CHAPTER 2

Statistics of One Variable

Specific Expectations	Section
Locate data to answer questions of significance or personal interest, by searching well-organized databases.	2.2
Use the Internet effectively as a source for databases.	2.2
Demonstrate an understanding of the purpose and the use of a variety of sampling techniques.	2.3, 2.4
Describe different types of bias that may arise in surveys.	2.4
Illustrate sampling bias and variability by comparing the characteristics of a known population with the characteristics of samples taken repeatedly from that population, using different sampling techniques.	2.4, 2.5, 2.6
Organize and summarize data from secondary sources, using technology.	2.1, 2.2, 2.5, 2.6
Compute, using technology, measures of one-variable statistics (i.e., the mean, median, mode, range, interquartile range, variance, and standard deviation), and demonstrate an understanding of the appropriate use of each measure.	2.5, 2.6
Interpret one-variable statistics to describe characteristics of a data set.	2.5, 2.6
Describe the position of individual observations within a data set, using *z*-scores and percentiles.	2.6
Explain examples of the use and misuse of statistics in the media.	2.4
Assess the validity of conclusions made on the basis of statistical studies, by analysing possible sources of bias in the studies and by calculating and interpreting additional statistics, where possible.	2.5, 2.6
Explain the meaning and the use in the media of indices based on surveys.	2.2

In earlier times they had no statistics, and so they had to fall back on lies. Hence the huge exaggerations of primitive literature—giants or miracles or wonders! They did it with lies and we do it with statistics; but it is all the same.
—Stephen Leacock (1869–1944)

Facts are stubborn, but statistics are more pliable.
—Mark Twain (1835–1910)

Chapter Problem

Contract Negotiations

François is a young NHL hockey player whose first major-league contract is up for renewal. His agent wants to bargain for a better salary based on François' strong performance over his first five seasons with the team. Here are some of François' statistics for the past five seasons.

Season	Games	Goals	Assists	Points
1	20	3	4	7
2	45	7	11	18
3	76	19	25	44
4	80	19	37	56
5	82	28	36	64
Total	**303**	**76**	**113**	**189**

1. How could François' agent use these statistics to argue for a substantial pay increase for his client?

2. Are there any trends in the data that the team's manager could use to justify a more modest increase?

As these questions suggest, statistics could be used to argue both for and against a large salary increase for François. However, the statistics themselves are not wrong or contradictory. François' agent and the team's manager will, understandably, each emphasize only the statistics that support their bargaining positions. Such selective use of statistics is one reason why they sometimes receive negative comments such as the quotations above. Also, even well-intentioned researchers sometimes inadvertently use biased methods and produce unreliable results. This chapter explores such sources of error and methods for avoiding them. Properly used, statistical analysis is a powerful tool for detecting trends and drawing conclusions, especially when you have to deal with large sets of data.

Review of Prerequisite Skills

If you need help with any of the skills listed in purple below, refer to Appendix A.

1. Fractions, percents, decimals The following amounts are the total cost for the items including the 7% goods and services tax (GST) and an 8% provincial sales tax (PST). Determine the price of each item.

a) watch $90.85

b) CD $19.54

c) bicycle $550.85

d) running shoes $74.39

2. Fractions, percents, decimals

a) How much will Josh make if he receives an 8% increase on his pay of $12.50/h?

b) What is the net increase in Josh's take-home pay if the payroll deductions total 17%?

3. Fractions, percents, decimals What is the percent reduction on a sweater marked down from $50 to $35?

4. Fractions, percents, decimals Determine the cost, including taxes, of a VCR sold at a 25% discount from its original price of $219.

5. Mean, median, mode Calculate the mean, median, and mode for each set of data.

a) 22, 26, 28, 27, 26

b) 11, 19, 14, 23, 16, 26, 30, 29

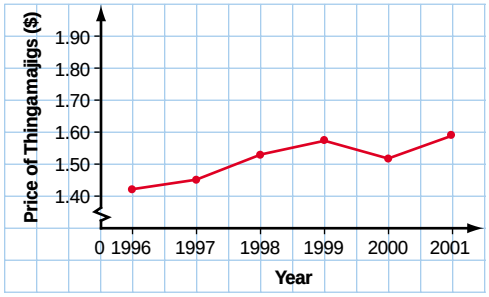
c) 10, 18, 30, 43, 18, 13, 10

d) 70, 30, 25, 52, 12, 70

e) 370, 260, 155, 102, 126, 440

f) 24, 32, 37, 24, 32, 38, 32, 36, 35, 42

6. Graphing data Consider the following graph, which shows the average price of thingamajigs over time.



a) What was the price of thingamajigs in 1996?

b) In what year did the price first rise above $1.50?

c) Describe the overall trend over the time period shown.

d) Estimate the percent increase in the price of thingamajigs from 1996 to 2001.

e) List the domain and range of these data.

7. Graphing data The table below gives the number of CDs sold at a music store on each day of the week for one week.

Day	Number of CDs Sold
Monday	48
Tuesday	52
Wednesday	44
Thursday	65
Friday	122
Saturday	152
Sunday	84

Display the data on a circle graph.

2.1 Data Analysis With Graphs

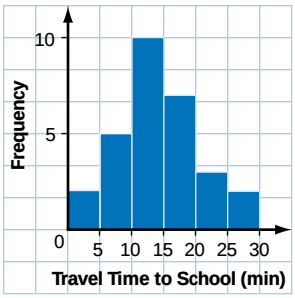
Statistics is the gathering, organization, analysis, and presentation of numerical information. You can apply statistical methods to almost any kind of data. Researchers, advertisers, professors, and sports announcers all make use of statistics. Often, researchers gather large quantities of data since larger samples usually give more accurate results. The first step in the analysis of such data is to find ways to organize, analyse, and present the information in an understandable form.

INVESTIGATE & INQUIRE: Using Graphs to Analyse Data

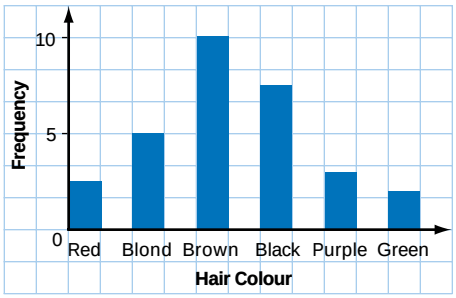
1. Work in groups or as a class to design a fast and efficient way to survey your class about a simple numerical variable, such as the students' heights or the distances they travel to school.
2. Carry out your survey and record all the results in a table.
3. Consider how you could organize these results to look for any trends or patterns. Would it make sense to change the order of the data or to divide them into groups? Prepare an organized table and see if you can detect any patterns in the data. Compare your table to those of your classmates. Which methods work best? Can you suggest improvements to any of the tables?
4. Make a graph that shows how often each value or group of values occurs in your data. Does your graph reveal any patterns in the data? Compare your graph to those drawn by your classmates. Which graph shows the data most clearly? Do any of the graphs have other advantages? Explain which graph you think is the best overall.
5. Design a graph showing the total of the frequencies of all values of the variable up to a given amount. Compare this cumulative-frequency graph to those drawn by your classmates. Again, decide which design works best and look for ways to improve your own graph and those of your classmates.

The unprocessed information collected for a study is called **raw data**. The quantity being measured is the **variable.** A **continuous variable** can have any value within a given range, while a **discrete variable** can have only certain separate values (often integers). For example, the height of students in your school is a continuous variable, but the number in each class is a discrete variable. Often, it is useful to know how frequently the different values of a variable occur in a set of data. **Frequency tables** and **frequency diagrams** can give a convenient overview of the distribution of values of the variable and reveal trends in the data.

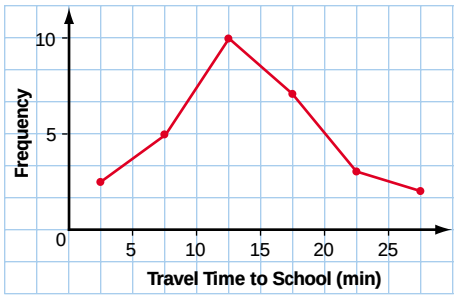
A **histogram** is a special form of **bar graph** in which the areas of the bars are proportional to the *frequencies* of the values of the variable. The bars in a histogram are connected and represent a continuous range of values. Histograms are used for variables whose values can be arranged in numerical order, especially continuous variables, such as weight, temperature, or travel time. Bar graphs can represent all kinds of variables, including the frequencies of separate categories that have no set order, such as hair colour or citizenship. A **frequency polygon** can illustrate the same information as a histogram or bar graph. To form a frequency polygon, plot frequencies versus variable values and then join the points with straight lines.



Histogram

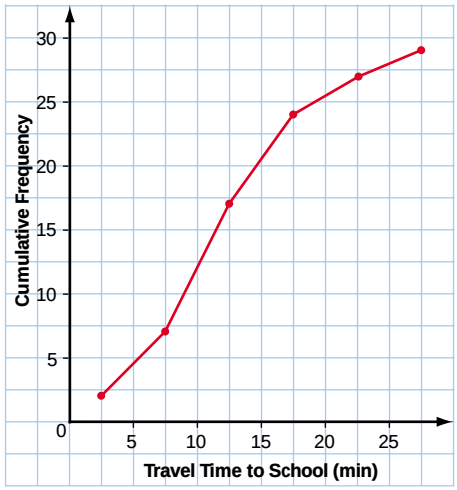


Bar Graph



Frequency Polygon

A **cumulative-frequency graph** or **ogive** shows the running total of the frequencies from the lowest value up.



WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

To learn more about histograms, visit the above web site and follow the links. Write a short description of how to construct a histogram.

Example 1 Frequency Tables and Diagrams

Here are the sums of the two numbers from 50 rolls of a pair of standard dice.

11	4	4	10	8	7	6	6	5	10	7	9	8	8
4	7	9	11	12	10	3	7	6	9	5	8	6	8
2	6	7	5	11	2	5	5	6	6	5	2	10	9
6	5	5	5	3	9	8	2						

a) Use a frequency table to organize these data.

b) Are any trends or patterns apparent in this table?

c) Use a graph to illustrate the information in the frequency table.

d) Create a cumulative-frequency table and graph for the data.

e) What proportion of the data has a value of 6 or less?

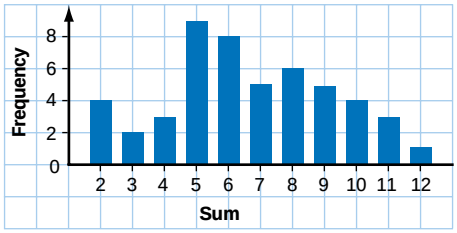
Solution

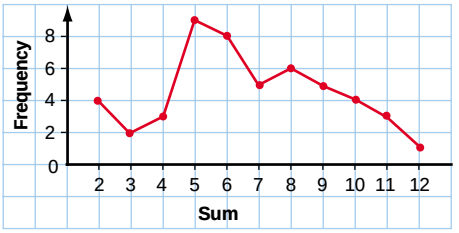
a) Go through the data and tally the frequency of each value of the variable as shown in the table on the right.

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

b) The table does reveal a pattern that was not obvious from the raw data. From the frequency column, notice that the middle values tend to be the most frequent while the high and low values are much less frequent.

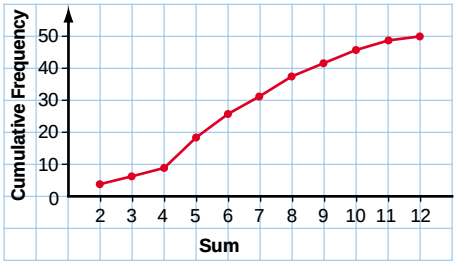
c) The bar graph or frequency polygon makes the pattern in the data more apparent.





d) Add a column for cumulative frequencies to the table. Each value in this column is the running total of the frequencies of each sum up to and including the one listed in the corresponding row of the sum column. Graph these cumulative frequencies against the values of the variable.

Sum	Tally	Frequency	Cumulative Frequency
2	||||	4	4
3	||	2	6
4	|||	3	9
5	𝍸 ||||	9	18
6	𝍸 |||	8	26
7	𝍸	5	31
8	𝍸 |	6	37
9	𝍸	5	42
10	||||	4	46
11	|||	3	49
12	|	1	50



e) From either the cumulative-frequency column or the diagram, you can see that 26 of the 50 outcomes had a value of 6 or less.

When the number of measured values is large, data are usually grouped into **classes** or **intervals**, which make tables and graphs easier to construct and interpret. Generally, it is convenient to use from 5 to 20 equal intervals that cover the entire **range** from the smallest to the largest value of the variable. The interval width should be an even fraction or multiple of the measurement unit for the variable. Technology is particularly helpful when you are working with large sets of data.

Example 2 Working With Grouped Data

This table lists the daily high temperatures in July for a city in southern Ontario.

Day	1	2	3	4	5	6	7	8	9	10	11
Temperature (°C)	27	25	24	30	32	31	29	24	22	19	21
Day	12	13	14	15	16	17	18	19	20	21	22
Temperature (°C)	25	26	31	33	33	30	29	27	28	26	27
Day	23	24	25	26	27	28	29	30	31		
Temperature (°C)	22	18	20	25	26	29	32	31	28		

a) Group the data and construct a frequency table, a histogram or frequency polygon, and a cumulative-frequency graph.

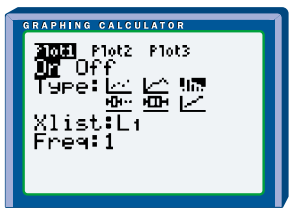
b) On how many days was the maximum temperature 25°C or less? On how many days did the temperature exceed 30°C?

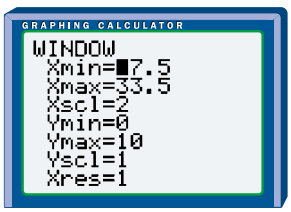
See Appendix B for more detailed information about technology functions and keystrokes.

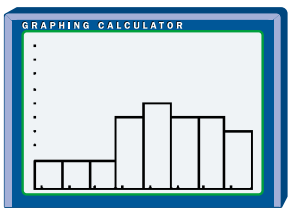
Solution 1 Using a Graphing Calculator

a) The range of the data is 33°C – 18°C = 15°C. You could use five 3-degree intervals, but then many of the recorded temperatures would fall on the interval boundaries. You can avoid this problem by using eight 2-degree intervals with the lower limit of the first interval at 17.5°C. The upper limit of the last interval will be 33.5°C.

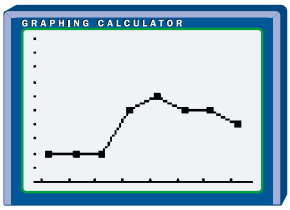
Use the STAT EDIT menu to make sure that lists L1 to L4 are clear, and then enter the temperature data into L1. Use STAT PLOT to turn on Plot1 and select the histogram icon. Next, adjust the window settings. Set Xmin and Xmax to the lower and upper limits for your intervals and set Xscl to the interval width. Ymin should be 0. Press GRAPH to display the histogram, then adjust Ymax and Yscl, if necessary.



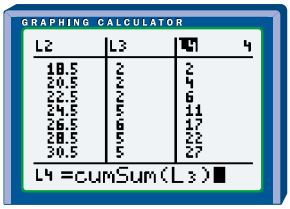


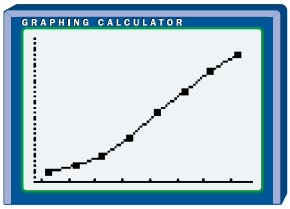


You can now use the **TRACE instruction** and the arrow keys to determine the tally for each of the intervals. Enter the midpoints of the intervals into L2 and the tallies into L3. Turn off Plot1 and set up Plot2 as an *x*-*y* line plot of lists L2 and L3 to produce a frequency polygon.



Use the **cumSum(function** from the LIST OPS menu to find the running totals of the frequencies in L3 and store the totals in L4. Now, an *x*-*y* line plot of L2 and L4 will produce a cumulative-frequency graph.





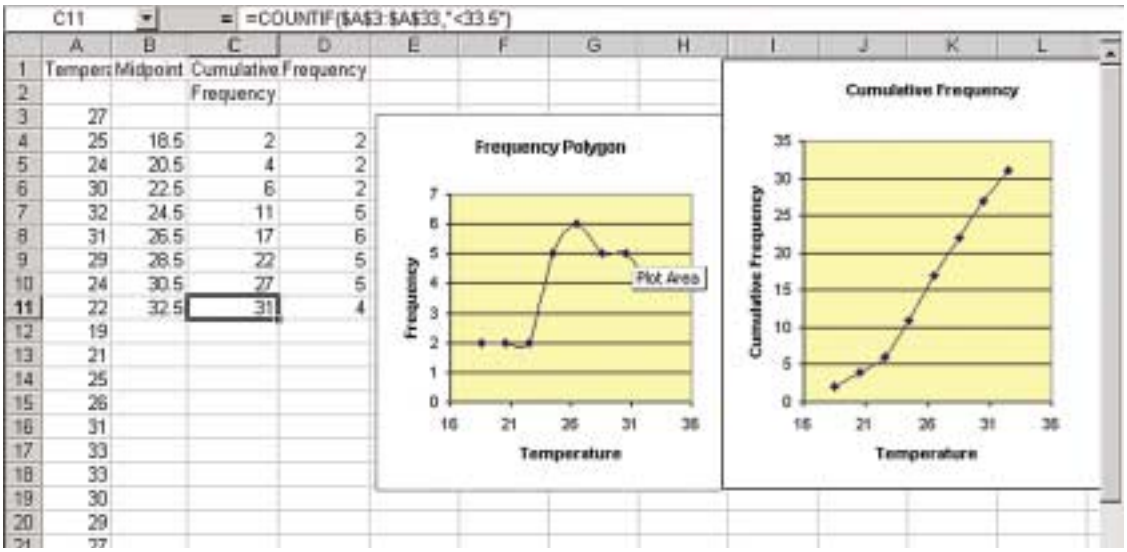
b) Since you know that all the temperatures were in whole degrees, you can see from the cumulative frequencies in L4 that there were 11 days on which the maximum temperature was no higher than 25°C. You can also get this information from the cumulative-frequency graph.

You cannot determine the exact number of days with temperatures over 30°C from the grouped data because temperatures from 29.5°C to 31.5°C are in the same interval. However, by interpolating the cumulative-frequency graph, you can see that there were about 6 days on which the maximum temperature was 31°C or higher.

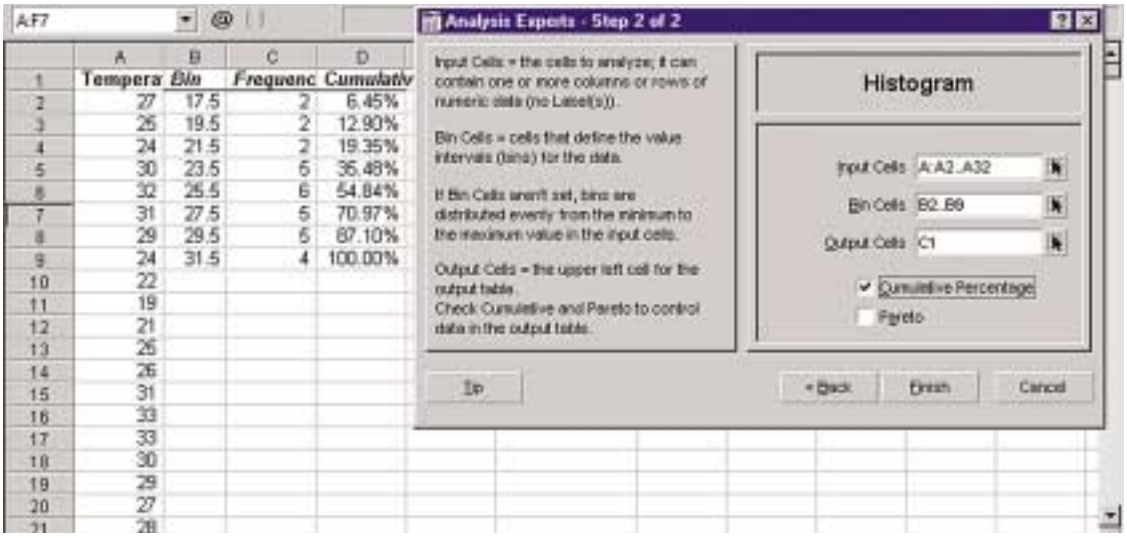
Solution 2 Using a Spreadsheet

a) Enter the temperature data into column A and the midpoints of the intervals into column B. Use the **COUNTIF function** in column C to tally the cumulative frequency for each interval. If you use **absolute cell referencing**, you can copy the formula down the column and then change just the upper limit in the counting condition. Next, find the frequency for each interval by finding the difference between its cumulative frequency and the one for the previous interval.

You can then use the **Chart feature** to produce a frequency polygon by graphing columns B and D. Similarly, charting columns B and C will produce a cumulative-frequency graph.



In Corel® Quattro® Pro, you can also use the Histogram tool in the Tools / Numeric Tools / Analysis menu to automatically tally the frequencies and cumulative frequencies.



b) As in the solution using a graphing calculator, you can see from the cumulative frequencies that there were 11 days on which the maximum temperature was no higher than 25°C. Also, you can estimate from the cumulative-frequency graph that there were 6 days on which the maximum temperature was 31°C or higher. Note that you could use the **COUNTIF function** with the raw data to find the exact number of days with temperatures over 30°C.

A **relative-frequency** table or diagram shows the frequency of a data group as a fraction or percent of the whole data set.

Project Prep

You may find frequency-distribution diagrams useful for your statistics project.

Example 3 Relative-Frequency Distribution

Here are a class' scores obtained on a data-management examination.

78	81	55	60	65	86	44	90
77	71	62	39	80	72	70	64
88	73	61	70	75	96	51	73
59	68	65	81	78	67		

a) Construct a frequency table that includes a column for relative frequency.

b) Construct a histogram and a frequency polygon.

c) Construct a relative-frequency histogram and a relative-frequency polygon.

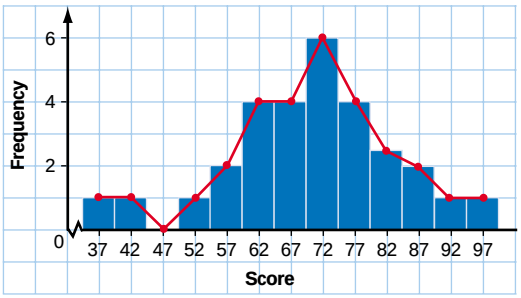
d) What proportion of the students had marks between 70% and 79%?

Solution

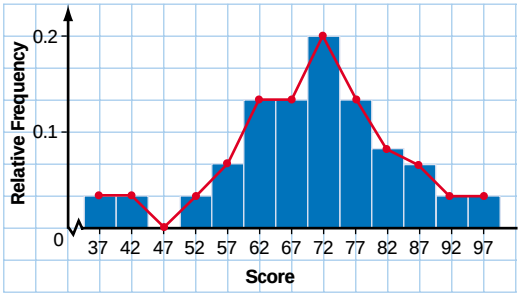
a) The lowest and highest scores are 39% and 96%, which give a range of 57%. An interval width of 5 is convenient, so you could use 13 intervals as shown here. To determine the relative frequencies, divide the frequency by the total number of scores. For example, the relative frequency of the first interval is $\frac{1}{30}$, showing that approximately 3% of the class scored between 34.5% and 39.5%.

Score (%)	Midpoint	Tally	Frequency	Relative Frequency
34.5–39.5	37	|	1	0.033
39.5–44.5	42	|	1	0.033
44.5–49.5	47	–	0	0
49.5–54.5	52	|	1	0.033
54.5–59.5	57	||	2	0.067
59.5–64.5	62	||||	4	0.133
64.5–69.5	67	||||	4	0.133
69.5–74.5	72	~~||||~~ |	6	0.200
74.5–79.5	77	||||	4	0.133
79.5–84.5	82	|||	3	0.100
84.5–89.5	87	||	2	0.067
89.5–94.5	92	|	1	0.033
94.5–99.5	97	|	1	0.033

b) The frequency polygon can be superimposed onto the same grid as the histogram.



c) Draw the relative-frequency histogram and the relative-frequency polygon using the same procedure as for a regular histogram and frequency polygon. As you can see, the only difference is the scale of the *y*-axis.



d) To determine the proportion of students with marks in the 70s, add the relative frequencies of the interval from 69.5 to 74.5 and the interval from 74.5 to 79.5:

$0.200 + 0.133 = 0.333$

Thus, 33% of the class had marks between 70% and 79%.

Categorical data are given labels rather than being measured numerically. For example, surveys of blood types, citizenship, or favourite foods all produce categorical data. **Circle graphs** (also known as **pie charts**) and **pictographs** are often used instead of bar graphs to illustrate categorical data.

Example 4 Presenting Categorical Data

The table at the right shows Canadians' primary use of the Internet in 1999.

Illustrate these data with

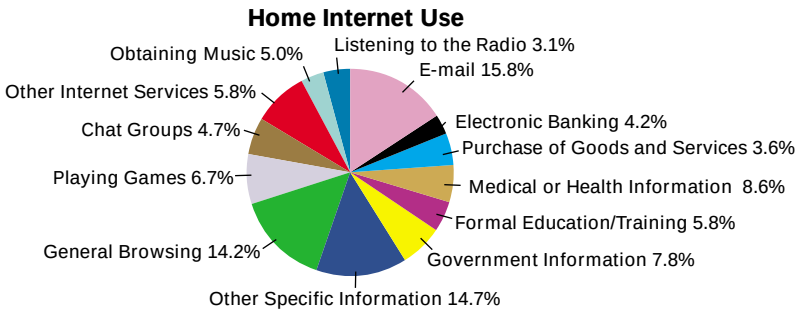
a) a circle graph

b) a pictograph

Primary Use	Households (%)
E-mail	15.8
Electronic banking	4.2
Purchase of goods and services	3.6
Medical or health information	8.6
Formal education/training	5.8
Government information	7.8
Other specific information	14.7
General browsing	14.2
Playing games	6.7
Chat groups	4.7
Other Internet services	5.8
Obtaining music	5.0
Listening to the radio	3.1

Solution

a)



b) There are numerous ways to represent the data with a pictograph. The one shown here has the advantages of being simple and visually indicating that the data involve computers.

Home Internet Use

E-mail	
Electronic Banking	
Purchase of Goods and Services	
Medical or Health Information	
Formal Education/Training	
Government Information	
Other Specific Information	
General Browsing	
Playing Games	
Chat Groups	
Other Internet Services	
Obtaining Music	
Listening to the Radio	

Each represents 2% of households.

You can see from the example above that circle graphs are good for showing the sizes of categories relative to the whole and to each other. Pictographs can use a wide variety of visual elements to clarify the data and make the graph more interesting. However, with both circle graphs and pictographs, the relative frequencies for the categories can be hard to read accurately. While a well-designed pictograph can be a useful tool, you will sometimes see pictographs with distorted or missing scales or confusing graphics.

Key Concepts

- Variables can be either continuous or discrete.
- Frequency-distribution tables and diagrams are useful methods of summarizing large amounts of data.
- When the number of measured values is large, data are usually grouped into classes or intervals. This technique is particularly helpful with continuous variables.
- A frequency diagram shows the frequencies of values in each individual interval, while a cumulative-frequency diagram shows the running total of frequencies from the lowest interval up.
- A relative-frequency diagram shows the frequency of each interval as a proportion of the whole data set.
- Categorical data can be presented in various forms, including bar graphs, circle graphs (or pie charts), and pictographs.

Communicate Your Understanding

1. **a)** What information does a histogram present?

b) Explain why you cannot use categorical data in a histogram.

2. **a)** What is the difference between a frequency diagram and a cumulative-frequency diagram?

b) What are the advantages of each of these diagrams?

3. **a)** What is the difference between a frequency diagram and a relative-frequency diagram?

b) What information can be easily read from a frequency diagram?

c) What information can be easily read from a relative-frequency diagram?

4. Describe the strengths and weaknesses of circle graphs and pictographs.

Practise

A

1. Explain the problem with the intervals in each of the following tables.

a)

Age (years)	Frequency
28–32	6
33–38	8
38–42	11
42–48	9
48–52	4

b)

Score (%)	Frequency
61–65	5
66–70	11
71–75	7
76–80	4
91–95	1

2. Would you choose a histogram or a bar graph with separated bars for the data listed below? Explain your choices.
   a) the numbers from 100 rolls of a standard die
   b) the distances 40 athletes throw a shot-put
   c) the ages of all players in a junior lacrosse league
   d) the heights of all players in a junior lacrosse league

3. A catering service conducted a survey asking respondents to choose from six different hot meals.

Meal Chosen	Number
Chicken cordon bleu	16
New York steak	20
Pasta primavera (vegetarian)	9
Lamb chop	12
Grilled salmon	10
Mushroom stir-fry with almonds (vegetarian)	5

   a) Create a circle graph to illustrate these data.
   b) Use the circle graph to determine what percent of the people surveyed chose vegetarian dishes.
   c) Sketch a pictograph for the data.
   d) Use the pictograph to determine whether more than half of the respondents chose red-meat dishes.

4. a) Estimate the number of hours you spent each weekday on each of the following activities: eating, sleeping, attending class, homework, a job, household chores, recreation, other.
   b) Present this information using a circle graph.
   c) Present the information using a pictograph.

Apply, Solve, Communicate

5. The examination scores for a biology class are shown below.

68	77	91	66	52	58	79	94	81
60	73	57	44	58	71	78	80	54
87	43	61	90	41	76	55	75	49

   a) Determine the range for these data.
   b) Determine a reasonable interval size and number of intervals.
   c) Produce a frequency table for the grouped data.
   d) Produce a histogram and frequency polygon for the grouped data.
   e) Produce a relative-frequency polygon for the data.
   f) Produce a cumulative-frequency polygon for the data.
   g) What do the frequency polygon, the relative-frequency polygon, and the cumulative-frequency polygon each illustrate best?

B

6. **a)** Sketch a bar graph to show the results you would expect if you were to roll a standard die 30 times.

b) Perform the experiment or simulate it with software or the random-number generator of a graphing calculator. Record the results in a table.

c) Produce a bar graph for the data you collected.

d) Compare the bar graphs from a) and c). Account for any discrepancies you observe.

7. **Application** In order to set a reasonable price for a "bottomless" cup of coffee, a restaurant owner recorded the number of cups each customer ordered on a typical afternoon.

2	1	2	3	0	1	1	1	2	2
1	3	1	4	2	0	1	2	3	1

a) Would you present these data in a grouped or ungrouped format? Explain your choice.

b) Create a frequency table and diagram.

c) Create a cumulative-frequency diagram.

d) How can the restaurant owner use this information to set a price for a cup of coffee? What additional information would be helpful?

8. **Application** The list below shows the value of purchases, in dollars, by 30 customers at a clothing store.

55.40	48.26	28.31	14.12	88.90	34.45
51.02	71.87	105.12	10.19	74.44	29.05
43.56	90.66	23.00	60.52	43.17	28.49
67.03	16.18	76.05	45.68	22.76	36.73
39.92	112.48	81.21	56.73	47.19	34.45

a) Would you present these data in a grouped or ungrouped format? Explain your choice.

b) Create a frequency table and diagram.

c) Create a cumulative-frequency diagram.

d) How might the store owner use this information in planning sales promotions?

9. The speeds of 24 motorists ticketed for exceeding a 60-km/h limit are listed below.

75	72	66	80	75	70	71	82
69	70	72	78	90	75	76	80
75	96	91	77	76	84	74	79

a) Construct a frequency-distribution table for these data.

b) Construct a histogram and frequency polygon.

c) Construct a cumulative-frequency diagram.

d) How many of the motorists exceeded the speed limit by 15 km/h or less?

e) How many exceeded the speed limit by over 20 km/h?

10. **Communication** This table summarizes the salaries for François' hockey 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) Reorganize these data into appropriate intervals and present them in a frequency table.

b) Create a histogram for these data.

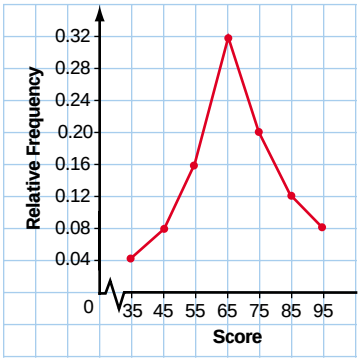
c) Identify and explain any unusual features about this distribution.

11. Communication

a) What is the sum of all the relative frequencies for any set of data?

b) Explain why this sum occurs.

12. The following relative-frequency polygon was constructed for the examination scores for a class of 25 students. Construct the frequency-distribution table for the students' scores.



13. Inquiry/Problem Solving The manager of a rock band suspects that MP3 web sites have reduced sales of the band's CDs. A survey of fans last year showed that at least 50% had purchased two or more of the band's CDs. A recent survey of 40 fans found they had purchased the following numbers of the band's CDs.

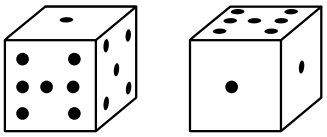
2	1	2	1	3	1	4	1	0	1
0	2	4	1	0	5	2	3	4	1
2	1	1	1	3	1	0	5	4	2
3	1	1	0	2	2	0	0	1	3

Does the new data support the manager's theory? Show the calculations you made to reach your conclusion, and illustrate the results with a diagram.

C

14. Inquiry/Problem Solving

a) What are the possible outcomes for a roll of two "funny dice" that have faces with the numbers 1, 1, 3, 5, 6, and 7?



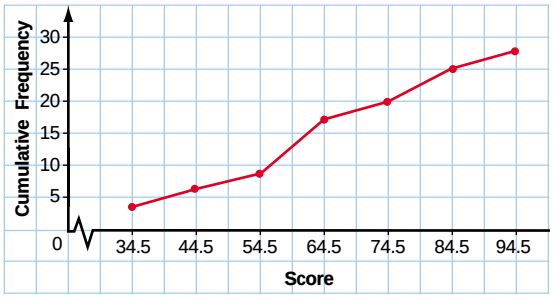
b) Sketch a relative-frequency polygon to show the results you would expect if these dice were rolled 100 times.

c) Explain why your graph has the shape it does.

d) Use software or a graphing calculator to simulate rolling the funny dice 100 times, and draw a relative-frequency polygon for the results.

e) Account for any differences between the diagrams in parts b) and d).

15. This cumulative-frequency diagram shows the distribution of the examination scores for a statistics class.



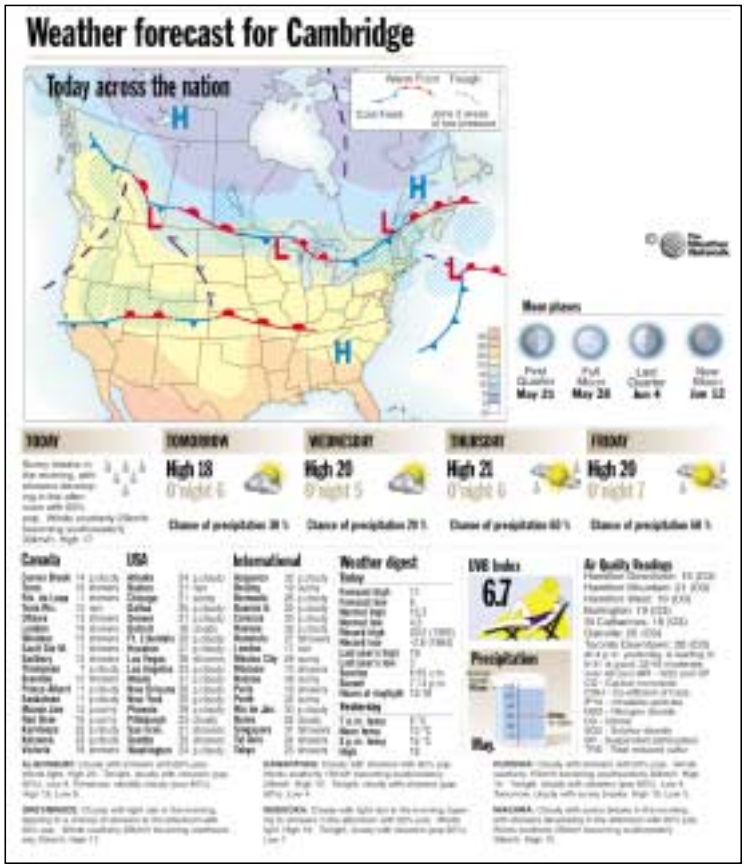
a) What interval contains the greatest number of scores? Explain how you can tell.

b) How many scores fall within this interval?

16. Predict the shape of the relative-frequency diagram for the examination scores of a first-year university calculus class. Explain why you chose the shape you did. Assume that students enrolled in a wide range of programs take this course. State any other assumptions that you need to make.

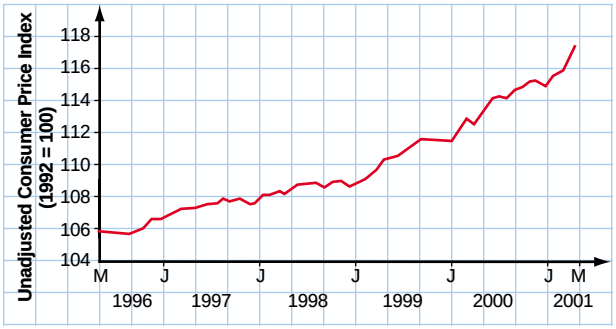
2.2 Indices

In the previous section, you used tables and graphs of frequencies to summarize data. Indices are another way to summarize data and recognize trends. An **index** relates the value of a variable (or group of variables) to a base level, which is often the value on a particular date. The base level is set so that the index produces numbers that are easy to understand and compare. Indices are used to report on a wide variety of variables, including prices and wages, ultraviolet levels in sunlight, and even the readability of textbooks.



INVESTIGATE & INQUIRE: Consumer Price Index

The graph below shows Statistics Canada's consumer price index (CPI), which tracks the cost of over 600 items that would be purchased by a typical family in Canada. For this chart, the base is the cost of the same items in 1992.



1. What trend do you see in this graph? Estimate the annual rate of increase.
2. Estimate the annual rate of increase for the period from 1992 to 1996. Do you think the difference between this rate and the one from 1996 to 2001 is significant? Why or why not?
3. What was the index value in February of 1998? What does this value tell you about consumer prices at that time?

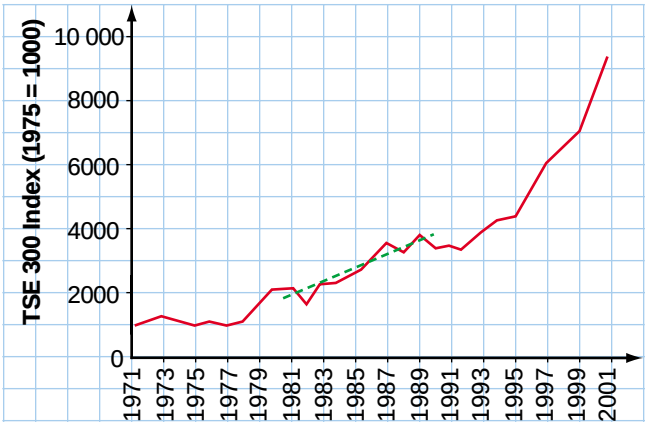
4. What would be the best way to estimate what the consumer price index will be in May of 2003? Explain your reasoning.
5. Explain how the choice of the vertical scale in the graph emphasizes changes in the index. Do you think this emphasis could be misleading? Why or why not?

The best-known Canadian business index is the S&P/TSX Composite Index, managed for the Toronto Stock Exchange by Standard & Poor's Corporation. Introduced in May, 2002, this index is a continuation of the TSE 300 Composite Index®, which goes back to 1977. The S&P/TSX Composite Index is a measure of the total market value of the shares of over 200 of the largest companies traded on the Toronto Stock Exchange. The index is the current value of these stocks divided by their total value in a base year and then multiplied by a scaling factor. When there are significant changes (such as takeovers or bankruptcies) in any of the companies in the index, the scaling factor is adjusted so that the values of the index remain directly comparable to earlier values. Note that the composite index weights each company by the total value of its shares (its market capitalization) rather than by the price of the individual shares. The S&P/TSX Composite Index usually indicates trends for major Canadian corporations reasonably well, but it does not always accurately reflect the overall Canadian stock market.

Time-series graphs are often used to show how indices change over time. Such graphs plot variable values versus time and join the adjacent data points with straight lines.

Example 1 Stock Market Index

The following table shows the TSE 300 Composite Index® from 1971 to 2001.



a) What does the notation "1975 = 1000" mean?

b) By what factor did the index grow over the period shown?

c) Estimate the rate of growth of the index during the 1980s.

Solution

a) The notation indicates that the index shows the stock prices *relative* to what they were in 1975. This 1975 base has been set at 1000. An index value of 2000 would mean that overall the stocks of the 300 companies in the index are selling for twice what they did in 1975.

b) From the graph, you can see that the index increased from about 1000 in 1971 to about 10 000 in 2001. Thus, the index increased by a factor of approximately 10 over this period.

c) To estimate the rate of growth of the index during the 1980s, approximate the time-series graph with a straight line during that 10-year interval. Then, calculate the slope of the line.

$$m = \frac{\text{rise}}{\text{run}}$$
$$\doteq \frac{3700 - 1700}{10}$$
$$= 200$$

The TSE 300 Composite Index® rose about 200 points a year during the 1980s.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

For more information on stock indices, visit the above web site and follow the links. Write a brief description of the rules for inclusion in the various market indices.

Statistics Canada calculates a variety of carefully researched economic indices. For example, there are price indices for new housing, raw materials, machinery and equipment, industrial products, and farm products. Most of these indices are available with breakdowns by province or region and by specific categories, such as agriculture, forestry, or manufacturing. Statisticians, economists, and the media make extensive use of these indices. (See section 1.3 for information on how to access Statistics Canada data.)

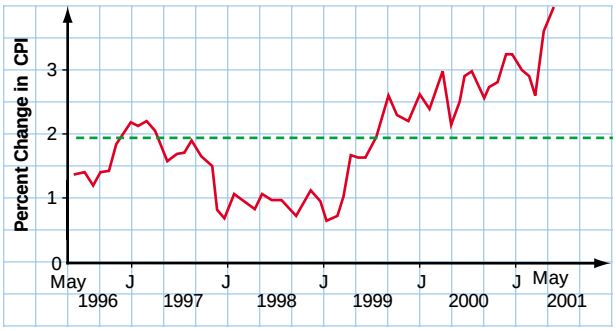
The **consumer price index (CPI)** is the most widely reported of these economic indices because it is an important measure of inflation. **Inflation** is a general increase in prices, which corresponds to a decrease in the value of money. To measure the average change in retail prices across Canada, Statistics Canada monitors the retail prices of a set of over 600 goods and services including food, shelter, clothing, transportation, household items, health and personal care, recreation and education, and alcohol and tobacco products. These items are representative of purchases by typical Canadians and are weighted according to estimates of the total amount Canadians spend on each item. For example, milk has a weighting of 0.69% while tea has a weighting of only 0.06%.

Data in Action

Statistics Canada usually publishes the consumer price index for each month in the third week of the following month. Over 60 000 price quotations are collected for each update.

Example 2 Consumer Price Index

The following graph shows the amount by which the consumer price index changed since the same month of the previous year.



a) What does this graph tell you about changes in the CPI from 1996 to 2001?

b) Estimate the mean annual change in the CPI for this period.

Solution

a) Note that the graph above shows the annual changes in the CPI, unlike the graph on page 104, which illustrates the value of the CPI for any given month. From the above graph, you can see that the annual change in the CPI varied between 0.5% and 4% from 1996 to 2001. Overall, there is an upward trend in the annual change during this period.

b) You can estimate the mean annual change by drawing a horizontal line such that the total area between the line and the parts of the curve above it is approximately equal to the total area between the line and the parts of the curve below it. As shown above, this line meets the y-axis near 2%.

Thus, the mean annual increase in the CPI was roughly 2% from 1996 to 2001.

Project Prep

If your statistics project examines how a variable changes over time, a time-series graph may be an effective way to illustrate your findings.

The consumer price index and the cost of living index are not quite the same. The cost of living index measures the cost of maintaining a constant standard of living. If consumers like two similar products equally well, their standard of living does not change when they switch from one to the other. For example, if you like both apples and pears, you might start buying more apples and fewer pears if the price of pears went up while the price of apples was unchanged. Thus, your cost of living index increases less than the consumer price index does.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

For more information about Statistics Canada indices, visit the above web site and follow the links to Statistics Canada.

Indices are also used in many other fields, including science, sociology, medicine, and engineering. There are even indices of the clarity of writing.

Example 3 Readability Index

The Gunning fog index is a measure of the readability of prose. This index estimates the years of schooling required to read the material easily.

Gunning fog index = 0.4(average words per sentence + percent "hard" words)

where "hard" words are all words over two syllables long except proper nouns, compounds of easy words, and verbs whose third syllable is *ed* or *es*.

a) Calculate the Gunning fog index for a book with an average sentence length of 8 words and a 20% proportion of hard words.

b) What are the advantages and limitations of this index?

Solution

a) Gunning fog index = 0.4(8 + 20)
= 11.2

The Gunning fog index shows that the book is written at a level appropriate for readers who have completed grade 11.

b) The Gunning fog index is easy to use and understand. It generates a grade-level rating, which is often more useful than a readability rating on an arbitrary scale, such as 1 to 10 or 1 to 100. However, the index assumes that bigger words and longer sentences always make prose harder to read. A talented writer could use longer words and sentences and still be more readable than writers who cannot clearly express their ideas. The Gunning fog index cannot, of course, evaluate literary merit.

Project Prep

You may want to use an index to summarize and compare sets of data in your statistics project.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

Visit the above web site to find a link to a readability-index calculator. Determine the reading level of a novel of your choice.

Key Concepts

- An index can summarize a set of data. Indices usually compare the values of a variable or group of variables to a base value.
- Indices have a wide variety of applications in business, economics, science, and other fields.
- A time-series graph is a line graph that shows how a variable changes over time.
- The consumer price index (CPI) tracks the overall price of a representative basket of goods and services, making it a useful measure of inflation.

Communicate Your Understanding

1. What are the key features of a time-series graph?

2. **a)** Name three groups who would be interested in the new housing price index.
b) How would this information be important for each group?

3. Explain why the consumer price index is not the same as the cost of living index.

Practise

A

1. Refer to the consumer price index graph on page 104.
a) By how many index points did the CPI increase from January, 1992 to January, 2000?
b) Express this increase as a percent.
c) Estimate what an item that cost
i) $7.50 in 1992 cost in April, 1998
ii) $55 in August, 1997 cost in May, 2000

Apply, Solve, Communicate

2. **a)** Explain why there is a wide variety of items in the CPI basket.
b) Is the percent increase for the price of each item in the CPI basket the same? Explain.

B

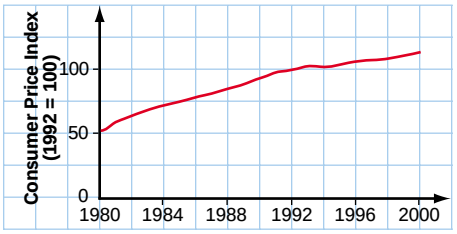
3. Refer to the graph of the TSE 300 Composite Index® on page 105.
a) When did this index first reach five times its base value?
b) Estimate the growth rate of the index from 1971 to 1977. What does this growth rate suggest about the Canadian economy during this period?
c) During what two-year period did the index grow most rapidly? Explain your answer.
d) Could a straight line be a useful mathematical model for the TSE 300 Composite Index®? Explain why or why not.

4. Communication
a) Define inflation.
b) In what way do the consumer price index and the new housing price index provide a measure of inflation?

c) How would you expect these two indices to be related?

d) Why do you think that they would be related in this way?

5. **Application** Consider the following time-series graph for the consumer price index.



a) Identify at least three features of this graph that are different from the CPI graph on page 104.

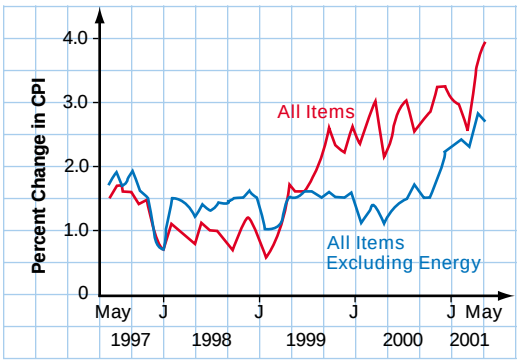
b) Explain two advantages that the graph shown here has over the one on page 104.

c) Explain two disadvantages of the graph shown here compared to the one on page 104.

d) Estimate the year in which the CPI was at 50.

e) Explain the significance of the result in part d) in terms of prices in 1992.

6. **Application** The following graph illustrates the CPI both with and without energy price changes.



a) How is this graph different from the one on page 107?

b) Describe how the overall trend in energy costs compares to that of the CPI for the period shown.

c) What insight is gained by removing the energy component of the CPI?

d) Estimate the overall increase in the energy-adjusted CPI for the period shown.

e) Discuss how your result in part d) compares to the value found in part b) of Example 2.

7. François' agent wants to bargain for a better salary based on François' statistics for his first five seasons with the team.

Chapter Problem

a) Produce a time-series graph for François' goals, assists, and points over the past five years.

b) Calculate the mean number of goals, assists, and points per game played during each of François' five seasons.

c) Generate a new time-series graph based on the data from part b).

d) Which time-series graph will the agent likely use, and which will the team's manager likely use during the contract negotiations? Explain.

e) Explain the method or technology that you used to answer parts a) to d).

8. Aerial surveys of wolves in Algonquin Park produced the following estimates of their population density.

Year	Wolves/100 km²
1988–89	4.91
1989–90	2.47
1990–91	2.80
1991–92	3.62
1992–93	2.53
1993–94	2.23
1994–95	2.82
1995–96	2.75
1996–97	2.33
1997–98	3.04
1998–99	1.59

a) Using 1988–89 as a base, construct an index for these data.

b) Comment on any trends that you observe.

9. Use Statistics Canada web sites or other sources to find statistics for the following and describe any trends you notice.

a) the population of Canada

b) the national unemployment rate

c) the gross domestic product

10. Inquiry / Problem Solving

a) Use data from E-STAT or other sources to generate a time-series graph that shows the annual number of crimes in Canada for the period 1989–1999. If using E-STAT, look in the Nation section under Justice / Crimes and Offences.

b) Explain any patterns that you notice.

c) In what year did the number of crimes peak?

d) Suggest possible reasons why the number of crimes peaked in that year. What other statistics would you need to confirm whether these reasons are related to the peak in the number of crimes?

11. a) Use data from E-STAT or other sources to generate a time-series graph that shows the number of police officers in Canada for the period 1989–1999. If using E-STAT, look in the Nation section under Justice / Police services.

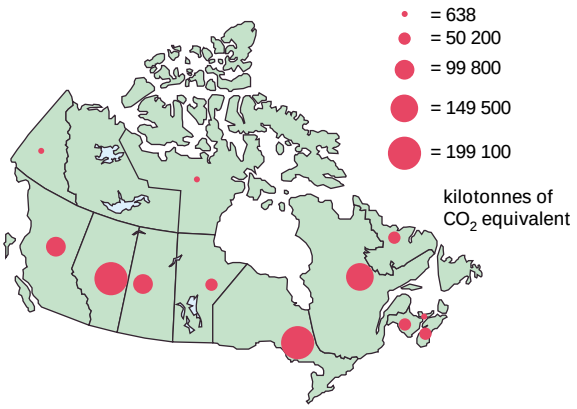
b) In what ways are the patterns in these data similar to the patterns in the data in question 10? In what ways are the patterns different?

c) In what year did the number of police officers peak?

d) Explain how this information could affect your answer to part d) of question 10.

12. Communication Use the Internet, a library, or other resources to research two indices not discussed in this section. Briefly describe what each index measures, recent trends in the index, and any explanation or rationale for these trends.

13. Inquiry / Problem Solving The pictograph below shows total greenhouse-gas emissions for each province and territory in 1996.



a) Which two provinces have the highest levels of greenhouse-gas emissions?

b) Are the diameters or areas of the circles proportional to the numbers they represent? Justify your answer.

c) What are the advantages and disadvantages of presenting these data as a pictograph?

d) Which provinces have the highest levels of greenhouse-gas emissions per geographic area?

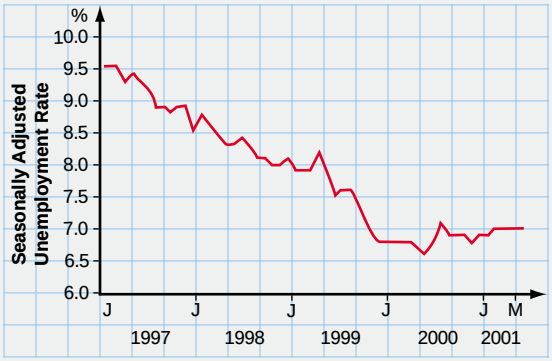
e) Is your answer to part d) what you would have expected? How can you account for such relatively high levels in these areas?

f) Research information from E-STAT or other sources to determine the greenhouse-gas emissions per person for each province.

ACHIEVEMENT CHECK

Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application

14. The graph below shows the national unemployment rate from January, 1997, to June, 2001.



a) Describe the overall trend for the period shown.

b) When did the unemployment rate reach its lowest level?

c) Estimate the overall unemployment rate for the period shown.

d) Explain what the term *seasonally adjusted* means.

e) Who is more likely to use this graph in an election campaign, the governing party or an opposing party? Explain.

f) How might an opposing party produce a graph showing rising unemployment without changing the data? Why would they produce such a graph?

C

15. A *Pareto chart* is a type of frequency diagram in which the frequencies for categorical data are shown by connected bars arranged in descending order of frequency. In a random survey, commuters listed their most common method of travelling to the downtown of a large city.

a) Construct a Pareto chart for these data.

b) Describe the similarities and differences between a Pareto chart and other frequency diagrams.

Method	Number of Respondents
Automobile: alone	26
Automobile: car pool	35
Bus/Streetcar	52
Train	40
Bicycle/Walking	13

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

For more information about Pareto charts, visit the above web site and follow the links. Give two examples of situations where you would use a Pareto chart. Explain your reasoning.

16. Pick five careers of interest to you.

a) Use resources such as CANSIM II, E-STAT, newspapers, or the Internet to obtain information about entry-level income levels for these professions.

b) Choose an effective method to present your data.

c) Describe any significant information you discovered.

17. a) Research unemployment data for Ontario over the past 20 years.

b) Present the data in an appropriate form.

c) Conduct additional research to account for any trends or unusual features of the data.

d) Predict unemployment trends for both the short term and the long term. Explain your predictions.

2.3 Sampling Techniques

Who will win the next federal election? Are Canadians concerned about global warming? Should a Canadian city bid to host the next Olympic Games? Governments, political parties, advocacy groups, and news agencies often want to know the public's opinions on such questions. Since it is not feasible to ask every citizen directly, researchers often survey a much smaller group and use the results to estimate the opinions of the entire population.

INVESTIGATE & INQUIRE: Extrapolating From a Sample

1. Work in groups or as a class to design a survey to determine the opinions of students in your school on a subject such as favourite movies, extra-curricular activities, or types of music.
2. Have everyone in your class answer the survey.
3. Decide how to categorize and record the results. Could you refine the survey questions to get results that are easier to work with? Explain the changes you would make.
4. How could you organize and present the data to make it easier to recognize any patterns? Can you draw any conclusions from the data?
5. a) Extrapolate your data to estimate the opinions of the entire school population. Explain your method.
   b) Describe any reasons why you think the estimates in part a) may be inaccurate.
   c) How could you improve your survey methods to get more valid results?

In statistics, the term **population** refers to all individuals who belong to a group being studied. In the investigation above, the population is all the students in your school, and your class is a **sample** of that population. The population for a statistical study depends on the kind of data being collected.

Example 1 Identifying a Population

Identify the population for each of the following questions.

a) Whom do you plan to vote for in the next Ontario election?

b) What is your favourite type of baseball glove?

c) Do women prefer to wear ordinary glasses or contact lenses?

Solution

a) The population consists of those people in Ontario who will be eligible to vote on election day.

b) The population would be just those people who play baseball. However, you might want to narrow the population further. For example, you might be interested only in answers from local or professional baseball players.

c) The population is all women who use corrective lenses.

Once you have identified the population, you need to decide how you will obtain your data. If the population is small, it may be possible to survey the entire group. For larger populations, you need to use an appropriate sampling technique. If selected carefully, a relatively small sample can give quite accurate results.

The group of individuals who actually have a chance of being selected is called the **sampling frame.** The sampling frame varies depending on the sampling technique used. Here are some of the most commonly used sampling techniques.

Simple Random Sample

In a **simple random sample**, every member of the population has an equal chance of being selected and the selection of any particular individual does not affect the chances of any other individual being chosen. Choosing the sample randomly reduces the risk that selected members will not be representative of the whole population. You could select the sample by drawing names randomly or by assigning each member of the population a unique number and then using a random-number generator to determine which members to include.

Systematic Sample

For a **systematic sample**, you go through the population sequentially and select members at regular intervals. The sample size and the population size determine the sampling interval.

$$\text{interval} = \frac{\text{population size}}{\text{sample size}}$$

For example, if you wanted the sample to be a tenth of the population, you would select every tenth member of the population, starting with one chosen randomly from among the first ten in sequence.

Example 2 Designing a Systematic Sample

A telephone company is planning a marketing survey of its 760 000 customers. For budget reasons, the company wants a sample size of about 250.

a) Suggest a method for selecting a systematic sample.

b) What expense is most likely to limit the sample size?

Solution

a) First, determine the sampling interval.

$$\text{interval} = \frac{\text{population size}}{\text{sample size}}$$
$$= \frac{760\,000}{250}$$
$$= 3040$$

The company could randomly select one of the first 3040 names on its list of customers and then choose every 3040th customer from that point on. For simplicity, the company might choose to select every 3000th customer instead.

b) The major cost is likely to be salaries for the staff to call and interview the customers.

Stratified Sample

Sometimes a population includes groups of members who share common characteristics, such as gender, age, or education level. Such groups are called **strata**. A **stratified sample** has the same proportion of members from each stratum as the population does.

Example 3 Designing a Stratified Sample

Before booking bands for the school dances, the students' council at Statsville High School wants to survey the music preferences of the student body. The following table shows the enrolment at the school.

Grade	Number of Students
9	255
10	232
11	209
12	184
Total	880

a) Design a stratified sample for a survey of 25% of the student body.

b) Suggest other ways to stratify this sample.

Solution

a) To obtain a stratified sample with the correct proportions, simply select 25% of the students in each grade level as shown on the right.

Grade	Number of Students	Relative Frequency	Number Surveyed
9	255	0.29	64
10	232	0.26	58
11	209	0.24	52
12	184	0.21	46
Total	880	1.00	220

b) The sample could be stratified according to gender or age instead of grade level.

Other Sampling Techniques

Cluster Sample: If certain groups are likely to be representative of the entire population, you can use a random selection of such groups as a **cluster sample**. For example, a fast-food chain could save time and money by surveying all its employees at randomly selected locations instead of surveying randomly selected employees throughout the chain.

Multi-Stage Sample: A **multi-stage sample** uses several levels of random sampling. If, for example, your population consisted of all Ontario households, you could first randomly sample from all cities and townships in Ontario, then randomly sample from all subdivisions or blocks within the selected cities and townships, and finally randomly sample from all houses within the selected subdivisions or blocks.

Voluntary-Response Sample: In a **voluntary-response sample,** the researcher simply invites any member of the population to participate in the survey. The results from the responses of such surveys can be skewed because the people who choose to respond are often not representative of the population. Call-in shows and mail-in surveys rely on voluntary-response samples.

Convenience Sample: Often, a sample is selected simply because it is easily accessible. While obviously not as random as some of the other techniques, such convenience samples can sometimes yield helpful information. The investigation at the beginning of this section used your class as a convenience sample.

Key Concepts

- A carefully selected sample can provide accurate information about a population.
- Selecting an appropriate sampling technique is important to ensure that the sample reflects the characteristics of the population. Randomly selected samples have a good chance of being representative of the population.
- The choice of sampling technique will depend on a number of factors, such as the nature of the population, cost, convenience, and reliability.

Communicate Your Understanding

1. What are the advantages and disadvantages of using a sample to estimate the characteristics of a population?
2. Discuss whether a systematic sample is a random sample.
3. a) Explain the difference between stratified sampling and cluster sampling.
   b) Suggest a situation in which it would be appropriate to use each of these two sampling techniques.

Practise

A

1. Identify the population for each of the following questions.
   a) Who should be the next president of the students' council?
   b) Who should be next year's grade-10 representative on the student council?
   c) What is the your favourite soft drink?
   d) Which Beatles song was the best?
   e) How effective is a new headache remedy?
2. Classify the sampling method used in each of the following scenarios.
   a) A radio-show host invites listeners to call in with their views on banning smoking in restaurants.
   b) The Heritage Ministry selects a sample of recent immigrants such that the proportions from each country of origin are the same as for all immigrants last year.
   c) A reporter stops people on a downtown street to ask what they think of the city's lakefront.
   d) A school guidance counsellor arranges interviews with every fifth student on the alphabetized attendance roster.
   e) A statistician conducting a survey randomly selects 20 cities from across Canada, then 5 neighbourhoods from each of the cities, and then 3 households from each of the neighbourhoods.
   f) The province randomly chooses 25 public schools to participate in a new fundraising initiative.
3. What type(s) of sample would be appropriate for
   a) a survey of engineers, technicians, and managers employed by a company?
   b) determining the most popular pizza topping?
   c) measuring customer satisfaction for a department store?

Apply, Solve, Communicate

B

4. Natasha is organizing the annual family picnic and wants to arrange a menu that will appeal to children, teens, and adults. She estimates that she has enough time to survey about a dozen people. How should Natasha design a stratified sample if she expects 13 children, 8 teens, and 16 adults to attend the picnic?

5. **Communication** Find out, or estimate, how many students attend your school. Describe how you would design a systematic sample of these students. Assume that you can survey about 20 students.

6. The newly elected Chancellor of the Galactic Federation is interested in the opinions of all citizens regarding economic conditions in the galaxy. Unfortunately, she does not have the resources to visit every populated planet or to send delegates to them. Describe how the Chancellor might organize a multi-stage sample to carry out her survey.

7. **Communication** A community centre chooses 15 of its members at random and asks them to have each member of their families complete a short questionnaire.

a) What type of sample is the community centre using?

b) Are the 15 community-centre members a random sample of the community? Explain.

c) To what extent are the family members randomly chosen?

8. **Application** A students' council is conducting a poll of students as they enter the cafeteria.

a) What sampling method is the student council using?

b) Discuss whether this method is appropriate for surveying students' opinions on

i) the new mural in the cafeteria

ii) the location for the graduation prom

c) Would another sampling technique be better for either of the surveys in part b)?

9. **Application** The host of a call-in program invites listeners to comment on a recent trade by the Toronto Maple Leafs. One caller criticizes the host, stating that the sampling technique is not random. The host replies: "So what? It doesn't matter!"

a) What sampling technique is the call-in show using?

b) Is the caller's statement correct? Explain.

c) Is the host's response mathematically correct? Why or why not?

C

10. Look in newspapers and periodicals or on the Internet for an article about a study involving a systematic, stratified, cluster, or multi-stage sample. Comment on the suitability of the sampling technique and the validity of the study. Present your answer in the form of a brief report. Include any suggestions you have for improving the study.

11. **Inquiry/Problem Solving** Design a data-gathering method that uses a combination of convenience and systematic sampling techniques.

12. **Inquiry/Problem Solving** Pick a professional sport that has championship playoffs each year.

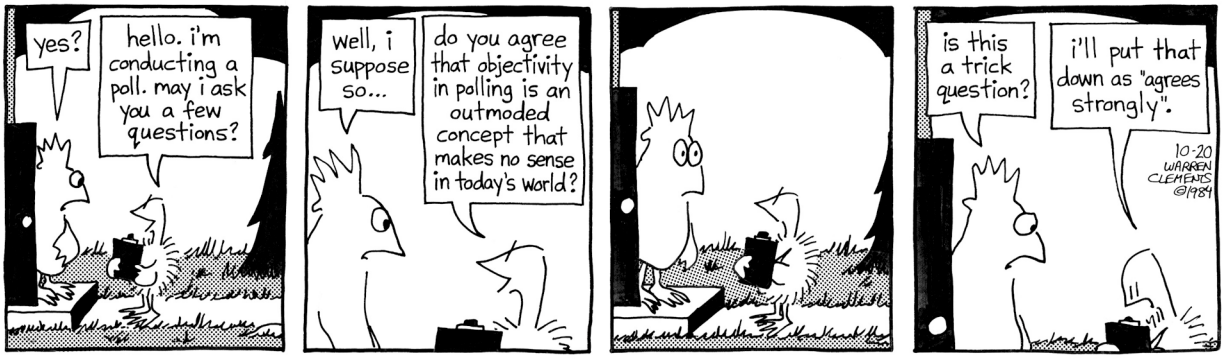
a) Design a multi-stage sample to gather your schoolmates' opinions on which team is likely to win the next championship.

b) Describe how you would carry out your study and illustrate your findings.

c) Research the media to find what the professional commentators are predicting. Do you think these opinions would be more valid than the results of your survey? Why or why not?

2.4 Bias in Surveys

The results of a survey can be accurate only if the sample is representative of the population and the measurements are objective. The methods used for choosing the sample and collecting the data must be free from **bias**. Statistical bias is any factor that favours certain outcomes or responses and hence systematically skews the survey results. Such bias is often unintentional. A researcher may inadvertently use an unsuitable method or simply fail to recognize a factor that prevents a sample from being fully random. Regrettably, some people deliberately bias surveys in order to get the results they want. For this reason, it is important to understand not only how to use statistics, but also how to recognize the misuse of statistics.



INVESTIGATE & INQUIRE: Bias in a Survey

1. What sampling technique is the pollster in this cartoon likely to be using?
2. What is wrong with his survey methods? How could he improve them?
3. Do you think the bias in this survey is intentional? Why or why not?
4. Will this bias seriously distort the results of the survey? Explain your reasoning.
5. What point is the cartoonist making about survey methods?
6. Sketch your own cartoon or short comic strip about data management.

Sampling bias occurs when the sampling frame does not reflect the characteristics of the population. Biased samples can result from problems with either the sampling technique or the data-collection method.

Example 1 Sampling Bias

Identify the bias in each of the following surveys and suggest how it could be avoided.

a) A survey asked students at a high-school football game whether a fund for extra-curricular activities should be used to buy new equipment for the football team or instruments for the school band.

b) An aid agency in a developing country wants to know what proportion of households have at least one personal computer. One of the agency's staff members conducts a survey by calling households randomly selected from the telephone directory.

Solution

a) Since the sample includes only football fans, it is not representative of the whole student body. A poor choice of sampling technique makes the results of the survey invalid. A random sample selected from the entire student body would give unbiased results.

b) There could be a significant number of households without telephones. Such households are unlikely to have computers. Since the telephone survey excludes these households, it will overestimate the proportion of households that have computers. By using a telephone survey as the data-collection method, the researcher has inadvertently biased the sample. Visiting randomly selected households would give a more accurate estimate of the proportion that have computers. However, this method of data collection would be more time-consuming and more costly than a telephone survey.

Non-response bias occurs when particular groups are under-represented in a survey because they choose not to participate. Thus, non-response bias is a form of sampling bias.

Example 2 Non-Response Bias

A science class asks every fifth student entering the cafeteria to answer a survey on environmental issues. Less than half agree to complete the questionnaire. The completed questionnaires show that a high proportion of the respondents are concerned about the environment and well-informed about environmental issues. What bias could affect these results?

Solution

The students who chose not to participate in the survey are likely to be those least interested in environmental issues. As a result, the sample may not be representative of all the students at the school.

To avoid non-response bias, researchers must ensure that the sampling process is truly random. For example, they could include questions that identify members of particular groups to verify that they are properly represented in the sample.

Measurement bias occurs when the data-collection method consistently either under- or overestimates a characteristic of the population. While random errors tend to cancel out, a consistent measurement error will skew the results of a survey. Often, measurement bias results from a data-collection process that affects the variable it is measuring.

Example 3 Measurement Bias

Identify the bias in each of the following surveys and suggest how it could be avoided.

a) A highway engineer suggests that an economical way to survey traffic speeds on an expressway would be to have the police officers who patrol the highway record the speed of the traffic around them every half hour.

b) As part of a survey of the "Greatest Hits of All Time," a radio station asks its listeners: Which was the best song by the Beatles?

i) Help! **ii)** Nowhere Man

iii) In My Life **iv)** Other:

c) A poll by a tabloid newspaper includes the question: "Do you favour the proposed bylaw in which the government will dictate whether you have the right to smoke in a restaurant?"

Solution

a) Most drivers who are speeding will slow down when they see a police cruiser. A survey by police cruisers would underestimate the average traffic speed. Here, the data-collection method would systematically decrease the variable it is measuring. A survey by unmarked cars or hidden speed sensors would give more accurate results.

b) The question was intended to remind listeners of some of the Beatles' early recordings that might have been overshadowed by their later hits. However, some people will choose one of the suggested songs as their answer even though they would not have thought of these songs without prompting. Such **leading questions** usually produce biased results. The survey would more accurately determine listeners' opinions if the question did not include any suggested answers.

c) This question distracts attention from the real issue, namely smoking in restaurants, by suggesting that the government will infringe on the respondents' rights. Such **loaded questions** contain wording or information intended to influence the respondents' answers. A question with straightforward neutral language will produce more accurate data. For example, the question could read simply: "Should smoking in restaurants be banned?"

Response bias occurs when participants in a survey deliberately give false or misleading answers. The respondents might want to influence the results unduly, or they may simply be afraid or embarrassed to answer sensitive questions honestly.

Project Prep

When gathering data for your statistics project, you will need to ensure that the sampling process is free from bias.

Example 4 Response Bias

A teacher has just explained a particularly difficult concept to her class and wants to check that all the students have grasped this concept. She realizes that if she asks those who did not understand to put up their hands, these students may be too embarrassed to admit that they could not follow the lesson. How could the teacher eliminate this response bias?

Solution

The teacher could say: "This material is very difficult. Does anyone want me to go over it again?" This question is much less embarrassing for students to answer honestly, since it suggests that it is normal to have difficulty with the material. Better still, she could conduct a survey that lets the students answer anonymously. The teacher could ask the students to rate their understanding on a scale of 1 to 5 and mark the ratings on slips of paper, which they would deposit in a box. The teacher can then use these ballots to decide whether to review the challenging material at the next class.

As the last two examples illustrate, careful wording of survey questions is essential for avoiding bias. Researchers can also use techniques such as follow-up questions and guarantees of anonymity to eliminate response bias. For a study to be valid, all aspects of the sampling process must be free from bias.

Key Concepts

- Sampling, measurement, response, and non-response bias can all invalidate the results of a survey.
- Intentional bias can be used to manipulate statistics in favour of a certain point of view.
- Unintentional bias can be introduced if the sampling and data-collection methods are not chosen carefully.
- Leading and loaded questions contain language that can influence the respondents' answers.

Communicate Your Understanding

1. Explain the difference between a measurement bias and a sampling bias.
2. Explain how a researcher could inadvertently bias a study.
3. Describe how each of the following might use intentional bias
   - **a)** the media
   - **b)** a marketing department
   - **c)** a lobby group

Practise

A

1. Classify the bias in each of the following scenarios.
   - **a)** Members of a golf and country club are polled regarding the construction of a highway interchange on part of their golf course.
   - **b)** A group of city councillors are asked whether they have ever taken part in an illegal protest.
   - **c)** A random poll asks the following question: "The proposed casino will produce a number of jobs and economic activity in and around your city, and it will also generate revenue for the provincial government. Are you in favour of this forward-thinking initiative?"
   - **d)** A survey uses a cluster sample of Toronto residents to determine public opinion on whether the provincial government should increase funding for the public transit.

Apply, Solve, Communicate

2. For each scenario in question 1, suggest how the survey process could be changed to eliminate bias.
3. Communication Reword each of the following questions to eliminate the measurement bias.
   - **a)** In light of the current government's weak policies, do you think that it is time for a refreshing change at the next federal election?
   - **b)** Do you plan to support the current government at the next federal election, in order that they can continue to implement their effective policies?
   - **c)** Is first-year calculus as brutal as they say?
   - **d)** Which of the following is your favourite male movie star?
     - **i)** Al Pacino
     - **ii)** Keanu Reeves
     - **iii)** Robert DeNiro
     - **iv)** Jack Nicholson
     - **v)** Antonio Banderas
     - **vi)** Other:
   - **e)** Do you think that fighting should be eliminated from professional hockey so that skilled players can restore the high standards of the game?

B

4. Communication
   - **a)** Write your own example of a leading question and a loaded question.
   - **b)** Write an unbiased version for each of these two questions.

ACHIEVEMENT CHECK

Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application

5. A school principal wants to survey data-management students to determine whether having computer Internet access at home improves their success in this course.

a) What type of sample would you suggest? Why? Describe a technique for choosing the sample.

b) The following questions were drafted for the survey questionnaire. Identify any bias in the questions and suggest a rewording to eliminate the bias.

i) Can your family afford high-speed Internet access?

ii) Answer the question that follows your mark in data management.
Over 80%: How many hours per week do you spend on the Internet at home?
60–80%: Would home Internet access improve your mark in data management ?
Below 60%: Would increased Internet access at school improve your mark in data management?

c) Suppose the goal is to convince the school board that every data-management student needs daily access to computers and the Internet in the classroom. How might you alter your sampling technique to help achieve the desired results in this survey? Would these results still be statistically valid?

6. **Application** A talk-show host conducts an on-air survey about re-instituting capital punishment in Canada. Six out of ten callers voice their support for capital punishment. The next day, the host claims that 60% of Canadians are in favour of capital punishment. Is this claim statistically valid? Explain your reasoning.

C

7. **a)** Locate an article from a newspaper, periodical, or Internet site that involves a study that contains bias.

b) Briefly describe the study and its findings.

c) Describe the nature of the bias inherent in the study.

d) How has this bias affected the results of the study?

e) Suggest how the study could have eliminated the bias.

8. **Inquiry / Problem Solving** Do you think that the members of Parliament are a representative sample of the population? Why or why not?

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, μ, 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.

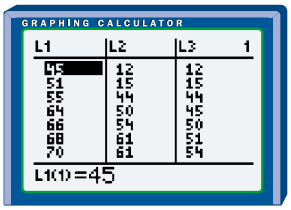
The median is often a better measure of central tendency than the mean for small data sets that contain outliers. For larger data sets, the effect of outliers on the mean is less significant.

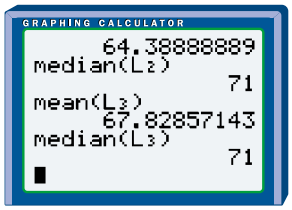
Example 2 Comparing Samples to a Population

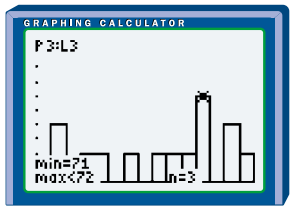
Compare the measures of central tendency for each class in Example 1 to those for all the students who wrote the physics examination.

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**.



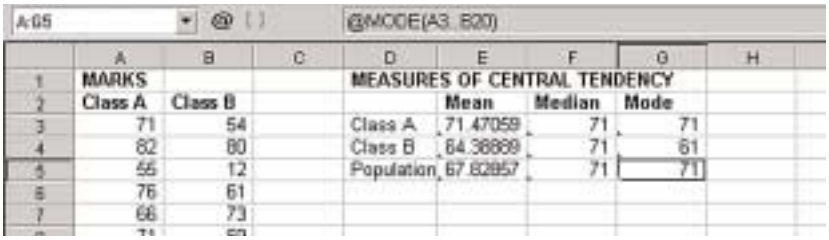




Note that the mean for class A overestimates the population mean, while the mean for class B underestimates it. The measures of central tendency for class A are reasonably close to those for the whole population of students who wrote the physics examination, but the two sets of measures are not identical. 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						

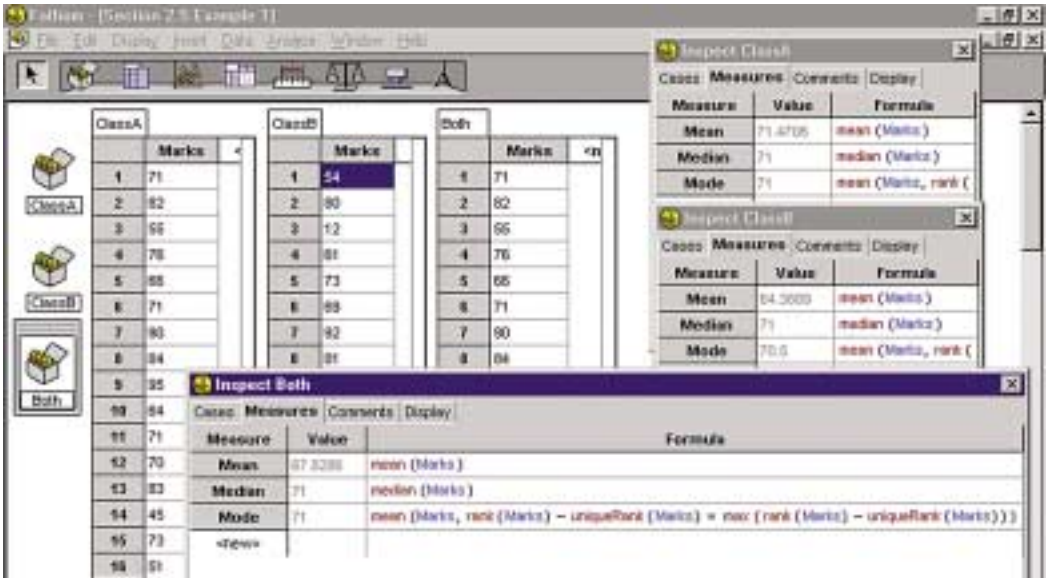
Solution 3 Using Fathom™

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% Thinking, inquiry, problem solving (TIPS) 25% Communication (C) 15% Application (A) 25%	Term mark 70% 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.

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

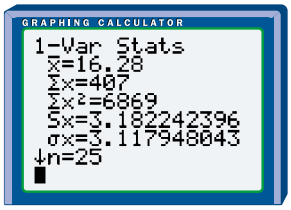
A veterinarian has collected data on the life spans of a rare breed of cats.

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

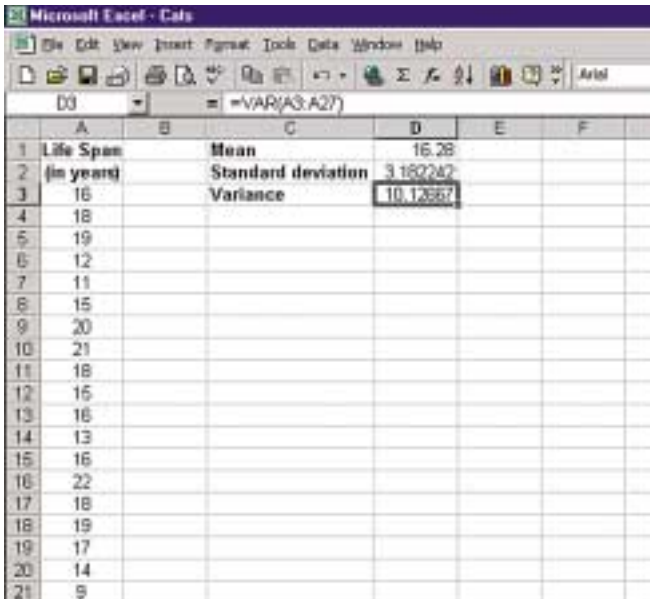
Use the ClrList command to make sure list L1 is clear, then enter the data into it. Use the 1-Var Stats command from the STAT CALC menu to calculate a set of statistics including the mean and the standard deviation. Note that the calculator displays both a sample standard deviation, Sx, and a population 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.

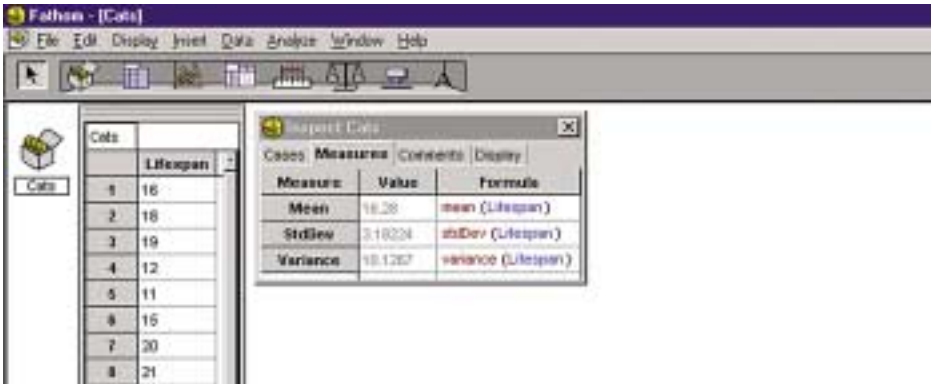
Solution 2 Using a Spreadsheet

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,

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

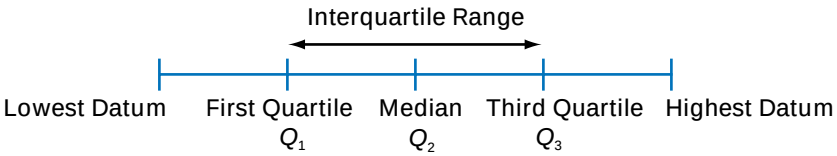
where f_i is the frequency for a given interval and m_i is the midpoint of the interval. However, calculating standard deviations from raw, ungrouped data will give more accurate results.

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

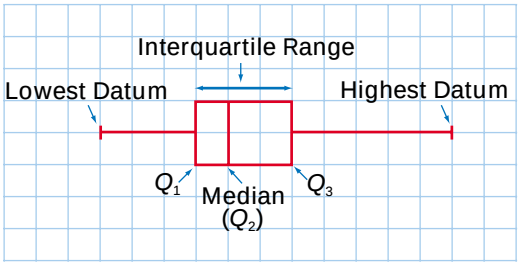
Quartiles divide a set of ordered data into four groups with equal numbers of values, just as the median divides data into two equally sized groups. The three "dividing points" are the first quartile (Q_1), the median (sometimes called the second quartile or Q_2), and the third quartile (Q_3). Q_1 and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



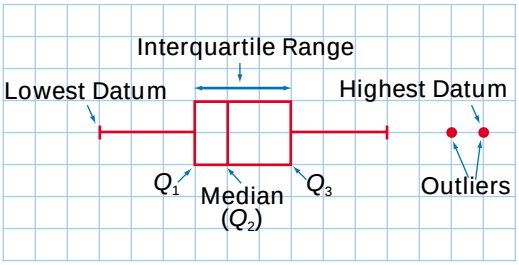
Recall that when there are an even number of data, you take the midpoint between the two middle values as the median. If the number of data below the median is even, Q_1 is the midpoint between the two middle values in this half of the data. Q_3 is determined in a similar way.

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, $Q_1 = 3$.

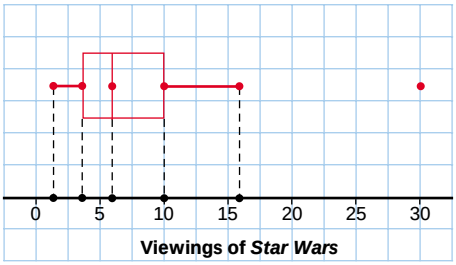
Similarly, the two middle values of the upper half of the data are both 10, so $Q_3 = 10$.

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.

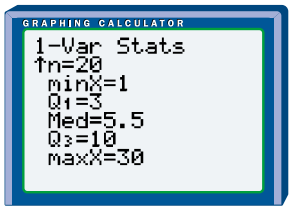
b) The value of 30 at the end of the ordered data is clearly an outlier. Therefore, a modified box-and-whisker plot will best illustrate this set of data.



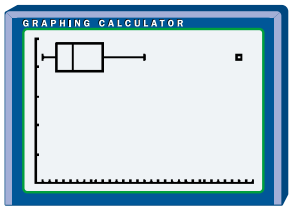
c) Comparing the two surveys shows that the median number of viewings is higher this year and the data are somewhat less spread out.

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.



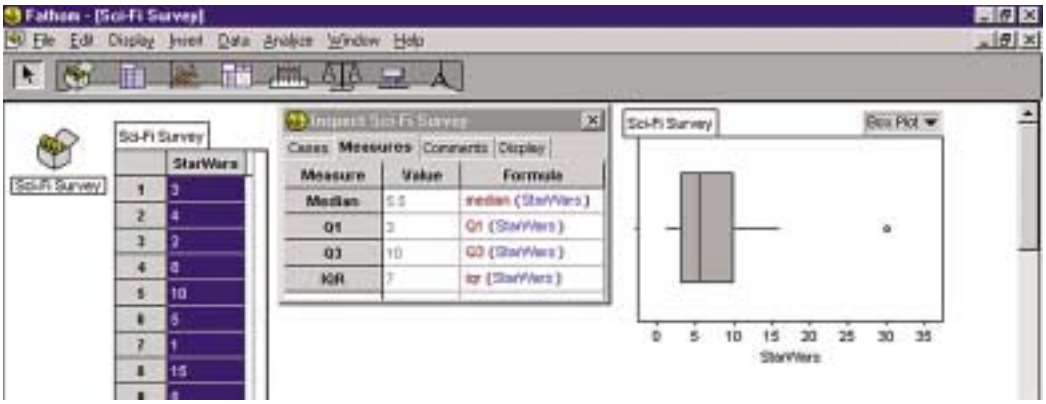
b) Use **STAT PLOT** to select a modified box plot of your list. Press GRAPH to display the box-and-whisker plot and adjust the **window settings**, if necessary.



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,
$z = \frac{x - \mu}{\sigma}$	$z = \frac{x - \bar{x}}{s}$

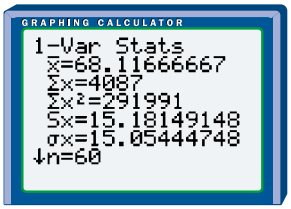
Variable values below the mean have negative z-scores, values above the mean have positive z-scores, and values equal to the mean have a zero z-score. Chapter 8 describes z-scores in more detail.

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.

Review of Key Concepts

2.1 Data Analysis With Graphs

Refer to the Key Concepts on page 100.

1. The following data show monthly sales of houses by a real-estate agency.

6	5	7	6	8	3	5	4	6
7	5	9	5	6	6	7		

a) Construct an ungrouped frequency table for this distribution.

b) Create a frequency diagram.

c) Create a cumulative-frequency diagram.

2. A veterinary study recorded the masses in grams of 25 kittens at birth.

240	300	275	350	280	260	320
295	340	305	280	265	300	275
315	285	320	325	275	270	290
245	235	305	265			

a) Organize these data into groups.

b) Create a frequency table and histogram.

c) Create a frequency polygon.

d) Create a relative-frequency diagram.

3. A class of data-management students listed their favourite board games.

Game	Frequency
Pictionary®	10
Chess	5
Trivial Pursuit®	8
MONOPOLY®	3
Balderdash®	6
Other	4

a) What type of data does this table show? Explain your reasoning.

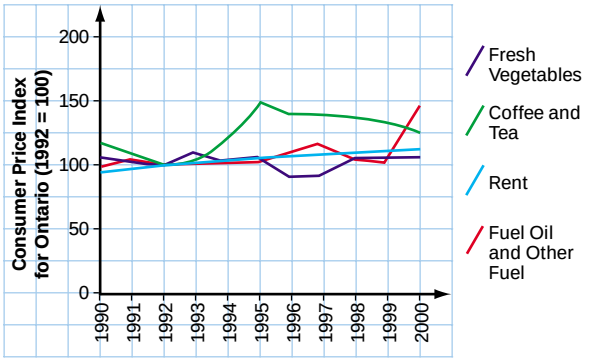
b) Graph these data using an appropriate format.

c) Explain why you chose the type of graph you did.

2.2 Indices

Refer to the Key Concepts on page 109.

The following graph shows four categories from the basket of goods and services used to calculate the consumer price index.



4. a) What is this type of graph called?

b) Which of the four categories had the greatest increase during the period shown?

c) Why do all four graphs intersect at 1992?

d) Which category was

i) the most volatile?

ii) the least volatile?

e) Suggest reasons for this difference in volatility.

5. a) If a tin of coffee cost $5.99 in 1992, what would you expect it to cost in

i) 1995?

ii) 1990?

b) What rent would a typical tenant pay in 2000 for an apartment that had a rent of $550 per month in 1990?

c) What might you expect to pay for broccoli in 2000, if the average price you paid in 1996 was $1.49 a bunch?

2.3 Sampling Techniques

Refer to the Key Concepts on page 116.

6. **a)** Explain the difference between a stratified sample and a systematic sample.

b) Describe a situation where a convenience sample would be an appropriate technique.

c) What are the advantages and disadvantages of a voluntary-response sample?

7. Suppose you are conducting a survey that you would like to be as representative as possible of the entire student body at your school. However, you have time to visit only six classes and to process data from a total of 30 students.

a) What sampling technique would you use?

b) Describe how you would select the students for your sample.

8. Drawing names from a hat and using a random-number generator are two ways to obtain a simple random sample. Describe two other ways of selecting a random sample.

2.4 Bias in Surveys

Refer to the Key Concepts on page 122.

9. Identify the type of bias in each of the following situations and state whether the bias is due to the sampling technique or the method of data collection.

a) A survey asks a group of children whether or not they should be allowed unlimited amounts of junk food.

b) A teachers asks students to raise their hands if they have ever told a harmless lie.

c) A budding musician plays a new song for family members and friends to see if it is good enough to record professionally.

d) Every fourth person entering a public library is asked: "Do you think Carol Shields should receive the Giller prize for her brilliant and critically acclaimed new novel?"

10. For each situation in question 9, suggest how the statistical process could be changed to remove the bias.

2.5 Measures of Central Tendency

Refer to the Key Concepts on page 133.

11. **a)** Determine the mean, median, and mode for the data in question 1.

b) Which measure of central tendency best describes these data? Explain your reasoning.

12. **a)** Use your grouped data from question 2 to estimate the mean and median masses for the kittens.

b) Determine the actual mean and median masses from the raw data.

c) Explain any differences between your answers to parts a) and b).

13. **a)** For what type of "average" will the following statement always be true?

"There are as many people with below-average ages as there are with above-average ages. "

b) Is this statement likely to be true for either of the other measures of central tendency discussed in this chapter? Why or why not?

14. Angela is applying to a university engineering program that weights an applicant's eight best grade-12 marks as shown in the following table.

Subjects	Weighting
Calculus, chemistry, geometry and discrete mathematics, physics	3
Computer science, data management, English	2
Other	1

Angela's grade-12 final marks are listed below.

Subject	Mark	Subject	Mark
Calculus	95	Computer science	84
English	89	Chemistry	90
Geometry and discrete mathematics	94	Mathematics of data management	87
Physical education	80	Physics	92

a) Calculate Angela's weighted average.

b) Calculate Angela's unweighted average.

c) Explain why the engineering program would use this weighting system.

15. Describe three situations where the mode would be the most appropriate measure of central tendency.

2.6 Measures of Spread

Refer to the Key Concepts on page 147.

16. **a)** Determine the standard deviation, the interquartile range, and the semi-interquartile range for the data in question 1.

b) Create a box-and-whisker plot for these data.

c) Are there any outliers in the data? Justify your answer.

17. **a)** Explain why you cannot calculate the semi-interquartile range if you know only the difference between either Q_3 and the median or median and Q_1.

b) Explain how you could determine the semi-interquartile range if you did know both of the differences in part a).

18. **a)** For the data in question 2, determine

i) the first and third quartiles

ii) the 10th, 25th, 75th, and 90th percentiles

b) Would you expect any of the values in part a) to be equal? Why or why not?

19. The scores on a precision-driving test for prospective drivers at a transit company have a mean of 100 and a standard deviation of 15.

a) Determine the z-score for each of the following raw scores.

i) 85 **ii)** 135 **iii)** 100 **iv)** 62

b) Determine the raw score corresponding to each of the following z-scores.

i) 1 **ii)** −2 **iii)** 1.5 **iv)** −1.2

20. Dr. Simba's fourth-year class in animal biology has only 12 students. Their scores on the midterm examination are shown below.

50 71 65 54 84 69 82
67 52 52 86 85

a) Calculate the mean and median for these data. Compare these two statistics.

b) Calculate the standard deviation and the semi-interquartile range. Compare these statistics and comment on what you notice.

c) Which measure of spread is most suitable for describing this data set? Explain why.

Chapter Test

ACHIEVEMENT CHART				
Category	Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application
Questions	**All**	**10, 11**	**4, 6, 7, 8, 9, 11**	**5, 6, 11**

Use the following set of data-management final examination scores to answer questions 1 through 5.

92	48	59	62	66	98	70	70	55	63
70	97	61	53	56	64	46	69	58	64

1. **a)** Group these data into intervals and create a frequency table.

b) Produce a frequency diagram and a frequency polygon.

c) Produce a cumulative-frequency diagram.

2. Determine the

a) three measures of central tendency

b) standard deviation and variance

c) interquartile and semi-interquartile ranges

3. **a)** Produce a modified box-and-whisker plot for this distribution.

b) Identify any outliers.

c) Identify and explain any other unusual features of this graph.

4. Explain which of the three measures of central tendency is most appropriate to describe this distribution of marks and why the other two measures are not appropriate.

5. Students with scores above the 90th percentile receive a book prize.

a) How many students will receive prizes?

b) What are these students' scores?

6. An interview committee graded three short-listed candidates for a management position as shown below. The scores are on a scale of 1 to 5, with 5 as the top score.

Criterion	Weight	Clarise	Pina	Steven
Education	2	3	3	4
Experience	2	4	5	3
Interpersonal skills	3	3	3	5
First interview	1	5	4	3

Who should the committee hire based on these data? Justify your choice.

7. Describe the type of sample used in each of the following scenarios.

a) A proportionate number of boys and girls are randomly selected from a class.

b) A software company randomly chooses a group of schools in a particular school district to test a new timetable program.

c) A newspaper prints a questionnaire and invites its readers to mail in their responses.

d) A telephone-survey company uses a random-number generator to select which households to call.

e) An interviewer polls people passing by on the street.

8. A group of 8 children in a day-care centre are to be interviewed about their favourite games. Describe how you would select a systematic sample if there are 52 children at the centre.

9. **a)** Identify the bias in the following surveys and explain the effect it could have on their results.

i) Parents of high-school students were asked: "Do you think that students should be released from school a half hour early on Friday, free to run around and get into trouble?"

ii) Audience members at an investment workshop were asked to raise their hands if they had been late with a bill payment within the last six months.

iii) A random survey of corporate executives asked: "Do you favour granting a cable-television licence for a new economics and business channel?"

b) Suggest how to eliminate the bias in each of the surveys in part a).

10. A mutual-fund company proudly advertises that all of its funds have "first-quartile performance." What mathematical errors has the company made in this advertisement?

ACHIEVEMENT CHECK

Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application

11. The graph below shows the stock price for an Ontario technology company over a one-month period in 2001.

a) When did the stock reach its lowest value during the period shown? Suggest a possible reason for this low point.

b) Compare the percent drop in stock price from September 1 to September 8 to the drop during the following week.

c) Sketch a new graph and provide a commentary that the company could use to encourage investors to buy the company's stock.