# 2.5 Measures of Central Tendency

It is often convenient to use a central value to summarize a set of data. People frequently use a simple arithmetic average for this purpose. However, there are several different ways to find values around which a set of data tends to cluster. Such values are known as **measures of central tendency**.

## INVESTIGATE & INQUIRE: Not Your Average Average

François is a NHL hockey player whose first major-league contract is up for renewal. His agent is bargaining with the team's general manager.

*Agent*: Based on François' strong performance, we can accept no less than the team's average salary.

*Manager*: Agreed, François deserves a substantial increase. The team is willing to pay François the team's average salary, which is $750 000 a season.

*Agent*: I'm certain that we calculated the average salary to be $1 000 000 per season. You had better check your arithmetic.

*Manager*: There is no error, my friend. Half of the players earn $750 000 or more, while half of the players receive $750 000 or less. $750 000 is a fair offer.

This table lists the current salaries for the team.

| 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 |

1. From looking at the table, do you think the agent or the manager is correct? Explain why.

2. Find the mean salary for the team. Describe how you calculated this amount.
3. Find the median salary. What method did you use to find it?
4. Were the statements by François' agent and the team manager correct?
5. Explain the problem with the use of the term *average* in these negotiations.

In statistics, the three most commonly used measures of central tendency are the mean, median, and mode. Each of these measures has its particular advantages and disadvantages for a given set of data.

A **mean** is defined as the sum of the values of a variable divided by the number of values. In statistics, it is important to distinguish between the mean of a population and the mean of a sample of that population. The sample mean will approximate the actual mean of the population, but the two means could have different values. Different symbols are used to distinguish the two kinds of means: The Greek letter mu, $\mu$, represents a population mean, while $\bar{x}$, read as "$x$-bar," represents a sample mean. Thus,

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} \quad \text{and} \quad \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$= \frac{\sum x}{N} \qquad\qquad = \frac{\sum x}{n}$$

where $\sum x$ is the sum of all values of $X$ in the population or sample, $N$ is the number of values in the entire population, and $n$ is the number of values in a sample. Note that $\sum$, the capital Greek letter sigma, is used in mathematics as a symbol for "the sum of." If no limits are shown above or below the sigma, the sum includes all of the data.

Usually, the mean is what people are referring to when they use the term *average* in everyday conversation.

The **median** is the middle value of the data when they are ranked from highest to lowest. When there is an even number of values, the median is the midpoint between the two middle values.

The **mode** is the value that occurs most frequently in a distribution. Some distributions do not have a mode, while others have several.

Some distributions have **outliers,** which are values distant from the majority of the data. Outliers have a greater effect on means than on medians. For example, the mean and median for the salaries of the hockey team in the investigation have substantially different values because of the two very high salaries for the team's star players.

## Example 1 Determining Mean, Median, and Mode

Two classes that wrote the same physics examination had the following results.

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class A | 71 | 82 | 55 | 76 | 66 | 71 | 90 | 84 | 95 | 64 | 71 | 70 | 83 | 45 | 73 | 51 | 68 |
| Class B | 54 | 80 | 12 | 61 | 73 | 69 | 92 | 81 | 80 | 61 | 75 | 74 | 15 | 44 | 91 | 63 | 50 | 84 |

**a)** Determine the mean, median, and mode for each class.

**b)** Use the measures of central tendency to compare the performance of the two classes.

**c)** What is the effect of any outliers on the mean and median?

### Solution

**a)** For class A, the mean is

$$\bar{x} = \frac{\Sigma x}{n}$$
$$= \frac{71 + 82 + \ldots + 68}{17}$$
$$= \frac{1215}{17}$$
$$= 71.5$$

When the marks are ranked from highest to lowest, the middle value is 71. Therefore, the median mark for class A is 71. The mode for class A is also 71 since this mark is the only one that occurs three times.

Similarly, the mean mark for class B is $\frac{54 + 80 + \ldots + 84}{18} = 64.4$. When the marks are ranked from highest to lowest, the two middle values are 69 and 73, so the median mark for class B is $\frac{69 + 73}{2} = 71$. There are two modes since the values 61 and 80 both occur twice. However, the sample is so small that all the values occur only once or twice, so these modes may not be a reliable measure.

**b)** Although the mean score for class A is significantly higher than that for class B, the median marks for the two classes are the same. Notice that the measures of central tendency for class A agree closely, but those for class B do not.

**c)** A closer examination of the raw data shows that, aside from the two extremely low scores of 15 and 12 in class B, the distributions are not all that different. Without these two outlying marks, the mean for class B would be 70.1, almost the same as the mean for class A. Because of the relatively small size of class B, the effect of the outliers on its mean is significant. However, the values of these outliers have no effect on the median for class B. Even if the two outlying marks were changed to numbers in the 60s, the median mark would not change because it would still be greater than the two marks.

**WEB CONNECTION**

www.mcgrawhill.ca/links/MDM12

For more information about means, medians, and modes, visit the above web site and follow the links. For each measure, give an example of a situation where that measure is the best indicator of the centre of the data.

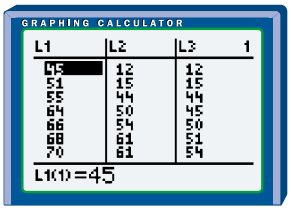
The median is often a better measure of central tendency than the mean for small data sets that contain outliers. For larger data sets, the effect of outliers on the mean is less significant.

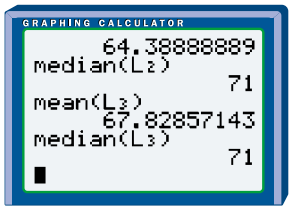
### Example 2 Comparing Samples to a Population

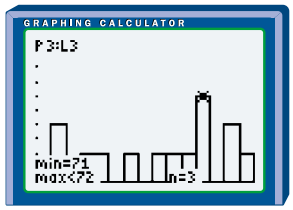
Compare the measures of central tendency for each class in Example 1 to those for all the students who wrote the physics examination.

#### *Solution 1 Using a Graphing Calculator*

Use the STAT EDIT menu to check that lists L1 and L2 are clear. Then, enter the data for class A in L1 and the data for class B in L2. Next, use the **augment( function** from the LIST OPS menu to combine L1 and L2, and store the result in L3.
You can use the **mean(** and **median( functions** from the LIST MATH menu to find the mean and median for each of the three lists. You can also find these measures by using the **1-Var Stats command** from the STAT CALC menu. To find the modes, sort the lists with the **SortA( function** from the LIST OPS menu, and then scroll down through the lists to find the most frequent values. Alternatively, you can use **STAT PLOT** to display a histogram for each list and read the $x$-values for the tallest bars with the **TRACE instruction**.



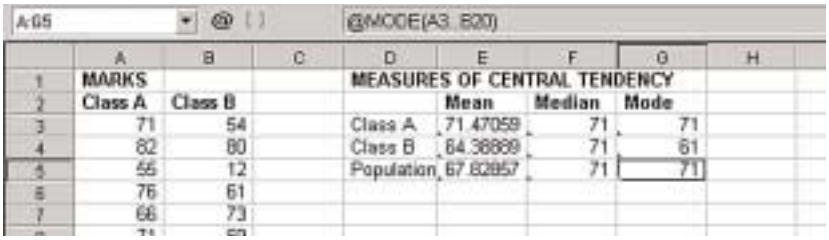




Note that the mean for class A overestimates the population mean, while the mean for class B underestimates it. The measures of central tendency for class A are reasonably close to those for the whole population of students who wrote the physics examination, but the two sets of measures are not identical. Because both of the low-score outliers happen to be in class B, it is a less representative sample of the population.

#### *Solution 2 Using a Spreadsheet*

Enter the data for class A and class B in separate columns. The AVG and MEAN functions in Corel® Quattro® Pro will calculate the **mean** for any range of cells you specify, as will the AVERAGE function in Microsoft® Excel.
In both spreadsheets, you can use the MEDIAN, and MODE functions to find the **median** and **mode** for each class and for the combined data for both classes. Note that all these functions ignore any blank cells in a specified range. The MODE function reports only one mode even if the data have two or more modes.



| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | MARKS | | | MEASURES OF CENTRAL TENDENCY | | | | |
| 2 | Class A | Class B | | | Mean | Median | Mode | |
| 3 | 71 | 54 | | Class A | 71.47059 | 71 | 71 | |
| 4 | 82 | 80 | | Class B | 64.38889 | 71 | 61 | |
| 5 | 55 | 12 | | Population | 67.82857 | 71 | 71 | |
| 6 | 76 | 61 | | | | | | |
| 7 | 66 | 73 | | | | | | |

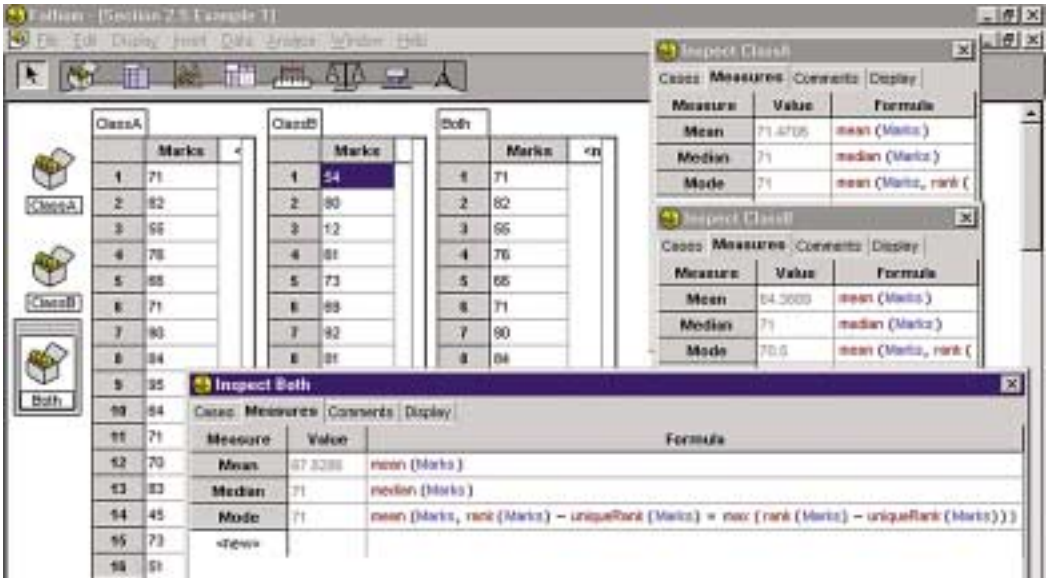
### *Solution 3 Using Fathom™*

Drag the **case table** icon to the workspace and name the attribute for the first column Marks. Enter the data for class A and change the name of the **collection** from Collection1 to ClassA. Use the same method to enter the marks for class B into a collection called ClassB. To create a collection with the combined data, first open another **case table** and name the **collection** Both. Then, go back to the class A **case table** and use the Edit menu to select all cases and then copy them. Return to the Both **case table** and select Paste Cases from the Edit menu. Copy the cases from the class B table in the same way.

Now, right-click on the class A **collection** to open the **inspector**. Click the Measures tab, and create Mean, Median, and Mode measures. Use the Edit Formula menu to enter the formulas for these measures. Use the same procedure to find the **mean**, **median**, and **mode** for the other two collections. Note from the screen below that Fathom™ uses a complicated formula to find modes. See the Help menu or the Fathom™ section of Appendix B for details.

**Project Prep**

In your statistics project, you may find measures of central tendency useful for describing your data.



Chapter 8 discusses a method for calculating how representative of a population a sample is likely to be.

Sometimes, certain data within a set are more significant than others. For example, the mark on a final examination is often considered to be more important than the mark on a term test for determining an overall grade for a course. A **weighted mean** gives a measure of central tendency that reflects the relative importance of the data:

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum_i w_i x_i}{\sum_i w_i}$$

where $\sum_i w_i x_i$ is the sum of the weighted values and $\sum_i w_i$ is the sum of the various weighting factors.

Weighted means are often used in calculations of indices.

### Example 3 Calculating a Weighted Mean

The personnel manager for Statsville Marketing Limited considers five criteria when interviewing a job applicant. The manager gives each applicant a score between 1 and 5 in each category, with 5 as the highest score. Each category has a weighting between 1 and 3. The following table lists a recent applicant's scores and the company's weighting factors.

| Criterion | Score, $x_i$ | Weighting Factor, $w_i$ |
|---|---|---|
| Education | 4 | 2 |
| Job experience | 2 | 2 |
| Interpersonal skills | 5 | 3 |
| Communication skills | 5 | 3 |
| References | 4 | 1 |

**a)** Determine the weighted mean score for this job applicant.

**b)** How does this weighted mean differ from the unweighted mean?

**c)** What do the weighting factors indicate about the company's hiring priorities?

**Solution**

**a)** To compute the weighted mean, find the sum of the products of each score and its weighting factor.

$$\begin{aligned}\bar{x}_w &= \frac{\sum_i w_i x_i}{\sum_i w_i} \\ &= \frac{2(4) + 2(2) + 3(5) + 3(5) + (1)4}{2 + 2 + 3 + 3 + 1} \\ &= \frac{46}{11} \\ &= 4.2\end{aligned}$$

Therefore, this applicant had a weighted-mean score of approximately 4.2.

**b)** The unweighted mean is simply the sum of unweighted scores divided by 5.

$$\begin{aligned}\bar{x} &= \frac{\sum x}{n} \\ &= \frac{4 + 2 + 5 + 5 + 4}{5} \\ &= 4\end{aligned}$$

Without the weighting factors, this applicant would have a mean score of 4 out of 5.

**c)** Judging by these weighting factors, the company places a high importance on an applicant's interpersonal and communication skills, moderate importance on education and job experience, and some, but low, importance on references.

When a set of data has been grouped into intervals, you can approximate the mean using the formula

$$\mu \doteq \frac{\sum_i f_i m_i}{\sum_i f_i} \qquad \bar{x} \doteq \frac{\sum_i f_i m_i}{\sum_i f_i}$$

where $m_i$ is the midpoint value of an interval and $f_i$ the frequency for that interval.

You can estimate the median for grouped data by taking the midpoint of the interval within which the median is found. This interval can be found by analysing the cumulative frequencies.

## Example 4 Calculating the Mean and Median for Grouped Data

A group of children were asked how many hours a day they spend watching television. The table at the right summarizes their responses.

| Number of Hours | Number of Children, $f_i$ |
|---|---|
| 0–1 | 1 |
| 1–2 | 4 |
| 2–3 | 7 |
| 3–4 | 3 |
| 4–5 | 2 |
| 5–6 | 1 |

**a)** Determine the mean and median number of hours for this distribution.

**b)** Why are these values simply approximations?

### Solution

**a)** First, find the midpoints and cumulative frequencies for the intervals. Then, use the midpoints and the frequencies for the intervals to calculate an estimate for the mean.

| Number of Hours | Midpoint, $x_i$ | Number of Children, $f_i$ | Cumulative Frequency | $f_i x_i$ |
|---|---|---|---|---|
| 0–1 | 0.5 | 1 | 1 | 0.5 |
| 1–2 | 1.5 | 4 | 5 | 6 |
| 2–3 | 2.5 | 7 | 12 | 17.5 |
| 3–4 | 3.5 | 3 | 15 | 10.5 |
| 4–5 | 4.5 | 2 | 17 | 9 |
| 5–6 | 5.5 | 1 | 18 | 5.5 |
| | | $\sum_i f_i = 18$ | | $\sum_i f_i x_i = 49$ |

$$\bar{x} \doteq \frac{\sum_i f_i x_i}{\sum_i f_i}$$
$$= \frac{49}{18}$$
$$= 2.7$$

Therefore, the mean time the children spent watching television is approximately 2.7 h a day.

To determine the median, you must identify the interval in which the middle value occurs. There are 18 data values, so the median is the mean of the ninth and tenth values. According to the cumulative-frequency column, both of these occur within the interval of 2–3 h. Therefore, an approximate value for the median is 2.5 h.

**b)** These values for the mean and median are approximate because you do not know where the data lie within each interval. For example, the child whose viewing time is listed in the first interval could have watched anywhere from 0 to 60 min of television a day. If the median value is close to one of the boundaries of the interval, then taking the midpoint of the interval as the median could give an error of almost 30 min.

## Key Concepts

- The three principal measures of central tendency are the mean, median, and mode. The measures for a sample can differ from those for the whole population.
- The mean is the sum of the values in a set of data divided by the number of values in the set.
- The median is the middle value when the values are ranked in order. If there are two middle values, then the median is the mean of these two middle values.
- The mode is the most frequently occurring value.
- Outliers can have a dramatic effect on the mean if the sample size is small.
- A weighted mean can be a useful measure when all the data are not of equal significance.
- For data grouped into intervals, the mean and median can be estimated using the midpoints and frequencies of the intervals.

## Communicate Your Understanding

**1.** Describe a situation in which the most useful measure of central tendency is
**a)** the mean **b)** the median **c)** the mode

**2.** Explain why a weighted mean would be used to calculate an index such as the consumer price index.

**3.** Explain why the formula $\bar{x} \doteq \frac{\sum_i f_i m_i}{\sum_i f_i}$ gives only an approximate value for the mean for grouped data.

## Practise

A

**1.** For each set of data, calculate the mean, median, and mode.
**a)** 2.4 3.5 1.9 3.0 3.5 2.4 1.6 3.8 1.2 2.4 3.1 2.7 1.7 2.2 3.3
**b)** 10 15 14 19 18 17 12 10 14 15 18 20 9 14 11 18

**2.** **a)** List a set of eight values that has no mode.
**b)** List a set of eight values that has a median that is not one of the data values.
**c)** List a set of eight values that has two modes.
**d)** List a set of eight values that has a median that is one of the data values.

## Apply, Solve, Communicate

**3.** Stacey got 87% on her term work in chemistry and 71% on the final examination. What will her final grade be if the term mark counts for 70% and the final examination counts for 30%?

**4.** **Communication** Determine which measure of central tendency is most appropriate for each of the following sets of data. Justify your choice in each case.

**a)** baseball cap sizes

**b)** standardized test scores for 2000 students

**c)** final grades for a class of 18 students

**d)** lifetimes of mass-produced items, such as batteries or light bulbs

## B

**5.** An interviewer rates candidates out of 5 for each of three criteria: experience, education, and interview performance. If the first two criteria are each weighted twice as much as the interview, determine which of the following candidates should get the job.

| Criterion | Nadia | Enzo | Stephan |
|---|---|---|---|
| Experience | 4 | 5 | 5 |
| Education | 4 | 4 | 3 |
| Interview | 4 | 3 | 4 |

**6.** Determine the effect the two outliers have on the mean mark for all the students in Example 2. Explain why this effect is different from the effect the outliers had on the mean mark for class B.

**7.** **Application** The following table shows the grading system for Xabbu's calculus course.

| Term Mark | Overall Mark |
|---|---|
| Knowledge and understanding (K/U) 35%<br>Thinking, inquiry, problem solving (TIPS) 25%<br>Communication (C) 15%<br>Application (A) 25% | Term mark 70%<br>Final examination 30% |

**a)** Determine Xabbu's term mark if he scored 82% in K/U, 71% in TIPS, 85% in C, and 75% in A.

**b)** Determine Xabbu's overall mark if he scored 65% on the final examination.

**8.** **Application** An academic award is to be granted to the student with the highest overall score in four weighted categories. Here are the scores for the three finalists.

| Criterion | Weighting | Paulo | Janet | Jamie |
|---|---|---|---|---|
| Academic achievement | 3 | 4 | 3 | 5 |
| Extra-curricular activities | 2 | 4 | 4 | 4 |
| Community service | 2 | 2 | 5 | 3 |
| Interview | 1 | 5 | 5 | 4 |

**a)** Calculate each student's mean score without considering the weighting factors.

**b)** Calculate the weighted-mean score for each student.

**c)** Who should win the award? Explain.

**9.** Al, a shoe salesman, needs to restock his best-selling sandal. Here is a list of the sizes of the pairs he sold last week. This sandal does not come in half-sizes.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 7 | 6 | 8 | 7 | 10 | 5 | 10 | 7 | 9 |
| 11 | 4 | 6 | 7 | 10 | 10 | 7 | 8 | 10 | 7 |
| 9 | 7 | 10 | 4 | 7 | 7 | 10 | 11 | | |

**a)** Determine the three measures of central tendency for these sandals.

**b)** Which measure has the greatest significance for Al? Explain.

**c)** What other value is also significant?

**d)** Construct a histogram for the data. What might account for the shape of this histogram?

**10.** **Communication** Last year, the mean number of goals scored by a player on Statsville's soccer team was 6.

**a)** How many goals did the team score last year if there were 15 players on the team?

**b)** Explain how you arrived at the answer for part a) and show why your method works.

**11.** **Inquiry/Problem Solving** The following table shows the salary structure of Statsville Plush Toys, Inc. Assume that salaries exactly on an interval boundary have been placed in the higher interval.

| Salary Range ($000) | Number of Employees |
|---|---|
| 20–30 | 12 |
| 30–40 | 24 |
| 40–50 | 32 |
| 50–60 | 19 |
| 60–70 | 9 |
| 70–80 | 3 |
| 80–90 | 0 |
| 90–100 | 1 |

**a)** Determine the approximate mean salary for an employee of this firm.

**b)** Determine the approximate median salary.

**c)** How much does the outlier influence the mean and median salaries? Use calculations to justify your answer.

**12.** **Inquiry/Problem Solving** A group of friends and relatives get together every Sunday for a little pick-up hockey. The ages of the 30 regulars are shown below.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 22 | 28 | 32 | 45 | 48 | 19 | 20 | 52 | 50 | 21 |
| 30 | 46 | 21 | 38 | 45 | 49 | 18 | 25 | 23 | 46 |
| 51 | 24 | 39 | 48 | 28 | 20 | 50 | 33 | 17 | 48 |

**a)** Determine the mean, median, and mode for this distribution.

**b)** Which measure best describes these data? Explain your choice.

**c)** Group these data into six intervals and produce a frequency table.

**d)** Illustrate the grouped data with a frequency diagram. Explain why the shape of this frequency diagram could be typical for such groups of hockey players.

**e)** Produce a cumulative-frequency diagram.

**f)** Determine a mean, median, and mode for the grouped data. Explain any differences between these measures and the ones you calculated in part a).

**13.** The **modal interval** for grouped data is the interval that contains more data than any other interval.

**a)** Determine the modal interval(s) for your data in part d) of question 12.

**b)** Is the modal interval a useful measure of central tendency for this particular distribution? Why or why not?

**14.** **a)** Explain the effect outliers have on the median of a distribution. Use examples to support your explanation.

**b)** Explain the effect outliers have on the mode of a distribution. Consider different cases and give examples of each.

C

**15.** The harmonic mean is defined as $\left(\sum_i \frac{1}{nx_i}\right)^{-1}$, where $n$ is the number of values in the set of data.

**a)** Use a harmonic mean to find the average price of gasoline for a driver who bought $20 worth at 65¢/L last week and another $20 worth at 70¢/L this week.

**b)** Describe the types of calculations for which the harmonic mean is useful.

**16.** The geometric mean is defined as $\sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$, where $n$ is the number of values in the set of data.

**a)** Use the geometric mean to find the average annual increase in a labour contract that gives a 4% raise the first year and a 2% raise for the next three years.

**b)** Describe the types of calculations for which the geometric mean is useful.