. For example, the median of $\{1,2,3,4,5\}$ is 3. For data set A, the values from least to greatest are

$$2, 2, 2, 5, 6, 7$$

However, there isn't a single number in the middle that we can consider the median. That's because there's an even number of numbers in the data set. When that's the case, the median is the average of the two middle numbers, 2 and 5:

$$\text{Median of data set A} = \frac{2+5}{2} = \boxed{3.5}$$

Now what if there were 100 numbers in data set A? How would we determine the median then? Well, if we're looking for the middle number, it would make sense to take half of 100 to get 50, which designates the 50th number. But is the 50th number really the median? Probably not since there's an even number of numbers. Maybe it's the average of the 50th and 51st numbers? Or is it the average of the 49th and 50th numbers? With large data sets, it's hard to tell.

Here's the foolproof technique I use to find the median of any data set: **regardless of whether the data set has an odd or even number of numbers, we always "go up."** So for a data set of 100 numbers, we divide by 2 to get 50 and "go up" to 51. Since we have two whole numbers, 50 and 51, the median is the average the 50th and 51st numbers. For a list of 101 numbers, we divide by 2 to get 50.5 and "go up" to 51. Since we only have one whole number, 51, the median is the 51st number.

This technique may seem strange but it works. Just for reassurance, let's test it out on a data set of 3 numbers, for which the 2nd number is obviously the median, and a data set of 4 numbers, for which the average of the 2nd and 3rd numbers is the median. For a data set of 3 numbers, we take half of 3 to get 1.5, and then "go up" to 2, which confirms that the 2nd number is the median. For a data set of 4 numbers, we take half of 4 to get 2, and then "go up" to 3. The whole numbers 2 and 3 confirm that the average of the 2nd and 3rd numbers is the median.

K Education & Technology

The **mode** of a data set is the value that appears most often in the data set. For data set A, it's $\boxed{2}$.

The **range** of a data set is the difference between the largest value in the data set and the smallest value:

$$\text{Range of data set A} = 7 - 2 = \boxed{5}$$

The **standard deviation** of a data set is a measure of how spread out the values in the data set are—in other words, the extent to which they "deviate" from the mean. The standard deviation is smaller when more numbers are closer to the mean. The standard deviation is greater when more numbers are spread out further away from the mean. For example, data set A,

$$2, 2, 2, 5, 6, 7$$

would have a higher standard deviation than the data set

$$5, 5, 5, 5, 6, 7$$

because this second data set has values that are closer together overall. It turns out that the standard deviation of data set A is 2.28 and the standard deviation of this second data set is 0.83. Don't worry about how we got these values—you'll never be asked to calculate standard deviations on the SAT. You'll only be asked to compare them, as we just did.

When assessing the standard deviation of a data set, it's important to note that the actual values in the data set don't matter, only their spacing in relation to one another. For example, consider the following two data sets:

Data set 1: 100, 101, 102, 103 Data set 2: 5, 10, 15, 20

Data set 1 has a smaller standard deviation than data set 2 even though its values are much larger. That's because the values in data set 1 are closer together (spaced one unit apart) whereas the values in data set 2 are more spread out (spaced five units apart).

EXAMPLE 1:

Data set A: 26, 32, 38, 44, 50, 56

Data set B: 26, 26, 32, 38, 44, 50, 56

The values in data sets A and B are shown. Which statement correctly compares the means of data sets A and B?

A) The mean of data set B is greater than the mean of data set A.
B) The mean of data set B is less than the mean of data set A.
C) The mean of data set B is equal to the mean of data set A.
D) There is not enough information to compare the means of the two data sets.

When a new value is added to a data set, it brings the mean up if it is greater than the existing mean and pulls the mean down if it is less than the existing mean.

Here, data sets A and B are identical except for the additional 26 in data set B. Because 26 is the minimum value in the data set (definitely lower than the mean), adding another 26 only serves to pull the mean down. Therefore, the mean of data set B must be less than the mean of data set A. Answer $\boxed{(B)}$. As proof, the mean of data set A is $\frac{26+32+38+44+50+56}{6} = 41$, whereas the mean of data set B is $\frac{26+26+32+38+44+50+56}{7} \approx 38.86$.

K Education & Technology

EXAMPLE 2: The mean weight of a group of pandas is 200 pounds. A new panda weighing 230 pounds is added to the group, increasing the mean weight of the entire group to 205 pounds. How many pandas were in the original group?

When dealing with word problems involving means or averages, think in terms of sums or totals. You can always find the sum of a data set by multiplying its mean by the number of values.

First, let the number of pandas in the original group be x. These x pandas weigh a total of $200x$ pounds. After the new panda is added, the $x+1$ pandas in the expanded group weigh a new total of $205(x+1)$ pounds.

Since the new panda weighs 230 pounds,

$$\begin{aligned} 200x + 230 &= 205(x+1) \\ 200x + 230 &= 205x + 205 \\ -5x &= -25 \\ x &= 5 \end{aligned}$$

Therefore, there were $\boxed{5}$ pandas in the original group.

EXAMPLE 3:

Value	Frequency
5	6
8	5
9	2
12	4

The frequency table summarizes the values in a data set. Which of the following statements correctly compares the mean and the median of this data set?

A) The mean is 2 greater than the median.
B) The mean is 1 greater than the median.
C) The mean is equal to the median.
D) The mean is 1 less than the median.

The **frequency** of a value is the number of times that value appears in a data set. The frequency table in this example indicates that there are six 5's, five 8's, two 9's, and four 12's in the data set:

$$5, 5, 5, 5, 5, 5, 8, 8, 8, 8, 8, 9, 9, 12, 12, 12, 12$$

We can calculate the sum of all these values by multiplying each value by its frequency and summing the results:

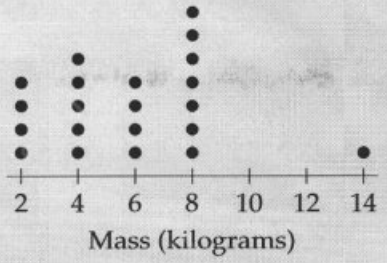
$$\text{Sum of the data set} = (5 \times 6) + (8 \times 5) + (9 \times 2) + (12 \times 4) = 136$$

Since there are $6+5+2+4 = 17$ values in the data set, the mean is $\frac{136}{17} = 8$.

Now let's use the technique discussed earlier to find the median. The median of this data set is the $17 \div 2 = 8.5 \rightarrow$ 9th value going from smallest to largest. Counting up to 9th value in the frequency table, we count past the six 5's and land within the five 8's, so the 9th value is 8. Thus, the median is 8 and is equal to the mean. Answer $\boxed{(C)}$.

K Education & Technology

EXAMPLE 4:



The masses, in kilograms, of 21 rock samples were recorded and summarized in the dot plot shown. It was found that the sample with the greatest recorded mass actually has a mass that is 4 kilograms more than what was recorded. If the record is corrected, which of the following measures of the data will increase?

I. The mean
II. The median
III. The range

A) I only B) I and III only C) II and III only D) I, II, and III

In a dot plot, the number of dots above a value represents the number of times that value appears in the data set (i.e. its frequency). A dot plot conveys the same information as a frequency table but allows us to more easily visualize how the values are spread out. In the given dot plot, the greatest recorded mass is 14 kilograms. In statistics, this data point is called an **outlier**, an extreme value that is further away from the rest of the data.

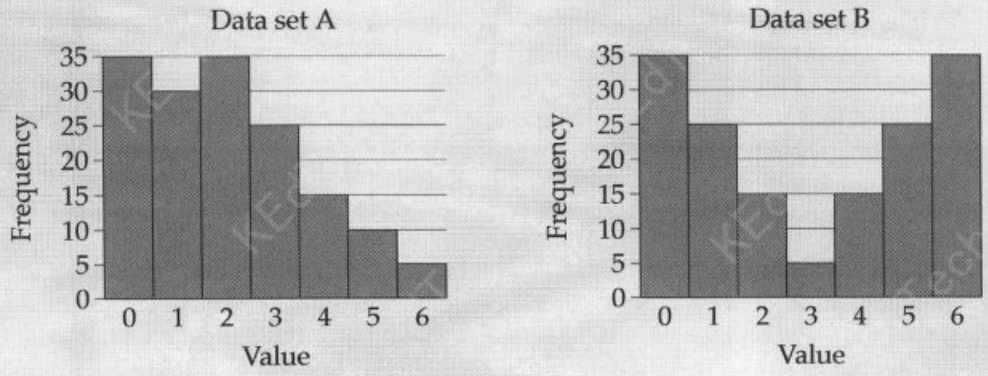
Because every value is taken into account when calculating the mean, the mean is easily affected by changes to the data. In fact, increasing any of the values in a data set increases the mean because the sum of the values in the data set increases in lockstep. Therefore, replacing the outlier value of 14 with an even greater value of 18 will undoubtedly increase the mean.

The median is less affected by changes to the data, especially changes to outliers. In this example, there are 21 values, so the median is the $21 \div 2 = 10.5 \rightarrow 11$th value, which is 6. Changing the value of the outlier does not change the location or value of the median. Therefore, the median will remain the same.

The range is the difference between a data set's maximum and minimum value. Because the value of 14 is also the maximum of the data set, increasing this value to 18 will increase the maximum, thereby increasing the range. In the end, the mean and the range will increase, but the median will remain the same. Answer $\boxed{(B)}$.

K Education & Technology

EXAMPLE 5:



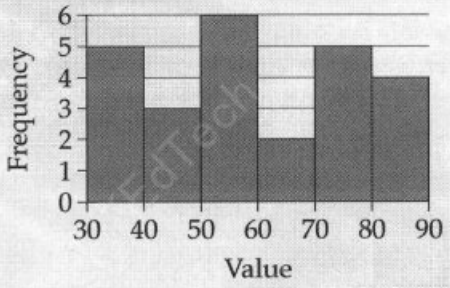
Data sets A and B are summarized in the bar graphs shown. Each data set consists of 155 values. Which of the following statements correctly compares the standard deviation of data set A and the standard deviation of data set B?

A) The standard deviation of data set A is greater than the standard deviation of data set B.
B) The standard deviation of data set A is less than the standard deviation of data set B.
C) The standard deviation of data set A is equal to the standard deviation of data set B.
D) There is not enough information to compare the standard deviations.

The bar graph for data set B is symmetrical. Because the lower values and the higher values are weighted equally, the mean of data set B is equal to its median of 3. We can estimate the mean of data set A to be roughly 2 since there are many more values from 0 to 2 than from 3 to 6.

Now most of the values in data set B are either at the low end or the high end, spread out away from its mean of 3. The values in data set A, on the other hand, are clustered towards the low end, where the mean also resides. Because the values in data set A are, on the whole, closer together and more tightly packed around their mean than those in data set B, the standard deviation of data set A is less than the standard deviation of data set B. Answer $\boxed{(B)}$.

EXAMPLE 6:



The histogram summarizes a data set consisting of 25 values. The first bar represents the frequency of values greater than or equal to 30, but less than 40. The second bar represents the frequency of values greater than or equal to 40, but less than 50, and so on. How many values in the data set are greater than or equal to 50 but less than 80?

A **histogram** is a bar chart that groups data values into intervals. When it's relevant, histogram questions on the SAT will clarify whether each interval includes its starting or ending value. In this example, there are 6 values in the data set that are greater than or equal to 50 but less than 60, 2 values greater than or equal to 60 but less than 70, and 5 values greater than or equal to 70 but less than 80. Therefore, there must be

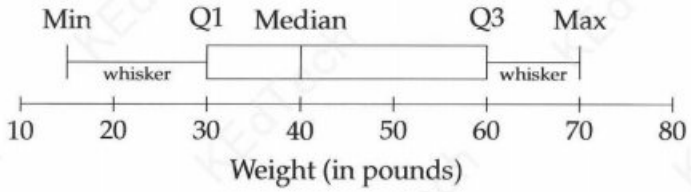
$6 + 2 + 5 = \boxed{13}$ values greater than or equal to 50 but less than 80.

K Education & Technology

Box plots

Like dot plots and histograms, box plots are just another way of visualizing numerical data. Five statistical measures of a data set are needed to construct a box plot: the minimum, the first quartile, the median, the third quartile, and the maximum.

Here's an example of a box plot that summarizes the weights of 30 tortoises:



The left and right sides of the "box" represent the first and third quartiles (Q1 and Q3), respectively, and the line segment inside the box designates the median. Line segments are drawn from the left and right sides of the box to the minimum and the maximum. These segments are sometimes called **whiskers**.

From the box plot, we can see that the minimum weight among the 30 tortoises is 15 pounds, the first quartile is 30 pounds, the median is 40 pounds, the third quartile is 60 pounds, and the maximum is 70 pounds.

The **first quartile**, also known as the lower quartile or 25th percentile, is the value for which 25% of the data set is less than. The **third quartile**, also known as the upper quartile or 75th percentile, is the value for which 75% of the data set is less than. So in this example, a first quartile of 30 pounds means that 25% of the tortoises represented in the data set weigh less than 30 pounds. A third quartile of 60 pounds means that 75% of the tortoises represented in the data set weigh less than 60 pounds.

To use another example, if your score on an exam is equal to the first quartile, then you scored better than 25% of the people who took the exam. If your score on an exam is equal to the third quartile, then you scored better than 75% of the people who took the exam.

Note that the **median** can be defined as the 50th percentile, the value for which 50% of the data set is less than.

Box plot questions on the SAT mainly test your ability to read a box plot. You will not need to know how to calculate the first and third quartiles from a raw data set.

K Education & Technology

EXERCISE 1: Answers for this exercise start on page 407.

1

Data set P: 41, 43, 45, 45, 47, 47, 51, 51, 55

Data set Q: 33, 33, 36, 38, 38, 40, 40, 40

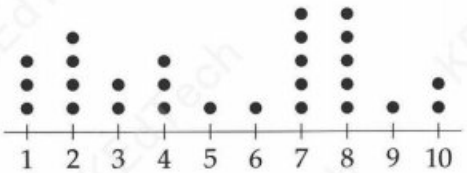
The data values for data sets P and Q are shown. The median of data set P is how much greater than the median of data set Q?

2

Value	Frequency
17	15
19	21
23	27
28	33

What is the range of the data set represented in the frequency table?

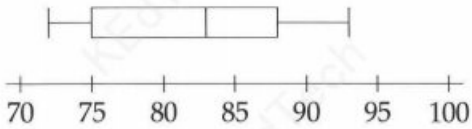
3



A data set consisting of 27 values is represented in the dot plot shown. What is the median of this data set?

4

Final Exam Scores
(out of 100 points)



The box plot summarizes the final exam scores of 26 students in a math class. Based on the box plot, which of the following best estimates the number of points by which the median score exceeds the lowest individual score?

A) 3
B) 5
C) 8
D) 11

5

The mean height of the 14 students in a particular class is 63 inches. The mean height of the 21 students in another class is 68 inches. If these two classes are combined, what is the mean height, in inches, of the students in the combined class?

A) 64.5
B) 65
C) 66
D) 66.5

K Education & Technology

6

A food company hires an independent research agency to determine its product's shelf life, the length of time it may be stored before it expires. Using a random sample of 40 units of the product, the research agency finds that the product's shelf life has a range of 3 days. Which of the following must be true about the units in the sample?

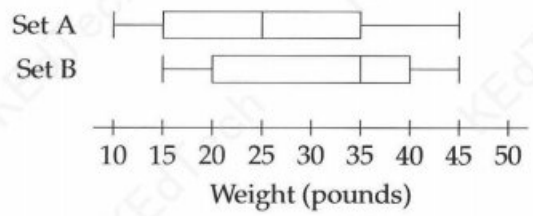
A) All the units expired within 3 days.

B) The unit with the longest shelf life took 3 days longer to expire than the unit with the shortest shelf life.

C) The mean shelf life of the units is 3 more than the median.

D) The median shelf life of the units is 3 more than the mean.

7



The box plots summarize the weights, in pounds, of two sets of kettlebells. Which of the following statements must be true?

A) The number of kettlebells in set A is equal to the number of kettlebells in set B.

B) The median kettlebell weight in set A is equal to the median kettlebell weight in set B.

C) The range of kettlebell weights in set A is equal to the range of kettlebell weights in set B.

D) The maximum kettlebell weight in set A is equal to the maximum kettlebell weight in set B.

8

Data set A: 30, 24, 61, 37

Data set B: 77, 63, 114, x

Data set A and data set B each consist of 4 integers. If the mean of data set B is twice the mean of data set A, what is the value of x ?

9

Low temperature (°F)	Frequency
60	3
61	4
63	4
67	10
70	7

The table gives the distribution of daily low temperatures for a city over 28 days. What is the median low temperature, in degrees Fahrenheit (°F), of the city for these 28 days?

10

For which of the following data sets is the mean different than the median?

A) 3, 3, 3, 13, 13, 13, 23, 23, 23

B) 4, 4, 4, 4, 8, 16, 16, 16, 16

C) 7, 7, 7, 7, 7, 7, 7, 7, 7

D) 8, 16, 24, 32, 40, 48, 56, 64, 72

K Education & Technology

11

Data set A

Value	Frequency
30	11
50	11
70	11

The frequency table shown represents data set A. Which of the following frequency tables represents a data set that has a mean greater than the mean of data set A?

A)

Value	Frequency
30	12
50	12
70	12

B)

Value	Frequency
30	9
50	10
70	11

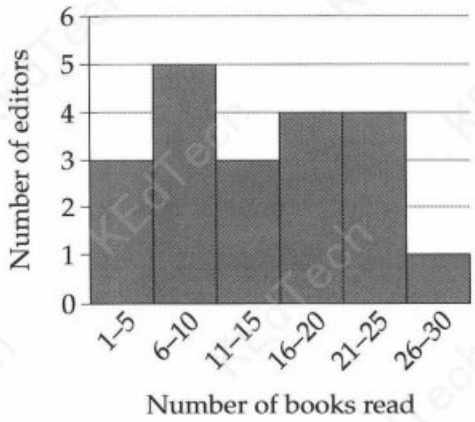
C)

Value	Frequency
30	12
50	11
70	10

D)

Value	Frequency
30	13
50	11
70	13

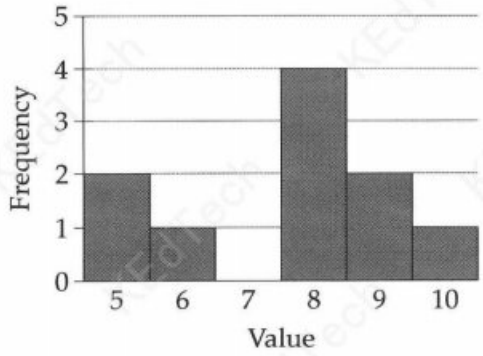
12



The histogram summarizes the number of books read last year by 20 editors at a publishing company. Which of the following could be the median number of books read by the 20 editors?

A) 9
B) 12
C) 18
D) 24

13



The bar graph shows the frequency of each value in a certain data set. Which of the following correctly gives the mean and the range of this data set?

A) Mean = 7.6, Range = 4
B) Mean = 7.6, Range = 5
C) Mean = 8.2, Range = 4
D) Mean = 8.2, Range = 5

K Education & Technology

14

91, 96, 88, 91, 84

The list shown gives the values in a data set. If a value of 100 is added to the data set, which of the following correctly describes how the mean and the median of the data set will change?

A) The mean and the median will increase.
B) The mean and the median will decrease.
C) The mean will increase, and the median will remain the same.
D) The mean will decrease, and the median will remain the same.

15

Data set A

Value	Frequency
10	2
11	3
12	7
13	3
14	9
15	5

Data set B

Value	Frequency
13	2
14	3
15	7
16	3
17	9
18	5

The tables show the frequencies of the values for data sets A and B. Which of the following statements best compares the means and the standard deviations of data sets A and B?

A) The mean of data set B is equal to the mean of data set A, and the standard deviation of data set B is equal to the standard deviation of data set A.
B) The mean of data set B is equal to the mean of data set A, and the standard deviation of data set B is greater than the standard deviation of data set A.
C) The mean of data set B is greater than the mean of data set A, and the standard deviation of data set B is equal to the standard deviation of data set A.
D) The mean of data set B is greater than the mean of data set A, and the standard deviation of data set B is greater than the standard deviation of data set A.

16

Value	Frequency
24	4
27	6
30	2
33	4
36	2

Which of the following statements correctly compares the mean and the median of the data set summarized in the frequency table?

A) The mean is 2 greater than the median.
B) The mean is 1.5 greater than the median.
C) The mean is equal to the median.
D) The mean is 1 less than the median.

17

Quiz	1	2	3	4	5	6	7
Score	87	75	90	83	98	87	91

The table shows Jay's scores on seven math quizzes. Which of the following are true about his scores?

I. The mode is greater than the median.
II. The median is greater than the mean.
III. The range is greater than 20.

A) II only
B) III only
C) II and III
D) I, II, and III

K Education & Technology