3.2 Measures of Central Tendency

The purpose of this section is to enable you to describe a set of numeric data using a single value. The value you calculate will describe the *centre* of the set of data.

There are various ways to describe this notion of centre. For example, a car dealer might claim that the average selling price of a two-year-old used car is $14 500. What does this average tell you? Does this average take into account the number of each type of car sold? What if the dealer doesn't sell very many used cars, but occasionally sells a very expensive car? The influence an **outlier** such as this can have must be taken into consideration.

outlier—an element of a data set that is very different from the others

It is important to know what kind of average is being used. You are already familiar with the mean, the median, and the mode. The examples that follow outline situations in which each measure of central tendency is most useful.

mean—a measure of central tendency found by dividing the sum of all the data by the number of pieces of data

THE MEAN

The formula for the mean of a set of values, x_i, is given as

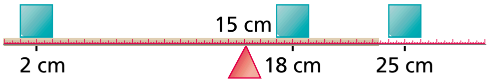
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{or} \quad \bar{x} = \frac{\sum x}{n}$$

The Greek letter $\sum$ (sigma) indicates that all the individual values of x_i from $i = 1$ to n are added together. The sum is divided by the number of values in the set, n.

The mean of the set of values {5, 7, 9, 10, 50} is calculated as

$$\frac{5 + 7 + 9 + 10 + 50}{5} = 16.2$$

In what way does this value describe the centre of the set of values? The value reported by this calculation may be better understood by examining the following situation involving weights and balances.



If three identical blocks were placed on a ruler as shown—one at the 2-cm mark, one at the 18-cm mark, and one at the 25-cm mark—where should the triangular fulcrum be placed so that the ruler and blocks balance?

The location of the fulcrum is called the *centre of gravity* and is the same as the arithmetic mean; in this case, 15 cm. At this location, the block on the left is 13 cm from the fulcrum and the two blocks on the right are 3 cm and 10 cm from the fulcrum, respectively. The sum of the distances from the fulcrum is the

deviation—the difference between a data value and the mean

same on the left and right side. In statistics, this distance is called the **deviation** from the mean. If you consider distances to the left of the fulcrum as negative, then the mean is the value that makes the sum of the deviations from the mean equal to zero.

WEIGHTED MEAN

Suppose that the individual blocks from the previous situation have different masses: the block at 2 cm is 10 g, the block at 18 cm is 15 g, and the block at 25 cm is 12 g. Where would you place the fulcrum now?

The simplest way to approach this is to imagine that you have a total of 37 identical 1-g blocks: 10 stacked at 2 cm, 15 stacked at 18 cm, and 12 stacked at 25 cm. To find the average, multiply the number of blocks by the distance to determine the sum, and then divide by the total number of blocks.

Distance	Number	Distance × Number
2	10	20
18	15	270
25	12	300
Total	37	590

The weighted mean is calculated as $\bar{x} = \frac{590}{37} \doteq 15.9$.

Therefore, you would place the fulcrum at 15.9 cm.

In general, the weighted mean can be calculated as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} \quad \text{or} \quad \frac{\sum xw}{w}$$

where x_i represents each data value and w_i represents its weight, or frequency.

Example 1 Calculating Weighted Means

A teacher weights student marks in her final calculation as follows: Knowledge and Understanding, 25%; Application, 15%; Problem Solving, 20%; Communication, 10%; and the Final Exam, 30%. A student's marks in these categories are 78, 75, 80, and 85, respectively. The final exam, has not, as yet, been written.

(a) Calculate the student's term mark before the final exam.

(b) What mark must the student achieve on the final exam to earn a final grade of 82?

Solution 1 *Without technology*

(a) Using the formula

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

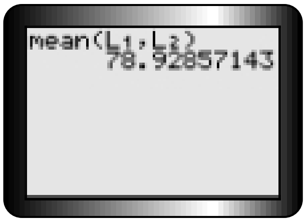
the term mark is calculated as

$$t = \frac{78 \times (0.25) + 75 \times (0.15) + 80 \times (0.20) + 85 \times (0.10)}{(0.25 + 0.15 + 0.20 + 0.10)}$$

$$t \doteq 78.9$$

Solution 2 *Using a TI-83 Plus calculator*

(a) First, enter the four term values in $\mathbf{L_1}$ and the weights in $\mathbf{L_2}$. Next, return to the home screen and press 2nd STAT > > 3 to retrieve the mean function. Enter the list names by pressing 2nd 1 and 2nd 2 separated by a comma, and then press ENTER.



```
mean(L1,L2)
            78.92857143
```

(b) The term mark of 78.9 is worth a total of 70% and the final exam is worth 30%. The final exam mark, E, may be calculated algebraically as

$$0.70(78.9) + 0.30E = 82$$

$$E = \frac{82 - 0.70(78.9)}{0.3}$$

$$E \doteq 89.2$$

USING GROUPED DATA

Suppose your data have already been organized into a frequency table with a class interval not equal to 1. You no longer have actual data values, so you must then use the midpoint of each class to estimate a mean weighted by the frequency.

Example 2 Finding the Mean for Grouped Data

A sample of car owners was asked how old they were when they got their first car. The results were then reported in a frequency distribution. Calculate the mean.

Age	16–20	21–25	26–30	31–35	36–40
Frequency	10	18	12	8	2

Solution

Finding the average of grouped data is the same as finding a weighted average; that is, using the interval midpoint as the data value.

Age	Frequency, f	Midpoint (Age), m	$f \times m$
16–20	10	18	$10 \times 18 = 180$
21–25	18	23	$18 \times 23 = 414$
26–30	12	28	$12 \times 28 = 336$
31–35	8	33	$8 \times 33 = 264$
36–40	2	38	$2 \times 38 = 76$

The mean can now be calculated as

$$\bar{x} = \frac{\sum(f \times m)}{\sum f}$$

$$\bar{x} = \frac{180 + 414 + 336 + 264 + 76}{10 + 18 + 12 + 8 + 2}$$

$$\bar{x} = 25.4$$

median—the middle value of an ordered data set

THE MEDIAN

The median value is the middle data point in an ordered set dividing the set into two sets of equal size. If the set has an even number of data points, then the median is halfway between the two middle-most values.

Think about Finding the Median

Does it matter whether the data are arranged in descending or ascending order?

Example 3 Finding the Median

Monthly rents downtown and in the suburbs are collected from the classified section of a newspaper. Calculate the median rent in each district.

Downtown: \$850, \$750, \$1225, \$1000, \$800, \$1100, \$3200
Suburbs: \$750, \$550, \$900, \$585, \$220, \$625, \$500, \$800

Solution

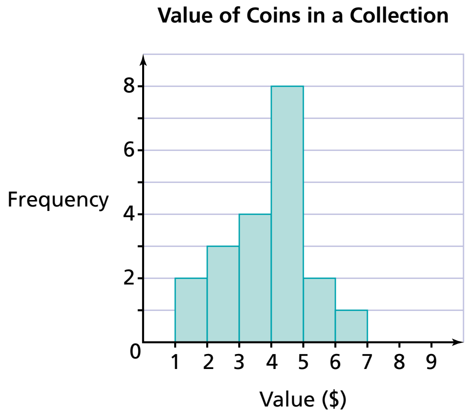
Downtown: The set 750, 800, 850, 1000, 1100, 1225, 3200 has seven elements, so the median is the 4th element. The median rent downtown is \$1000/month.

Suburbs: The ordered set {220, 500, 550, 585, 625, 750, 800, 900} has eight elements, so the median is halfway between the 4th and 5th elements; in this case, halfway between 585 and 625. The median rent in the suburbs is \$605/month.

mode—the most frequent value or interval

THE MODE

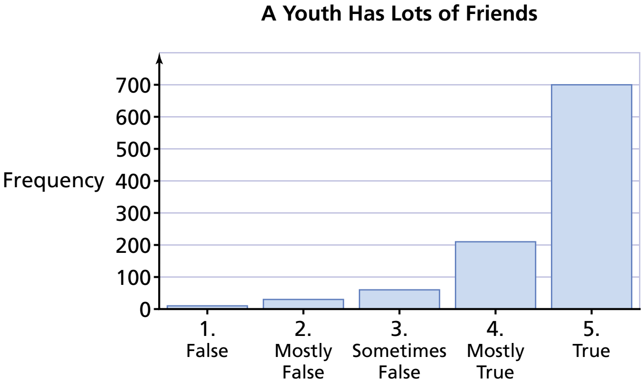
The mode is simply the most frequent value or range of values in a data set. It is easy to determine the mode from a histogram as it is the highest column. In the histogram shown here, the modal interval is \$4 to \$5.



If there are two or more measurements that occur most often, the data set is called bimodal (or trimodal or multimodal). If no measurement is repeated, the data set has no mode.

Example 4 Analyzing Qualitative Data

The graph to the right represents the results when Ontario youths were asked if they have a lot of friends. Which measure of central tendency can be used to represent these data?



Solution

The mode is the only appropriate measure of central tendency for qualitative data such as this. The modal interval is the one where the youths answered “true.” Since the mean and median depend on quantitative data that can be measured numerically, it is not meaningful to calculate a mean or median for these data.

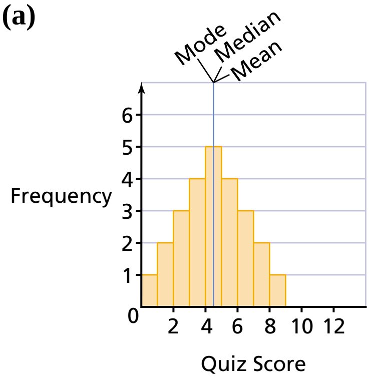
Example 5 Exploring Distributions and Central Tendency

Compare the following data sets. What is the relationship between the shape of the distribution and the mean, median, and mode?

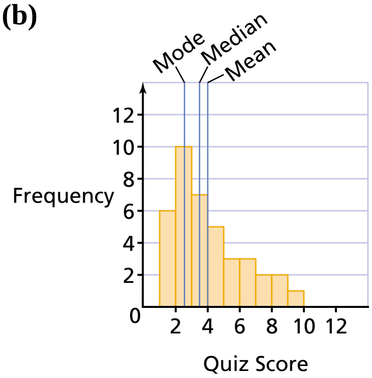
right skew—a distribution where the mean is skewed to the right (median < mean)

left skew—a distribution where the mean is skewed to the left (mean < median)

(a)



(b)



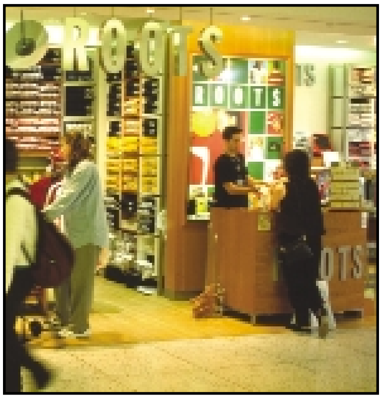
Solution

In part (a), the distribution is mound-shaped and symmetric. The mean, median, and mode are all equal. In part (b), the distribution is skewed right. Notice that mode < median < mean. When the distribution is skewed left, you would typically find that mean < median < mode. The outliers affect the mean more than they do the median and mode. In addition, notice that the median is typically between the mean and mode for non-symmetric distributions.

WHAT MEASURE SHOULD YOU USE?

While there can be no single rule governing which measure of central tendency you should use to describe a set of data, take the following into consideration:

- Outliers will affect the mean the most. If data contain outliers, use the median to avoid misrepresenting the data.
- If the data are strongly skewed, the median may best represent the central tendency of the data.
- If the data are roughly symmetric, the mean and the median will be close, so either is appropriate.
- If the data are not numeric (e.g., colour) or if the frequency of the data is more important than the value (e.g., shoe size), then the mode should be used.



Example 6 Measuring Central Tendency

Describe the central tendency of each of the following monthly incomes for six salespeople working on commission using the most appropriate measure.

(a) January: \$1241, \$1499, \$2020, \$1371, \$1622, \$1853

(b) February: \$1529, \$0, \$2127, \$1933, \$1686, \$1893

(c) March: \$1712, \$2540, \$1392. The remaining three salespeople received \$1000 in holiday pay.

Solution

(a) Mean $\frac{\$1241 + \$1499 + \$2020 + \$1371 + \$1622 + \$1853}{6} = \$1601$

Median $\frac{\$1499 + \$1622}{2} = \$1560.50$

Mode None as no repeated data

The values in this set are evenly distributed. Both the mean and median provide a good measure of the central tendency.

(b) Mean $\frac{\$1529 + \$0 + \$2127 + \$1933 + \$1686 + \$1893}{6} = \$1528$

Median $\frac{\$1686 + \$1893}{2} = \$1789.50$

Mode None as no repeated data

The one salesperson who earned nothing represents an outlier, which makes it appear as if most commissions are down in February. On the contrary, every other salesperson improved on the previous month's results. The median income figure represents a more accurate measure of central tendency.

(c) Mean $\frac{\$1712 + \$2540 + \$1392 + \$1000 + \$1000 + \$1000}{6} \doteq \$1440.67$

Median $\frac{\$1000 + \$1392}{2} = \$1196$

Mode \$1000

Because three of the six salespeople received the same amount—\$1000 in holiday pay—the mode most effectively captures this month's results.

KEY IDEAS

measures of central tendency—values that describe the centre of a body of data

outlier—an element of a data set that is very different; affects the mean more than the median or the mode

mean—measure of central tendency found by dividing the sum of all the data by the number of elements; calculated as $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

deviation—difference between a data value and the mean; sum of the mean deviations is 0

weighted mean—calculated by multiplying all the data values by their weights and dividing by the sum of the weights: $\overline{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$

median—middle value in an ordered data set; the most appropriate measure of central tendency when outliers are present

mode—most frequently occurring value; highest rectangle on a histogram; only measure of central tendency for qualitative data

distributions—when skewed right, median < mean; when skewed left, mean < median

3.2 Exercises

A

1. Knowledge and Understanding Use technology to calculate the mean, median, and mode of the following data sets.

(a) marks on a set of tests {66, 65, 72, 78, 93, 70, 68, 64}

(b) monthly rent (\$){625, 750, 800, 650, 725, 850, 625, 650, 625, 1250}

(c) survey responses (1 = never, 2 = sometimes, 3 = often, 4 = always) {1, 2, 3, 4, 3, 3, 4, 3, 2, 3, 3, 2, 3, 2, 1, 2, 3, 4, 3, 3, 2, 3, 2, 3, 2, 3, 3}

(d) waiting time, in minutes, at a fast-food restaurant {5, 5.5, 6.5, 7, 7.5, 7, 7, 5, 6.5, 5, 5, 8.5, 0.5, 4.5, 7}

(e) points scored by a basketball player {12, 15, 8, 12, 15, 10, 3, 14, 15}

(f) daily sales totals (\$) {0, 0, 0, 17 000, 0, 0, 28 455, 0, 0, 41 590}

2. Communication Of the three measures calculated in Question 1, which is the most appropriate for each situation? Why?

3. Use the relative location of the mean, median, and mode calculated in Question 1 to describe the sets as symmetric, skewed left, or skewed right.

4. Hakim's Shoes reported the following sales results:

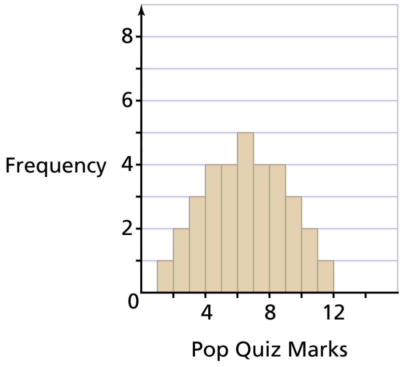
Size	4	5	6	7	8	9	10
Frequency	5	11	15	18	19	13	7

(a) Calculate the mean, median, and mode shoe size.
(b) Which measure of central tendency is most appropriate? Why?

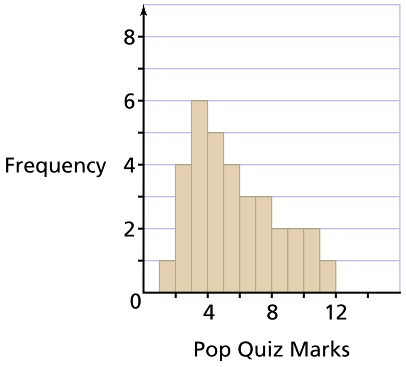
5. Match each distribution with its mean, median, and mode.

(a) mean: 6.2 median: 6 mode: 3.8
(b) mean: 6 median: 6 mode: 10, 2
(c) mean: 6 median: 6 mode: 6
(d) mean: 8.1 median: 8.5 mode: 10

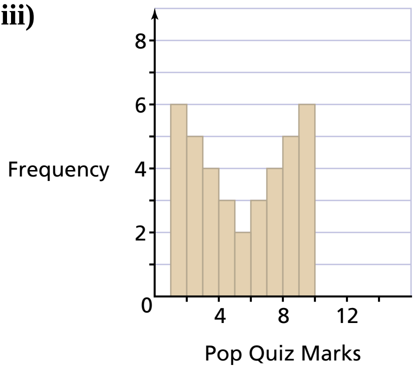
(i)



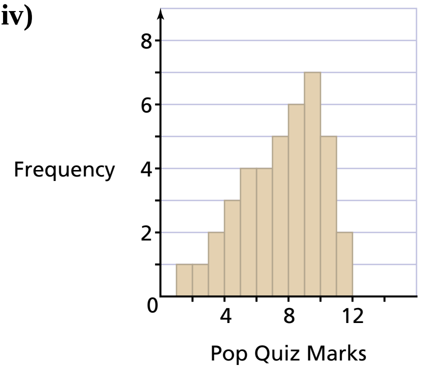
(ii)



(iii)



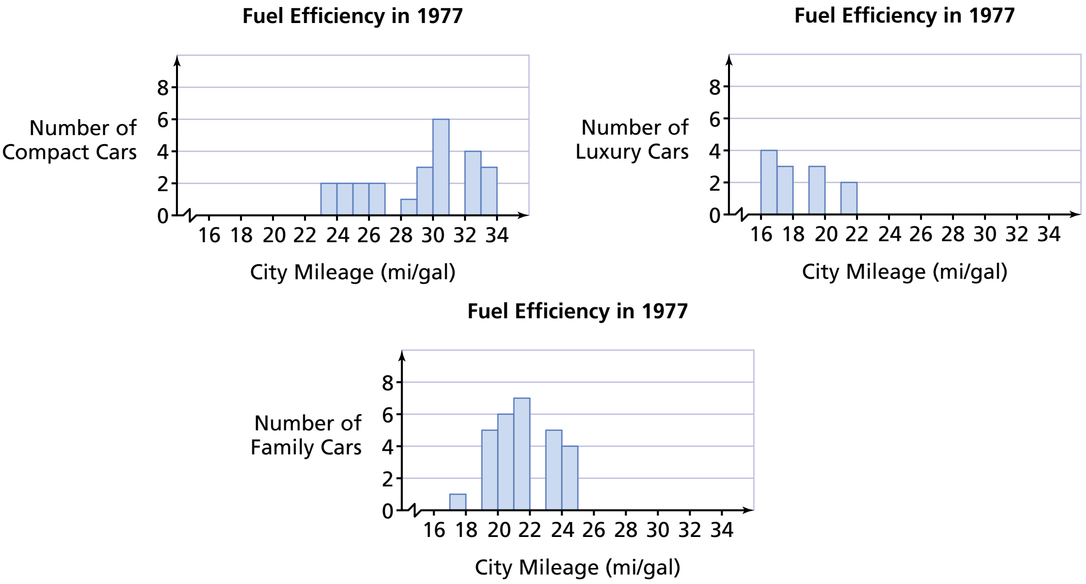
(iv)



B

6. For each of the following, determine if the argument is valid. Explain.
   - **(a)** A sales team has a mean monthly sales record of $12 500. Therefore, half the team members sold more than that.
   - **(b)** The mean mark of one class is 65, while the mean mark of another class is 75. Therefore, the mean of the two classes is 70.
   - **(c)** The mean of four whole numbers is 5. Therefore, the maximum value must have been 20.
   - **(d)** My median monthly expense total is $500, so my total for the year must have been $6000.
   - **(e)** The mean salary before a 10% raise was $30 000. Therefore, the mean salary after is $33 000.
   - **(f)** The median salary before a 10% raise was $30 000. Therefore, the median salary after is $33 000.
   - **(g)** A survey shows that 80% of all salaries were below the mean. Therefore, there must be a mistake.
   - **(h)** The most popular type of music sold in a store is classical. Therefore, more than half the sales are of classical music.

7. Highway fuel efficiency for cars in 1977 is shown below.



Source: Data have been extracted from Fathom Dynamic Statistics™, Key Curriculum Press.

   - **(a)** Estimate the mean and median values for each type of car.
   - **(b)** Explain the relative location of the mean and median in each case.
   - **(c)** Open the file **cars1** on the textbook CD and calculate the actual values for each type.

8. Create a data set of at least five values that has the following properties.
   - **(a)** The mean, median, and mode are all equal to 10.
   - **(b)** The median is 5 and the mean is greater than 10.
   - **(c)** The mean is 5 and the median is greater than 10.

9. **Knowledge and Understanding** Calculate the mean temperature for the two Canadian cities given below.

Calgary												
	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
Temp (°C)	–10.2	–8.0	–3.4	4.1	9.6	13.5	16.4	15.3	10.5	5.5	–2.5	–7.2

Source: Environment Canada

Ottawa												
	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
Temp (°C)	–11.0	–10.1	–3.6	5.1	12.8	18.2	20.6	19.3	14.7	8.1	0.7	–7.9

Source: Environment Canada

10. The mean of one class is 65 and the mean of another class is 75. Explain the steps you would need to take to calculate a combined class average.

11. A pair of dice is rolled numerous times. The sum of the dice, as well as the frequency, is recorded. Calculate the mean, median, and mode for the results.

Sum	2	3	4	5	6	7	8	9	10	11	12
Frequency	2	3	5	7	9	11	8	7	4	2	1

12. Jasmine records the dates on 125 pennies. Estimate the mean and find the median and modal interval of the pennies.

Date	1990–1999	1980–1989	1970–1979	1960–1969
Frequency	56	42	21	6

13. A student's term mark is 75. The term mark counts for 70% of the final mark. What mark must the student achieve on the exam to earn a final mark of

(a) at least 70? **(b)** 75? **(c)** 85?

C **14.** **Thinking, Inquiry, Problem Solving**

(a) Create a data set of at least six values that has the following.

(i) a mean and a median of 10

(ii) a mean of 10 and a median of 12

(b) Write a short paragraph to explain why the median is more resistant to outliers than the mean is.

15. A data set is made up of four values $\{a, b, c, d\}$. Write an expression for (i) the mean of the four values and (ii) the median of the four values.

(a) $a < b < c < d$.

(b) The four values are multiplied by k.

(c) The four values are increased by adding the constant p.

ADDITIONAL ACHIEVEMENT CHART QUESTIONS

Contract negotiations between a union and the management of a local company have recently begun. The chart to the right represents the distribution of salaries.

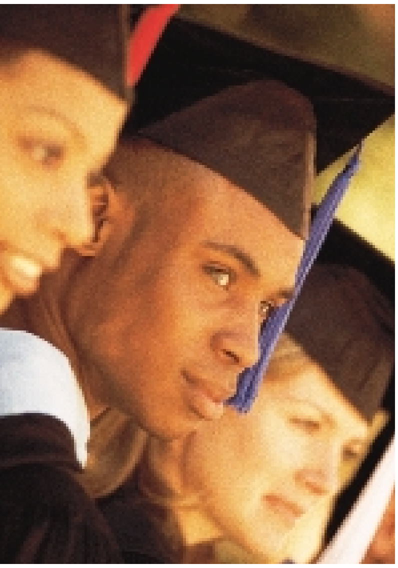
Salary ($)	Number of Employees
18 000–20 999	4
21 000–23 999	16
24 000–26 999	14
27 000–29 999	7
30 000–32 999	3
33 000–35 999	0
36 000–38 999	0
39 000–41 999	0
42 000–44 999	2
45 000–47 999	0
48 000–50 999	1

16. **Knowledge and Understanding** Calculate the median and modal salary interval.

17. **Application** Calculate the weighted mean salary.

18. **Thinking, Inquiry, Problem Solving** Create an employee proposal using one measure of central tendency to justify a salary increase.

19. **Communication** Which measure is fair? Which measure of central tendency would management use to describe current salary levels? Why?



Chapter Problem

Comparing Marks

Justin would like to analyze the overall performance of his school's university applicants by looking at the central tendency. Refer to the data on page 140.

CP5. What measures of central tendency would be most appropriate to describe the performance of the students applying to university?

CP6. Use technology to calculate the median and mean for the set of marks in the table.

CP7. What can you conclude about the distribution of average marks by comparing the two calculated values?