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%.

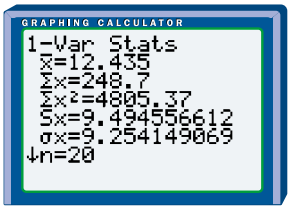
$$\begin{aligned} P(X > 6) &= P\left(Z > \frac{6 - 12.4}{9.5}\right) \\ &= P(Z > -0.674) \\ &= 1 - P(Z < -0.674) \\ &\doteq 1 - 0.25 \\ &= 0.75 \end{aligned}$$

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 $s \doteq 9.49$.

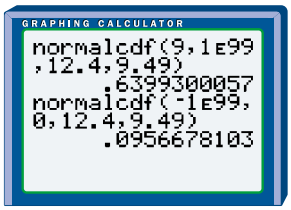
Recall that, since the data is a sample, you should use the value of Sx rather than σx.

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 $x = 9$.

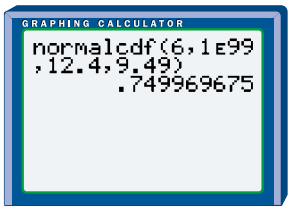
Therefore, use the **normalcdf(function** as shown on the screen on the right.

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.

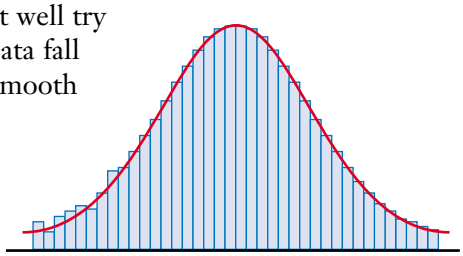
	A	B	C	D	E	F	G	H
1	Year	Return (%)						
2	1	7.2		mean	12.435			
3	2	12.3		Sx	9.494557			
4	3	17.1						
5	4	17.9		x	9	0	6	
6	5	10.8		P(X<x)	0.358766	0.095149	0.248962891	
7	6	19.3		P(X>x)	0.641244		0.751037109	
8	7	12.2						
9	8	-13.1						
10	9	20.2						
11	10	18.6						
12	11	6.4						
13	12	27						
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?

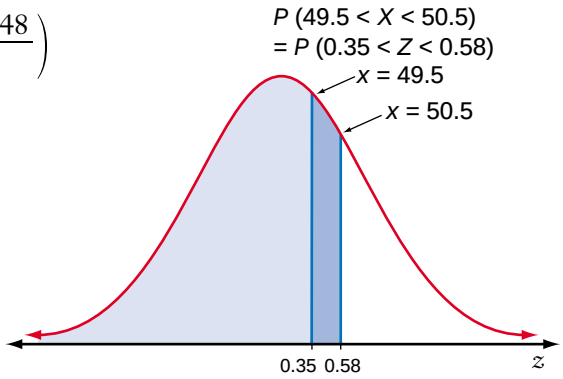
c) Each filling machine produces 130 000 boxes per shift. How many of these will lie within the acceptable range?

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

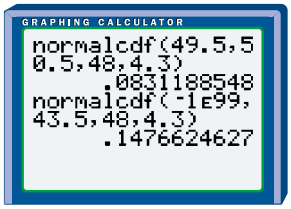
$$P(43.5 < X < 54.5) = P\left(\frac{43.5 - 48}{4.3} < Z < \frac{54.5 - 48}{4.3}\right)$$
$$= P(-1.05 < Z < +1.51)$$
$$= P(Z < +1.51) - P(Z < -1.05)$$
$$= 0.9345 - 0.1469$$
$$= 0.788$$

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.

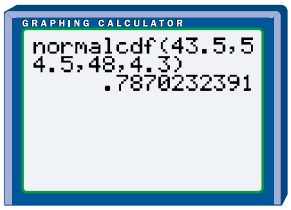
Solution 2: Using a Graphing Calculator

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.