CHAPTER 8

The Normal Distribution

Specific Expectations	Section
Interpret one-variable statistics to describe the characteristics of a data set.	8.1, 8.2, 8.3
Organize and summarize data from secondary sources.	8.1, 8.2, 8.3
Identify situations that give rise to common distributions.	8.1, 8.2, 8.4, 8.5, 8.6
Interpret probability statements, including statements about odds, from a variety of sources.	8.1, 8.2, 8.3, 8.4, 8.5, 8.6
Assess the validity of some simulation results by comparing them with the theoretical probabilities, using the probability concepts developed in the course.	8.2
Describe the position of individual observations within a data set, using *z*-scores and percentiles.	8.2
Demonstrate an understanding of the properties of the normal distribution.	8.2, 8.3, 8.4, 8.5, 8.6
Make probability statements about normal distributions.	8.2, 8.3, 8.4, 8.5, 8.6
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.	8.3
Assess the validity of conclusions made on the basis of statistical studies.	8.3, 8.5, 8.6
Determine probabilities, using the binomial distribution.	8.4, 8.5, 8.6

Chapter Problem

The Restless Earth

The table shows the number of major earthquakes around the world from 1900 through 1999.

1. Can you predict the number of earthquakes around the world next year?
2. Is it possible to quantify how accurate your prediction is likely to be?

In this chapter, you will develop the skills required to answer these questions as well as others involving statistical predictions. You will learn more about probability distributions, including continuous distributions. In particular, this chapter introduces the normal distribution, one of the most common and important probability distributions. Also, you will analyse statements about probabilities made in the media, statistical studies, and a wide variety of other applications.

Major Earthquakes (7.0 or Greater on the Richter Scale)

Year	Frequency	Year	Frequency	Year	Frequency
1900	13	1934	22	1968	30
1901	14	1935	24	1969	27
1902	8	1936	21	1970	29
1903	10	1937	22	1971	23
1904	16	1938	26	1972	20
1905	26	1939	21	1973	16
1906	32	1940	23	1974	21
1907	27	1941	24	1975	21
1908	18	1942	27	1976	25
1909	32	1943	41	1977	16
1910	36	1944	31	1978	18
1911	24	1945	27	1979	15
1912	22	1946	35	1980	18
1913	23	1947	26	1981	14
1914	22	1948	28	1982	10
1915	18	1949	36	1983	15
1916	25	1950	39	1984	8
1917	21	1951	21	1985	15
1918	21	1952	17	1986	6
1919	14	1953	22	1987	11
1920	8	1954	17	1988	8
1921	11	1955	19	1989	7
1922	14	1956	15	1990	12
1923	23	1957	34	1991	11
1924	18	1958	10	1992	23
1925	17	1959	15	1993	16
1926	19	1960	22	1994	15
1927	20	1961	18	1995	25
1928	22	1962	15	1996	22
1929	19	1963	20	1997	20
1930	13	1964	15	1998	16
1931	26	1965	22	1999	23
1932	13	1966	19		
1933	14	1967	16		

Review of Prerequisite Skills

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

1. Graphing exponential functions

a) Graph the following functions:

i) $y = 10^{-x}$

ii) $y = 10^{-2x}$

iii) $y = 10^{-\frac{x}{2}}$

b) Determine the y-intercepts of each function in part a).

c) For $x > 0$, which function has

i) the largest area under its curve?

ii) the smallest area under its curve?

2. Representing data (Chapter 2) During July, a local theatre recorded the following numbers of patrons per day over a 30-day period.

102	116	113	132	128	117	156	182
183	171	160	140	154	160	122	187
185	158	112	145	168	187	117	108
171	171	156	163	168	182		

a) Construct a histogram of these data.

b) Determine the mean and standard deviation of these data.

c) Construct a box-and-whisker plot of these data.

3. Summary measures for data (Chapter 2) Fifteen different cars were tested for stopping distances at a speed of 20 km/h. The results, in metres, are given below.

15	18	18	20	22	24	24	26
18	26	24	16	23	24	30	

a) Find the mean, median, and mode of these data.

b) Construct a box-and-whisker plot of these data.

c) Find the standard deviation of these data.

4. Summary measures for data (Chapter 2) Analyse the following data, which represent the numbers of e-mail messages received by 30 executives on a Wednesday.

22	14	12	9	54	12	16	12	14
49	10	14	8	21	31	37	28	36
22	9	33	59	31	41	19	28	52
22	7	24						

5. Summary measures for data (Chapter 2) An insurance bureau listed the ratio of registered vehicles to cars stolen for selected towns and cities in Ontario. The results for a recent year were as follows:

50	38	53	56	69	90	94	88
58	68	78	89	89	52	50	70
83	98	91	90	90	84	80	70
83	89	79	75	78	73	92	105
100							

Analyse these data and write a brief report of your findings.

6. Scatter plots The table below gives the depreciation (loss in value) of a new car for each year of ownership. Construct a scatter plot of the depreciation of this car.

Year of Ownership	Depreciation by End of Year
1	30%
2	20%
3	18%
4	15%
5	15%
6–9	10% per year
10	5%

7. Scatter plots The table below gives data on the planets in our solar system.

Planet	Mean Distance From the Sun (AU)	Time for One Revolution (years)
Mercury	0.387	0.241
Venus	0.723	0.615
Earth	1.000	1.000
Mars	1.523	1.881
Jupiter	5.203	11.861
Saturn	9.541	29.457
Uranus	19.190	84.008
Neptune	30.086	164.784
Pluto	39.507	248.350

a) Construct a scatter plot of these data.

b) Construct a line of best fit for these data.

c) Explore whether a curve would be a better approximation for this relationship.

d) Find out what Kepler's third law states, and investigate it using the data provided.

8. ***Z*-scores (section 2.6)** The mean age of the viewers of a popular quiz show is 38.3 years with a standard deviation of 12.71 years.

a) What is the z-score of a 25-year-old viewer?

b) What is the z-score of a 70-year-old viewer?

c) What is the z-score of a 40-year-old viewer?

d) What age range is within 1 standard deviation of the mean?

e) What age range is within 2 standard deviations of the mean?

f) What age range is within 1.3 standard deviations of the mean?

9. ***Z*-scores (section 2.6)** On a recent grade-11 mathematics contest, the mean score was 57.9 with a standard deviation of 11.6. On the grade-10 mathematics contest written at the same time, the mean score was 61.2 with a standard deviation of 11.9. Gavin scored 84.3 on the grade-11 contest and his sister, Patricia, scored 86.2 on the grade-10 contest. Explain why Gavin's results could be considered better than his sister's.

10. Binomial distribution (Chapter 7) According to Statistics Canada, about 85% of all Canadian households own a VCR. Twelve people were selected at random.

a) What is the probability that exactly 8 of them own a VCR?

b) What is the probability that no more than 8 of them own a VCR?

c) What is the probability that at least 8 of them own a VCR?

d) How many of them would you expect to own a VCR?

11. Geometric distribution (section 7.3) If you program a calculator to generate random integers between –4 and 10, what is

a) the probability that the first zero will not occur until the 6th number generated?

b) the probability that the first odd number will be generated within the first 5 numbers?

c) the expected waiting time before a negative number is generated?

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

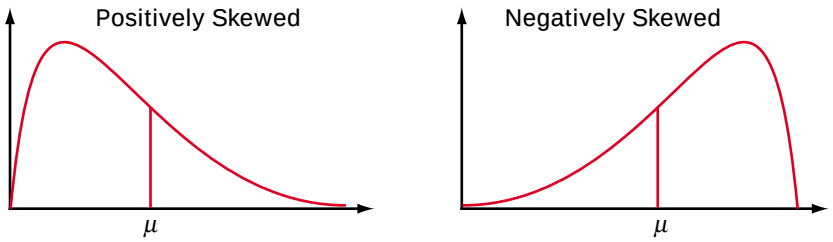
Manufacturers often compile reliability data to help them predict demand for repair services and to improve their products. The table below gives the failure rates for a model of computer printer during its first four years of use.

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

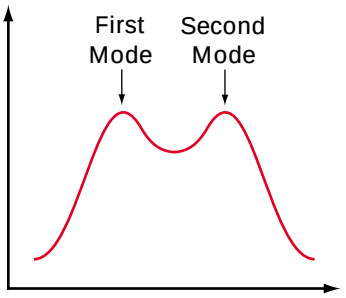
1. Construct a scatter plot of these data. Use the midpoint of each interval. Sketch a smooth curve that is a good fit to the data.

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.

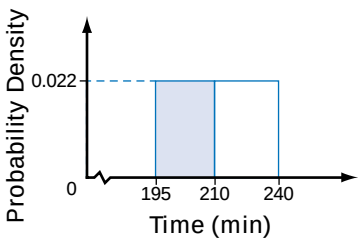
Many distribution curves can be modelled with equations that allow the areas to be calculated rather than estimated. The simplest such curve is the uniform distribution.

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 $\frac{1}{240 - 195} = 0.022$.



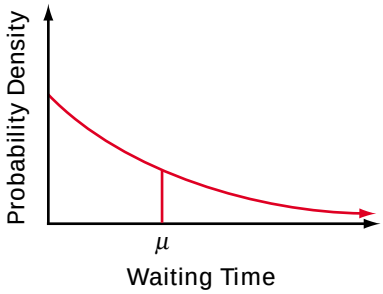
The probability that the drive will take less than 210 min will be the area under the probability graph to the left of 210. This area is a rectangle. So,

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

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.

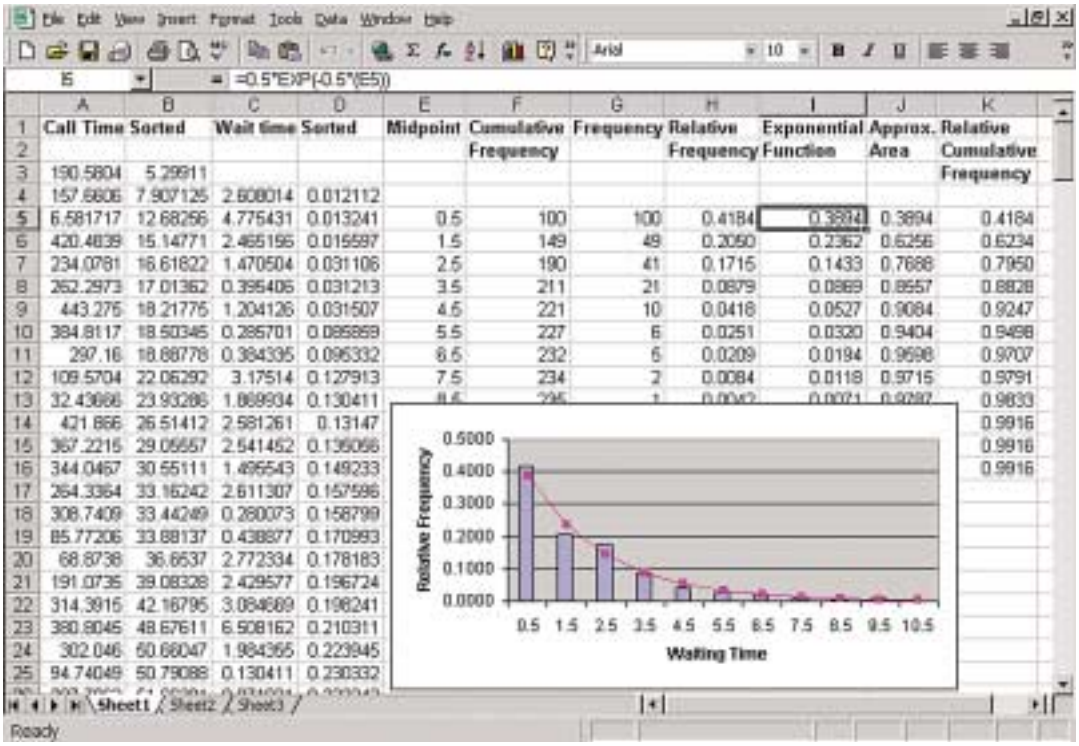
b) Draw the graph of $y = ke^{-kx}$ on your relative frequency plot. Comment on the fit of this curve to the data.

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

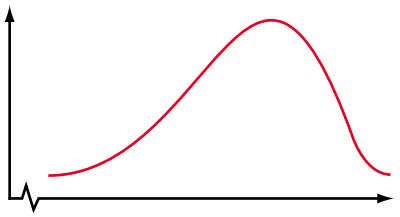
1. Explain the terms *discrete*, *continuous*, *symmetric*, *positively skewed*, and *negatively skewed*.
2. Give at least two examples of data that might result in
   a) a bimodal distribution
   b) an exponential distribution
   c) a positively-skewed distribution
3. Are summary measures such as mean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

Apply, Solve, Communicate

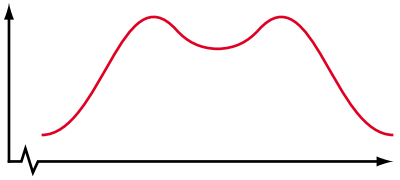
A

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

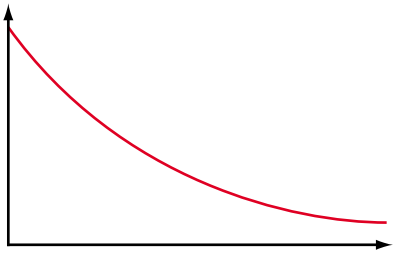
i)



ii)



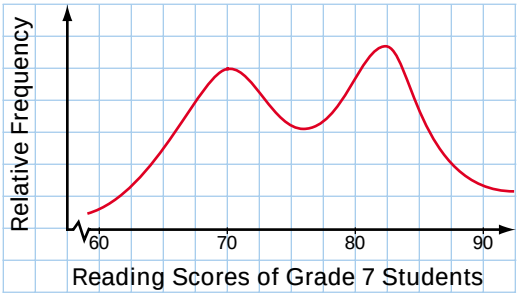
iii)



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.

8.2 Properties of the Normal Distribution

Many physical quantities like height and mass are distributed symmetrically and unimodally about the mean. Statisticians observe this bell curve so often that its mathematical model is known as the **normal distribution**.

Perhaps the most remarkable thing about the normal distribution is that a single mathematical formulation turns out to be the best model for statistical data from so many diverse sources. Like the exponential distribution in section 8.1, all normal curves can be described with a single form of equation. This equation can be used to calculate probabilities in a wide range of contexts. For example, in manufacturing, normal-distribution theory is used to design quality-control processes. The physical, social, and psychological sciences all make extensive use of normal models.

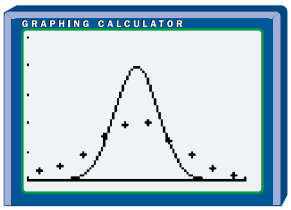
INVESTIGATE & INQUIRE: Plotting the Bell Curve

Rene 41® nickel superalloy, used to make components of jet and rocket engines, is a mixture of several metals, including molybdenum. The nominal percent of molybdenum by mass for this alloy is 9.75%. Dynajet Ltd. is considering two alloy suppliers for its line of jet-engine components. An analysis of the molybdenum content in 100 samples from supplier A and 75 samples from supplier B produced the following results:

Molybdenum (parts per 10 000)	Supplier A Frequency	Relative Frequency	Supplier B Frequency	Relative Frequency
970–971	3	0.030	1	0.013
971–972	5	0.050	2	0.027
972–973	9	0.090	9	0.120
973–974	15	0.150	15	0.200
974–975	19	0.190	20	0.267
975–976	20	0.200	16	0.213
976–977	14	0.140	8	0.107
977–978	9	0.090	3	0.040
978–979	4	0.040	1	0.013
979–980	2	0.020	0	0

1. Draw or generate a relative-frequency graph for each supplier. Using a graphing calculator or Fathom™, plot the relative frequencies versus the mid-interval values for each supplier. Use the same axis scales for each data set.

2. Fit a normal curve to each data set. Use parameters $\mu = 975$ and $\sigma = 1$ to start or use your own choice of μ and σ. If using a graphing calculator, enter the **normalpdf(function** in the **Y= editor** as Y1=normalpdf(X,975,1). Using Fathom™, open the Graph menu, select Plot Function, and use the **normalDensity() function** to enter normalDensity(*x*,975,1). Adjust the parameters μ and σ to improve the fit as much as possible.



3. a) Using the techniques of section 2.1, create a table like the one below for each data set. Calculate the sample mean amount $\bar{x}$ of molybdenum for each supplier.

Supplier A				
Molybdenum (parts per 10 000)	Frequency (f)	Mid-Interval Value (x)	$f \times x$	$f \times x^2$
970–971	3	970.5	2912	2.826×10^6
971–972				
…				
Total	$n = \sum f$ $= 100$		$\sum fx$	$\sum fx^2$

   b) Compare the sample mean values $\bar{x}_A$ and $\bar{x}_B$ with the means μ_A and μ_B of the normal curves you fitted to each data set.

4. a) Calculate the sample standard deviation for each data set, using the formula $s = \sqrt{\dfrac{\sum fx^2 - n(\bar{x})^2}{n-1}}$, the **stdDev(function** of the graphing calculator, or the **standard deviation function** in Fathom™.

   b) Compare these standard deviations with the standard deviations of the normal curves you fitted to each data set.

5. The product-development team at DynaJet Ltd. decide that alloy will be used only if its molybdenum content falls in the range 972.5 to 979.5 parts per 10 000. Based on your curve-fitting analysis, which supplier would you advise DynaJet to consider first? Which supplier would you expect to be more expensive?

Project Prep

You will need an understanding of the normal distribution to complete your probability distributions project.

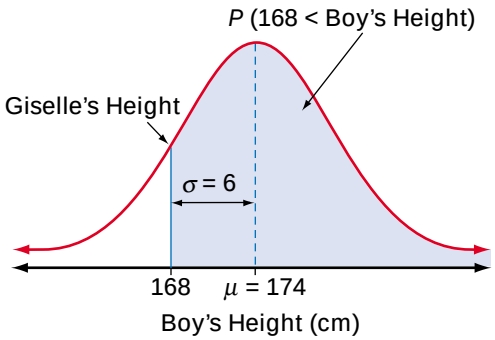
A population that follows a normal distribution can be completely described by its mean, μ, and standard deviation, σ. The symmetric, unimodal form of a normal distribution makes both the mode and median equal to the mean. As you saw in the investigation, the smaller the value of σ, the more the data cluster about the mean, so the narrower the bell shape. Larger values of σ correspond to more dispersion and a wider bell shape.

Example 1 Predictions From a Normal Model

Giselle is 168 cm tall. In her high school, boys' heights are normally distributed with a mean of 174 cm and a standard deviation of 6 cm. What is the probability that the first boy Giselle meets at school tomorrow will be taller than she is?

Solution

You need to find the probability that the height of the first boy Giselle meets falls in a certain range. The normal distribution is a continuous curve, so the probability is the area under the appropriate part of the probability-distribution curve.



You need to find $P(168 < \text{boy's height})$. For this distribution, $168 = \mu - \sigma$. So, you can use the fact that, for *any* normally distributed random variable X, $P(\mu - \sigma < X < \mu - \sigma) \doteq 68\%$.

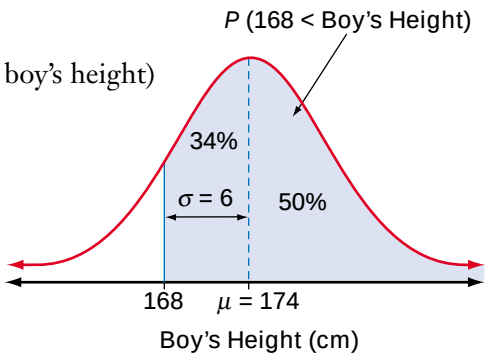
Break $P(168 < \text{boy's height})$ up as

$P(168 < \text{boy's height}) = P(168 < \text{boy's height} < 174) + P(174 > \text{boy's height})$

Since the heights of boys are normally distributed with mean $\mu = 174$ cm and standard deviation $\sigma = 6$ cm, 68% of the boys' heights lie within 1σ (6 cm) of the mean. So, the range 168 cm to 180 cm will contain 68% of the data. Because the normal distribution is symmetric, the bottom half of this range, 168 cm to 174 cm (the mean value), contains $\frac{68\%}{2} = 34\%$ of the data.

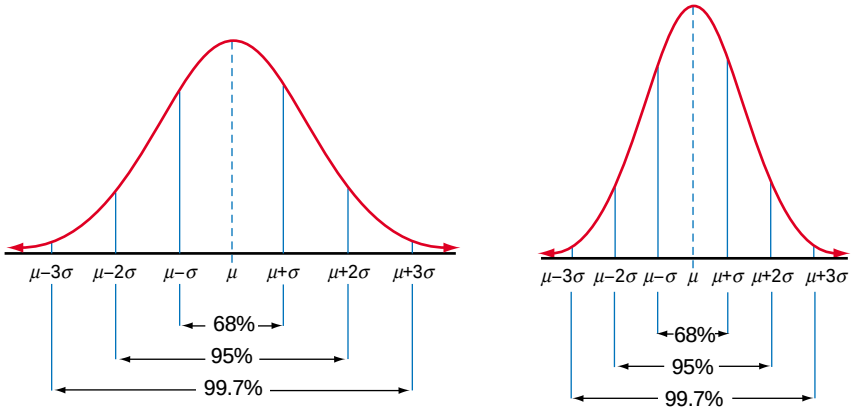
Therefore,

$$\begin{aligned} P(168 < \text{boy's height}) &= P(168 < \text{boy's height} < 174) + P(174 > \text{boy's height}) \\ &\doteq 34\% + 50\% \\ &= 84\% \end{aligned}$$



So, the probability that Giselle will meet a boy taller than 168 cm is approximately 84%.

Your work on normal distributions so far should suggest a pattern of highly predictable properties. For example, although different normal distributions have different standard deviations, the value of σ scales the distribution curve in a simple, regular way. Doubling the value of σ, for example, is equivalent to a stretch of factor 2 in the horizontal direction about the line $x = \mu$, and a stretch of factor 0.5 in the vertical direction about the x-axis. The area under the curve between the values $\mu - \sigma$ and $\mu + \sigma$ is multiplied by $2 \times 0.5 = 1$, and therefore remains at approximately 68%. This fact was used in the solution of Example 1. These normal curves show how the areas are the same between each corresponding set of verticals.



Notice that, for any normal distribution X,

- approximately 68% of the data values of X will lie within the range $\mu - \sigma$ and $\mu + \sigma$
- approximately 95% of the data values of X will lie within the range $\mu - 2\sigma$ and $\mu + 2\sigma$
- approximately 99.7% of the data values of X (almost all of them) will lie within the range $\mu - 3\sigma$ and $\mu + 3\sigma$

An understanding of these properties is vital for correctly interpreting statistics. For example, designers and administrators of psychometric tests need to understand the structure of normal distributions.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

To learn more about the normal distribution, visit the above web site and follow the links. Write a brief description of how changes to the mean and standard deviation affect the graph of the normal distribution.

Equations and Probabilities for Normal Distributions

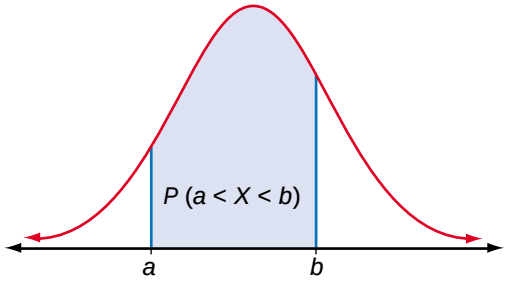
The curve of a normal distribution with mean μ and standard deviation σ is given by the equation

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

A graphing calculator uses this equation when you select the normalpdf(function from the DISTR menu. The initials *pdf* stand for **probability density function**. Statisticians use this term for the "curve equation" of any continuous probability distribution.

As with the exponential distributions you studied in section 8.1, the family resemblance of all normal distributions is reflected in the form of the equations for these distributions. The bracketed expression $\left(\frac{x-\mu}{\sigma}\right)$ creates a peak at $x = \mu$, the centre of the characteristic bell shape. This exponent also scales the function horizontally by a factor of σ. Using calculus, it can be shown that the vertical scaling factor of $\frac{1}{\sigma\sqrt{2\pi}}$ ensures that the area under the whole curve is equal to 1.

Probabilities for normally distributed populations, as for other continuous distributions, are equal to areas under the distribution curve. The area under the curve from $x = a$ to $x = b$ gives the probability $P(a < X < b)$ that a data value X will lie between the values a and b.



No simple formula exists for the areas under normal distribution curves. Instead, these areas have to be calculated using a more accurate version of "counting grid squares." However, you can simplify such calculations considerably by using z-scores.

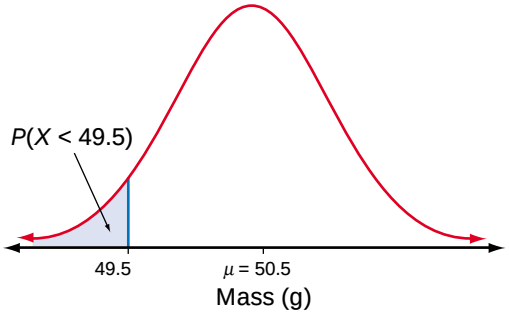
Recall from section 2.6 that the z-score for a value of the random variable is $z = \frac{x-\mu}{\sigma}$. The distribution of the z-scores of a normally distributed variable is a normal distribution with mean 0 and standard deviation 1. This particular distribution is often called the **standard normal distribution**. Areas under this normal curve are known to a very high degree of accuracy and can be printed in table format, for easy reference. A table of these areas can be found on pages 606 and 607.

Example 2 Normal Probabilities

All That Glitters, a sparkly cosmetic powder, is machine-packaged in a process that puts approximately 50 g of powder in each package. The actual masses have a normal distribution with mean 50.5 and standard deviation 0.6. The manufacturers want to ensure that each package contains at least 49.5 g of powder. What percent of packages do not contain this much powder?

Solution 1: Using a Normal Distribution Table

Let X be the mass of powder in a package. The probability required is $P(X < 49.5)$. This probability is equal to the shaded area under the normal curve shown at the right.

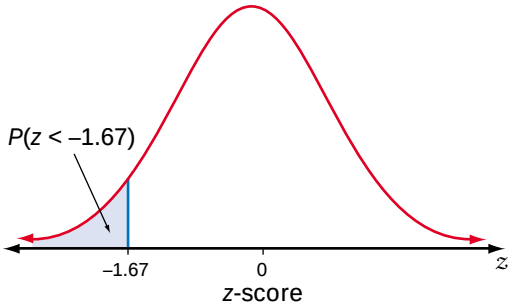


To calculate $P(X < 49.5)$, first find the corresponding z-score.

$$z = \frac{x - \mu}{\sigma}$$
$$= \frac{49.5 - 50.5}{0.6}$$
$$= -1.67$$

Use the table of Areas Under the Normal Distribution Curve on pages 606 and 607 to find the probability that a standard normal variable is less than this z-score. The table gives an area of 0.0475 for $z = -1.67$, so

$$P(X < 49.5) = P(Z < -1.67)$$
$$\doteq 0.0475$$



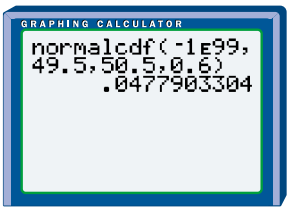
So, 4.8% of packages contain less than 49.5 g of powder.

Solution 2: Using a Graphing Calculator

If X is the mass of powder in a package, the required probability is $P(X < 49.5)$. A probability of this form, $P(X < x)$, is called a **cumulative probability**. From the DISTR menu, select the normalcdf(function. The initials *cdf* stand for cumulative density function.

The syntax for this function is normalcdf(*lower bound*, *upper bound*, μ, σ).

You need all the area to the left of the line $x = 49.5$, so the lower bound should be as far to the left along the x-axis as your calculator will go. Therefore, enter normalcdf(⁻1E99, 49.5, 50.5, 0.6).



So, 4.8% of packages contain less than 49.5 g of powder.

As described in section 2.6, a z-score is the number of standard deviations a value lies above the mean. Negative z-scores represent values that lie below the mean. You can use z-scores to rank any set of data, using the standard deviation as a unit of measure. Thus, in Example 2, the mass 49.5 g has a z-score of -1.67, or 1.67 standard deviations below the mean. As well as allowing the probability to be read from the normal distribution table, converting 49.5 g to a z-score gives you a useful measure of where this value lies in the distribution.

Example 3 Standardized Test Scores

To qualify for a special program at university, Sharma had to write a standardized test. The test had a maximum score of 750, with a mean score of 540 and a standard deviation of 70. Scores on this test were normally distributed. Only those applicants scoring above the third quartile (the top 25%) are admitted to the program. Sharma scored 655 on this test. Will she be admitted to the program?

Solution 1: Using Z-Scores

One way to answer this question is to calculate Sharma's z-score.

$$z = \frac{x - \mu}{\sigma}$$
$$= \frac{655 - 540}{70}$$
$$\doteq 1.64$$

So, Sharma's score lies 1.64 standard deviations above the mean. The percent of area lying more than one standard deviation above or below the mean is $100\% - 68\% = 32\%$. Half of this area, or 16%, lies *above* $\mu + \sigma$. Therefore, Sharma's z-score of 1.64 places her well above the third quartile.

Solution 2: Using Percentiles

Just like a raw data set, a probability distribution can be analysed in terms of percentiles. The 80th percentile of a normal distribution of X, for example, is the value of x such that

$$P(X \leq x) = 80\%$$

To discover whether Sharma's score is in the top 25%, calculate $P(X < 655)$ using the **normalcdf(function** on a graphing calculator or $P(Z < 1.64)$ using the table of Areas Under the Normal Distribution Curve on pages 606 and 607.

$$P(X < 655) \doteq 0.9498.$$

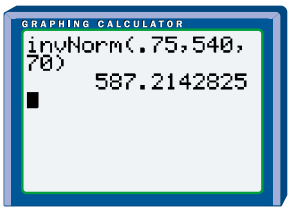
Sharma's score is, therefore, just below the 95th percentile for this test. At least 94% of all those who wrote this test scored lower than Sharma.

Solution 3: Using a Cut-Off Score

Find the score which represents the lower limit of the top 25% of all scores (the third quartile). This means you need to find a raw score, Q_3, for which the cumulative probability is

$$P(X < Q_3) = 0.75$$

You can use the **invNorm(function** in the DISTR menu on a graphing calculator to find Q_3.



So, a score of 587 or above qualifies as third quartile (top 25%). Sharma will be accepted to the program, since her score of 655 easily exceeds 587.

Key Concepts

- The normal distribution models many quantities that vary symmetrically about a mean value. This distribution has a characteristic bell shape.
- The standard normal distribution is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. Probabilities for this distribution are given in the table of Areas Under the Normal Distribution Curve on pages 606 and 607 and can be used if a graphing calculator is not available. The table gives probabilities of the form $P(Z < a)$.
- A z-score, $z = \frac{x - \mu}{\sigma}$, indicates the number of standard deviations a value lies from the mean. A z-score also converts a particular normal distribution to the standard normal distribution, so z-scores can be used with the areas under the normal distribution curve to find probabilities.
- Z-scores, percentiles, and cut-off scores are all useful techniques for analysing normal distributions.

Communicate Your Understanding

1. Explain how to find the probability that the next student through the classroom door will be at most 172 cm tall, if heights for your class are normally distributed with a mean of 167 cm and a standard deviation of 2.5 cm.
2. Ron works in quality control for a company that manufactures steel washers. The diameter of a washer is normally distributed with a mean of 8 mm and a standard deviation of 1 mm. Ron checked a washer and found that its diameter was 9.5 mm. Should Ron be surprised by this result? Explain your answer to Ron.
3. Suppose the probability for a certain outcome based on a test score is 0.1. What information do you need in order to find the raw score for this outcome? Explain how to find the raw score.

Practise

A

1. Copy and complete the chart below, assuming a normal distribution for each situation.

Mean, μ	Standard Deviation, σ	Probability
12	3	$P(X < 9) =$
30	5	$P(X < 25) =$
5	2.2	$P(X > 6) =$
245	18	$P(233 < X < 242) =$

Apply, Solve, Communicate

B

2. Michael is 190 cm tall. In his high school, heights are normally distributed with a mean of 165 cm and a standard deviation of 20 cm. What is the probability that Michael's best friend is shorter than he is?
3. **Application** Testing has shown that new CD players have a mean lifetime of 6.2 years. Lifetimes for these CD players are normally distributed and have a standard deviation of 1.08 years. If the company offers a 5-year warranty on parts and labour, what percent of CD players will fail before the end of the warranty period?
4. A class of 135 students took a final examination in mathematics. The mean score on the examination was 68% with a standard deviation of 8.5%. Determine the percentile rank of each of the following people:
   a) Joey, who scored 78%
   b) Shaheed, who got 55%
   c) Michelle, who was very happy with her mark of 89%
5. **Communication** There have been some outstanding hitters in baseball. In 1911, Ty Cobb's batting average was 0.420. In 1941, Ted Williams batted 0.406. George Brett's 0.390 average in 1980 was one of the highest since Ted Williams. Batting averages have historically been approximately normally distributed, with means and standard deviations as shown below.

Decade	Mean	Standard Deviation
1910s	0.266	0.0371
1940s	0.267	0.0326
1970s–1980s	0.261	0.0317

Compute z-scores for each of these three outstanding hitters. Can you rank the three hitters? Explain your answer.

6. To run without causing damage, diameters of engine crankshafts for a certain car model must fall between 223.92 mm and 224.08 mm. Crankshaft diameters are normally distributed, with a mean of 224 mm and a standard deviation of 0.03 mm. What percent of these crankshafts are likely to cause damage?

7. The daily discharge of lead from a mine's tailings is normally distributed, with a mean level of 27 mg/L and a standard deviation of 14 mg/L. On what proportion of days will the daily discharge exceed 50 mg/L?

8. **Application** The success rate of shots on goal in an amateur hockey league's games is normally distributed, with a mean of 56.3% and a standard deviation of 8.1%.

a) In what percent of the games will fewer than 40% of shots on goal score?

b) In what percent of the games will the success rate be between 50% and 60%?

c) What is the probability that a game will have a success rate of more than 66% of shots?

9. The daily sales by Hank's Hot Dogs have a mean of $572.50 and a standard deviation of $26.10.

a) What percent of the time will the daily sales be greater than $564?

b) What percent of the time will the daily sales be greater than $600?

10. **Inquiry/Problem Solving** The average heights of teenage girls are normally distributed, with a mean of 157 cm and a standard deviation of 7 cm.

a) What is the probability that a teenage girl's height is greater than 170 cm?

b) What range of heights would occur about 90% of the time?

C

11. **Inquiry/Problem Solving** The results of a university examination had a mean of 55 and a standard deviation of 13. The professor wishes to "bell" the marks by converting all the marks so that the mean is now 70, with a standard deviation of 10.

a) Develop a process or formula to perform this conversion.

b) What would happen to a mark of 80 after the conversion?

c) What would happen to a mark of 40 after the conversion?

12. The heights of 16-month-old oak seedlings are normally distributed with a mean of 31.5 cm and a standard deviation of 10 cm. What is the range of heights between which 75% of the seedlings will grow?

13. The coach of a track team can send only the top 5% of her runners to a regional track meet. For the members of her team, times for a 1-km run are normally distributed with a mean of 5.6 min and a standard deviation of 0.76 min. What is the cut-off time to determine which members of the team qualify for the regional meet?

8.3 Normal Sampling and Modelling

Many statistical studies take sample data from an underlying normal population. As you saw in the investigation on page 422, the distribution of the sample data reflects the underlying distribution, with most values clustered about the mean in an approximate bell shape. Therefore, if a population is believed or expected to be normally distributed, predictions can be made from a sample taken from that population. As you will see, this predictive process is most reliable when the sample size is large.

Example 1 **Investment Returns**

The annual returns from a particular mutual fund are believed to be normally distributed. Erin is considering investing in this mutual fund. She obtained a sample of 20 years of historic returns, which are listed in the table below.

Year	Return (%)	Year	Return (%)
1	7.2	11	6.4
2	12.3	12	27.0
3	17.1	13	14.5
4	17.9	14	25.2
5	10.8	15	−0.5
6	19.3	16	2.4
7	12.2	17	16.7
8	−13.1	18	12.8
9	20.2	19	2.9
10	18.6	20	18.8

a) Determine the mean and standard deviation of these data.

b) Assuming the data are normally distributed, what is the probability that an annual return will be

i) at least 9%?

ii) negative?

c) Out of the next ten years, how many years should Erin expect to show returns greater that 6%? What assumptions are necessary to answer this question?

Solution 1: Using a Normal Distribution Table

a) Using the formulas for the sample mean and sample standard deviation,

$$\bar{x} = \frac{\sum x}{n}$$
$$= \frac{248.7}{20}$$
$$= 12.435$$

$$s = \sqrt{\frac{\sum x^2 - n(\bar{x})^2}{n-1}}$$
$$= \sqrt{\frac{4805.37 - 20 \times (12.435)^2}{19}}$$
$$= 9.49$$

The mean of the data is $\bar{x} \doteq 12.4$ and the standard deviation is $s = 9.49$.

b) **i)** Find the z-score of 9.

$$z = \frac{x - \bar{x}}{s}$$
$$= \frac{9 - 12.4}{9.49}$$
$$= -0.36$$

Then, use the table of Areas Under the Normal Distribution Curve on pages 606 and 607 to find the probability.

$$P(X \geq 9) = P(Z \geq -0.36)$$
$$= 1 - P(Z \leq +0.36)$$
$$= 1 - 0.3594$$
$$\doteq 0.64$$

The probability of at least a 9% return is 0.64, or 64%.

ii) $$P(X < 0) = P\left(Z < \frac{0 - 12.4}{9.49}\right)$$
$$= P(Z < -1.31)$$
$$= 0.0951$$

The probability of a negative return is approximately 10%.

c) First, find the probability of a return greater than 6%.

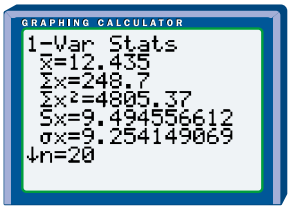
$$\begin{aligned} P(X > 6) &= P\left(Z > \frac{6 - 12.4}{9.5}\right) \\ &= P(Z > -0.674) \\ &= 1 - P(Z < -0.674) \\ &\doteq 1 - 0.25 \\ &= 0.75 \end{aligned}$$

In any given year, there is a 75% probability of a return greater than 6%. Therefore, Erin can expect such a return in seven or eight years out of the next ten years. This prediction depends on the assumptions that the return data are normally distributed, and that this distribution does not change over the next ten years.

Solution 2: Using a Graphing Calculator

a) To find the mean and standard deviation, enter the returns in L1.

Use the **1-Var Stats command** from the STAT CALC menu to obtain the following information.



From the calculator, the mean is $\bar{x} \doteq 12.4$ and the standard deviation is $s \doteq 9.49$.

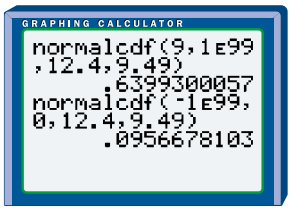
Recall that, since the data is a sample, you should use the value of Sx rather than σx.

b) Since the underlying population is normally distributed, use a normal distribution with a mean of 12.4 and a standard deviation of 9.49 to make predictions about the population.

i) $P(X \geq 9)$ is the area under the normal curve to the right of $x = 9$.

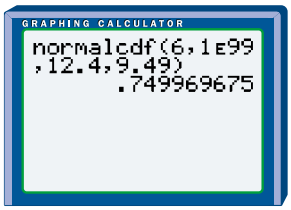
Therefore, use the **normalcdf(function** as shown on the screen on the right.

This screen shows the probability of a return of at least 9% as 0.64, or 64%.



ii) For the area to the left of $x = 0$, use the **normalcdf(function** as shown on the screen on the right.

The probability of a negative return is approximately 10%.



c) You can use the **normalcdf(function** to find the probability of a return greater than 6% and then proceed as in Solution 1.

Solution 3: Using a Spreadsheet

a) Copy the table into a spreadsheet starting at cell A1 and ending at cell B21. In cells E2 and E3, respectively, calculate the **mean** and **standard deviation** using the AVERAGE function and the STDEV function in Microsoft® Excel or by selecting Tools / Numeric Tools / Analysis... / Descriptive Statistics in Corel® Quattro® Pro.

b) **i)** You can use the **NORMDIST function** to find the cumulative probability for a result up to a given value. Subtract this probability from 1 to find the probability of an annual return of at least 9%:

E6: =NORMDIST(9,E2,E3,TRUE)
E7: =1-E6

From cell E7, you can see that $P(X \geq 9) \doteq 0.64$.

ii) Copy the **NORMDIST function** and change the value for X to 0 to find that there is about a 10% probability that next year's returns will be negative.

c) Copy the formula again and change the value for X to 6. The **NORMDIST function** will calculate the probability of an annual return of up to 6%. Subtracting this probability from 1 gives the probability of an annual return of greater than 6% (see cell G7).

$$P(X \geq 6) = 1 - P(X < 6))$$
$$\doteq 0.75$$

So, Erin should expect returns greater than 6% in seven or eight out of the next ten years.

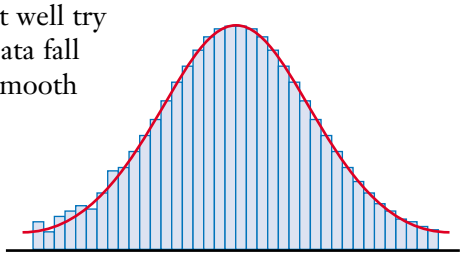
	A	B	C	D	E	F	G	H
1	Year	Return (%)						
2	1	7.2		mean	12.435			
3	2	12.3		Sx	9.494557			
4	3	17.1						
5	4	17.9		x	9	0	6	
6	5	10.8		P(X<x)	0.358766	0.095149	0.248962891	
7	6	19.3		P(X>x)	0.641244		0.751037109	
8	7	12.2						
9	8	-13.1						
10	9	20.2						
11	10	18.6						
12	11	6.4						
13	12	27						
14	13	14.5						
15	14	25.2						
16	15	-0.5						
17	16	2.4						
18	17	16.7						
19	18	12.8						
20	19	2.9						
21	20	18.8						
22								

Sheet1 / Sheet2 / Sheet3 /

Ready

Normal Models for Discrete Data

All the examples of normal distributions you have seen so far have modelled continuous data. There are many situations, however, where discrete data can also be modelled as normal distributions. For instance, the earthquake data presented in the Chapter Problem are discrete, but a statistician might well try a normal model for them. If the data set is reasonably large, and the data fall into a symmetric, unimodal bell shape, it makes sense to try fitting a smooth normal curve to them. Just as with the continuous investment data in Example 1, the normal model can then be used to make predictions.



Example 2 Candy Boxes

A company produces boxes of candy-coated chocolate pieces. The number of pieces in each box is assumed to be normally distributed with a mean of 48 pieces and a standard deviation of 4.3 pieces. Quality control will reject any box with fewer than 44 pieces. Boxes with 55 or more pieces will result in excess costs to the company.

a) What is the probability that a box selected at random contains exactly 50 pieces?

b) What percent of the production will be rejected by quality control as containing too few pieces?

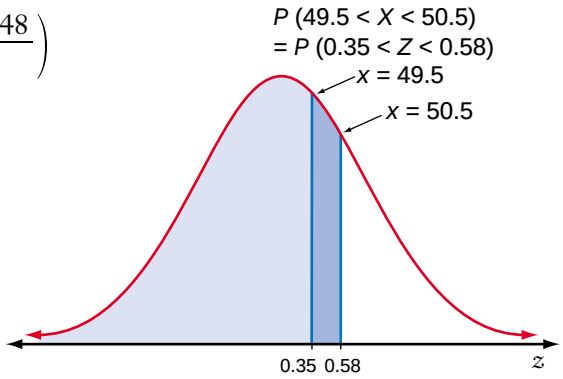
c) Each filling machine produces 130 000 boxes per shift. How many of these will lie within the acceptable range?

d) If you owned this company, what conclusions might you reach about your current production process?

Solution 1: Using a Normal Distribution Table

a) For a continuous distribution, the probabilities are for ranges of values. For example, all probabilities listed in the table of Areas Under the Normal Distribution Curve on pages 606 and 607 are of the form $P(Z < z)$, not $P(Z = z)$. Since a normal model is being used, discrete values such as "50 chocolates" have to be treated as though they were continuous. The simplest way is to calculate the value $P(49.5 < X < 50.5)$, treating a value of 50 chocolates as "between 49.5 and 50.5 chocolates." This technique, called **continuity correction**, enables predictions to be made about discrete quantities using a normal model.

$$P(49.5 < X < 50.5) = P\left(\frac{49.5 - 48}{4.3} < Z < \frac{50.5 - 48}{4.3}\right)$$
$$= P(0.35 < Z < 0.58)$$
$$= P(Z < 0.58) - P(Z < 0.35)$$
$$= 0.7190 - 0.6368$$
$$= 0.082$$



The probability that a box selected at random contains exactly 50 pieces is 0.082, or 8.2%.

b) A box is rejected by quality control if it has fewer than 44 pieces. A box with exactly 44 pieces is accepted, a box with exactly 43 pieces is not. With continuity correction, therefore, the probability required is $P(X < 43.5)$.

$$P(X < 43.5) = P\left(Z < \frac{43.5 - 48}{4.3}\right)$$
$$= P(Z < -1.05)$$
$$= 0.147$$

Approximately 14.7% of the production will be rejected by quality control as containing too few pieces.

c) The probability of a box being in the acceptable range of 44 to 54 pieces inclusive is

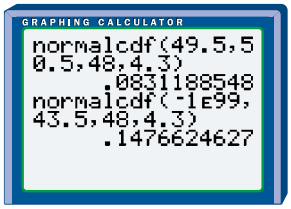
$$P(43.5 < X < 54.5) = P\left(\frac{43.5 - 48}{4.3} < Z < \frac{54.5 - 48}{4.3}\right)$$
$$= P(-1.05 < Z < +1.51)$$
$$= P(Z < +1.51) - P(Z < -1.05)$$
$$= 0.9345 - 0.1469$$
$$= 0.788$$

Thus, out of 130 000 boxes, approximately $130\,000 \times 0.788$ or 102 000 boxes, to the nearest thousand, will be within the acceptable range.

d) Clearly there are too many rejects with the current process. The packaging process should be adjusted to reduce the standard deviation and get a more consistent number of pieces in each box. If such improvements are not possible, you might have to raise the price of each box to cover the cost of the high number of rejected boxes.

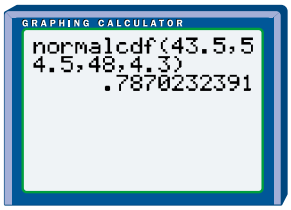
Solution 2: Using a Graphing Calculator

a) To find $P(X = 50)$ using a graphing calculator, apply continuity correction and calculate $P(49.5 < X < 50.5)$ using the **normalcdf(function**. Thus, the probability of a box containing exactly 50 pieces is approximately 0.083, or 8.3%.



b) You need to find $P(X < 43.5)$. Again use the **normalcdf(function**. Approximately 14.8% of the production will be rejected for having too few candies.

c) From the calculator, $P(43.5 < X < 54.5) = 0.787$. So, out of 130 000 boxes, $130\,000 \times 0.787$ or 102 000 boxes, to the nearest thousand, will lie within the acceptable range.



d) See Solution 1.

Key Concepts

- For a sample from a normal population,
  - the distribution of frequencies in the sample data tends to follow the same bell-shaped curve as the underlying distribution
  - the sample mean, $\bar{x}$, and sample standard deviation, s, provide estimates of the underlying parameters, μ and σ
  - the larger the sample from a normal population, the more reliably the sample data will reflect the underlying population
- Discrete data can sometimes be modelled by a normal distribution. Continuity correction should be used to calculate probabilities with these models.

Communicate Your Understanding

1. Why do you think it may be dangerous to make predictions about a population based on a single random sample from that population?
2. Give an example of a probability calculation that involves a continuity correction. Explain, using a sketch graph, why the continuity correction is needed in your example.

Apply, Solve, Communicate

B

Use appropriate technology for these problems. Assume that all the data are normally distributed.

1. A police radar unit measured the speeds, in kilometres per hour, of 70 cars travelling along a straight stretch of highway in Ontario. The speed limit on this highway is 100 km/h. The speeds of the 70 cars are listed below.

115	95	95	103	91	105	124	92
111	128	112	128	113	103	105	114
116	120	107	108	118	103	113	110
108	119	114	111	94	92	118	111
103	118	104	103	118	114	115	95
126	106	92	120	122	112	100	129
120	130	115	96	111	97	98	115
141	114	118	117	104	105	107	103
122	98	117	110	113	95		

a) Calculate the mean and standard deviation of these data.

b) What is the probability that a car travelling along this stretch of highway is speeding?

2. Application A university surveyed 50 graduates from its engineering program to determine entry-level salaries. The results are listed below.

$30 400	$31 458	$31 338	$30 950	$33 560
$33 378	$32 250	$32 254	$32 000	$29 547
$32 228	$31 050	$29 074	$36 943	$33 830
$29 549	$30 838	$29 746	$31 116	$30 477
$39 708	$28 730	$34 802	$29 522	$33 582
$40 728	$33 570	$35 495	$36 416	$33 627
$29 639	$28 525	$34 169	$30 965	$33 912
$27 485	$34 299	$33 500	$30 477	$27 028
$40 829	$33 294	$28 528	$32 428	$31 526
$38 953	$36 246	$37 239	$28 469	$27 385

a) Calculate the mean and standard deviation of these data.

b) What is the probability that a graduate of this program will have an entry-level salary below $30 000?

3. Communication A local grocery store wants to obtain a profile of its typical customer. As part of this profile, the dollar values of purchases for 30 shoppers were recorded. The results are listed below.

$65.53	$57.11	$75.45	$53.73	$32.44
$68.85	$85.48	$65.60	$73.67	$73.11
$73.06	$56.51	$44.70	$101.77	$82.25
$45.30	$93.25	$62.47	$39.98	$68.45
$69.79	$56.90	$53.16	$65.09	$81.70
$88.95	$52.63	$68.22	$101.63	$64.45

a) Calculate the mean and standard deviation of these data.

b) What is the probability that a typical shopper's purchase is more than $60?

c) What is the probability that a typical shopper's purchase is less than $50?

d) Does the grocery store need to collect more data? Give reasons for your answer.

For questions 4 through 7 you will need access to the E-STAT database.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

To connect to E-STAT, go to the above web site and follow the links.

4. In E-STAT, access the People/Labour/Job Search section.

a) Download the monthly help wanted index data from 1991–2001 for Canada and Ontario.

b) Make a histogram for the data for Canada. Do these data appear to be normally distributed?

c) Calculate the mean and standard deviation of the data for Canada and the data for Ontario.

d) Do your calculations show that it was easier to find a job in Ontario than in the rest of Canada during this period?

5. From E-STAT, access the Inflation data table.

a) Download the table into a spreadsheet or Fathom™.

b) Calculate the mean and standard deviation of the data.

c) What is the probability that the inflation rate in a year was less than 3%?

6. From E-STAT, access the Greenhouse Gas Emissions data table.

a) Download the table into a spreadsheet or Fathom™.

b) Calculate the mean and standard deviation of the data.

c) Use these data to formulate and solve two questions involving probability.

7. Inquiry/Problem Solving From E-STAT, access a data table on an area of interest to you.

a) Download the table into a spreadsheet or Fathom™.

b) Use these data to formulate and solve two questions involving probability.

8. Babe Ruth played for the New York Yankees from 1920 to 1934. The list below gives the number of home runs he hit each year during that time.

54	59	35	41	46	25	47	60
54	46	49	46	41	34	22	

a) Calculate the mean and standard deviation of these data.

b) Estimate the probability that he would have hit more than 46 home runs if he had played another season for the Yankees.

9. Application The weekly demand for laser printer cartridges at Office Oasis is normally distributed with a mean of 350 cartridges and a standard deviation of 10 cartridges. The store has a policy of avoiding stockouts (having no product on hand). The manager decides that she wants the chance of a stockout in any given week to be at most 5%. How many cartridges should the store carry each week to meet this policy?

10. Application The table gives estimates of wolf population densities and population growth rates for the wolf population in Algonquin Park.

Year	Wolves/100 km²	Population Growth Rate
1988–89	4.91	
1989–90	2.47	−0.67
1990–91	2.80	0.12
1991–92	3.62	0.26
1992–93	2.53	−0.36
1993–94	2.23	−0.13
1994–95	2.82	0.24
1995–96	2.75	−0.02
1996–97	2.33	−0.17
1997–98	3.04	0.27
1998–99	1.59	−0.65

a) Group the population densities into intervals and make a frequency diagram. Do these data appear to be normally distributed?

b) Use the same method to determine whether the growth rate data appear to be normally distributed.

c) Is it possible that you would change your answer to part b) if you had a larger set of data? Explain why or why not.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

To learn more about the decline in the wolf population in Algonquin Park, visit the above web site and follow the links.

11. Suppose the earthquake data given on page 411 are approximately normally distributed. Estimate the probability that the number of earthquakes in a given year will be greater than 30. What assumptions do you have to make for your estimate?

ACHIEVEMENT CHECK

Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application

12. A soft-drink manufacturer runs a bottle-filling machine, which is designed to pour 355 mL of soft drink into each can it fills. Overfilling costs money, but underfilling may result in unhappy consumers and lost sales. The quality-control inspector measured the volume of soft drink in 25 cans randomly selected from the filling machine. The results are shown below.

351.82	349.52	354.15	351.57	347.91
350.08	357.55	351.43	350.24	354.58
351.18	354.86	350.76	349.11	360.16
353.08	347.60	356.41	350.62	349.50
352.12	349.80	348.86	345.07	353.60

a) Calculate the mean and standard deviation of these data.

b) What is the probability that a can holds between 352 mL and 356 mL of soft drink?

c) Should the manufacturer adjust the filling machine? Justify your answer.

C

13. **Inquiry / Problem Solving** Given a chronological sequence of data, statistical fluctuations from day to day or year to year are sometimes reduced if you group or combine the data into longer periods.

a) Copy and complete the following table, using the data from Example 1 on page 432. Explain how each entry in the third column is calculated.

Year	Return (%)	Five-Year Return (%)
1	7.2	
2	12.3	
3	17.1	
4	17.9	
5	10.8	
6	19.3	
7	…	

b) Find the sample mean and standard deviation of the data in the third column. Compare these with the sample mean and standard deviation you found for the yearly returns in Example 1. Are the 5-year returns normally distributed Is there an advantage to longer-term investment in this fund?

c) Make a similar study of the earthquake data on page 411.

Normal Probability Plots

If it is not known whether the underlying population is normally distributed, you can use a graphing calculator or software to construct a normal probability plot of the sample data. A **normal probability plot** graphs the data according to the probabilities you would expect if the data are normal, using z-scores. If the plot is approximately linear (a straight line), the underlying population can be assumed to be normally distributed.

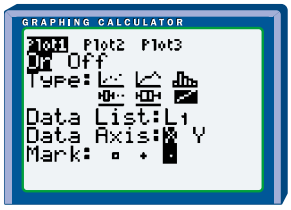
Using a Graphing Calculator

A toy tricycle comes with this label: "Easy-To-Assemble. An adult can complete this assembly in 20 min or less." Thirty-six adults were asked to complete the assembly of a tricycle, and record their times. Here are the results:

16	10	20	22	19	14	30	22	12	24	28	11	17	13	18	19	17	21
29	22	16	28	21	15	26	23	24	20	8	17	21	32	18	25	22	20

1. Using a graphing calculator, enter these data in L1. Find the mean and standard deviation of the data.

2. Make a normal probability plot of the data. Using STAT PLOT, select 1:Plot1, and the settings shown below.



Based on the plot, are assembly times normally distributed?

3. **a)** What is the probability that an adult can complete this assembly in 20 min or less?

b) What proportion of adults should complete this assembly within 15 to 30 min?

Using Fathom™

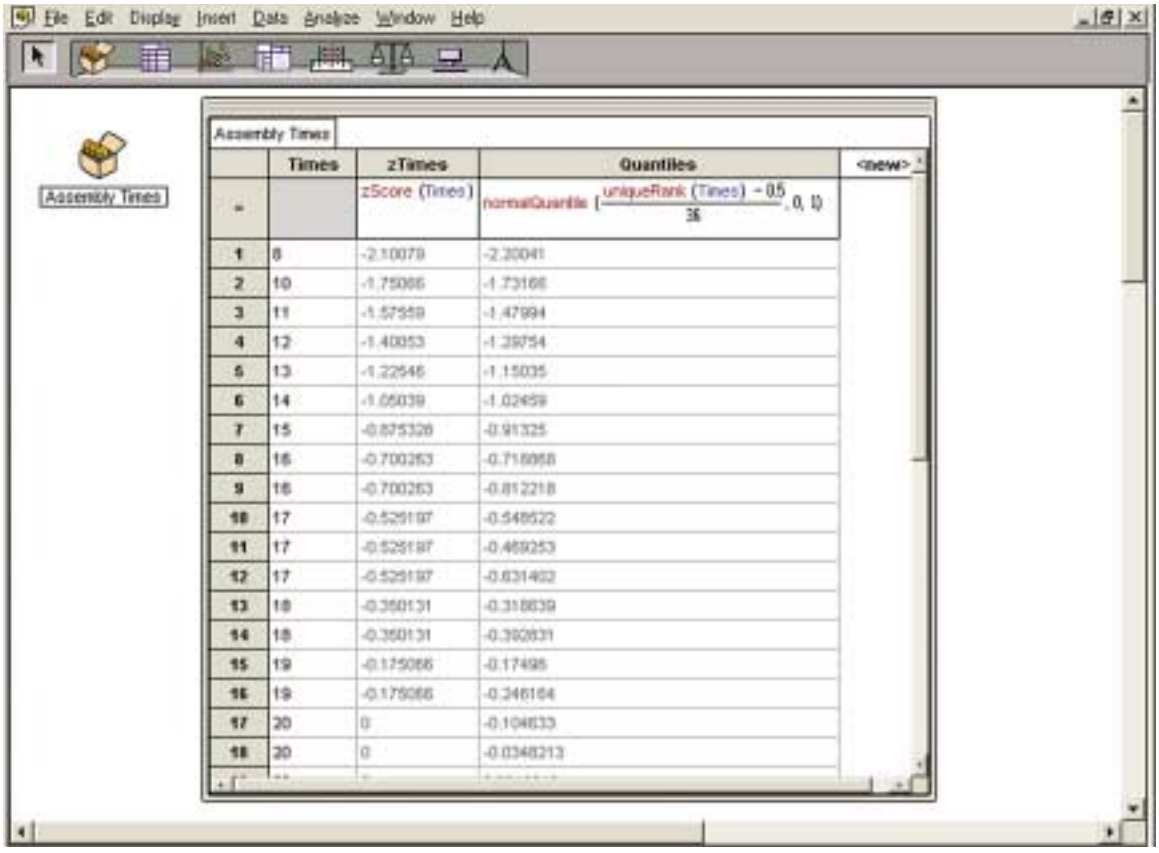
4. **a)** Open Fathom™, and open a new document if necessary. Drag a new collection box to the workspace. Rename the collection Assembly Times, and create 36 new cases.

b) Drag a new case table to the workspace. Name the first column Times, the second column zTimes, and the third column Quantiles.

c) **i)** Enter the time data in the first column. **Sort** it in ascending order.

ii) Edit the formula in the second column to **zScore**(Times). This will calculate the z-scores for the data.

iii) Edit the formula in the third column to **normalQuantile**((uniqueRank(Times) – 0.5)/36, 0, 1). This formula will calculate the z-scores of the quantiles corresponding to each entry in the Times column. The **uniqueRank( )** function returns the "row number" of the *sorted* data. Note that most of the quantile z-scores in the screen below are different from the z-scores for the corresponding data.



	Times	zTimes	Quantiles
1	8	-2.10079	-2.20041
2	10	-1.75066	-1.73166
3	11	-1.57559	-1.47994
4	12	-1.40053	-1.28754
5	13	-1.22546	-1.15035
6	14	-1.05039	-1.02459
7	15	-0.875326	-0.91325
8	16	-0.700263	-0.716868
9	16	-0.700263	-0.612218
10	17	-0.525197	-0.548522
11	17	-0.525197	-0.469253
12	17	-0.525197	-0.631402
13	18	-0.350131	-0.318839
14	18	-0.350131	-0.392831
15	19	-0.175066	-0.17495
16	19	-0.175066	-0.246164
17	20	0	-0.104833
18	20	0	-0.0348213

d) Drag a new graph to the workspace. Drag the Times title to the horizontal axis, and the Quantiles title to the vertical axis to generate a normal probability plot. Calculate the linear correlation coefficient for Times and Quantiles and comment on how near to linear this graph is. Are the data normally distributed?

e) Double click on the collection to open the **inspector**. Choose the Measures tab. Create four measures: Mean, StdDev, P20orLess, and P15to30. Use the **mean**, **standard deviation**, and **normalCumulative** functions to calculate the mean, the standard deviation, and the answers to question 3.

5. For each of the questions 1 to 6, 8, and 12 of section 8.3, pages 439 to 441, use a normal probability plot to determine how close to a normal distribution each data set is.

6. Let $x_1, x_2, \ldots x_n$ be a set of data, ranked in increasing order so that $x_1 \leq x_2 \leq \ldots \leq x_n$.
For $i = 1, 2, \ldots, n$, define the **quantile** z_i by $P(Z < z_i) = \dfrac{(i - 0.5)}{n}$,
where Z is a standard normal distribution (mean 0, standard deviation 1).

a) For a data set of your choice, plot a graph of z_i against x_i. Remember to sort the x-values into increasing order. Use the **invNorm(function** on your graphing calculator, or the table of Areas Under the Normal Distribution Curve on pages 606 and 607, to calculate the z-values. Notice that these quantile z-values are different from the z-scores in earlier sections.

b) Compare this graph with the normal probability plot for the data set. Explain your findings.

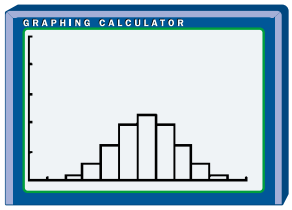
c) Explain why, if the data are normally distributed, a graph of z_i against x_i should be close to a straight line.

8.4 Normal Approximation to the Binomial Distribution

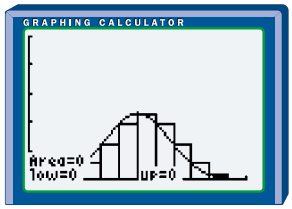
The normal distribution is a continuous distribution. Many real-life situations involve discrete data, such as surveys of people or testing of units produced on an assembly line. As you saw in section 8.3, these situations can often be modelled by a normal distribution. If the discrete data have a binomial probability distribution and certain simple conditions are met, the normal distribution makes a very good approximation. This approximation allows the probabilities of value ranges to be calculated more easily than with the binomial formulas.

INVESTIGATE & INQUIRE: Approximating a Binomial Distribution

1. On your graphing calculator, enter the integers from 0 to 12 in L1. This list represents the number of successes in 12 trials.
2. With the cursor on L2, enter the **binompdf(function**. Use 12 for the number of trials and 0.5 for the probability of success. The calculator will place into L2 the binomial probabilities for each number of successes from L1.
3. Use **STAT PLOT** to construct a histogram using L1 as Xlist and L2 as Freq. Your result should look like the screen on the right.



4. Now, construct a normal distribution approximation for this binomial distribution. From the DISTR menu, select the **ShadeNorm(function** and use 0, 0, 6, and $\sqrt{3}$ as the parameters. The normal approximation should now appear, superimposed on the binomial histogram.



5. Investigate the effect of changing the binomial probability of success, p. Use p values of 0.3, 0.1, 0.95, and 0.7. Keep the number of trials at 12 and repeat steps 2 through 4 for each value of p. In step 4, use the **ShadeNorm(function** and enter 0, 0, $12p$, $\sqrt{12pq}$ for each value of p. You will have to adjust the **window settings** for some of these situations.

6. For each different p value in step 5, make a subjective estimate of how good an approximation the normal distribution is for the underlying binomial distribution. Summarize your findings.
7. Statistical theory states that for a binomial distribution with n trials and probability of success p, a normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$ is a reasonable approximation as long as both np and nq are greater than 5. Recall that $q = 1 - p$ is the probability of failure. Do your results meet these criteria?

Example 1 Bank Loans

A bank found that 24% of its loans to new small businesses become delinquent. If 200 small businesses are selected randomly from the bank's files, what is the probability that at least 60 of them are delinquent? Compare the results from the normal approximation with the results from the calculations using a binomial distribution.

Solution 1: Using a Normal Distribution Table or Graphing Calculator

Here, $np = (200)(0.24)$ and $nq = (200)(0.76)$
$= 48$ $= 152$
> 5 > 5

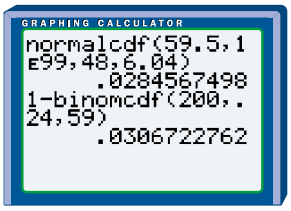
Therefore, the normal approximation should be reasonable.

$\mu = np$ $\sigma = \sqrt{npq}$
$= (200)(0.24)$ $= \sqrt{(200)(0.24)(0.76)}$
$= 48$ $= 6.04$

Using the normal approximation with continuity correction, and referring to the table of Areas Under the Normal Distribution Curve on pages 606 and 607, the required probability is

$$P(X > 59.5) = P\left(Z > \frac{59.5 - 48}{6.04}\right)$$
$$= P(Z > 1.90)$$
$$= 1 - P(Z < 1.90)$$
$$= 0.029$$

Using a graphing calculator, this normal approximation probability can be calculated using the function normalcdf(59.5,1E99,48,6.04). Alternatively, the binomial probability can be calculated as 1–binomcdf(200,.24,59). The calculator screen shows the results of the two methods.



The normal approximation gives a result of 0.0285. The binomial cumulative density function gives 0.0307. The results of the two methods are close enough to be equal to the nearest percent. So, the probability that at least 60 of the 200 loans are delinquent is approximately 3%.

Solution 2: Using Fathom™

Open Fathom, and open a new document if necessary. Drag a new **collection** box to the workspace. Rename it Loans.

Double click on the collection to open the **inspector**. Choose the Measures tab.
Create two measures: Binomial and Normal.
Edit the Binomial formula to 1 – binomialCumulative(59,200,.24).
Edit the Normal formula to 1 – normalCumulative(59.5,48,6.04).
These functions give the same probability values as the equivalent graphing calculator functions.

Example 2 Market Share

QuenCola, a soft-drink company, knows that it has a 42% market share in one region of the province. QuenCola's marketing department conducts a blind taste test of 70 people at the local mall.

a) What is the probability that fewer than 25 people will choose QuenCola?

b) What is the probability that exactly 25 people will choose QuenCola?

Solution 1 Using a Normal Distribution Table

a) For this binomial distribution,

$np = 70(0.42)$ and $nq = 70(0.58)$
$= 29.4$ $= 40.6$
> 5 > 5

Therefore, you can use the normal approximation.

$\mu = np$ $\sigma = \sqrt{npq}$
$= (70)(0.42)$ $= \sqrt{(70)(0.42)(0.58)}$
$= 29.4$ $= 4.13$

Using the normal approximation with continuity correction,

$P(X < 24.5) = P(Z < -1.19)$
$= 0.117$

So, there is a 12% probability that fewer than 25 of the people surveyed will choose QuenCola.

b) The probability of exactly 25 people choosing QuenCola can also be calculated using the normal approximation.

$P(24.5 < X < 25.5) = P(-1.19 < Z < -0.94)$
$= P(Z < -0.94) - P(Z < -1.19)$
$= 0.1736 - 0.1170$
$= 0.057$

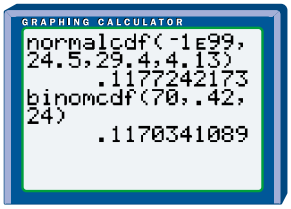
The probability that exactly 25 people will choose QuenCola is approximately 5.7%.

Solution 2: Using a Graphing Calculator

a) To find the probability that 24 or fewer people will choose QuenCola, you need to use a parameter of 24.5 for the normal approximation, but 24 for the cumulative binomial function. The normalcdf(and binomcdf(functions differ by less than 0.001.

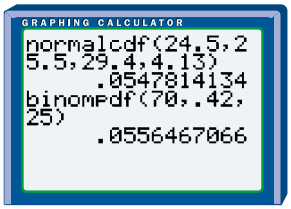
$P(X_{\text{normal}} < 24.5) = 0.118$
$P(X_{\text{binomial}} \leq 24) = 0.117$

There is a 12% probability that fewer than 25 of the people surveyed will choose this company's product.



b) The normalcdf(and binompdf(functions return the following values.

$P(24.5 < X_{\text{normal}} < 25.5) = 0.0548$
$P(X_{\text{binomial}} = 25) = 0.0556$



Here, the normal approximation is a bit smaller than the value calculated using the actual binomial distribution. The manual calculation gave a slightly higher value because of rounding. The probability that exactly 25 of the people surveyed will choose this company's product is 5.6%.

Key Concepts

- A discrete binomial probability distribution can be approximated with a continuous normal distribution as long as np and nq are both greater than 5.
- To approximate the mean and standard deviation, the values $\mu = np$ and $\sigma = \sqrt{npq}$ are used.
- As with other discrete data, continuity correction should be applied when approximating a binomial distribution by a normal distribution.

Communicate Your Understanding

1. The probability of an airline flight arriving on time is 88%. Explain how to use the normal approximation to find the probability that at least 350 of a random sample of 400 flights arrived on time.
2. Construct an example of a binomial distribution for which np is less than 5, and for which the normal distribution is *not* a good approximation. Then, show that the condition $np > 5$ and $nq > 5$ is needed.
3. Has technology reduced the usefulness of the normal approximation to a binomial distribution? Justify your answer.

Practise

A

1. For which of the binomial distributions listed below is the normal distribution a reasonable approximation?
   a) $n = 60, p = 0.4$ **b)** $n = 45, p = 0.1$
   c) $n = 80, p = 0.1$ **d)** $n = 30, p = 0.8$

B

2. Copy the table. Use the normal approximation to complete the table.

Sample Size, n	Probability of Success, p	μ	σ	Probability
60	0.4			$P(X < 22) =$
200	0.7			$P(X < 160) =$
75	0.6			$P(X > 50) =$
250	0.2			$P(X > 48) =$
1000	0.8			$P(780 < X < 840) =$
90	0.65			$P(52 < X < 62) =$
100	0.36			$P(X = 40) =$
3000	0.52			$P(X = 1650) =$

Apply, Solve, Communicate

Use the normal approximation to the binomial distribution, unless otherwise indicated.

3. It is estimated that 62% of television viewers "channel surf" during commercials. A market-research firm surveyed 1500 television viewers. What is the probability that at least 950 of them were channel surfing?
4. **Application** Salespeople sometimes advertise their products by telephoning strangers. Only about 1.5% of these "cold calls" result in a sale. Toni makes cold calls 8 h per day for 5 days. The average time for a cold call is 90 s. What is the probability that Toni gets at least 30 new customers for the week?
5. A theatre found that 7% of people who purchase tickets for a play do not show up. If the theatre's capacity is 250 people, what is the probability that there are fewer than 20 "no shows" for a sold-out performance?
6. A magazine reported that 18% of car drivers use a cellular phone while driving. In a survey of 200 drivers, what is the probability that exactly 40 of them will use a cellular phone while driving? Compare the results of using the binomial distribution and the normal approximation.
7. The human-resources manager at a company knows that 34% of the workforce belong to a union. If she randomly surveys 50 employees, what is the probability that exactly 30 of them do not belong to a union? Compare the results of using the binomial distribution with the results of using the normal approximation.

8. **Application** A recent survey of a gas-station's customers showed that 68% paid with credit cards, 29% used debit cards, and only 3% paid with cash. During her eight-hour shift as cashier at this gas station, Serena had a total of 223 customers.

a) What is the probability that

i) at least 142 customers used a credit card?

ii) fewer than 220 customers paid with credit or debit cards?

b) What is the expected number of customers who paid Serena with cash?

9. A computer-chip manufacturer knows that 72% of the chips produced are defective. Suppose 3000 chips are produced every hour.

a) What is the probability that

i) at least 800 chips are acceptable?

ii) exactly 800 chips are acceptable?

b) Compare the results of using the binomial distribution with those found using the normal approximation.

10. Calculate the probability that 200 rolls of two dice rolls will include

a) more than 30 sums of 5

b) between 30 and 40, inclusive, sums of 5

11. On some busy streets, diamond lanes are reserved for taxis, buses, and cars with three or more passengers. It is estimated that 20% of cars travelling in a certain diamond lane have fewer than three passengers. Sixty cars are selected at random.

a) Use the normal approximation to find the probability that

i) fewer than 10 cars have fewer than three passengers

ii) at least 15 cars have fewer than three passengers

b) Compare these results with those found using the binomial distribution.

c) How would the results compare if 600 cars were selected?

ACHIEVEMENT CHECK

Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application

12. The probability of winning a large plush animal in the ring-toss game at the Statsville School Fair is 8%.

a) Find the probability of winning in at least 10% of 300 games, using

i) a binomial distribution

ii) a normal distribution

b) Predict how the probabilities of winning at least 50 times in 500 games will differ from the answers in part a). Explain your prediction.

c) Verify your prediction in part b) by calculating the probabilities using both distributions. Do your calculations support your predictions?

d) When designing the game, one student claims that having $np > 3$ and $nq > 3$ is a sufficient test for the normal approximation. Another student claims that np and nq both need to be over 10. Whom would you agree with and why?

C

13. **Inquiry / Problem Solving**

a) A newspaper knows that 64% of the households in a town are subscribers. If 50 households are surveyed randomly, how many of these households should the newspaper expect to be subscribers?

b) The marketing manager for the newspaper has asked you for an upper and lower limit for the number of subscribers likely to be in this sample. Find upper and lower bounds of a range that has a 90% probability of including this number.

8.5 Repeated Sampling and Hypothesis Testing

Repeated Sampling

When you draw a sample from a population, you often use the sample mean, $\bar{x}$, as an estimate of the population mean, μ, and the sample standard deviation, s, as an estimate of the population standard deviation, σ. However, the statistics for a single sample may differ radically from those of the underlying population. Statisticians try to address this problem by repeated sampling. Do additional samples improve the accuracy of the estimate?

INVESTIGATE & INQUIRE: Simulating Repeated Sampling

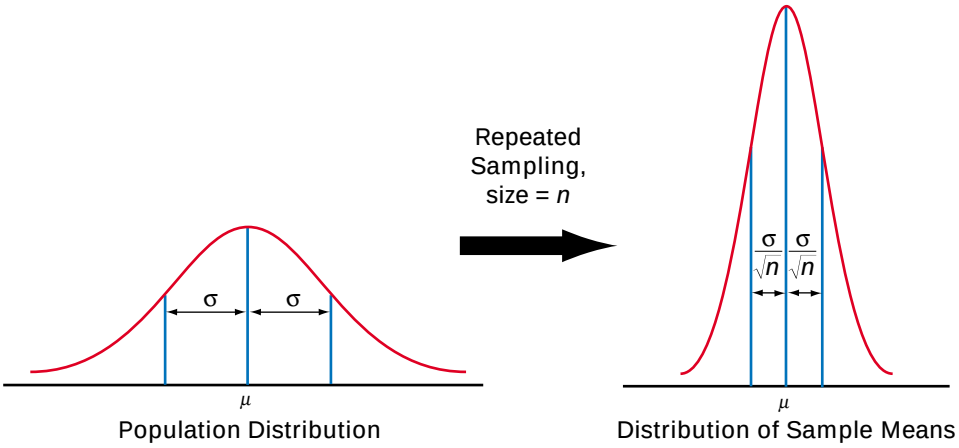
Simulate drawing samples of size 100 from a normally distributed population with mean $\mu = 10$ and standard deviation $\sigma = 5$. After 20 samples, examine the mean of the sample means.

The steps below outline a method using a graphing calculator. However, you can also simulate repeated sampling with a spreadsheet or statistical software such as Fathom™. See section 1.4 and the technology appendix descriptions of the software functions you could use.

1. Use the **mode settings** to set the number of decimal places to 2. Using the STAT EDIT menu, check that L1 and L2 are clear.
2. Place the cursor on L1. From the MATH PRB menu, select the **randNorm(function** and enter 10 as the mean, 5 as the standard deviation, and 100 as the number of trials. From the STAT CALC menu, select the **1-Var Stats command** to find the mean of the 100 random values in L1. Enter this mean in L2.
3. Repeat step 2 twenty times. You should then have 20 entries in L2. Each of these entries is the sample mean, $\bar{x}$, of a random sample of 100 drawn from a population with a mean $\mu = 10$ and a standard deviation $\sigma = 5$.
4. Use the **1-Var Stats command** to find the mean and standard deviation of L2.
5. Construct a histogram for L2. What does the shape of the histogram tell you about the distribution of the sample means?

When repeated samples of the same size are drawn from a normal population, the sample means will be normally distributed with a mean equal to the population mean μ.

The distribution of sample means will have a standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, where n is the sample size. Notice that if $n = 1$, the samples are single data points and $\sigma_{\bar{x}} = \sigma$. As n increases, however, $\sigma_{\bar{x}}$ decreases, so the distribution of sample means becomes more tightly grouped around the true mean. Usually a sample size of at least 30 is sufficient to give a reasonably accurate estimate of μ.



Compare, from your investigation results, the mean from L2 with the true population mean, $\mu = 10$. Then, compare the standard deviation of L2 with $\frac{5}{\sqrt{100}} = 0.5$. How close are your results to these theoretical values?

Example 1 Tire Wear

The tires on a rental-car fleet have lifetimes that are normally distributed with a mean life of 64 000 km and a standard deviation of 4800 km. Every week a mechanic checks the tires on ten randomly selected cars.

a) What are the mean and standard deviation of these samples?

b) How likely is the mechanic to find a sample mean of 62 500 km or less?

Solution

a) The mean of the sample means will be

$$\mu_{\bar{x}} = \mu$$
$$= 64\,000$$

The sample size is 40, assuming all four tires on each of the ten cars are checked. Therefore, the standard deviation of the sample means will be

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
$$= \frac{4800}{\sqrt{40}}$$
$$\doteq 759$$

b) To find the z-score for a sample mean of 62 500 km, you need to use the mean and standard deviation of the sample means. Thus,

$$P(\bar{x} < 63\ 500) = P\left(Z < \frac{62\ 500 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)$$
$$\doteq P\left(Z < \frac{62\ 500 - 64\ 000}{759}\right)$$
$$= P(Z < -1.98)$$
$$= 0.0239$$

You can find this value easily using the table on pages 606 and 607 or the **normalcdf(function** on a graphing calculator. There is a 2.4% probability that the mechanic will find a sample mean of 62 500 km or less.

Common sense suggests that the mean of a good-sized random sample from a population gives a better estimate of the population mean than a single value taken randomly from the population. As you have just seen, this idea can be quantified: sample means from a normal population are distributed with much smaller standard deviations than single data points from the population. The table below compares the two distributions.

Statistic	Sample Size	Mean	Standard Deviation
Single data point, x	1	μ	σ
Sample mean, $\bar{x}$	n	μ	$\frac{\sigma}{\sqrt{n}}$

Hypothesis Testing

Statisticians use the standard deviation of the sample mean to quantify the reliability of statistical studies and conclusions. Often, it is not possible to determine with certainty whether a statement is true or false. However, it is possible to test the strength of the statement, based on a sample. This procedure is called a **hypothesis test**.

Consider this scenario. A large candy manufacturer produces chocolate bars with labels stating "45 g net weight." A company spokesperson claims that the masses are normally distributed with a mean of 45 g and a standard deviation of 2 g. A mathematics class decides to check the truth of the 45 g label. They purchase and weigh 30 bars. The mean mass is 44.5 g. Is this evidence enough to challenge the company's claim?

You can construct a hypothesis test to investigate the 45-g net-weight claim, using these steps.

- *Step 1* State the hypothesis being challenged, **null hypothesis** (null means no change in this context). The null hypothesis is usually denoted H_0. So, for the chocolate bars, H_0: $\mu = 45$.
- *Step 2* State the alternative hypothesis H_1 (sometimes called H_a). You suspect that the mean mass may be lower than 45 g; so, H_1: $\mu < 45$.
- *Step 3* Establish a decision rule. How strong must the evidence be to reject the null hypothesis? If, for a normal distribution with a population mean of 45 g, the probability of a sample with a mean of 44.5 g is *very small*, then getting such a sample would be strong evidence that the actual population mean is not 45 g. The **significance level**, α, is the probability threshold that you choose for deciding whether the observed results are rare enough to justify rejecting H_0. For example, if $\alpha = 0.05$, you are willing to be wrong 5% of the time.
- *Step 4* Conduct an experiment. For the chocolate bars, you weigh the sample of 30 bars.
- *Step 5* Assume H_0 is true. Calculate the probability of obtaining the results of the experiment given this assumption. If the standard deviation for chocolate bar weights is 2 g, then

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{30}} = 0.365$$

and

$$P(\bar{x} < 44.5) = P\left(Z < \frac{44.5 - 45}{0.365}\right) \doteq P(Z < -1.37) = 0.0853$$

The probability of a sample mean of 44.5 or less, given a sample of size 30 from an underlying normal distribution with a mean of 45, is 8.5%.

- *Step 6* Compare this probability to the significance level, α: 8.5% is greater than 5%.
- *Step 7* Accept H_0 if the probability is greater than the significance level. Such probabilities show that the sample result is not sufficiently rare to support an alternative hypothesis. If the probability is less than the significance level reject H_0 and instead accept H_1. So, for the chocolate bars, you would accept H_0.
- *Step 8* Draw a conclusion. How reliable is the company's claim? In this case, the statistical evidence is slightly too weak to refute the company's claim.

Project Prep

You will be asked to conduct a hypothesis test when you complete your probability distributions project.

A significance level of 5% is the same as saying you want a **confidence level** of 95% that, if you should reject H_0, you are making the correct decision. Thus, with the chocolate bars, you can be only about 91% confident that H_1 is true. For a simple survey, a confidence level of 90% may be adequate, so you could choose to accept H_1 at that level. For more important decisions, such as the effect on humans of a new drug, a significance level of 1% or even less, corresponding to a confidence level of 99% or more, would be appropriate. The importance of hypothesis testing is that it allows a *quantifiable* check of such claims about value or effectiveness.

Example 2 Drug Effectiveness

A drug company tested a new drug on 250 pigs with swine flu. Historically, 20% of pigs contracting swine flu die from the disease. Of the 250 pigs treated with the new drug, 215 recovered. Make a hypothesis test of the drug's effectiveness, with a significance level of $\alpha = 1\%$.

a) Determine whether you can model this study with a normal distribution.

b) Set up a null hypothesis and an alternative hypothesis for the value of μ in the normal approximation.

c) Which hypothesis corresponds to the drug being effective? Explain.

d) Conduct the hypothesis test.

e) Can the company claim that its new drug is effective?

Solution

a) The data are discrete, there are only two possible outcomes, and successive trials are independent. So, the distribution is binomial with $n = 250$ and $p = 0.2$. Both np and nq are greater than 5. You can use the normal approximation to structure your test. So,

$$\begin{aligned}\mu &= np \\ &= 250 \times 0.2 \\ &= 50\end{aligned}$$

$$\begin{aligned}\sigma &= \sqrt{npq} \\ &= \sqrt{250(0.2)(0.8)} \\ &= \sqrt{40} \\ &\doteq 6.32\end{aligned}$$

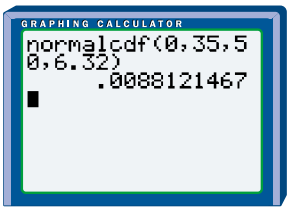
b) H_0: probability of death = 0.2 (the drug has no significant effect)

H_1: probability of death < 0.2 (the drug has some effect)

c) If the drug is effective, the probability of death will be reduced from the historical value of 20%, or 0.2. So, the alternate hypothesis, H_1, corresponds to the drug being effective.

d) Use software or a graphing calculator to find the probability that 35 or fewer pigs would die, using μ and σ from above.

$P(X < 35) \doteq 0.0088$



This probability is less than the significance level of $\alpha = 0.01$, indicating that the observed result would be a very rare event if the drug has no effect. Therefore, reject H_0.

e) The result of the hypothesis test is to accept H_1. The drug appears to have some effect on recovery rate, at the 1% significance level.

Key Concepts

- Sample means $\bar{x}$ from a normal population with mean μ and standard deviation σ are also normally distributed, with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, where n is the sample size.
- Hypothesis tests assume the truth of the null hypothesis H_0 and investigate an alternative hypothesis H_1.
- The significance level α is the probability of error which the researcher is willing to accept.
- The confidence level $1 - \alpha$ is the probability that the decision is correct.
- If the probability of an experimental outcome is greater than the significance level, accept H_0.
- If the probability of an experimental outcome is less than the significance level, reject H_0.

Communicate Your Understanding

1. A researcher performed a hypothesis test, getting a result of $P(X < x) = 0.08$.
   a) Should the researcher accept or reject H_0 if $\alpha = 10\%$? Explain your answer.
   b) Should the researcher accept or reject H_0 if $\alpha = 5\%$? Explain your answer.
2. Outline the steps in a hypothesis test to determine whether getting 13 heads in 20 coin tosses is sufficient evidence to show that the coin is biased.
3. List at least four situations where hypothesis tests might be used. State H_0 and H_1 for each situation.

Practice

A

1. Copy and complete the table below.

Poplulation Mean	Population Standard Deviation	Sample Size	Mean of Sample Means	Standard Deviation of Sample Means
20	6	49		
12	4	25		
5	2	36		
40	8	100		
8.4	3.2	68		
17.6	10.4	87		
73.9	21.4	250		

B

2. For each situation, test the significance of the experimental results, given H_0 and H_1.

	Sample Size	Number of Successes	α	H_0	H_1
a)	50	23	10%	$p = 0.4$	$p > 0.4$
b)	200	55	5%	$p = 0.3$	$p < 0.3$
c)	250	175	1%	$p = 0.68$	$p > 0.68$
d)	40	8	10%	$p = 0.15$	$p > 0.15$
e)	400	80	1%	$p = 0.15$	$p > 0.15$

Apply , Solve, Communicate

3. **Application** A machine makes steel bearings with a mean diameter of 39 mm and a standard deviation of 3 mm. The bearing diameters are normally distributed. A quality-control technician found that in a sample of 50 bearings the mean diameter was 44 mm. Test the significance of this result with a significance level of 10%. Decide whether the machine needs to be adjusted.
4. **Application** A newspaper stated that 70% of the population supported a particular candidate's position on health care. In a random survey of 50 people, 31 agreed with the candidate's position. Test the significance of this result with a confidence level of 90%. Should the newspaper print a correction?
5. **Communication** A new drug will not be considered for acceptance by Health Canada unless it causes serious side effects in less than 0.01% of the population. In a trial with 80 000 people, 9 suffered serious side effects. Test the significance of this result with $\alpha = 0.01$. Do you recommend that this drug be accepted by Health Canada? Explain your answer.
6. A certain soft-drink manufacturer claims that its product holds 28% of the market. In a blind taste test, 13 out of 60 people chose this product. Does this test support or refute the soft drink manufacturer's claim? Choose a significance level you feel is appropriate for this situation.

7. **Inquiry / Problem Solving** An insurance company claims that 38% of automobile accidents occur within 5 km of home. The company examined 400 recent accidents and found that 120 occurred within 5 km of the driver's home. Does this result support or refute the company's claim? Choose a significance level and justify your choice.

8. **Inquiry / Problem Solving** A student-loan program claims that the average loan per student per year is $7500. Dana investigated this statement by asking 50 students about this year's student loan. The mean of the results was $5800. What additional information does Dana need to test the significance of this result? What significance level would be appropriate here? Why?

9. In finance, the *strong form* of the efficient-market hypothesis states that studying financial information about stocks is a waste of time, since all public and private information that might affect the stock's price is already reflected in the price of the stock. However, a study of 450 stocks found that only about 8% had price movements that could be accounted for in this way. At what significance level could you accept the strong form of the efficient-market hypothesis?

10. Does advertising influence behaviour? Before a recent advertising campaign, a children's breakfast cereal held 8% of the market. After the campaign, 18 families out of a sample of 200 families indicated they purchased the cereal. Was the advertising campaign a success? Select a confidence level you feel is appropriate for this situation.

11. Researchers often use repeated samples to test a hypothesis. Why do you think they use this method? Outline some of its advantages and disadvantages.

ACHIEVEMENT CHECK

Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application

12. A new medication is designed to lower cholesterol. The cholesterol level in a group of patients is normally distributed with a mean of 6.15 mmol/L and a standard deviation of 1.35 mmol/L. A sample of 40 people used the medication for 30 days, after which their mean cholesterol level was 5.87 mmol/L. The drug company wants to have a 95% confidence level that the drug is effective before releasing it.

a) Should the company release the drug?

b) The drug appears to have more effect if used for longer periods of time. What mean cholesterol level would the test group have to reach for the company to be confident about releasing the drug? Support your answer with mathematical calculations.

13. With a graphing calculator, you can use the **Z-Test instruction** to test a value for the mean of a normal distribution, based on a sample data set. Enter the first 20 years of the earthquake data on page 411. Use the **Z-Test instruction**. Switch the input to Data (as opposed to Stats). Choose and enter test values of μ_0 and σ, and also choose an alternative hypothesis. What information does this test give you?

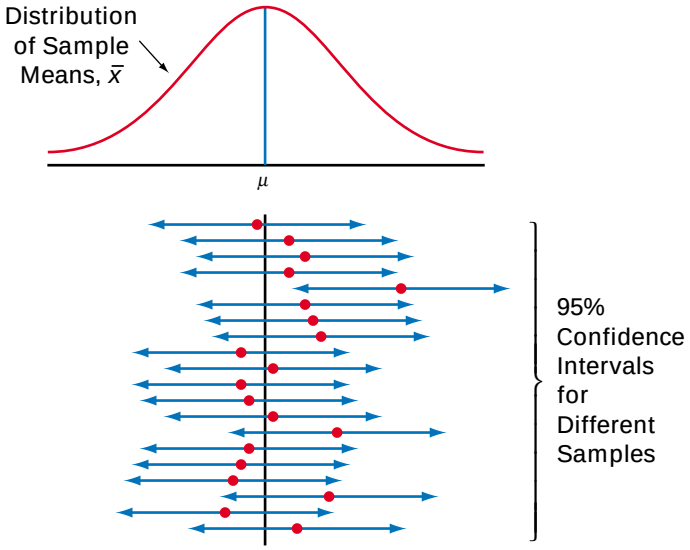
8.6 Confidence Intervals

Governments often commission polls to gauge support for new initiatives. The polling organization surveys a small number of people and estimates support in the entire population based on the sample results. Opinion polls printed in newspapers often include a note such as "These results are accurate to within ±3%, 19 times in 20." In a statement like this, the figure ±3% is a margin of error, and the phrase "19 times in 20" is a confidence level of 0.95 or 95%. The statement means that, if 43% of the sample supported an initiative, there is a 95% probability (you can be 95% confident) that between 40% and 46% of the population supports the initiative. The range 40% to 46% is an example of a confidence interval. The probability of error for this finding is $1 - 0.95 = 0.05$ or 5%.

In such surveys, you do not know the population mean, μ. However, you can determine **confidence intervals**, ranges of values within which μ is likely to fall. These intervals are centered on the sample mean, $\bar{x}$, and their widths depend on the confidence level, $1 - \alpha$. For example, a 95% confidence interval has a 0.95 probability of including μ. In a normal distribution, μ is as likely to lie above the confidence interval as below it, so

$$P\left(\bar{x} - z_{0.975}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{0.975}\frac{\sigma}{\sqrt{n}}\right) = 0.95,$$

where $P(Z < z_{0.975}) = 0.975$
$= 97.5\%$



A $(1-\alpha)$ or $(1-\alpha)\times 100\%$ confidence interval for μ, given a population standard deviation σ and a sample of size n with sample mean $\bar{x}$, represents the range of values

$$\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

The table gives a list of common confidence levels and their associated z-scores.

Confidence Level	Tail size, $\frac{\alpha}{2}$	z-score, $z_{\frac{\alpha}{2}}$
90%	0.05	1.645
95%	0.025	1.960
99%	0.005	2.576

Project Prep

You will need to construct a confidence interval when you complete your probability distributions project.

Example 1 Drying Times

A paint manufacturer knows from experience that drying times for latex paints have a standard deviation of $\sigma = 10.5$ min. The manufacturer wants to use the slogan: "Dries in T minutes" on its advertising. Twenty test areas of equal size are painted and the mean drying time, $\bar{x}$, is found to be 75.4 min. Find a 95% confidence level for the actual mean drying time of the paint. What would be a reasonable value for T?

Solution 1: Using Pencil and Paper

For a 95% confidence level, the acceptable probability of error, or significance level, is $\alpha = 5\%$, so $z_{\frac{\alpha}{2}} = z_{0.975} = 1.960$. Substituting into the formula gives

$$\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

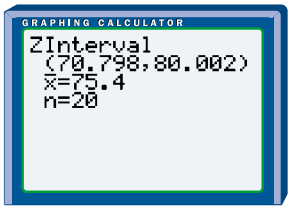
$$75.4 - (1.960)\left(\frac{10.5}{\sqrt{20}}\right) < \mu < 75.4 + (1.960)\left(\frac{10.5}{\sqrt{20}}\right)$$

$$70.8 < \mu < 80.0$$

So, the manufacturer can be 95% confident that the actual mean drying time for the paint lies between 70.8 min and 80.0 min. It would be reasonable to advertise: "Dries in 80 minutes."

Solution 2: Using a Graphing Calculator

Use the Z-Interval instruction in the STAT TESTS menu. Make sure the input is set to Stats. Enter the population parameter $\sigma = 10.5$ and the sample parameters $\bar{x} = 75.4$ and $n = 20$. Set the confidence level to 0.95. Select Calculate and press ENTER to find the required interval.



Often, you want to know the *proportion* of a population that have a particular opinion or characteristic. This proportion is simply p, the probability of success in the binomial distribution. When data are expressed in terms of proportions, the confidence interval formula becomes

$$\hat{p} - z_{\frac{\alpha}{2}}\frac{\sqrt{pq}}{\sqrt{n}} < p < \hat{p} + z_{\frac{\alpha}{2}}\frac{\sqrt{pq}}{\sqrt{n}},$$

where $\hat{p}$ is the proportion in the sample. This sample proportion is an estimate of the population proportion just as the sample mean, $\bar{x}$, is an estimate of the population mean, μ.

For many polls, the population proportion is not known. In fact, the purpose of the polls is to estimate this parameter. Since $p \doteq \hat{p}$ and $q \doteq 1 - \hat{p}$, you can *estimate* a confidence interval using the formula

$$\hat{p} - z_{\frac{\alpha}{2}}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} < p < \hat{p} + z_{\frac{\alpha}{2}}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

Example 2 Municipal Elections

Voter turnout in municipal elections is often very low. In a recent election, the mayor got 53% of the voters, but only about 1500 voters turned out.

a) Construct a 90% confidence interval for the proportion of people who support the mayor.

b) Comment on any assumptions you have to make for your calculation.

Solution

a) In this case, you want to find a confidence interval for a *proportion* of the population in a binomial distribution. Here, p is the proportion of the population who support the mayor, so you can use the election results to estimate this proportion:

$$\begin{aligned} p &\doteq \hat{p} \\ &= 0.53 \end{aligned} \qquad \text{and} \qquad \begin{aligned} q &= 1 - p \\ &\doteq 1 - 0.53 \\ &= 0.47 \end{aligned}$$

These estimated values gives a 90% confidence interval of

$$\hat{p} - z_{\frac{\alpha}{2}}\frac{\sqrt{pq}}{\sqrt{n}} < p < \hat{p} + z_{\frac{\alpha}{2}}\frac{\sqrt{pq}}{\sqrt{n}}$$

$$0.53 - 1.645\frac{\sqrt{0.53(0.47)}}{\sqrt{1500}} < p < 0.53 + 1.645\frac{\sqrt{0.53(0.47)}}{\sqrt{1500}}$$

$$0.51 < p < 0.55$$

So the mayor can be 90% confident of having the support of between 51% and 55% of the population.

b) You have to assume that the people who voted are representative of whole population. This assumption might not be valid because the people who take the trouble to vote are likely to be the ones most interested in municipal affairs.

Sample Sizes and Margin of Error

Given a sample of size n, from a normal population with standard deviation σ, you can use the sample mean to construct a confidence interval. You can express this confidence interval in terms of its central value and width. For example, suppose a sample of bolts has a 95% confidence interval of 7.51 mm $< \mu <$ 7.55 mm for the diameters. You can express this interval as a 95% confident estimate of 7.53 mm $\pm$ 0.02 mm for the mean diameter μ. Sometimes, however, a statistician might first decide on the confidence interval width, or **margin of error**, required, and then use this value to calculate the minimum sample size necessary to achieve this width. Opinion polls and other surveys are constructed in this way.

The width of the confidence interval $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ is $w = 2z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$. Solving this equation for the sample size, n, gives

$$n = \left(\frac{2z_{\frac{\alpha}{2}}\sigma}{w} \right)^2$$

If the pollsters know (or have a good estimate of) the population standard deviation, σ, they can use this formula to find the sample size they require for a survey to have a specified margin of error.

Example 3 Sample Size for Quality-Control Testing

Suppose the diameters of the bolts mentioned above have a standard deviation of $\sigma = 0.1$. How large a sample would you need to be 90% confident that the mean diameter is 7.53 mm $\pm$0.01 mm?

Solution

For a 90% confidence level, $z_{\frac{\alpha}{2}} = 1.645$. Substituting the known values into the equation for n gives

$$n = \left(\frac{2z_{\frac{\alpha}{2}}\sigma}{w}\right)^2$$
$$= \left(\frac{2(1.645)(0.1)}{(0.02)}\right)^2$$
$$\doteq 271$$

You would need a sample of about 270 bolts.

You can use a similar method to find the sample sizes required for surveys involving population proportions. The margin of error for a proportion is $w = 2z_{\frac{\alpha}{2}}\frac{\sqrt{pq}}{\sqrt{n}}$, so $n = \left(\frac{2z_{\frac{\alpha}{2}}\sqrt{pq}}{w}\right)^2$. This formula simplifies to

$$n = 4pq\left(\frac{z_{\frac{\alpha}{2}}}{w}\right)^2$$

Example 4 Sample Size for a Poll

A recent survey indicated that 82% of secondary-school students graduate within five years of entering grade 9. This result is considered accurate within plus or minus 3%, 19 times in 20. Estimate the sample size in this survey.

Solution

The result describes a confidence interval with a margin of error $w = 6\%$, and the confidence level is "19 times out of 20" or 95%, giving $\alpha = 0.05$.

Here, as in Example 2, you have a binomial distribution with data expressed as proportions. You can use the survey results to estimate p. Since 82% of the students in the survey graduated,

$p \doteq \hat{p}$ and $q = 1 - p$
$= 0.82$ $\doteq 1 - 0.82$
$= 0.18$

Substituting into the formula for n,

$$n = 4pq\left(\frac{z_{\frac{\alpha}{2}}}{w}\right)^2$$
$$\doteq 4(0.82)(0.18)\left(\frac{1.960}{0.06}\right)^2$$
$$\doteq 630$$

So, to obtain the stated level of accuracy and confidence, approximately 630 people would need to be surveyed.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

To learn more about confidence intervals, visit the above web site and follow the links.

Key Concepts

- A P% confidence interval for normally distributed data is given by $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$. Here, α is the acceptable probability of error, and $\alpha = (100 - P)\%$.
- A P% confidence interval for a population proportion, based on binomial data, is given by $\hat{p} \pm z_{\frac{\alpha}{2}} \frac{\sqrt{pq}}{\sqrt{n}}$.
- For a specified margin of error w, the required sample size is $n = \left(\frac{2 z_{\frac{\alpha}{2}} \sigma}{w} \right)^2$.

Communicate Your Understanding

1. How does the population distribution affect the distribution of the sample means?

2. **a)** Why is the z-score for 97.5% used to construct a 95% confidence interval? Support your answer with a sketched distribution.

b) Given that $z_{0.975} = 1.96$, approximately how many standard deviations wide is a 95% confidence interval?

3. To obtain the desired margin of error, an investigator must sample 2000 people. List at least three possible problems the investigator may encounter.

4. Interpret the following headline using confidence intervals: "4 out of every 13 Canadians think that the government should subsidize professional sports teams in Canada. These results are considered accurate within plus or minus 4%, nine times out of ten."

Practice

A

1. Construct the following confidence intervals.

	Confidence Level	$\bar{x}$	σ	Sample Size
a)	90%	15	3	36
b)	95%	30	10	75
c)	99%	6.4	2.5	60
d)	90%	30.6	8.7	120
e)	95%	41.8	12.6	325
f)	99%	4.25	0.86	44

2. Interpret each of the following statements using confidence intervals.

a) In a recent survey, 42% of high school graduates indicated that they expected to earn over $100 000 per year by the time they retire. This survey is considered accurate within plus or minus 3%, 19 times in 20.

b) A survey done by the incumbent MP indicated that 48% of decided voters said they would vote for him again in the next election. The result is considered accurate within plus or minus 5%, nine times in ten.

c) According to a market research firm, 28% of teenagers will purchase the latest CD by the rock band Drench. The result is considered accurate within ±4%, 11 out of 15 times.

Apply, Solve, Communicate

B

3. **Application** A large water pipeline is being constructed to link a town with a fresh water aquifer. A construction supervisor measured the diameters of 40 pipe segments and found that the mean diameter was 25.5 cm. In the past, pipe manufactured by the same company have had a standard deviation of 7 mm. Determine a 95% confidence interval for the mean diameter of the pipe segments.

4. **Application** A study of 55 patients with low-back pain reported that the mean duration of the pain was 17.6 months, with a standard deviation of 5.1 months. Assuming that the duration of this problem is normally distributed in the population, determine a 99% confidence interval for the mean duration of low-back pain in the population.

5. The Statsville school board surveyed 70 parents on the question: "Should school uniforms be instituted at your school?" 28% of the respondents answered "Yes." Construct a 90% confidence interval for the proportion of Statsville parents who want school uniforms.

6. The most popular name for pleasure boats is "Serenity." A survey of 200 000 boat owners found that 12% of their boats were named "Serenity." Determine a 90% confidence interval for the proportion of pleasure boats carrying this name.

7. The football coach wants an estimate of the physical-fitness level of 44 players trying out for the varsity team. He counted the number of sit-ups done in 2 min by each player. Here are the results:

38	95	86	63	68	73	26	43
90	30	71	100	92	57	71	67
47	56	68	61	61	92	83	50
66	51	87	64	80	58	60	103
14	39	88	75	60	87	70	66
95	26	75	61				

Construct a 90% confidence interval for the mean number of sit-ups that varsity football players are capable of performing.

8. The students in a school environment club are concerned that recycling efforts are failing in smaller communities. The amount of waste recycled in 33 towns with under 5000 households is given below:

12%	10%	12%	11%	13%	12%	18%
10%	33%	3%	10%	15%	12%	18%
20%	24%	18%	5%	13%	12%	14%
25%	17%	22%	11%	12%	26%	20%
17%	22%	11%	20%	30%		

a) Construct a 95% confidence interval for the mean percent of waste recycled in towns with under 5000 households.

b) **Communication** Write a letter to the mayors of towns with under 5000 households outlining the results.

9. A market-research firm asked 300 people about their shampoo-purchasing habits. Fifty-five people said they bought S'No Flakes. Determine a 95% confidence interval for the percent of people who purchase S'No Flakes.

10. A manufacturing company wants to estimate the average number of sick days its employees take per year. A pilot study found the standard deviation to be 2.5 days. How large a sample must be taken to obtain an estimate with a maximum error of 0.5 day and a 90% confidence level?

11. An industrial-safety inspector wishes to estimate the average noise level, in decibels (dB), on a factory floor. She knows that the standard deviation is 8 dB. She wants to be 90% confident that the estimate is correct to within ±2 dB. How many noise-level measurements should she take?

12. An ergonomics advisor wants to estimate the percent of computer workers who experience carpal-tunnel syndrome. An initial survey of 50 workers found three cases of the syndrome. To be 99% confident of an accuracy of ±2%, how many workers must the advisor survey?

13. **a)** Obtain a survey result from your local newspaper that contains accuracy information. Determine the sample size. State explicitly any assumptions you needed to make.

b) Try to find a survey result from your local newspaper that gives the sample size. Use this information to estimate the standard deviation for the population surveyed.

14. Take 12 random samples of ten data from the earthquake table on page 411. Use these samples to estimate a 90% confidence interval for the mean number of major earthquakes you should expect this year.

15. **Inquiry/Problem Solving** The table below gives a sample of population growth rates of wolves in Algonquin Park.

Year	Population Growth Rate
1989–90	–0.67
1990–91	0.12
1991–92	0.26
1992–93	–0.36
1993–94	–0.13
1994–95	0.24
1995–96	–0.02
1996–97	–0.17
1997–98	0.27
1998–99	–0.65

a) Use these samples to estimate the mean and standard deviation for the growth rate of the wolf population in Algonquin Park. Explain your results.

b) Assuming that the growth rates are normally distributed, estimate the probability that the growth rate for the wolf population is less than zero.

c) A population is in danger of extinction if its population growth rate is –0.05 or less. A study based on these samples claimed that there is a 71% probability that this wolf population is in danger of extinction. Is the study correct?

d) Construct a 90% confidence interval for the true population growth rate of wolves in Algonquin Park.

16. A social scientist wants to estimate the average salary of office managers in a large city. She wants to be 95% confident that her estimate is correct. Assume that the salaries are normally distributed and that $\sigma = \$1050$. How large a sample must she take to obtain the desired information and be accurate within $200?

17. **Communication** An opinion pollster determines that a sample of 1500 people should give a margin of error of 3% at the 95% confidence level 19 times in 20. The pollster decides that an efficient way to find a representative sample of 1500 people is to conduct the poll at Pearson International Airport. Discuss, in terms of techniques such as stratified sampling, how representative the poll will be.

ACHIEVEMENT CHECK

Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application

18. Emilio has played ten rounds of golf at the Statsville course this season. His mean score is 80 and the standard deviation is 4. Assume Emilio's golf scores are normally distributed.

a) Find the 95% confidence interval of Emilio's mean golf score.

b) Predict how the confidence interval would change if the standard deviation of the golf scores was 8 instead of 4. Explain your reasoning.

c) Find the 95% confidence interval of Emilio's mean golf score if the standard deviation was 8 instead of 4. Does the answer support your prediction? Explain.

d) Emilio's most recent golf score at the course is 75. He claims that his game has improved and this latest score should determine whether he qualifies for entry in the Statsville tournament. Should the tournament organizers accept his claim? Justify your answer mathematically.

C

19. Given a sample of size n with mean $\bar{x}$, the population mean μ can be estimated via the 95% confidence interval defined by the probability

$$P\left(\bar{x} - z_{0.975}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{0.975}\frac{\sigma}{\sqrt{n}}\right) = 0.95 \quad (1)$$

If a value μ_0 is assumed for μ, a 5% significance level hypothesis test H_0: $\mu = \mu_0$: against H_1: $\mu \neq \mu_0$ can be performed on the sample mean $\bar{x}$, using the probability

$$P\left(\mu_0 - z_{0.975}\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu_0 + z_{0.975}\frac{\sigma}{\sqrt{n}}\right) = 0.95 \quad (2)$$

a) Show that the probability in equation (1) can be rearranged as

$$P\left(|\bar{x} - \mu| < z_{0.975}\frac{\sigma}{\sqrt{n}}\right)$$

b) Show that the probability in equation (2) can also be rearranged into a similar form.

c) Use parts a) and b) to prove that, if μ_0 is in the 95% confidence interval defined by equation (1), then H_0 will be accepted at the 5% significance level, but if μ_0 is not in this confidence interval, H_0 will be rejected and H_1 accepted.

20. Suppose you are designing a poll on a subject for which you have no information on what people's opinions are likely to be. What sample size should you use to ensure a 90% confidence level that your results are accurate to plus or minus 2%? Explain your reasoning.

Review of Key Concepts

8.1 Continuous Probability Distributions

Refer to the Key Concepts on page 418.

1. Suppose the commuting time from Georgetown to downtown Toronto varies uniformly from 30 to 55 min, depending on traffic and weather conditions. Construct a graph of this distribution and use the graph to find

a) the probability that a trip takes 45 min or less

b) the probability that a trip takes more than 48 min

2. The lifetime of a critical component in microwave ovens is exponentially distributed with $k = 0.16$.

a) Sketch a graph of this distribution. Identify the distribution by name.

b) Calculate the approximate probability that this critical component will require replacement in less than five years.

3. Many people invest in the stock market by buying stocks recommended in investment newsletters. The table gives the annual returns after one year for 105 stocks recommended in investment newsletters.

Return (%)	Number of Stocks
−15	3
−12	1
−9	1
−6	1
−3	2
0	6
+3	10
+6	16
+9	22
+12	18
+15	9
+18	9
+21	4
+24	1
+27	1
+30	1

a) Construct a graph of these data.

b) Describe the shape of the graph. Use terms such as symmetric, skewed, or bimodal.

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

8.2 Properties of the Normal Distribution

Refer to the Key Concepts on page 429.

4. An electrician is testing the accuracy of resistors that have a nominal resistance of 15 Ω (ohms). He finds that the distribution of resistances is approximately normal with a mean of 15.08 Ω and a standard deviation of 1.52 Ω. What is the probability that

a) a resistor selected randomly has a resistance less than 13 Ω?

b) a resistor selected randomly has a resistance greater than 14.5 Ω?

c) a resistor selected randomly has a resistance between 13.8 Ω and 16.2 Ω?

5. The results of a blood test at a medical laboratory are normally distributed with $\mu = 60$ and $\sigma = 15$.

a) What is the probability that a blood test chosen randomly from these data has a score greater than 90?

b) What percent of these blood tests will have results between 50 and 80?

c) How low must a score be to lie in the lowest 5% of the results?

6. The lifetimes of a certain brand of photographic light are normally distributed with a mean of 210 h and a standard deviation of 50 h.

a) What is the probability that a particular light will last more than 250 h?

b) What percent of lights will need to be replaced within 235 h?

c) Out of 2000 lights, how many will have a lifetime between 200 h and 400 h?

8.3 Normal Sampling and Modelling

Refer to the Key Concepts on page 438.

7. The list below gives the age in months of 30 deer tagged in an Ontario provincial park last fall.

47	28	31	41	39	25	21	29	26	23
34	25	33	37	28	45	18	36	54	40
33	47	42	29	37	22	42	37	48	64

a) Use a method of your choice to assess whether these data are normally distributed. Explain your conclusion.

b) Find the mean and standard deviation of these data.

c) Determine the probability that a deer selected randomly from this sample is at least 30 months old. State and justify any assumptions you have made.

8. A quality-control inspector chose 20 bolt housings randomly from an assembly line. The interior diameters of these housings are listed in centimetres below.

2.29	2.23	2.48	2.24	2.40	2.23	2.37
2.33	2.37	2.31	2.31	2.26	2.26	2.18
2.33	2.31	2.30	2.30	2.24	2.34	

a) Find the mean and standard deviation for these data.

b) Assume the data are normally distributed. If the minimum acceptable diameter of these bolt housings is 2.25 cm, what proportion of the housings would be rejected as below this minimum?

9. Last year, Satsville High School ran nine classes of mathematics of data management, each with the same number of students. The class-average scores at year-end were as follows:

80.4	70.5	68.9	72.7	83.1	78.6	76.6
74.4	75.8					

a) Find the mean and standard deviation of these class-average scores.

b) How is the standard deviation of the class-average scores related to the underlying distribution of individual scores?

c) A study of the scores for all of the data-management students found that the scores were approximately normally distributed, with a mean of 75.7 and a standard deviation of 24.3. Given that all nine classes were the same size, find the most likely value for this class size. Explain your answer.

10. Use a normal approximation to find the probability that in a given year there will be more than 20 major earthquakes.

8.4 Normal Approximation to the Binomial Distribution

Refer to the Key Concepts on page 448.

11. A manufacturer of pencils has 60 dozen pencils randomly chosen from each day's production and checked for defects. A defect rate of 10% is considered acceptable.

a) Assuming that 10% of all the manufacturer's pencils are actually defective, what is the probability of finding 80 or more defective pencils in this sample?

b) If 110 pencils are found to be defective in today's sample, is it likely that the manufacturing process needs improvement? Explain your conclusion.

12. A store manager believes that 42% of her customers are repeat business (have visited her store within the last two weeks). Assuming she is correct, what is the probability that out of the next 500 customers, between 200 and 250 customers are repeat business?

8.5 Repeated Sampling and Hypothesis Testing

Refer to the Key Concepts on page 456.

13. An auto body repair shop plans its billing based on an average of 0.9 h to paint a car. The owner recently checked times for the last 50 cars painted and found that the average time for these cars was 1.2 h. She knows that the standard deviation is 0.4 h. Test the significance of this result at $\alpha = 0.10$.

14. A basketball coach claims that the average cost of basketball shoes is less than $80. He surveyed the costs of 36 pairs of basketball shoes in local stores and found the following prices:

$60	$50	$120	$110	$75	$110	$70
$40	$90	$65	$60	$85	$75	$80
$75	$80	$90	$45	$55	$70	$85
$85	$90	$90	$80	$50	$80	$85
$60	$70	$55	$95	$60	$45	$95
$70						

Test the coach's hypothesis at a 5% significance level.

15. A perfume company's long-term market share is estimated to be 6%. After an extensive advertising campaign, 11 out of 150 consumers surveyed claim to have purchased this company's perfume recently. Was the advertising campaign a success? Justify your assessment.

8.6 Confidence Intervals

Refer to the Key Concepts on page 464.

16. A study found that the average time it took for a university graduate to find a job was 5.4 months, with a standard deviation of 0.8 months. If a sample of 64 graduates were surveyed, determine a 95% confidence interval for the mean time to find a job.

17. In a survey of 200 households, 72 had central air-conditioning. Find a 90% confidence interval for the proportion of homes with central air-conditioning.

18. Here are the numbers of employees at 40 selected corporations in southern Ontario.

7685	11 778	11 370	9953	6200
900	2100	1270	1960	887
3100	7300	5400	3114	348
1650	400	873	195	173
725	3472	1570	256	895
120	347	40	2290	4236
850	540	164	285	12
390	60	713	175	213

a) Find a 90% confidence interval for the average number of employees at corporations in southern Ontario.

b) Comment on your findings. Does your confidence interval describe the sample data realistically? What problems exist with constructing a confidence interval for data of this sort?

19. A regional planner has been asked to estimate the average income of businesses in her region. She wants to be 90% confident of her conclusion. She sampled 40 companies and listed their net incomes (in thousands of dollars) as shown below:

84	49	3	133	85	4240	461	60
28	97	14	252	18	16	24	345
254	29	254	5	72	31	23	225
70	8	61	366	77	8	26	10
55	137	158	834	123	47	2	21

Estimate the average income of businesses in this region, with a 90% confidence level.

20. A survey of reading habits found that 63% of those surveyed said they regularly read at least part of a daily newspaper. The results are considered accurate within plus or minus 5%, 19 times in 20. Determine the number of people surveyed.

21. A market-research company found that 14.5% of those surveyed used Gleemodent toothpaste. The company states that the survey is accurate within ±4%, nine times out of ten. How many people did they survey?

22. A city's transportation department surveyed 50 students, 70 city residents, and 30 cyclists concerning their opinion on how well the city supported bicycling as an alternative means of transportation. The table below summarizes the survey results:

Rating	Students	Community	Cyclists
Excellent	0	4	0
Very Good	3	12	0
Good	23	17	8
Not So Good	15	19	8
Poor	7	11	11
Very Poor	2	7	3

a) Construct a 95% confidence interval for the proportion of students who feel that city's support is good or better.

b) Identify three other proportions which you feel are important. Construct confidence intervals for these proportions. Report your findings, together with reasons why these proportions may be important to the city's transportation department.

Chapter Test

ACHIEVEMENT CHART				
Category	Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application
Questions	All	3, 9, 10	5, 7, 9, 10	2, 3, 4, 5, 6, 7, 8, 10

1. Give a real-life example of data which could have the following probability distributions. Explain your answers.

a) normal distribution

b) uniform distribution

c) exponential distribution

d) bimodal distribution

2. The volume of orange juice in 2-L containers is normally distributed with a mean of 1.95 L and a standard deviation of 0.15 L.

a) What is the probability that a container chosen at random has a volume of juice between 1.88 L and 2.15 L?

b) If containers with less than 1.75 L are considered below standard, what proportion of juice containers would be rejected?

c) Out of 500 containers, how many have a volume greater than 2.2 L?

3. According to its label, a soft-drink can contains 500 mL. Currently, the filling machine is set so that the volume per can is normally distributed, with a mean of 502 mL and a standard deviation of 1.5 mL. If too many cans contain less than 500 mL, the company will lose sales. If many cans contain more than 504 mL, the company will incur excess costs. Does the company need to recalibrate its filling machine?

4. A farmer finds the mass of the yields of 20 trees in an orchard. Here are the results in kilograms.

16.89	7.77	7.26	14.05	10.85	15.69
12.95	7.92	16.12	9.06	5.71	6.11
5.95	9.25	8.09	8.02	10.43	9.42
9.19	6.86				

a) Find the mean and standard deviation of these data.

b) What is the probability that a tree selected randomly from this orchard has a yield greater than 10 kg? State any assumptions you make in this calculation.

5. A manufacturer of mixed nuts promises: "At least 20% cashews in every can." A consumer-research agency tests 150 cans of nuts and finds a mean of 22% cashews with a standard deviation of 1.5%. The proportions of cashews are normally distributed.

a) What is the probability that the population mean is less than 20% cashews?

b) Based on the sample, what proportion of cans have between 15% and 30% cashews?

c) The company must stop making their claim if more than 3% of the cans contain less than 20% cashews. Write a brief report to the company outlining whether they need a new motto.

d) Suggest a better motto for the company.

6. Approximately 85% of applicants get their G-1 driver's licence the first time they try the test. If 80 applicants try the test, what is the probability that more than 10 applicants will need to retake the test?

7. A car assembly line produces 1920 cars per shift. A defect rate of 3% is considered acceptable. From the production of one recent shift, 65 cars were found to be defective.

a) Find the probability of this occurrence.

b) Explain to the shift supervisor, who has not studied probability theory, what your answer means and whether changes will need to be made in the production process to reduce the number of defective cars.

8. Find a 95% confidence interval for the percent of voters who are likely to vote "Yes" in a referendum if, in a sample of 125 voters, 55 said they would vote "Yes."

9. A politician asks a polling firm to determine the likelihood that he will be re-elected. The polling firm reports that of decided voters, 48% indicated they would vote for him if an election were called today. The result is accurate within plus or minus 5%, 19 times in 20.

a) How many decided voters were polled?

b) Should the politician be worried about his chances of re-election? Justify your answer.

c) The polling company found a large number of undecided voters. Should this fact influence the conclusion? What action, if any, should the politician take regarding this large pool of undecided voters?

ACHIEVEMENT CHECK

Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application

10. Students in the first-year statistics course at Statsville College wrote a 100-point examination in which the grades were normally distributed. The mean is 60 and the standard deviation is 12. Students in the first-year calculus course also wrote a 100-point examination in which the grades were normally distributed. For the calculus examination, the mean is 68 and the standard deviation is 8.

a) Betty has a mark of 85% in statistics and Brianna has a mark of 85% in Calculus. Who has the higher standing relative to her classmates? Justify your answer.

b) The statistics professor has decided to bell the marks in order to match the mean and standard deviation of the calculus class. Find Betty's new mark.

c) Suppose a subgroup of the students in this school, those on the school's honour roll, were selected. Would you still expect the distribution of their examination scores to be normally distributed? Would the distribution have the same mean and standard deviation? Explain your reasoning.

Wrap-Up

Implementing Your Action Plan

1. Determine the criteria for classifying your chosen species as endangered.
2. Collect sample data. You will need at least three estimates for the population of the species. More estimates would be useful.
3. Formulate a hypothesis concerning the status of the species.
4. Perform a hypothesis test to decide whether the data indicate that the species is endangered.
5. Determine a confidence interval for the estimated population of the species.
6. Draw a conclusion about the status of the species. Does your analysis lead to any further conclusions? For example, can you determine whether a total ban on commercial exploitation of the species is justified?
7. Present your results, using appropriate technology.

Evaluating Your Project

1. Identify factors that could affect the validity of your conclusion, such as possible measurement errors or bias in your data.
2. Are there improvements you could make to the methods you used for this project?
3. If you were to update your analysis a year from now, do you think your conclusions would be different? Justify your answer.
4. Could you apply your data collection and analysis techniques to other endangered species?
5. Did your research or analysis suggest related topics that you would like to explore?

Suggested Resources

- Canadian Nature Federation
- Committee on the Status of Wildlife in Canada
- Sheldrick Wildlife Trust
- World Conservation Union (IUCN)

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

Visit the above web site and follow the links to learn more about endangered species.

Presentation

Choose the most appropriate method for presenting your findings. Your could use one or a combination of the following forms:

- a written report
- an oral presentation
- a computer presentation (using software such as Corel® Presentations™ or Microsoft® PowerPoint®)
- a web page
- a display board

See section 9.5 and Appendix D for ideas on how to prepare a presentation. Be sure to document the sources for your data.

Preparing for the Culminating Project

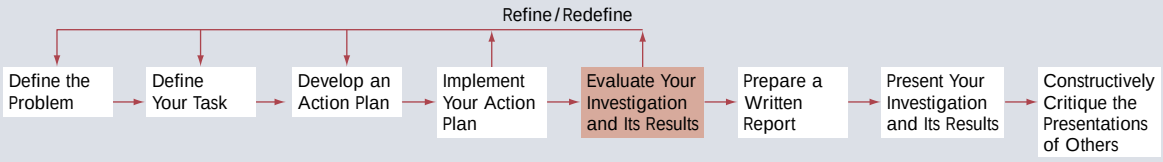
Applying Project Skills

In this probability distributions project, you have learned new skills and further developed some of the skills used in the earlier projects. Many of these skills will be vital for your culminating project:

- carrying out an action plan
- formulating a hypothesis
- using technology to collect data
- using sample data to test a hypothesis
- formulating and interpreting a confidence interval
- evaluating the quality of data
- critiquing research methodology
- presenting results using appropriate technology

Keeping on Track

At this time, your data analysis should be complete and you should have determined what conclusions you can draw from it. You should be ready to evaluate your culminating project and prepare a presentation of your results. Section 9.4 details questions you can use to guide your evaluation of your own work. Appendix D outlines techniques for presentations.



1. A lottery has sold 5000 000 tickets at $1.00 each. The prizes are shown in the table. Determine the expected value per ticket.

Prize	Number of Prizes
$1 000 000	1
$50 000	10
$500	100
$10	1000

2. The speed limit in a school zone is 40 km/h. A survey of cars passing the school shows that their speeds are normally distributed with a mean of 38 km/h and a standard deviation of 6 km/h.

a) What percent of cars passing the school are speeding?

b) If drivers receive speeding tickets for exceeding the posted speed limit by 10%, what is the probability that a driver passing the school will receive a ticket?

3. Determine the probability distribution for the number of heads that you could get if you flipped a coin seven times. Show your results with a table and a graph.

4. Suppose that 82.5% of university students use a personal computer for their studies. If ten students are selected at random, what is

a) the probability that exactly five use a personal computer?

b) the probability that at least six use a personal computer?

c) the expected number of students who use a personal computer?

5. Harvinder and Sean work in the quality-control department of a large electronics manufacturer that is having problems with its assembly line for producing CD players. The defective rate on this assembly line has gone up to 12%, and the department head wants to know the probability that a skid of 50 CD players will contain at least 4 defective units. Harvinder uses the binomial distribution to answer this question, while Sean uses the normal approximation. By what percent will Harvinder's answer exceed Sean's?

6. A box contains 15 red, 13 green, and 16 blue light bulbs. Bulbs are randomly selected from this box to replace all the bulbs in a string of 15 lights.

a) Design a simulation to estimate the expected number of each colour of light bulb in the string.

b) Calculate the theoretical probability of having exactly 5 red bulbs in the string.

c) What is the expected number of blue bulbs?

d) Would you expect your simulation to produce the same probability as you calculated in part b)? Why or why not?

7. A newspaper poll indicated that 70% of Canadian adults were in favour of the anti-terrorist legislation introduced in 2001. The poll is accurate within $\pm$ 4%, 19 times in 20. Estimate the number of people polled for this survey. Describe any assumptions you make about the sampling procedure.

8. The Ministry of Natural Resources conducted aerial surveys to estimate the number of wolves in Algonquin Park. Aerial surveys of 50 randomly selected 100-km^2 sections of the park had a mean of 1.67 wolves and a standard deviation of 0.32.

a) Determine a 95% confidence interval for the mean number of wolves per 100 km^2 in the park.

a) Describe any assumptions you made for your calculation in part a).

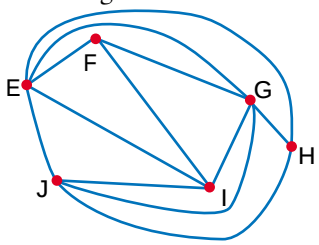
1. If the probability of rain tomorrow is 40%, what are the odds

a) that it will rain?

b) that it will not rain?

2. Describe each field you would include in a database for scheduling deliveries by an appliance dealer.

3. **a)** List the vertices of odd-degree in the following network.



b) Is this network

i) complete? **ii)** traceable? **iii)** planar?

4. At Burger Barn, a combination meal consists of a hamburger and any two side orders, which can be soup, French fries, onion rings, coleslaw, pie, or ice cream. How many combination meals are possible?

5. Make a table of the probability distribution for the days of the month in a leap year.

6. **a)** What is a Bernoulli trial?

b) What is the key difference between trials in a geometric probability distribution and those in a hypergeometric probability distribution?

7. Marc is making change for a $5 bill.

a) List the ways he can provide change using only $1 and $2 coins.

b) How many ways can he make change using only $1 coins and quarters?

c) Would the number of ways be the same if he used only $2 coins and quarters? Explain your reasoning.

d) How many ways can he make change using only $1 coins, $2 coins, and quarters?

8. Give an example of each of the following sampling techniques and list an advantage of using each type.

a) stratified sample

b) simple random sample

c) systematic sample

9. In the fairy tale Rumplestiltskin, the queen must guess the name *Rumplestiltskin*.

a) If she knows only what letters are in the name, how many different guesses could she make?

b) At one guess per minute, how long would it take to try all the possible arrangements?

10. A survey asked randomly chosen movie fans how many videos they rented in the last month.

Number of Videos	Frequency
0–1	2
2–3	5
4–5	17
6–7	24
8–9	21
10–11	9
12–13	1

a) Estimate the mean and median number of videos rented by those surveyed.

b) Draw a histogram and relative frequency polygon for the data.

c) Discuss any sources of inaccuracy in your calculations.

11. A group of data-entry clerks had the following results in a keyboarding test:

Speed (words/min)	44	62	57	28	46	71	50
Number of Errors	4	3	3	5	11	2	4

a) Create a scatter plot and classify the linear correlation.

b) Determine the correlation coefficient and the equation of the line of best fit.

c) Identify the outlier and repeat part b) without it. Use calculations and a graph to show how this affects the strength of the linear correlation and the line of best fit.

d) What does your analysis tell you about the relationship between speed and accuracy for this group of data-entry clerks?

12. A shopper observes that whenever the price of butter goes up, the price of cheese goes up also. Can the shopper conclude that the price of butter causes the increase in the price of cheese? If not, how would you account for the correlation in prices?

13. A market research company has a contract to determine the percent of adults in Ontario who want speed limits on expressways to be increased. The poll's results must be accurate within ±4%, 19 times in 20.

a) If a small initial sample finds opinions almost equally divided, how many people should the company survey?

b) Suggest a sampling method that would give reliable results.

14. How many ways can a bank of six jumpers on a circuit board be set if

a) each jumper can be either on or off?

b) each jumper can be off or connect either pins 1 and 2 or pins 2 and 3?

15. How many arrangements of the letters in the word *mathematics* begin with a vowel and end with a letter other than *h*?

16. In 2001, 78 books were nominated for the $25 000 Giller Award for Canadian fiction. How many different shortlists of 6 finalists could the jury select?

17. A marketing survey of consumers' soft-drink preferences collected the following data:

- 75 liked cola, 65 liked ginger ale, and 32 liked spring water.
- 43 liked cola and ginger ale.
- 13 liked cola and spring water.
- 15 liked ginger ale and spring water.
- 7 liked all three and 12 liked none of them.

a) How many people were surveyed?

b) How many liked only ginger ale?

c) How many liked only spring water and ginger ale?

d) How many liked only one of the choices?

e) How many liked exactly two of the choices?

18. In how many ways can a box of 18 different chocolates be evenly distributed among three people?

19. Use Pascal's triangle to

a) expand $\left(\frac{x}{3} + 3y\right)^5$

b) develop a formula expressing the sum of the first n natural numbers in terms of combinations

20. Naomi's favourite cereal includes a free mini-puck with the emblem of one of the 30 NHL hockey teams. If equal numbers of the different pucks are randomly distributed in the cereal boxes, what is the probability that Naomi will get a mini-puck for one of the 6 original NHL teams in any given box of cereal?

21. Steve has a bag containing five red, three green, six orange, and ten black jelly beans. Steve's favourites are the black licorice ones. He randomly selects eight jelly beans.

a) What is the probability that he will have at least four black ones?

b) What is the expected number of black jelly beans?

22. Of the 24 guests invited to Hannah's party, 12 are male and 15 have dark hair. If 7 of the females have dark hair, what is the probability that the first guest to arrive will either have dark hair or be a male?

23. When testing its new insect repellant, a company found that 115 people out of an experimental group of 200 got fewer than the mean number of mosquito bites reported by the 100 people in the control group. Is this evidence sufficient for the company to claim that its spray is effective at a

a) 5% significance level?

b) 99% confidence level?

24. A particular car dealer's records show that 16.5% of the cars it sold were red.

a) What is the probability that the first red car sold at the dealership on a given day will be the fifth car sold that day?

b) What is the probability that the first red car sold will be among the first five cars sold?

c) What is the expected waiting time before a red car is sold?

25. A town has three barbeque-chicken restaurants. In the past year, UltraChicken lost 20% of its customers to Churrasqueira Champion and 15% to Mac's Chicken, Churrasqueira Champion lost 10% of its customers to each of its two competitors, and Mac's Chicken lost 25% of its customers to UltraChicken and 30% to Churrasqueira Champion.

a) Predict the long-term market share for each of these three restaurants.

b) What assumptions must you make for your solution to part a)?

26. A battery maker finds that its production line has a 2.1% rate of defects.

a) What is the probability that the first defect found will be in the 20th battery tested?

b) What is the probability that there are no defects in the first 6 batteries tested?

c) What is the expected waiting time until a defective battery is tested?

27. Explain the difference between a leading question and a loaded question.

28. During the winter, 42% of the patients of a walk-in clinic come because of symptoms of the common cold or flu.

a) What is the probability that, of the 32 patients on one winter morning, exactly 10 had symptoms of the common cold or flu?

b) What is the expected number of patients who have symptoms of the common cold or flu?

29. The Ministry of Natural Resources is concerned that hunters are killing a large number of wolves that leave the park to follow deer. For this reason, the Ministry is considering a permanent ban on wolf hunting in the area bordering the park. Outline the data and statistical analysis that you would require to determine whether such a ban is justified.