8.1 Continuous Probability Distributions

Distributions like the binomial probability distribution and the hypergeometric distribution deal with discrete data. The possible values of the random variable are natural numbers, because they arise from counting processes (usually successful or unsuccessful trials). Many characteristics of a population such as the heights of human adults are continuous in nature, and have fractional or decimal values. Just as with discrete data, however, these continuous variables have statistical distributions. For some discrete quantities such as the earthquake data on page 411 a smooth, continuous model of their variation may be more useful than a bar graph. Continuous probability distributions, by contrast with those you studied in Chapter 7, allow fractional values and can be graphed as smooth curves.

INVESTIGATE & INQUIRE: Modelling Failure Rates

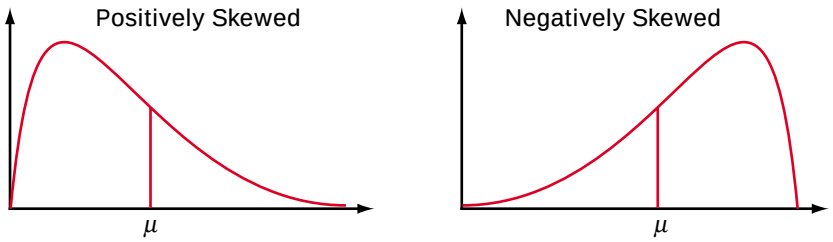
Manufacturers often compile reliability data to help them predict demand for repair services and to improve their products. The table below gives the failure rates for a model of computer printer during its first four years of use.

Age of Printer (Months)	Failure Rate (%)
0–6	4.5
6–12	2.4
12–18	1.2
18–24	1.5
24–30	2.2
30–36	3.1
36–42	4.0
42–48	5.8

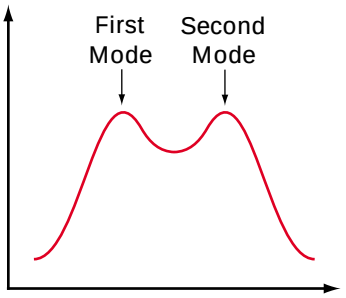
1. Construct a scatter plot of these data. Use the midpoint of each interval. Sketch a smooth curve that is a good fit to the data.

2. Describe the resulting probability distribution. Is this distribution symmetric?
3. Why do you think printer failure rates would have this shape of distribution?
4. Calculate the mean and standard deviation of the failure rates. How useful are these summary measures in describing this distribution? Explain.

In the investigation above, you modelled failure rates as a smooth curve, which allowed you to describe some of the features of the distribution. A distribution which is not symmetric may be **positively skewed** (tail pulled to the right) or **negatively skewed** (tail pulled to the left). For example, the number of children in a Canadian family has a positively-skewed distribution, because there is a relatively low modal value of two and an extended tail that represents a small number of significantly larger families.



Both kinds of skewed distributions are **unimodal**. The single "hump" is similar to the mode of a set of discrete values, which you studied in Chapter 2. A distribution with two "humps" is called **bimodal**. This distribution may occur when a population consists of two groups with different attributes. For example, the distribution of adult shoe sizes is bimodal because men tend to have larger feet than women do.



Modelling Distributions With Equations

Often you want to find the probability that a variable falls in a particular range of values. This kind of probability can be determined from the area under the distribution curve. The curve itself represents the **probability density**, the probability per unit of the continuous variable.

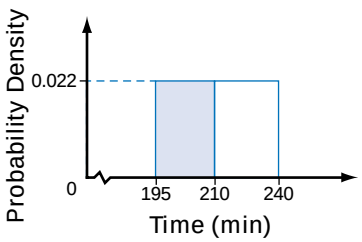
Many distribution curves can be modelled with equations that allow the areas to be calculated rather than estimated. The simplest such curve is the uniform distribution.

Example 1 Probabilities in a Uniform Distribution

The driving time between Toronto and North Bay is found to range evenly between 195 and 240 min. What is the probability that the drive will take less than 210 min?

Solution

The time distribution is uniform. This means that every time in the range is equally likely. The graph of this distribution will be a horizontal straight line. The total area under this line must equal 1 because all the possible driving times lie in the range 195 to 240 min. So, the height of the line is $\frac{1}{240 - 195} = 0.022$.



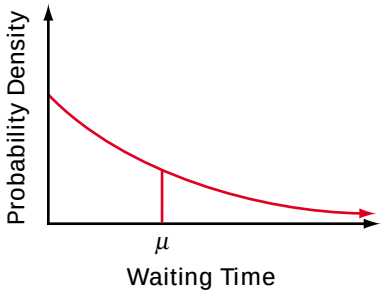
The probability that the drive will take less than 210 min will be the area under the probability graph to the left of 210. This area is a rectangle. So,

$$P(\text{driving time} \leq 210) = 0.022 \times (210 - 195)$$
$$= 0.33$$

The **exponential distribution** predicts the waiting times between consecutive events in any random sequence of events. The equation for this distribution is

$$y = ke^{-kx}$$

where $k = \frac{1}{\mu}$ is the number of events per unit time and $e \doteq 2.71828$.



The longer the average wait, the smaller the value of k, and the more gradually the graph slopes downward.

Notice that the smallest waiting times are the most likely. This distribution is similar to the discrete geometric distribution in section 7.3. Recall that the geometric distribution models the number of trials before a success. If you think of the event "receiving a phone call in a given minute" as successive trials, you can see that the exponential distribution is the continuous equivalent of the geometric distribution.

Example 2 Exponential Distribution

The average time between phone calls to a company switchboard is $\mu = 2$ min.

a) Simulate this process by generating random arrival times for an 8-h business day. Group the waiting times in intervals and plot the relative frequencies of the intervals.

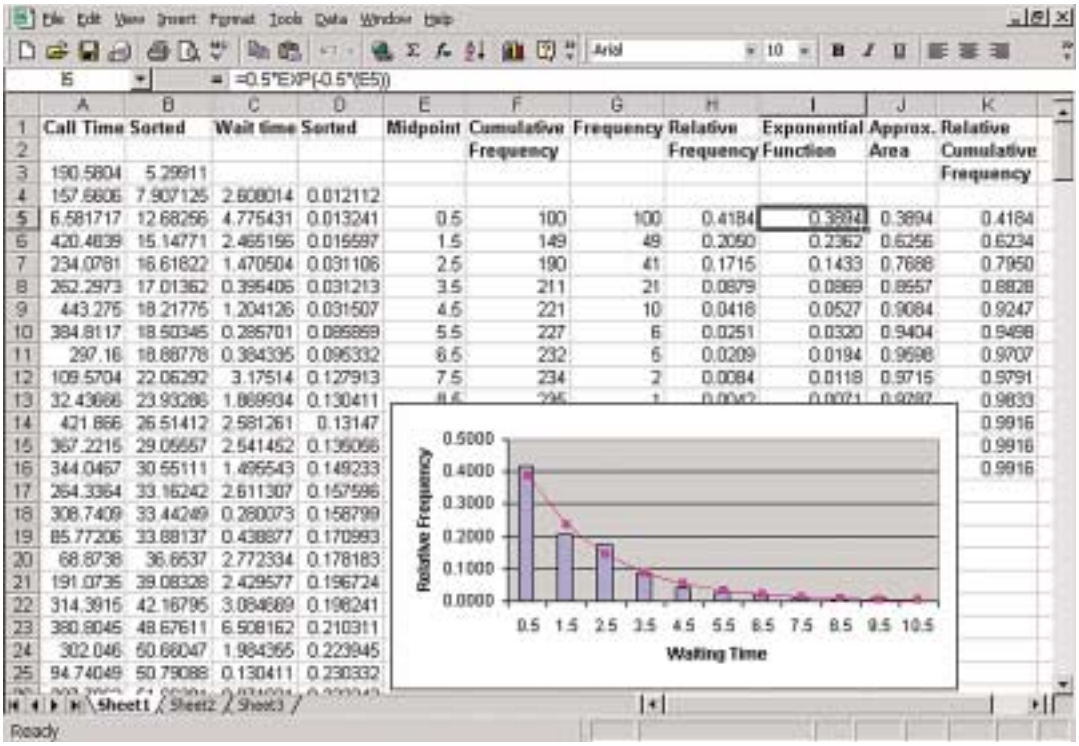
b) Draw the graph of $y = ke^{-kx}$ on your relative frequency plot. Comment on the fit of this curve to the data.

c) Calculate the probability that the time between two consecutive calls is less than 3 min.

Solution

a) Your random simulation should have a mean time between calls of 2 min. Over 8 h, you would expect about 240 calls. Use the **RAND function** of a spreadsheet to generate 240 random numbers between 0 and 480 in column A. These numbers simulate the times (in minutes) at which calls come in to the switchboard. Copy these numbers as values into column B so that you can sort them. Use the **Sort feature** to sort column B, then calculate the difference between each pair of consecutive numbers to find the waiting times between calls.

Next, copy the values for the waiting times into column D and sort them. In column F, use the **COUNTIF function** to cumulatively group the data into 1-min intervals. In column G, calculate the frequencies for the intervals by subtracting the cumulative frequency for each interval from that for the following interval. In column H, divide the frequencies by 239 to find the relative frequencies. Use the **Chart feature** to plot the relative frequencies.



	A	B	C	D	E	F	G	H	I	J	K
1	Call Time Sorted		Wait time Sorted		Midpoint	Cumulative Frequency	Frequency	Relative Frequency	Exponential Function	Approx. Area	Relative Cumulative Frequency
2											
3	190.5804	5.29911									
4	157.6606	7.907125	2.608014	0.012112							
5	6.581717	12.68256	4.775431	0.013241	0.5	100	100	0.4184	0.3894	0.3894	0.4184
6	420.4839	15.14771	2.465195	0.015597	1.5	149	49	0.2050	0.2362	0.6256	0.6234
7	234.0781	16.61822	1.470504	0.031106	2.5	190	41	0.1715	0.1433	0.7688	0.7950
8	262.2973	17.01362	0.395406	0.031213	3.5	211	21	0.0879	0.0869	0.8557	0.8828
9	443.275	18.21775	1.204126	0.031507	4.5	221	10	0.0418	0.0527	0.9084	0.9247
10	384.8117	18.50345	0.285701	0.095869	5.5	227	6	0.0251	0.0320	0.9404	0.9498
11	297.16	18.88778	0.384335	0.095332	6.5	232	5	0.0209	0.0194	0.9598	0.9707
12	109.5704	22.06292	3.17514	0.127913	7.5	234	2	0.0084	0.0118	0.9715	0.9791
13	32.43666	23.93286	1.869934	0.130411	8.5	235	1	0.0042	0.0071	0.9787	0.9833
14	421.866	26.51412	2.581261	0.13147							0.9916
15	367.2215	29.05657	2.541452	0.136056							0.9916
16	344.0457	30.55111	1.496543	0.149233							0.9916
17	264.3364	33.16242	2.611307	0.157596							
18	308.7409	33.44249	0.280073	0.158799							
19	85.77206	33.88137	0.438877	0.170993							
20	68.8738	36.6537	2.772334	0.178183							
21	191.0735	39.08328	2.429577	0.196724							
22	314.3915	42.16795	3.084689	0.198241							
23	380.8045	48.67611	6.508162	0.210311							
24	302.046	50.68047	1.984365	0.223945							
25	94.74049	50.79088	0.130411	0.230332							

b) Since $\mu = 2$ min, the exponential equation is $P(X < x) = 0.5e^{-0.5x}$. You can use the EXP function to calculate values for the midpoint of each interval, and then plot these data with the **Chart feature**.

The exponential model fits the simulation data reasonably well. The sample size is small enough that statistical fluctuations could account for some intervals not fitting the model as closely as the others do.

c) You can estimate the probability of waiting times of various lengths from the cumulative relative frequencies for the simulation. Calculate these values in column K by dividing the cumulative frequencies by 239, the number of data. Cell K7 shows the relative cumulative frequency for the third interval, which gives an estimate of about 0.8 for $P(X < 3)$.

Alternatively, you can calculate this probability from the area under the probability distribution curve, as you did with the uniform distribution. The area will be approximately equal to the sum of the values of the exponential function at the midpoints of the first four intervals times the interval width. Use the **SUM function** to calculate this sum in column J. As shown in cell J7, this method also gives an estimate of about 0.8.

The exponential distribution is not symmetric, and its mode is always zero. The statistical measure you need to know, however, is the mean, which cannot easily be seen from the shape of the curve, but which can be found easily if you know the equation of the distribution, since $\mu = \frac{1}{k}$.

Notice also that the y-intercept of the curve is equal to k, so you could easily estimate the mean from the graph.

What would a symmetric version of an exponential distribution look like? Many quantities and characteristics have such a distribution, which is sometimes called a "bell curve" because of its shape. The photograph and curve on page 414 suggest one common example, people's heights. In fact, the bell curve is the most frequently observed probability distribution, and you will explore its mathematical formulation throughout this chapter.

Key Concepts

- Continuous probability distributions allow for fractional or decimal values of the random variable.
- Distributions can be represented by a relative-frequency table, a graph, or an equation.
- Probabilities can be computed by finding the area under the curve within the appropriate interval.

Communicate Your Understanding

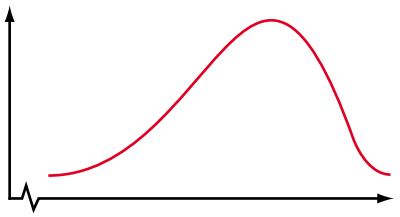
1. Explain the terms *discrete*, *continuous*, *symmetric*, *positively skewed*, and *negatively skewed*.
2. Give at least two examples of data that might result in
   a) a bimodal distribution
   b) an exponential distribution
   c) a positively-skewed distribution
3. Are summary measures such as mean and standard deviation always a good indicator of the probability distribution for a set of data? Justify your answer.

Apply, Solve, Communicate

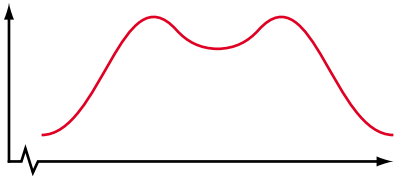
A

1. Match the following distribution curves to the random variables. Give reasons for your choices.
   a) waiting times between arrivals at a pizza outlet during lunchtime
   b) collar sizes in the adult population
   c) hours worked per week

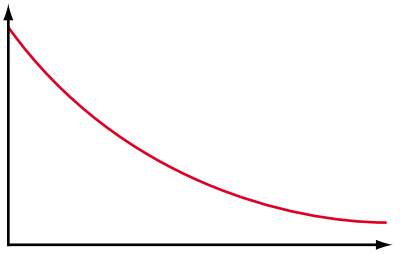
i)



ii)



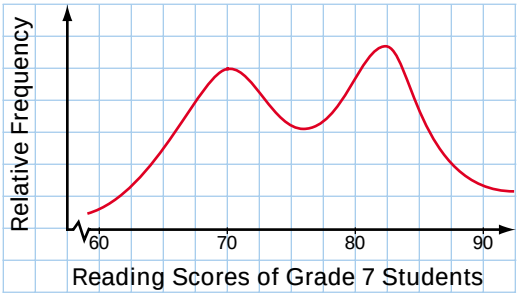
iii)



B

Use appropriate technology, wherever possible.

2. **Communication** The graph below shows a relative-frequency distribution for reading-test scores of grade 7 students in a school district.



   a) Estimate the mean for these data.
   b) Give a possible explanation for the shape of this distribution.

3. **Application** The growing season for farmers is the number of days from the last frost in the spring until the first frost in the fall. The growing seasons for some areas of Ontario are listed below.

179	145	156	141	178	148	244	192
181	142	202	220	218	217	156	211
201	175	162	179	165	196	173	188
135	182	166	169	152	160	161	210
148	137	149	176	165	171	198	136
129	128	180	202	220	203	200	201
169	164	184	217	152	192	189	164
203							

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

b) Describe the shape of this distribution.

c) From the graph, estimate the mean for these data.

d) Compute the mean and standard deviation of these data.

4. **Inquiry/Problem Solving** The following table gives the number of time-loss injuries in the Canadian agriculture sector in 1997.

Age	Number of Accidents
20–25	719
25–30	611
30–35	629
35–40	588
40–45	430
45–50	323
50–55	246
55–60	150
60–65	99

a) Construct a scatter plot of these data. Use the midpoint of each interval. Join the points with a smooth curve.

b) Describe the shape of this distribution.

c) Give possible reasons for the key features of this distribution.

d) Estimate the mean of this distribution. Explain what this mean tells you.

5. Choose a paragraph of at least seven sentences from a book of your choice.

a) Construct a tally chart and relative-frequency distribution of the number of letters per word for your selected passage.

b) Construct a smooth curve representing the probability distribution for these data.

c) Describe the characteristics of these data.

d) Estimate the mean of the data from your graph.

e) Calculate the mean of the data.

f) How might the mean of the data be used to estimate the reading level of the paragraph?

6. **Application** The lifetime of a species of housefly ranges uniformly from 18 h to 30 h.

a) Construct a graph of this distribution.

b) Determine the probability that a fly selected at random will live at least 28 h.

c) Determine the probability that a fly selected at random will live less than 24 h.

7. The following list gives the length of time in minutes between arrivals at the emergency room of a hospital.

8	20	5	4	2	1	4	18	15	2
25	10	2	5	4	5	1	6	2	2
5	2	1	5	1	2	3	15	1	1
6	10	15	2	2	4	2	3	4	5
1	2	1	35	6	8	3	5	8	5
2	5	7	1	6					

a) Construct a relative-frequency diagram for of these data and draw a smooth curve.

b) Is the exponential distribution a reasonable model for these data? Justify your answer.

c) Calculate the mean of these data.

d) Determine an equation for the probability distribution of these data.

8. Using calculus, it can be shown that the area under the exponential distribution curve is $P(X \leq x) = 1 - e^{-kx}$.

a) What is the maximum value of $P(X \leq x)$? For what value of x does this maximum occur?

b) Use this equation for $P(X \leq x)$ to determine the accuracy of the estimates for $P(X \leq 3)$ in Example 2.

c) Account for the difference between the estimates and the value calculated using the equation above.

d) How could you improve the accuracy of the estimates?

9. The manager of the local credit union knows that the average length of time it takes to serve one client is 3.5 min. Use the equation $y = ke^{-kx}$ to find the probability that

a) the next customer will be served in less than 3 min.

b) the time to serve the next customer will be greater than 5 min.

10. Chapter Problem

a) Construct a bar graph for the earthquake data on page 411.

b) Construct a relative-frequency bar graph for these data.

c) Calculate the mean and standard deviation for these data.

C

11. **Communication** A probability distribution has equation $y = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

a) Graph this distribution.

b) Describe the shape of this distribution.

c) Suggest a reason why the equation for this distribution includes the factor $\frac{1}{\sqrt{2\pi}}$.

12. **Inquiry/Problem Solving** Graphing calculators contain a number of different probability distributions. These distributions can be found in the DISTR menu. One interesting distribution is the chi-square distribution (symbol χ^2).

a) Paste the χ^2 pdf function into the Y= editor. The syntax is χ^2pdf(X,positive integer).

b) Investigate the shape of the graph for this distribution for several positive integers.

c) Write a brief report of your findings.