2.6 Measures of Spread

The measures of central tendency indicate the central values of a set of data. Often, you will also want to know how closely the data cluster around these centres.

INVESTIGATE & INQUIRE: Spread in a Set of Data

For a game of basketball, a group of friends split into two randomly chosen teams. The heights of the players are shown in the table below.

Falcons		Ravens	
Player	Height (cm)	Player	Height (cm)
Laura	183	Sam	166
Jamie	165	Shannon	163
Deepa	148	Tracy	168
Colleen	146	Claudette	161
Ingrid	181	Maria	165
Justiss	178	Amy	166
Sheila	154	Selena	166

1. Judging by the raw data in this table, which team do you think has a height advantage? Explain why.
2. Do the measures of central tendency confirm that the teams are mismatched? Why or why not?
3. Explain how the distributions of heights on the two teams might give one of them an advantage. How could you use a diagram to illustrate the key difference between the two teams?

The **measures of spread** or **dispersion** of a data set are quantities that indicate how closely a set of data clusters around its centre. Just as there are several measures of central tendency, there are also different measures of spread.

Standard Deviation and Variance

A **deviation** is the difference between an individual value in a set of data and the mean for the data.

For a population, deviation $= x - \mu$

For a sample, deviation $= x - \bar{x}$

The larger the size of the deviations, the greater the spread in the data. Values less than the mean have negative deviations. If you simply add up all the deviations for a data set, they will cancel out. You could use the sum of the absolute values of the deviations as a measure of spread. However, statisticians have shown that a root-mean-square quantity is a more useful measure of spread. The **standard deviation** is the square root of the mean of the squares of the deviations.

The lowercase Greek letter sigma, σ, is the symbol for the standard deviation of a population, while the letter s stands for the standard deviation of a sample.

Population standard deviation

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

Sample standard deviation

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

where N is the number of data in the population and n is the number in the sample.

Note that the formula for s has $n - 1$ in the denominator instead of n. This denominator compensates for the fact that a sample taken from a population tends to underestimate the deviations in the population. Remember that the sample mean, $\bar{x}$, is not necessarily equal to the population mean, μ. Since $\bar{x}$ is the central value of the sample, the sample data cluster closer to $\bar{x}$ than to μ. When n is large, the formula for s approaches that for σ.

Also note that the standard deviation gives greater weight to the larger deviations since it is based on the *squares* of the deviations.

The mean of the squares of the deviations is another useful measure. This quantity is called the **variance** and is equal to the square of the standard deviation.

Population variance

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

Sample variance

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Example 1 Using a Formula to Calculate Standard Deviations

Use means and standard deviations to compare the distribution of heights for the two basketball teams listed in the table on page 136.

Solution

Since you are considering the teams as two separate populations, use the mean and standard deviation formulas for populations. First, calculate the mean height for the Falcons.

$$\mu = \frac{\Sigma x}{N}$$
$$= \frac{1155}{7}$$
$$= 165$$

Next, calculate all the deviations and their squares.

Falcons	Height (cm)	Deviation, $x - \mu$	$(x - \mu)^2$
Laura	183	18	324
Jamie	165	0	0
Deepa	148	−17	289
Colleen	146	−19	361
Ingrid	181	16	256
Justiss	178	13	169
Sheila	154	−11	121
Sum	**1155**	**0**	**1520**

Now, you can determine the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$$
$$= \sqrt{\frac{1520}{7}}$$
$$= 14.7$$

Therefore, the Falcons have a mean height of 165 cm with a standard deviation of 14.7 cm.

Similarly, you can determine that the Ravens also have a mean height of 165 cm, but their standard deviation is only 2.1 cm. Clearly, the Falcons have a much greater spread in height than the Ravens. Since the two teams have the same mean height, the difference in the standard deviations indicates that the Falcons have some players who are taller than any of the Ravens, but also some players who are shorter.

If you were to consider either of the basketball teams in the example above as a sample of the whole group of players, you would use the formula for s to calculate the team's standard deviation. In this case, you would be using the sample to estimate the characteristics of a larger population. However, the teams are very small samples, so they could have significant random variations, as the difference in their standard deviations demonstrates.

For large samples the calculation of standard deviation can be quite tedious. However, most business and scientific calculators have built-in functions for such calculations, as do spreadsheets and statistical software.

See Appendix B for more detailed information about technology functions and keystrokes.

Example 2 Using Technology to Calculate Standard Deviations

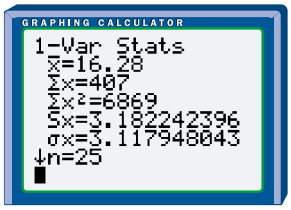
A veterinarian has collected data on the life spans of a rare breed of cats.

Life Spans (in years)													
16	18	19	12	11	15	20	21	18	15	16	13	16	22
18	19	17	14	9	14	15	19	20	15	15			

Determine the mean, standard deviation, and the variance for these data.

Solution 1 Using a Graphing Calculator

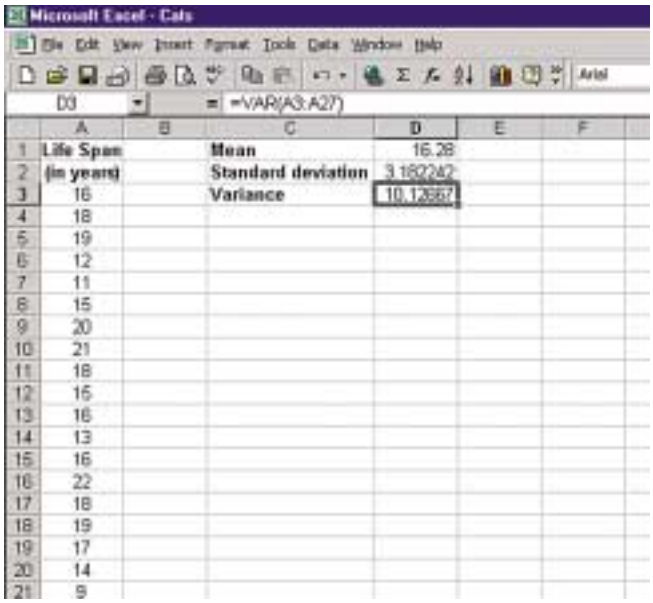
Use the ClrList command to make sure list L1 is clear, then enter the data into it. Use the 1-Var Stats command from the STAT CALC menu to calculate a set of statistics including the mean and the standard deviation. Note that the calculator displays both a sample standard deviation, Sx, and a population standard deviation, σx. Use Sx since you are dealing with a sample in this case. Find the variance by calculating the square of Sx.



The mean life span for this breed of cats is about 16.3 years with a standard deviation of 3.2 years and a variance of 10.1. Note that variances are usually stated without units. The units for this variance are years squared.

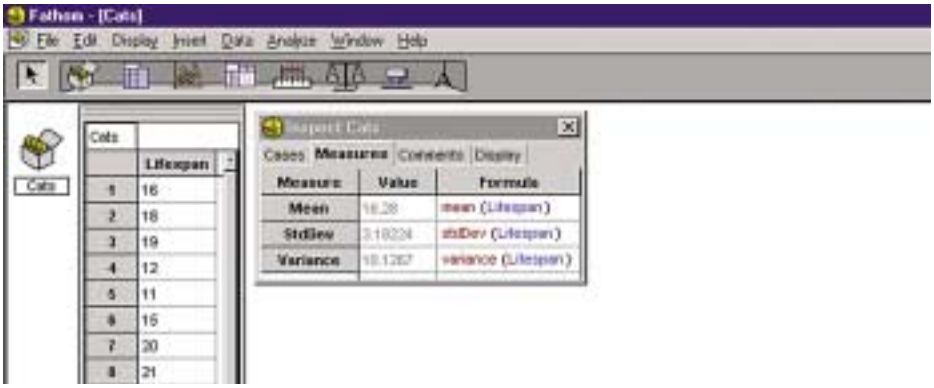
Solution 2 Using a Spreadsheet

Enter the data into your spreadsheet program. With Corel® Quattro® Pro, you can use the AVG, STDS, and VARS functions to calculate the mean, sample standard deviation, and sample variance. In Microsoft® Excel, the equivalent functions are AVERAGE, STDEV, and VAR.



Solution 3 Using Fathom™

Drag a new **case table** onto the workspace, name the attribute for the first column Lifespan, and enter the data. Right-click to open the **inspector**, and click the Measures tab. Create Mean, StdDev, and Variance measures and select the formulas for the **mean**, **standard deviation**, and **variance** from the Edit Formula/Functions/Statistical/One Attribute menu.



If you are working with grouped data, you can estimate the standard deviation using the following formulas.

For a population,

$$\sigma \doteq \sqrt{\frac{\Sigma f_i (m_i - \mu)^2}{N}}$$

For a sample,

$$s \doteq \sqrt{\frac{\Sigma f_i (m_i - \bar{x})^2}{n - 1}}$$

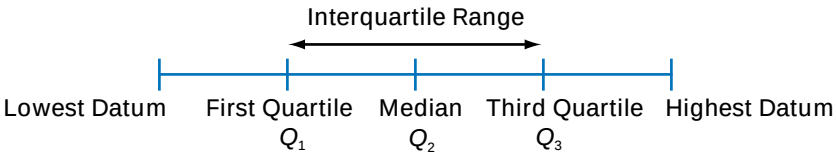
where f_i is the frequency for a given interval and m_i is the midpoint of the interval. However, calculating standard deviations from raw, ungrouped data will give more accurate results.

Project Prep

In your statistics project, you may wish to use an appropriate measure of spread to describe the distribution of your data.

Quartiles and Interquartile Ranges

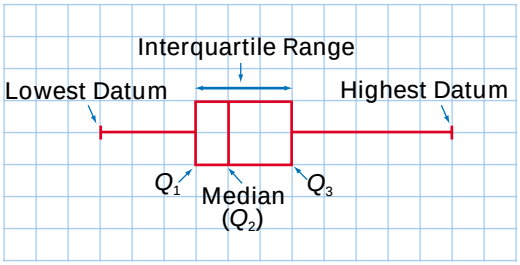
Quartiles divide a set of ordered data into four groups with equal numbers of values, just as the median divides data into two equally sized groups. The three "dividing points" are the first quartile (Q_1), the median (sometimes called the second quartile or Q_2), and the third quartile (Q_3). Q_1 and Q_3 are the medians of the lower and upper halves of the data.



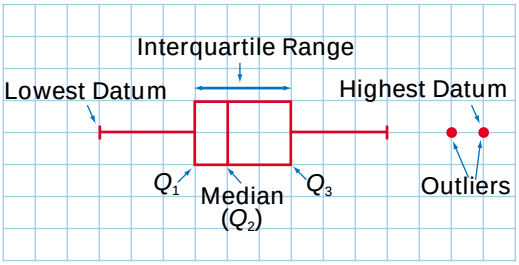
Recall that when there are an even number of data, you take the midpoint between the two middle values as the median. If the number of data below the median is even, Q_1 is the midpoint between the two middle values in this half of the data. Q_3 is determined in a similar way.

The **interquartile range** is $Q_3 - Q_1$, which is the range of the middle half of the data. The larger the interquartile range, the larger the spread of the central half of the data. Thus, the interquartile range provides a measure of spread. The **semi-interquartile range** is one half of the interquartile range. Both these ranges indicate how closely the data are clustered around the median.

A **box-and-whisker plot** of the data illustrates these measures. The box shows the first quartile, the median, and the third quartile. The ends of the "whiskers" represent the lowest and highest values in the set of data. Thus, the length of the box shows the interquartile range, while the left whisker shows the range of the data below the first quartile, and the right whisker shows the range above the third quartile.



A **modified box-and-whisker plot** is often used when the data contain outliers. By convention, any point that is at least 1.5 times the box length away from the box is classified as an outlier. A modified box-and-whisker plot shows such outliers as separate points instead of including them in the whiskers. This method usually gives a clearer illustration of the distribution.



Example 3 Determining Quartiles and Interquartile Ranges

A random survey of people at a science-fiction convention asked them how many times they had seen *Star Wars*. The results are shown below.

3	4	2	8	10	5	1	15	5	16	6	3	4	9	12	3	30	2	10	7

a) Determine the median, the first and third quartiles, and the interquartile and semi-interquartile ranges. What information do these measures provide?

b) Prepare a suitable box plot of the data.

c) Compare the results in part a) to those from last year's survey, which found a median of 5.1 with an interquartile range of 8.0.

Solution 1 Using Pencil and Paper

a) First, put the data into numerical order.

1	2	2	3	3	3	4	4	5	5	6	7	8	9	10	10	12	15	16	30

The median is either the middle datum or, as in this case, the mean of the two middle data:

$$\text{median} = \frac{5 + 6}{2}$$
$$= 5.5$$

The median value of 5.5 indicates that half of the people surveyed had seen *Star Wars* less than 5.5 times and the other half had seen it more than 5.5 times.

To determine Q_1, find the median of the lower half of the data. Again, there are two middle values, both of which are 3. Therefore, $Q_1 = 3$.

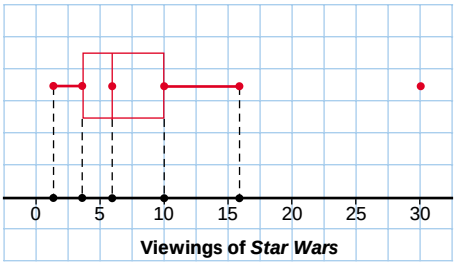
Similarly, the two middle values of the upper half of the data are both 10, so $Q_3 = 10$.

Since Q_1 and Q_3 are the boundaries for the central half of the data, they show that half of the people surveyed have seen *Star Wars* between 3 and 10 times.

$$Q_3 - Q_1 = 10 - 3$$
$$= 7$$

Therefore, the interquartile range is 7. The semi-interquartile range is half this value, or 3.5. These ranges indicate the spread of the central half of the data.

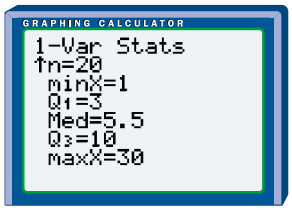
b) The value of 30 at the end of the ordered data is clearly an outlier. Therefore, a modified box-and-whisker plot will best illustrate this set of data.



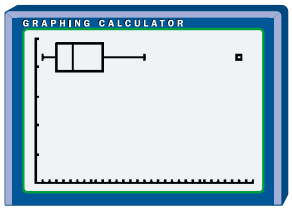
c) Comparing the two surveys shows that the median number of viewings is higher this year and the data are somewhat less spread out.

Solution 2 Using a Graphing Calculator

a) Use the STAT EDIT menu to enter the data into a list. Use the **1-Var Stats command** from the CALC EDIT menu to calculate the statistics for your list. Scroll down to see the values for the median, Q_1, and Q_3. Use the values for Q_1 and Q_3 to calculate the interquartile and semi-interquartile ranges.



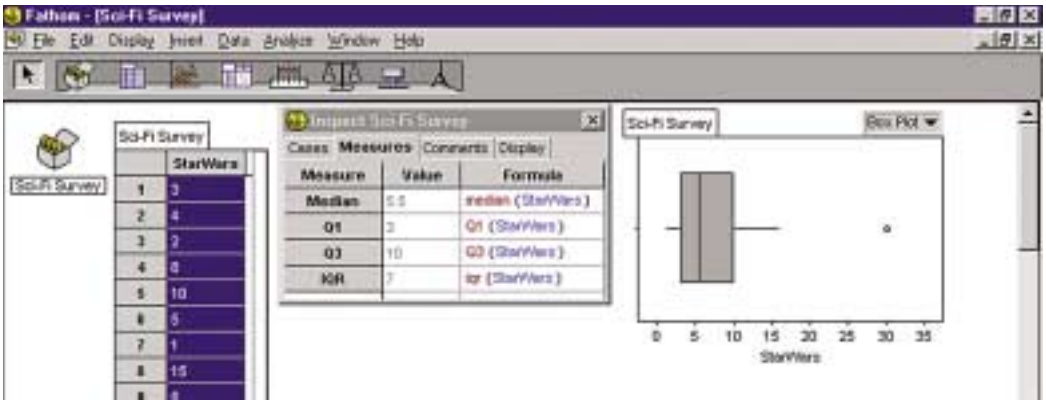
b) Use **STAT PLOT** to select a modified box plot of your list. Press GRAPH to display the box-and-whisker plot and adjust the **window settings**, if necessary.



Solution 3 Using Fathom™

a) Drag a new **case table** onto the workspace, create an attribute called StarWars, and enter your data. Open the **inspector** and create Median, Q1, Q3, and IQR measures. Use the Edit Formula/Functions/Statistical/One Attribute menu to enter the formulas for the **median**, **quartiles**, and **interquartile range**.

b) Drag the **graph icon** onto the workspace, then drop the StarWars attribute on the x-axis of the graph. Select Box Plot from the drop-down menu in the upper right corner of the graph.



Although a quartile is, strictly speaking, a single value, people sometimes speak of a datum being *within* a quartile. What they really mean is that the datum is in the quarter whose upper boundary is the quartile. For example, if a value x_1 is "within the first quartile," then $x_1 \leq Q_1$. Similarly, if x_2 is "within the third quartile," then the median $\leq x_2 \leq Q_3$.

Example 4 Classifying Data by Quartiles

In a survey of low-risk mutual funds, the median annual yield was 7.2%, while Q_1 was 5.9% and Q_3 was 8.3%. Describe the following funds in terms of quartiles.

Mutual Fund	Annual Yield (%)
XXY Value	7.5
YYZ Dividend	9.0
ZZZ Bond	7.2

Solution

The yield for the XXY Value fund was between the median and Q_3. You might see this fund described as being in the third quartile or having a third-quartile yield.

YYZ Dividend's yield was above Q_3. This fund might be termed a fourth- or top-quartile fund.

ZZZ Bond's yield was equal to the median. This fund could be described as a median fund or as having median performance.

Percentiles

Percentiles are similar to quartiles, except that percentiles divide the data into 100 intervals that have equal numbers of values. Thus, k percent of the data are less than or equal to kth percentile, P_k, and $(100 - k)$ percent are greater than or equal to P_k. Standardized tests often use percentiles to convert raw scores to scores on a scale from 1 to 100. As with quartiles, people sometimes use the term *percentile* to refer to the intervals rather than their boundaries.

Example 5 Percentiles

An audio magazine tested 60 different models of speakers and gave each one an overall rating based on sound quality, reliability, efficiency, and appearance. The raw scores for the speakers are listed in ascending order below.

35	47	57	62	64	67	72	76	83	90
38	50	58	62	65	68	72	78	84	91
41	51	58	62	65	68	73	79	86	92
44	53	59	63	66	69	74	81	86	94
45	53	60	63	67	69	75	82	87	96
45	56	62	64	67	70	75	82	88	98

a) If the Audio Maximizer Ultra 3000 scored at the 50th percentile, what was its raw score?

b) What is the 90th percentile for these data?

c) Does the SchmederVox's score of 75 place it at the 75th percentile?

Solution

a) Half of the raw scores are less than or equal to the 50th percentile and half are greater than or equal to it. From the table, you can see that 67 divides the data in this way. Therefore, the Audio Maximizer Ultra 3000 had a raw score of 67.

b) The 90th percentile is the boundary between the lower 90% of the scores and the top 10%. In the table, you can see that the top 10% of the scores are in the 10th column. Therefore, the 90th percentile is the midpoint between values of 88 and 90, which is 89.

c) First, determine 75% of the number of raw scores.

$60 \times 75\% = 45$

There are 45 scores less than or equal to the 75th percentile. Therefore, the 75th percentile is the midpoint between the 45th and 46th scores. These two scores are 79 and 81, so the 75th percentile is 80. The SchmederVox's score of 75 is below the 75th percentile.

Z-Scores

A ***z*-score** is the number of standard deviations that a datum is from the mean. You calculate the z-score by dividing the deviation of a datum by the standard deviation.

For a population,	For a sample,
$z = \frac{x - \mu}{\sigma}$	$z = \frac{x - \bar{x}}{s}$

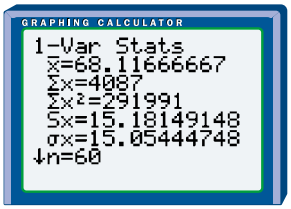
Variable values below the mean have negative z-scores, values above the mean have positive z-scores, and values equal to the mean have a zero z-score. Chapter 8 describes z-scores in more detail.

Example 6 Determining *Z*-Scores

Determine the z-scores for the Audio Maximizer Ultra 3000 and SchmederVox speakers.

Solution

You can use a calculator, spreadsheet, or statistical software to determine that the mean is 68.1 and the standard deviation is 15.2 for the speaker scores in Example 4.



Now, use the mean and standard deviation to calculate the z-scores for the two speakers.

For the Audio Maximizer Ultra 3000,

$$z = \frac{x - \bar{x}}{s}$$
$$= \frac{67 - 68.1}{15.2}$$
$$= -0.072$$

For the SchmederVox,

$$z = \frac{x - \bar{x}}{s}$$
$$= \frac{75 - 68.1}{15.2}$$
$$= 0.46$$

The Audio Maximizer Ultra 3000 has a z-score of -0.072, indicating that it is approximately 7% of a standard deviation below the mean. The SchmederVox speaker has a z-score of 0.46, indicating that it is approximately half a standard deviation above the mean.

Key Concepts

- The variance and the standard deviation are measures of how closely a set of data clusters around its mean. The variance and standard deviation of a sample may differ from those of the population the sample is drawn from.
- Quartiles are values that divide a set of ordered data into four intervals with equal numbers of data, while percentiles divide the data into 100 intervals.
- The interquartile range and semi-interquartile range are measures of how closely a set of data clusters around its median.
- The z-score of a datum is a measure of how many standard deviations the value is from the mean.

Communicate Your Understanding

1. Explain how the term *root-mean-square* applies to the calculation of the standard deviation.
2. Why does the semi-interquartile range give only an approximate measure of how far the first and third quartiles are from the median?
3. Describe the similarities and differences between the standard deviation and the semi-interquartile range.
4. Are the median, the second quartile, and the 50th percentile always equal? Explain why or why not.

Practise

A

1. Determine the mean, standard deviation, and variance for the following samples.

a) Scores on a data management quiz (out of 10 with a bonus question):

5	7	9	6	5	10	8	2
11	8	7	7	6	9	5	8

b) Costs for books purchased including taxes (in dollars):

12.55	15.31	21.98	45.35	19.81
33.89	29.53	30.19	38.20	

2. Determine the median, Q_1, Q_3, the interquartile range, and semi-interquartile range for the following sets of data.

a) Number of home runs hit by players on the Statsville little league team:

6	4	3	8	9	11	6	5	15

b) Final grades in a geography class:

88	56	72	67	59	48	81	62
90	75	75	43	71	64	78	84

3. For a recent standardized test, the median was 88, Q_1 was 67, and Q_3 was 105. Describe the following scores in terms of quartiles.

a) 8

b) 81

c) 103

4. What percentile corresponds to

a) the first quartile?

b) the median?

c) the third quartile?

5. Convert these raw scores to z-scores.

18	15	26	20	21

Apply, Solve, Communicate

B

6. The board members of a provincial organization receive a car allowance for travel to meetings. Here are the distances the board logged last year (in kilometres).

44	18	125	80	63	42	35	68	52
75	260	96	110	72	51			

a) Determine the mean, standard deviation, and variance for these data.

b) Determine the median, interquartile range, and semi-interquartile range.

c) Illustrate these data using a box-and-whisker plot.

d) Identify any outliers.

7. The nurses' union collects data on the hours worked by operating-room nurses at the Statsville General Hospital.

Hours Per Week	Number of Employees
12	1
32	5
35	7
38	8
42	5

a) Determine the mean, variance, and standard deviation for the nurses' hours.

b) Determine the median, interquartile range, and semi-interquartile range.

c) Illustrate these data using a box-and-whisker plot.

8. Application

a) Predict the changes in the standard deviation and the box-and-whisker plot if the outlier were removed from the data in question 7.

b) Remove the outlier and compare the new results to your original results.

c) Account for any differences between your prediction and your results in part b).

9. **Application** Here are the current salaries for François' team.

Chapter Problem

Salary ($)	Number of Players
300 000	2
500 000	3
750 000	8
900 000	6
1 000 000	2
1 500 000	1
3 000 000	1
4 000 000	1

a) Determine the standard deviation, variance, interquartile range, and semi-interquartile range for these data.

b) Illustrate the data with a modified box-and-whisker plot.

c) Determine the z-score of François' current salary of $300 000.

d) What will the new z-score be if François' agent does get him a million-dollar contract?

10. **Communication** Carol's golf drives have a mean of 185 m with a standard deviation of 25 m, while her friend Chi-Yan shoots a mean distance of 170 m with a standard deviation of 10 m. Explain which of the two friends is likely to have a better score in a round of golf. What assumptions do you have to make for your answer?

11. Under what conditions will Q_1 equal one of the data points in a distribution?

12. **a)** Construct a set of data in which $Q_1 = Q_3$ and describe a situation in which this equality might occur.

b) Will such data sets always have a median equal to Q_1 and Q_3? Explain your reasoning.

13. Is it possible for a set of data to have a standard deviation much smaller than its semi-interquartile range? Give an example or explain why one is not possible.

14. **Inquiry / Problem Solving** A business-travellers' association rates hotels on a variety of factors including price, cleanliness, services, and amenities to produce an overall score out of 100 for each hotel. Here are the ratings for 50 hotels in a major city.

39	50	56	60	65	68	73	77	81	87
41	50	56	60	65	68	74	78	81	89
42	51	57	60	66	70	74	78	84	91
44	53	58	62	67	71	75	79	85	94
48	55	59	63	68	73	76	80	86	96

a) What score represents

i) the 50th percentile?

ii) the 95th percentile?

b) What percentile corresponds to a rating of 50?

c) The travellers' association lists hotels above the 90th percentile as "highly recommended" and hotels between the 75th and 90th percentiles as "recommended." What are the minimum scores for the two levels of recommended hotels?

ACHIEVEMENT CHECK

Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application

15. **a)** A data-management teacher has two classes whose midterm marks have identical means. However, the standard deviations for each class are significantly different. Describe what these measures tell you about the two classes.

b) If two sets of data have the same mean, can one of them have a larger standard deviation and a smaller interquartile range than the other? Give an example or explain why one is not possible.

C

16. Show that $\sum(x - \bar{x}) = 0$ for any distribution.

17. **a)** Show that $s = \sqrt{\dfrac{n(\sum x^2) - (\sum x)^2}{n(n-1)}}$.

(Hint: Use the fact that $\sum x = n\bar{x}$.)

b) What are two advantages of using the formula in part a) for calculating standard deviations?

18. Communication The **midrange** of a set of data is defined as half of the sum of the highest value and the lowest value. The incomes for the employees of Statsville Lawn Ornaments Limited are listed below (in thousands of dollars).

28	34	49	22	50	31	55	32	73	21
63	112	35	19	44	28	59	85	47	39

a) Determine the midrange and interquartile range for these data.

b) What are the similarities and differences between these two measures of spread?

19. The **mean absolute deviation** of a set of data is defined as $\dfrac{\sum|x - \bar{x}|}{n}$, where $|x - \bar{x}|$ is the absolute value of the difference between each data point and the mean.

a) Calculate the mean absolute deviation and the standard deviation for the data in question 18.

b) What are the similarities and differences between these two measures of spread?

Career Connection

Statistician

Use of statistics today is so widespread that there are numerous career opportunities for statisticians in a broad range of fields. Governments, medical-research laboratories, sports agencies, financial groups, and universities are just a few of the many organizations that employ statisticians. Current trends suggest an ongoing need for statisticians in many areas.

A statistician is engaged in the collection, analysis, presentation, and interpretation of data in a variety of forms. Statisticians provide insight into which data are likely to be reliable and whether valid conclusions or predictions can be drawn from them. A research statistician might develop new statistical techniques or applications.

Because computers are essential for analysing large amounts of data, a statistician should possess a strong background in computers as well as mathematics. Many positions call for a minimum of a bachelor's or master's degree. Research at a university or work for a consulting firm usually requires a doctorate.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

For more information about a career as a statistician and other careers related to mathematics, visit the above web site and follow the links.