

## B.2 Measures of Central Tendency and Dispersion

### ▶ What you should learn

- Find and interpret the mean, median, and mode of a set of data.
- Determine the measure of central tendency that best represents a set of data.
- Find the standard deviation of a set of data.
- Create and use box-and-whisker plots.

### ▶ Why you should learn it

Measures of central tendency and dispersion provide a convenient way to describe and compare sets of data. For instance, in Exercise 36 on page B13, the mean and standard deviation are used to analyze the price of gold for the years 1984 through 2003.

## Mean, Median, and Mode

In many real-life situations, it is helpful to describe data by a single number that is most representative of the entire collection of numbers. Such a number is called a **measure of central tendency**. The most commonly used measures are as follows.

1. The **mean**, or **average**, of  $n$  numbers is the sum of the numbers divided by  $n$ .
2. The numerical **median** of  $n$  numbers is the middle number when the numbers are written in order. If  $n$  is even, the median is the average of the two middle numbers.
3. The **mode** of  $n$  numbers is the number that occurs most frequently. If two numbers tie for most frequent occurrence, the collection has two modes and is called **bimodal**.

### Example 1 ▶ Comparing Measures of Central Tendency



On an interview for a job, the interviewer tells you that the average annual income of the company's 25 employees is \$60,849. The actual annual incomes of the 25 employees are shown below. What are the mean, median, and mode of the incomes?

\$17,305,	\$478,320,	\$45,678,	\$18,980,	\$17,408,
\$25,676,	\$28,906,	\$12,500,	\$24,540,	\$33,450,
\$12,500,	\$33,855,	\$37,450,	\$20,432,	\$28,956,
\$34,983,	\$36,540,	\$250,921,	\$36,853,	\$16,430,
\$32,654,	\$98,213,	\$48,980,	\$94,024,	\$35,671

### Solution

The mean of the incomes is

$$\begin{aligned}\text{Mean} &= \frac{17,305 + 478,320 + 45,678 + 18,980 + \cdots + 35,671}{25} \\ &= \frac{1,521,225}{25} = \$60,849.\end{aligned}$$

To find the median, order the incomes as follows.

\$12,500,	\$12,500,	\$16,430,	\$17,305,	\$17,408,
\$18,980,	\$20,432,	\$24,540,	\$25,676,	\$28,906,
\$28,956,	\$32,654,	\$33,450,	\$33,855,	\$34,983,
\$35,671,	\$36,540,	\$36,853,	\$37,450,	\$45,678,
\$48,980,	\$94,024,	\$98,213,	\$250,921,	\$478,320

From this list, you can see that the median income is \$33,450. You can also see that \$12,500 is the only income that occurs more than once. So, the mode is \$12,500.



**CHECKPOINT** Now try Exercise 1.

In Example 1, was the interviewer telling you the truth about the annual incomes? Technically, the person was telling the truth because the average is (generally) defined to be the mean. However, of the three measures of central tendency *mean*: \$60,849 *median*: \$33,450 *mode*: \$12,500 it seems clear that the median is most representative. The mean is inflated by the two highest salaries.

## Choosing a Measure of Central Tendency

Which of the three measures of central tendency is the most representative? The answer is that it depends on the distribution of the data *and* the way in which you plan to use the data.

For instance, in Example 1, the mean salary of \$60,849 does not seem very representative to a potential employee. To a city income tax collector who wants to estimate 1% of the total income of the 25 employees, however, the mean is precisely the right measure.

### Example 2 ► Choosing a Measure of Central Tendency

Which measure of central tendency is the most representative of the data shown in each frequency distribution?

- a.
- |           |   |    |    |    |   |   |   |   |    |
|-----------|---|----|----|----|---|---|---|---|----|
| Number    | 1 | 2  | 3  | 4  | 5 | 6 | 7 | 8 | 9  |
| Frequency | 7 | 20 | 15 | 11 | 8 | 3 | 2 | 0 | 15 |
- b.
- |           |   |   |   |   |   |   |   |   |   |
|-----------|---|---|---|---|---|---|---|---|---|
| Number    | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Frequency | 9 | 8 | 7 | 6 | 5 | 6 | 7 | 8 | 9 |
- c.
- |           |   |   |   |   |   |   |   |   |   |
|-----------|---|---|---|---|---|---|---|---|---|
| Number    | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Frequency | 6 | 1 | 2 | 3 | 5 | 5 | 4 | 3 | 0 |

### Solution

- a. For these data, the mean is 4.23, the median is 3, and the mode is 2. Of these, the mode is probably the most representative measure.
- b. For these data, the mean and median are each 5 and the modes are 1 and 9 (the distribution is bimodal). Of these, the mean or median is the most representative measure.
- c. For these data, the mean is 4.59, the median is 5, and the mode is 1. Of these, the mean or median is the most representative measure.

 **CHECKPOINT** Now try Exercise 15.

## Variance and Standard Deviation

Very different sets of numbers can have the same mean. You will now study two **measures of dispersion**, which give you an idea of how much the numbers in a data set differ from the mean of the set. These two measures are called the *variance* of the set and the *standard deviation* of the set.

**Definitions of Variance and Standard Deviation**

Consider a set of numbers  $\{x_1, x_2, \dots, x_n\}$  with a mean of  $\bar{x}$ . The **variance** of the set is

$$v = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}$$

and the **standard deviation** of the set is  $\sigma = \sqrt{v}$  ( $\sigma$  is the lowercase Greek letter *sigma*).

The standard deviation of a data set is a measure of how much a typical number in the set differs from the mean. The greater the standard deviation, the more the numbers in the set *vary* from the mean. For instance, each of the following data sets has a mean of 5.

$$\{5, 5, 5, 5\}, \quad \{4, 4, 6, 6\}, \quad \text{and} \quad \{3, 3, 7, 7\}$$

The standard deviations of the data sets are 0, 1, and 2.

$$\sigma_1 = \sqrt{\frac{(5 - 5)^2 + (5 - 5)^2 + (5 - 5)^2 + (5 - 5)^2}{4}}$$

$$= 0$$

$$\sigma_2 = \sqrt{\frac{(4 - 5)^2 + (4 - 5)^2 + (6 - 5)^2 + (6 - 5)^2}{4}}$$

$$= 1$$

$$\sigma_3 = \sqrt{\frac{(3 - 5)^2 + (3 - 5)^2 + (7 - 5)^2 + (7 - 5)^2}{4}}$$

$$= 2$$

### Example 3 ▶ Estimations of Standard Deviation

Consider the three frequency distributions represented by the bar graphs in Figure B.4. Which data set has the smallest standard deviation? Which has the largest?

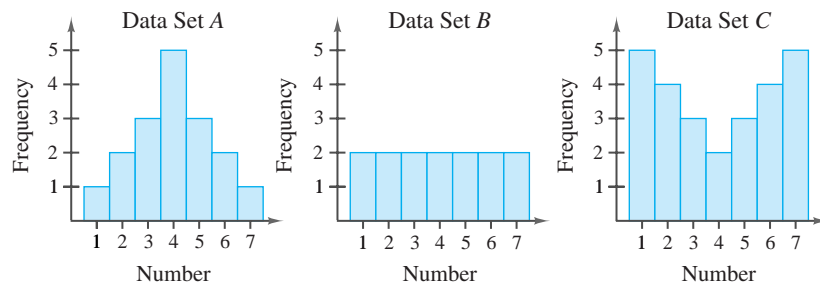


FIGURE B.4

### Solution

Of the three data sets, the numbers in data set A are grouped most closely to the center and the numbers in data set C are the most dispersed. So, data set A has the smallest standard deviation and data set C has the largest standard deviation.



Now try Exercise 17.

**Example 4 ▶ Finding Standard Deviation**

Find the standard deviation of each data set shown in Example 3.

**Solution**

Because of the symmetry of each bar graph, you can conclude that each has a mean of  $\bar{x} = 4$ . The standard deviation of data set A is

$$\begin{aligned}\sigma &= \sqrt{\frac{(-3)^2 + 2(-2)^2 + 3(-1)^2 + 5(0)^2 + 3(1)^2 + 2(2)^2 + (3)^2}{17}} \\ &\approx 1.53.\end{aligned}$$

The standard deviation of data set B is

$$\begin{aligned}\sigma &= \sqrt{\frac{2(-3)^2 + 2(-2)^2 + 2(-1)^2 + 2(0)^2 + 2(1)^2 + 2(2)^2 + 2(3)^2}{14}} \\ &= 2.\end{aligned}$$

The standard deviation of data set C is

$$\begin{aligned}\sigma &= \sqrt{\frac{5(-3)^2 + 4(-2)^2 + 3(-1)^2 + 2(0)^2 + 3(1)^2 + 4(2)^2 + 5(3)^2}{26}} \\ &\approx 2.22.\end{aligned}$$

These values confirm the results of Example 3. That is, data set A has the smallest standard deviation and data set C has the largest.

 **CHECKPOINT** Now try Exercise 19.

The following alternative formula provides a more efficient way to compute the standard deviation.

**Alternative Formula for Standard Deviation**

The standard deviation of  $\{x_1, x_2, \dots, x_n\}$  is

$$\sigma = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n} - \bar{x}^2}.$$

Because of lengthy computations, this formula is difficult to verify. Conceptually, however, the process is straightforward. It consists of showing that the expressions

$$\sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}}$$

and

$$\sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n} - \bar{x}^2}$$

are equivalent. Try verifying this equivalence for the set  $\{x_1, x_2, x_3\}$  with  $\bar{x} = (x_1 + x_2 + x_3)/3$ .

**Example 5 ▶ Using the Alternative Formula**

Use the alternative formula for standard deviation to find the standard deviation of the following set of numbers.

5, 6, 6, 7, 7, 8, 8, 8, 9, 10

**Solution**

Begin by finding the mean of the set, which is 7.4. So, the standard deviation is

$$\begin{aligned}\sigma &= \sqrt{\frac{5^2 + 2(6^2) + 2(7^2) + 3(8^2) + 9^2 + 10^2}{10} - (7.4)^2} \\ &= \sqrt{\frac{568}{10} - 54.76} \\ &= \sqrt{2.04} \\ &\approx 1.43.\end{aligned}$$

You can use the *one-variable statistics* feature of a graphing utility to check this result.

**CHECKPOINT**

Now try Exercise 27.

A well-known theorem in statistics, called *Chebychev's Theorem*, states that at least

$$1 - \frac{1}{k^2}$$

of the numbers in a distribution must lie within  $k$  standard deviations of the mean. So, at least 75% of the numbers in a data set must lie within two standard deviations of the mean, and at least 88.9% of the numbers must lie within three standard deviations of the mean. For most distributions, these percentages are low. For instance, in all three distributions shown in Example 3, 100% of the numbers lie within two standard deviations of the mean.

**Example 6 ▶ Describing a Distribution**

The table at the left shows the number of outpatient visits to hospitals (in millions) in each state and the District of Columbia in 2002. Find the mean and standard deviation of the data. What percent of the data values lie within two standard deviations of the mean? (Source: Health Forum)

**Solution**

Begin by entering the numbers into a graphing utility. Then use the *one-variable statistics* feature to obtain  $\bar{x} \approx 10.91$  and  $\sigma = 11.40$ . The interval that contains all numbers that lie within two standard deviations of the mean is

$$[10.91 - 2(11.40), 10.91 + 2(11.40)] \quad \text{or} \quad [-11.89, 33.71].$$

From the table you can see that all but two of the data values (96%) lie in this interval—all but the data values that correspond to the numbers of outpatient visits to hospitals in California and New York.

**CHECKPOINT**

Now try Exercise 36.

AK	1.3	MT	2.7
AL	10.0	NC	13.8
AR	4.8	ND	1.9
AZ	5.1	NE	3.5
CA	53.3	NH	3.0
CO	7.0	NJ	15.9
CT	6.6	NM	4.2
DC	1.5	NV	2.5
DE	1.9	NY	46.8
FL	22.5	OH	28.4
GA	12.5	OK	4.8
HI	2.0	OR	8.2
IA	9.3	PA	32.4
ID	2.4	RI	2.2
IL	26.5	SC	7.3
IN	14.1	SD	1.5
KS	5.7	TN	10.0
KY	8.6	TX	33.5
LA	10.7	UT	4.8
MA	19.0	VA	10.8
MD	6.5	VT	1.5
ME	3.7	WA	9.7
MI	25.9	WI	11.6
MN	8.9	WV	5.7
MO	14.9	WY	0.9
MS	4.0		

## Box-and-Whisker Plots

Standard deviation is the measure of dispersion that is associated with the mean. **Quartiles** measure dispersion associated with the median.

### Definition of Quartiles

Consider an ordered set of numbers whose median is  $m$ . The **lower quartile** is the median of the numbers that occur before  $m$ . The **upper quartile** is the median of the numbers that occur after  $m$ .

### Example 7 ► Finding Quartiles of a Data Set

Find the lower and upper quartiles for the data set.

34, 14, 24, 16, 12, 18, 20, 24, 16, 26, 13, 27

### Solution

Begin by ordering the data.

$\underbrace{12, 13, 14}_{1\text{st } 25\%}$ ,
  $\underbrace{16, 16, 18}_{2\text{nd } 25\%}$ ,
  $\underbrace{20, 24, 24}_{3\text{rd } 25\%}$ ,
  $\underbrace{26, 27, 34}_{4\text{th } 25\%}$

The median of the entire data set is 19. The median of the six numbers that are less than 19 is 15. So, the lower quartile is 15. The median of the six numbers that are greater than 19 is 25. So, the upper quartile is 25.



Now try Exercise 39(a).

Quartiles are represented graphically by a **box-and-whisker plot**, as shown in Figure B.5. In the plot, notice that five numbers are listed: the smallest number, the lower quartile, the median, the upper quartile, and the largest number. Also notice that the numbers are spaced proportionally, as though they were on a real number line.

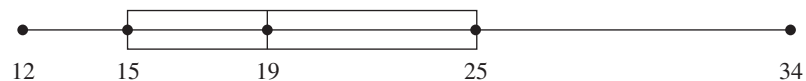


FIGURE B.5

The next example shows how to find quartiles when the number of elements in a data set is not divisible by 4.

**Example 8** ▶ Sketching Box-and-Whisker Plots

Sketch a box-and-whisker plot for each data set.

- 27, 28, 30, 42, 45, 50, 50, 61, 62, 64, 66
- 82, 82, 83, 85, 87, 89, 90, 94, 95, 95, 96, 98, 99
- 11, 13, 13, 15, 17, 18, 20, 24, 24, 27

**Solution**

- This data set has 11 numbers. The median is 50 (the sixth number). The lower quartile is 30 (the median of the first five numbers). The upper quartile is 62 (the median of the last five numbers). See Figure B.6.

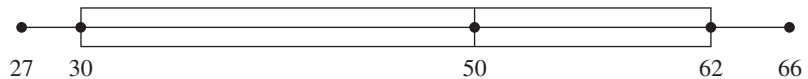


FIGURE B.6

- This data set has 13 numbers. The median is 90 (the seventh number). The lower quartile is 84 (the median of the first six numbers). The upper quartile is 95.5 (the median of the last six numbers). See Figure B.7.

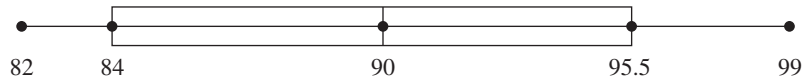


FIGURE B.7

- This data set has 10 numbers. The median is 17.5 (the average of the fifth and sixth numbers). The lower quartile is 13 (the median of the first five numbers). The upper quartile is 24 (the median of the last five numbers). See Figure B.8.

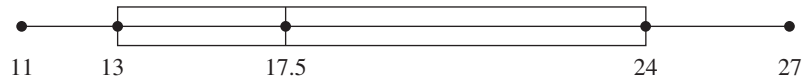


FIGURE B.8



Now try Exercise 41(b).

## B.2 Exercises

**VOCABULARY CHECK:** Fill in the blanks.

- A single number that is the most representative of a data set is called a \_\_\_\_\_ of \_\_\_\_\_ .
- The \_\_\_\_\_ of  $n$  numbers is the sum of the numbers divided by  $n$ .
- If there is an even number of data values in a data set, then the \_\_\_\_\_ is the average of the two middle numbers.
- If two numbers of a data set are tied for the most frequent occurrence, the collection has two \_\_\_\_\_ and is called \_\_\_\_\_.
- Two measures of dispersion associated with the mean are called the \_\_\_\_\_ and the \_\_\_\_\_ of a data set.
- \_\_\_\_\_ measure dispersion associated with the median.
- You can represent quartiles graphically by creating a \_\_\_\_\_ .

In Exercises 1–6, find the mean, median, and mode of the set of measurements.

1. 5, 12, 7, 14, 8, 9, 7    2. 30, 37, 32, 39, 33, 34, 32  
 3. 5, 12, 7, 24, 8, 9, 7    4. 20, 37, 32, 39, 33, 34, 32  
 5. 5, 12, 7, 14, 9, 7    6. 30, 37, 32, 39, 34, 32

7. **Reasoning** Compare your answers for Exercises 1 and 3 with those for Exercises 2 and 4. Which of the measures of central tendency is sensitive to extreme measurements? Explain your reasoning.

8. **Reasoning**

- (a) Add 6 to each measurement in Exercise 1 and calculate the mean, median, and mode of the revised measurements. How are the measures of central tendency changed?  
 (b) If a constant  $k$  is added to each measurement in a set of data, how will the measures of central tendency change?

9. **Electric Bills** A person had the following monthly bills for electricity. What are the mean and median of the collection of bills?

January	\$67.92	February	\$59.84
March	\$52.00	April	\$52.50
May	\$57.99	June	\$65.35
July	\$81.76	August	\$74.98
September	\$87.82	October	\$83.18
November	\$65.35	December	\$57.00

10. **Car Rental** A car rental company kept the following record of the numbers of miles a rental car was driven. What are the mean, median, and mode of the data?

Monday	410	Tuesday	260
Wednesday	320	Thursday	320
Friday	460	Saturday	150

11. **Families** A study was done on families having six children. The table shows the numbers of families in the study with the indicated numbers of girls. Determine the mean, median, and mode of this set of data.

Number of girls	0	1	2	3	4	5	6
Frequency	1	24	45	54	50	19	7

12. **Sports** A baseball fan examined the records of a favorite baseball player's performance during his last 50 games. The numbers of games in which the player had 0, 1, 2, 3, and 4 hits are recorded in the table.

Number of hits	0	1	2	3	4
Frequency	14	26	7	2	1

- (a) Determine the average number of hits per game.  
 (b) Determine the player's batting average if he had 200 at-bats during the 50-game series.

13. **Think About It** Construct a collection of numbers that has the following properties. If this is not possible, explain why it is not.

Mean = 6, median = 4, mode = 4

14. **Think About It** Construct a collection of numbers that has the following properties. If this is not possible, explain why it is not.

Mean = 6, median = 6, mode = 4

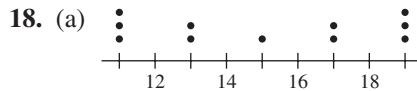
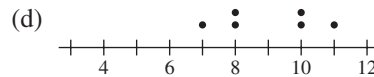
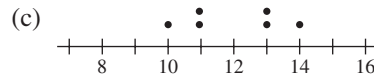
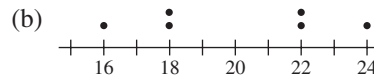
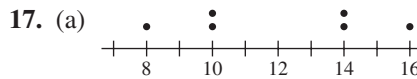
15. **Test Scores** A professor records the following scores for a 100-point exam.

- 99, 64, 80, 77, 59, 72, 87, 79, 92, 88,  
 90, 42, 20, 89, 42, 100, 98, 84, 78, 91

Which measure of central tendency best describes these test scores?

16. **Shoe Sales** A salesman sold eight pairs of men's black dress shoes. The sizes of the eight pairs were as follows:  $10\frac{1}{2}$ , 8, 12,  $10\frac{1}{2}$ , 10,  $9\frac{1}{2}$ , 11, and  $10\frac{1}{2}$ . Which measure (or measures) of central tendency best describes the typical shoe size for the data?

In Exercises 17 and 18, line plots of sets of data are given. Determine the mean and standard deviation of each set.




In Exercises 19–26, find the mean ( $\bar{x}$ ), variance ( $v$ ), and standard deviation ( $\sigma$ ) of the data set.

19. 4, 10, 8, 2                      20. 3, 15, 6, 9, 2  
 21. 0, 1, 1, 2, 2, 2, 3, 3, 4      22. 2, 2, 2, 2, 2, 2  
 23. 1, 2, 3, 4, 5, 6, 7            24. 1, 1, 1, 5, 5, 5  
 25. 49, 62, 40, 29, 32, 70      26. 1.5, 0.4, 2.1, 0.7, 0.8

In Exercises 27–32, use the alternative formula to find the standard deviation of the data set.

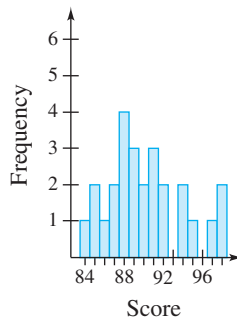
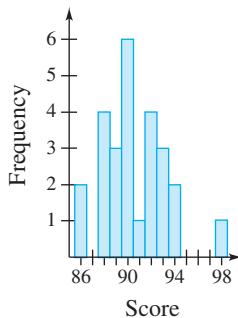
27. 2, 4, 6, 6, 13, 5  
 28. 10, 25, 50, 26, 15, 33, 29, 4  
 29. 246, 336, 473, 167, 219, 359  
 30. 6.0, 9.1, 4.4, 8.7, 10.4  
 31. 8.1, 6.9, 3.7, 4.2, 6.1  
 32. 9.0, 7.5, 3.3, 7.4, 6.0

33. **Reasoning** Without calculating the standard deviation, explain why the data set {4, 4, 20, 20} has a standard deviation of 8.  
 34. **Reasoning** If the standard deviation of a data set of numbers is 0, what does this imply about the set?  
 35. **Test Scores** An instructor adds five points to each student’s exam score. Will this change the mean or standard deviation of the exam scores? Explain.

-  36. **Price of Gold** The following data represents the average prices of gold (in dollars per fine ounce) for the years 1984 to 2003. Use a computer or graphing utility to find the mean, variance, and standard deviation of the data. What percent of the data lies within two standard deviations of the mean? (Source: U.S. Bureau of Mines and U.S. Geological Survey)

361,	318,	368,	478,	438,
383,	385,	363,	345,	361,
385,	386,	389,	332,	295,
280,	280,	272,	311,	350


37. **Think About It** The histograms represent the test scores of two classes of a college course in mathematics. Which histogram has the smaller standard deviation?



38. **Test Scores** The scores of a mathematics exam given to 600 science and engineering students at a college had a mean and standard deviation of 235 and 28, respectively. Use Chebychev’s Theorem to determine the intervals containing at least  $\frac{3}{4}$  and at least  $\frac{8}{9}$  of the scores. How would the intervals change if the standard deviation were 16?

In Exercises 39–42, (a) find the lower and upper quartiles of the data and (b) sketch a box-and-whisker plot for the data without the aid of a graphing utility.

39. 23, 15, 14, 23, 13, 14, 13, 20, 12  
 40. 11, 10, 11, 14, 17, 16, 14, 11, 8, 14, 20  
 41. 46, 48, 48, 50, 52, 47, 51, 47, 49, 53  
 42. 25, 20, 22, 28, 24, 28, 25, 19, 27, 29, 28, 21

 In Exercises 43–46, use a graphing utility to create a box-and-whisker plot for the data.

43. 19, 12, 14, 9, 14, 15, 17, 13, 19, 11, 10, 19  
 44. 9, 5, 5, 5, 6, 5, 4, 12, 7, 10, 7, 11, 8, 9, 9  
 45. 20.1, 43.4, 34.9, 23.9, 33.5, 24.1, 22.5, 42.4, 25.7, 17.4, 23.8, 33.3, 17.3, 36.4, 21.8  
 46. 78.4, 76.3, 107.5, 78.5, 93.2, 90.3, 77.8, 37.1, 97.1, 75.5, 58.8, 65.6

47. **Product Lifetime** A company has redesigned a product in an attempt to increase the lifetime of the product. The two sets of data list the lifetimes (in months) of 20 units with the original design and 20 units with the new design. Create a box-and-whisker plot for each set of data, and then comment on the differences between the plots.

*Original Design*

15.1	78.3	56.3	68.9	30.6
27.2	12.5	42.7	72.7	20.2
53.0	13.5	11.0	18.4	85.2
10.8	38.3	85.1	10.0	12.6

*New Design*

55.8	71.5	25.6	19.0	23.1
37.2	60.0	35.3	18.9	80.5
46.7	31.1	67.9	23.5	99.5
54.0	23.2	45.5	24.8	87.8