2.1 Data Analysis With Graphs

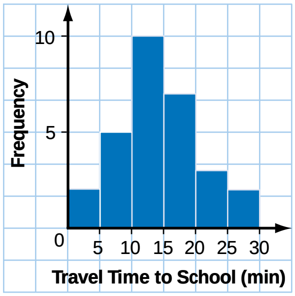
Statistics is the gathering, organization, analysis, and presentation of numerical information. You can apply statistical methods to almost any kind of data. Researchers, advertisers, professors, and sports announcers all make use of statistics. Often, researchers gather large quantities of data since 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

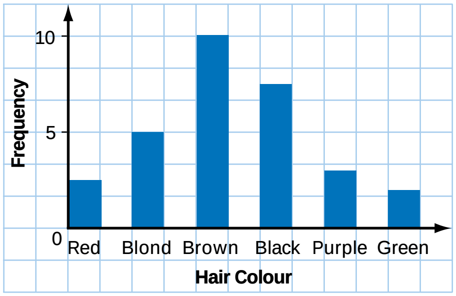
1. Work in groups or as a class to design a fast and efficient way to survey your class about a simple numerical variable, such as the students' heights or the distances they travel to school.
2. Carry out your survey and record all the results in a table.
3. Consider how you could organize these results 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.

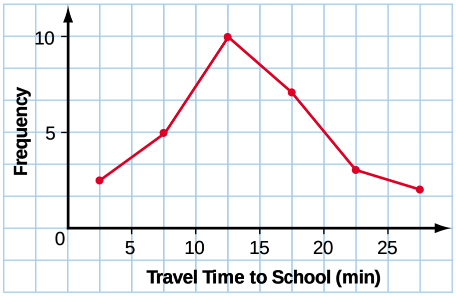
A **histogram** is a special form of **bar graph** in which the areas of the bars are proportional to the *frequencies* of the values of the variable. The bars in a histogram are connected and represent a continuous range of values. Histograms are used for variables whose values can be arranged in numerical 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

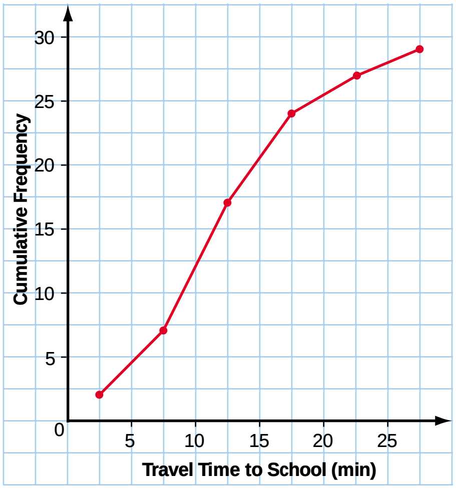


Bar Graph



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?

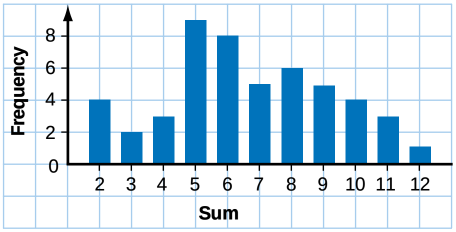
Solution

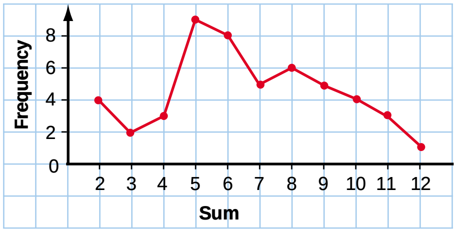
a) Go through the data and tally the frequency of each value of the variable as shown in the table on the right.

Sum	Tally	Frequency
2	ǀǀǀǀ	4
3	ǀǀ	2
4	ǀǀǀ	3
5	𝍸 ǀǀǀǀ	9
6	𝍸 ǀǀǀ	8
7	𝍸	5
8	𝍸 ǀ	6
9	𝍸	5
10	ǀǀǀǀ	4
11	ǀǀǀ	3
12	ǀ	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.

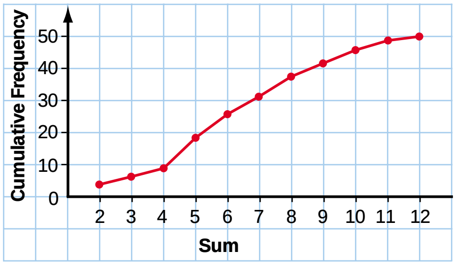
c) The bar graph or frequency polygon makes the pattern in the data more apparent.





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	ǀǀǀǀ	4	4
3	ǀǀ	2	6
4	ǀǀǀ	3	9
5	𝍸 ǀǀǀǀ	9	18
6	𝍸 ǀǀǀ	8	26
7	𝍸	5	31
8	𝍸 ǀ	6	37
9	𝍸	5	42
10	ǀǀǀǀ	4	46
11	ǀǀǀ	3	49
12	ǀ	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.

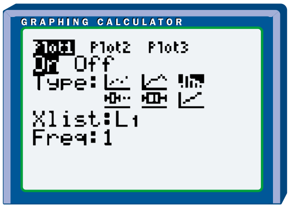
b) On how many days was the maximum temperature 25°C or less? On how many days did the temperature exceed 30°C?

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

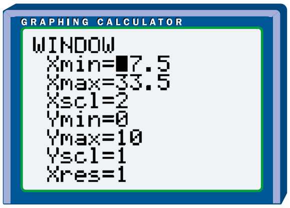
Solution 1 Using a Graphing Calculator

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.

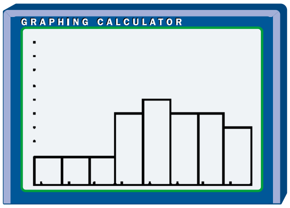
Use the STAT EDIT menu to make sure that lists L1 to L4 are clear, and then enter the temperature data into L1. Use STAT PLOT to turn on Plot1 and select the histogram icon. Next, adjust the window settings. Set Xmin and Xmax to the lower and upper limits for your intervals and set Xscl to the interval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



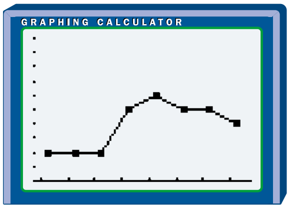
```
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Freq:1
```



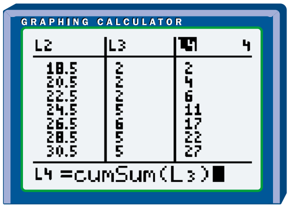
```
WINDOW
 Xmin=17.5
 Xmax=33.5
 Xscl=2
 Ymin=0
 Ymax=10
 Yscl=1
 Xres=1
```



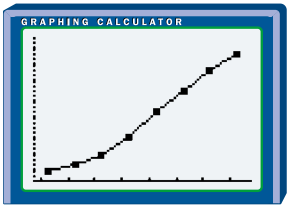
You can now use the TRACE instruction and the arrow keys to determine the tally for each of the intervals. Enter the midpoints of the intervals into L2 and the tallies into L3. Turn off Plot1 and set up Plot2 as an *x*-*y* line plot of lists L2 and L3 to produce a frequency polygon.



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.



```
L2     |L3    |L4     4
 18.5  | 2    | 2
 20.5  | 2    | 4
 22.5  | 2    | 6
 24.5  | 5    | 11
 26.5  | 6    | 17
 28.5  | 5    | 22
 30.5  | 5    | 27
L4 =cumSum(L3)
```



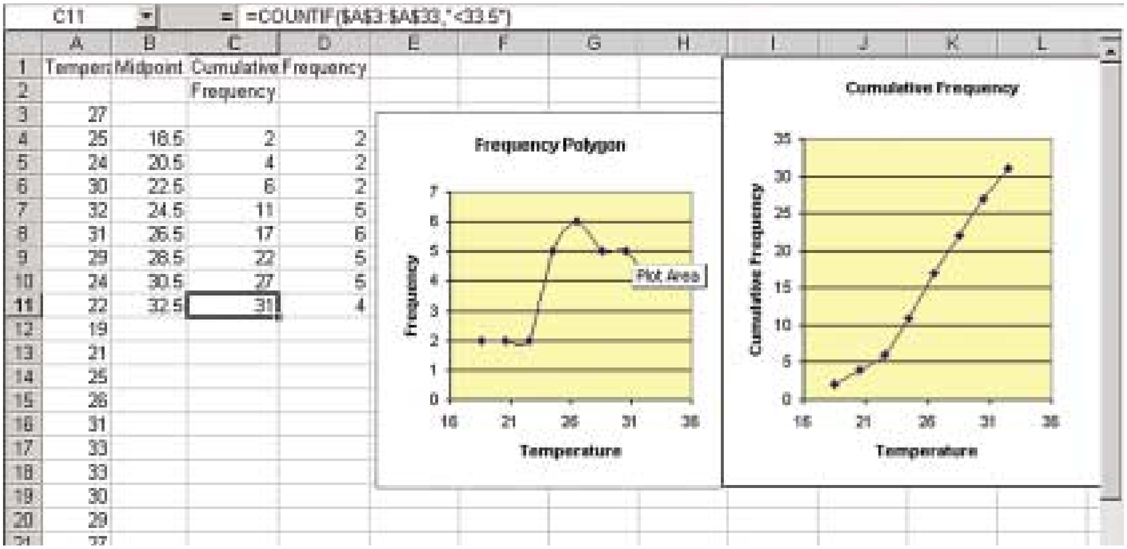
b) Since you know that all the temperatures were in whole degrees, you can see from the cumulative frequencies in L4 that there were 11 days on which the maximum temperature was no higher than 25°C. You can also get this information from the cumulative-frequency graph.

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.

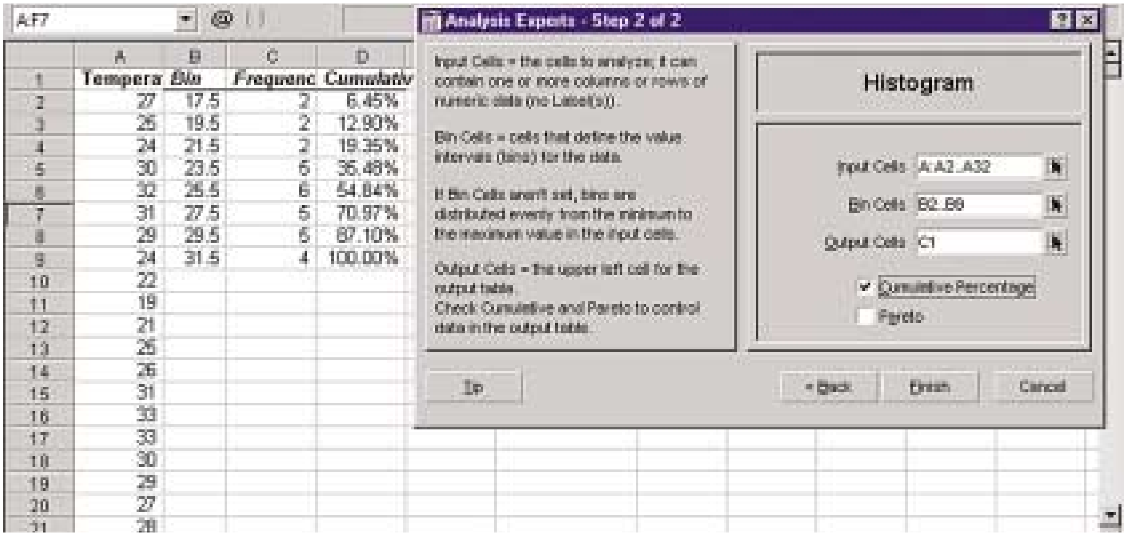
You can then use the Chart feature to produce a frequency polygon by graphing columns B and D. Similarly, charting columns B and C will produce a cumulative-frequency graph.



C11 =COUNTIF($A$3:$A$33,"<33.5")

	A	B	C	D
1	Temper:	Midpoint	Cumulative	Frequency
2			Frequency	
3	27			
4	25	18.5	2	2
5	24	20.5	4	2
6	30	22.5	6	2
7	32	24.5	11	5
8	31	26.5	17	6
9	29	28.5	22	5
10	24	30.5	27	5
11	22	32.5	31	4
12	19			
13	21			
14	25			
15	26			
16	31			
17	33			
18	33			
19	30			
20	29			

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.



A:F7

	A	B	C	D
1	Tempera	Bin	Frequenc	Cumulativ
2	27	17.5	2	6.45%
3	25	19.5	2	12.90%
4	24	21.5	2	19.35%
5	30	23.5	6	35.48%
6	32	25.5	6	54.84%
7	31	27.5	5	70.97%
8	29	29.5	5	87.10%
9	24	31.5	4	100.00%
10	22			
11	19			
12	21			
13	25			
14	26			
15	31			
16	33			
17	33			
18	30			
19	29			
20	27			
21	28			

Analysis Experts - Step 2 of 2

Input Cells = the cells to analyze; it can contain one or more columns or rows of numeric data (no Label(s)).

Bin Cells = cells that define the value intervals (bins) for the data.

If Bin Cells aren't set, bins are distributed evenly from the minimum to the maximum value in the input cells.

Output Cells = the upper left cell for the output table.
Check Cumulative and Pareto to control data in the output table.

Histogram

Input Cells: A:A2..A32

Bin Cells: B2..B9

Output Cells: C1

☑ Cumulative Percentage

☐ Pareto

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.

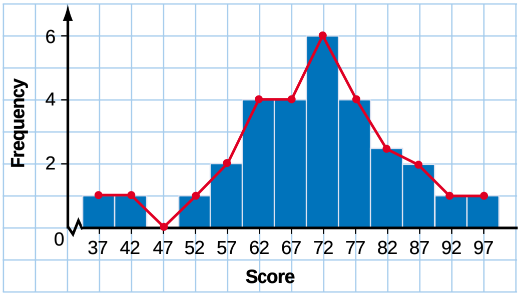
d) What proportion of the students had marks between 70% and 79%?

Solution

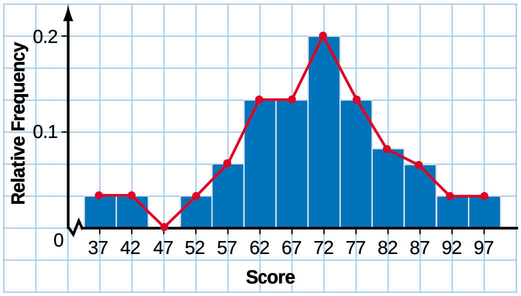
a) The lowest and highest scores are 39% and 96%, which give a range of 57%. An interval width of 5 is convenient, so you could use 13 intervals as shown here. To determine the relative frequencies, divide the frequency by the total number of scores. For example, the relative frequency of the first interval 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	𝍸 |	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.

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

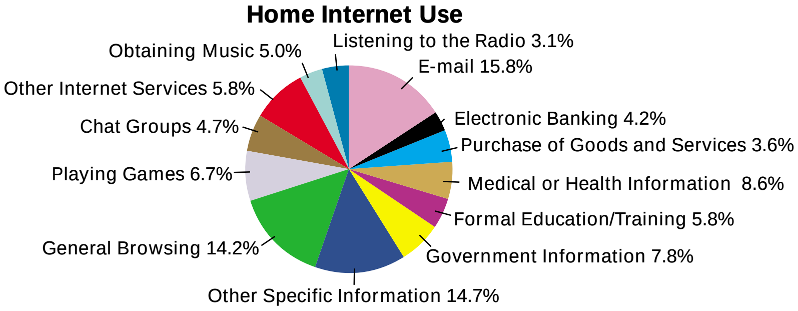
Illustrate these data with

a) a circle graph

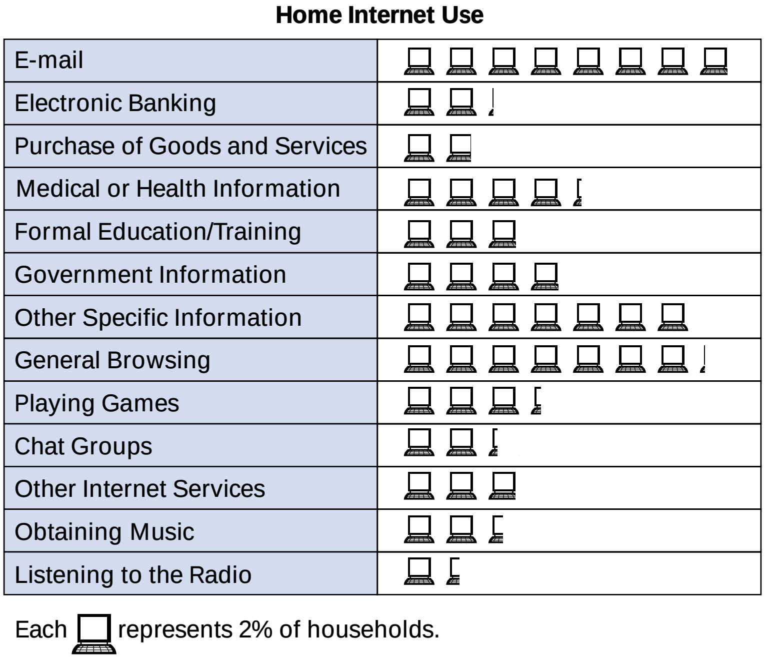
b) a pictograph

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.



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?

Chapter Problem

10. Communication This table summarizes the salaries for François' hockey 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

a) Reorganize these data into appropriate intervals and present them in a frequency table.

b) Create a histogram for these data.

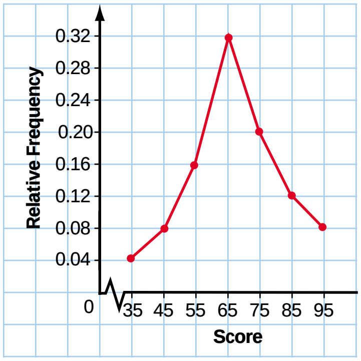
c) Identify and explain any unusual features about this distribution.

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.

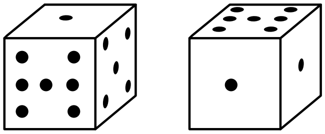
2	1	2	1	3	1	4	1	0	1
0	2	4	1	0	5	2	3	4	1
2	1	1	1	3	1	0	5	4	2
3	1	1	0	2	2	0	0	1	3

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?



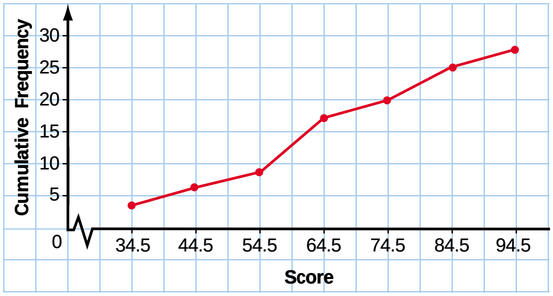
b) Sketch a relative-frequency polygon to show the results you would expect if these dice were rolled 100 times.

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.