5.4 The Normal Distribution

GOAL

Determine the properties of a normal distribution, and compare normally distributed data.

YOU WILL NEED

- calculator
- grid paper
- dice
- dice simulator

INVESTIGATE the Math

Many games require dice. For example, the game of Yacht requires five dice.

? What shape is the data distribution for the sum of the numbers rolled with dice, using various numbers of dice?

A. If you rolled a single die 50 000 times, what do you think the graph would look like?

B. Predict what the graph would look like if you rolled two dice 50 000 times.

C. With a partner, roll two dice 50 times. Record the sum for each roll in a frequency distribution table. Then draw a graph to represent the distribution of the data. Comment on the distribution of the data.

D. Combine the data for the entire class, and draw a graph. Comment on the distribution of the combined data.

E. Using a dice simulator, roll two dice 50 000 times. Compare the graph for this set of data with the graph you drew in part D.

F. Make a conjecture about what the graph would look like if you rolled three dice 50 000 times.

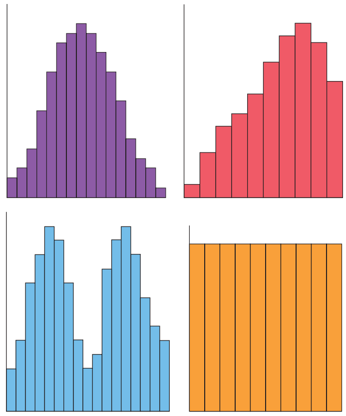
G. Using a dice simulator, roll three dice 50 000 times. What do you notice about the shape of the graph as the number of rolls increases? Draw a frequency polygon to represent the distribution of the data. Was your conjecture correct?

H. Make a conjecture about what the graph would look like if you rolled four dice 50 000 times. What would the graph look like for five dice rolled 50 000 times?

I. Using a dice simulator, roll four dice and then five dice 50 000 times. What do you notice about the shape of the graph as the number of rolls increases? Draw a frequency polygon to represent the distribution of the data for both four dice and five dice. Describe the shape of each polygon. Was your conjecture correct?

EXPLORE…

- Sometimes the distribution of data has a special shape. For example, the first graph below has one peak, so the shape has one mode. Describe the shape of each graph, and suggest a context that the graph could represent.



normal curve
A symmetrical curve that represents the normal distribution; also called a **bell curve**.

normal distribution
Data that, when graphed as a histogram or a frequency polygon, results in a unimodal symmetric distribution about the mean.

Reflecting

J. How does increasing the number of dice rolled each time affect the distribution of the data?

K. How does increasing the sample size affect the distribution of the data for three dice, four dice, and five dice?

L. What do you think a graph that represents 100 000 rolls of 10 dice would look like? Why do you think this shape is called a **normal curve**?

M. As the number of dice increases, the graph approaches a **normal distribution**. What does the line of symmetry in the graph represent?

APPLY the Math

EXAMPLE 1 Examining the properties of a normal distribution

Heidi is opening a new snowboard shop near a local ski resort. She knows that the recommended length of a snowboard is related to a person's height. Her research shows that most of the snowboarders who visit this resort are males, 20 to 39 years old. To ensure that she stocks the most popular snowboard lengths, she collects height data for 1000 Canadian men, 20 to 39 years old. How can she use the data to help her stock her store with boards that are the appropriate lengths?

Height (in.)	Frequency
61 or shorter	3
61–62	4
62–63	10
63–64	18
64–65	30
65–66	52
66–67	64
67–68	116
68–69	128
69–70	147
70–71	129
71–72	115
72–73	63
73–74	53
74–75	29
75–76	20
76–77	12
77–78	5
taller than 78	2

Heidi's Solution

$\bar{x} = 69.521$ in.
$\sigma = 2.987...$ in.

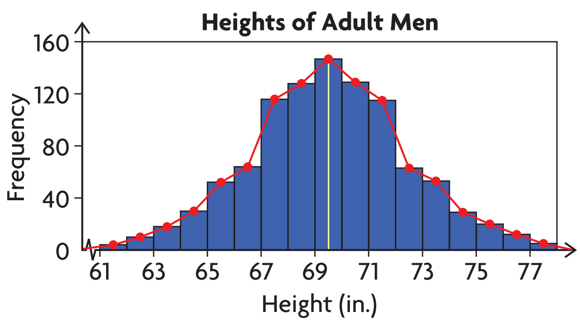
I decided to use 60.5 as the midpoint of the first interval and 78.5 as the midpoint of the last interval. I calculated the mean and standard deviation using a graphing calculator.

I examined the table. The median is the average of the 500th and the 501st height.

There are 425 heights less than or equal to 69 in. The median height is the height of the 73rd person in the 69–70 in. interval.

There are 147 heights in the 69–70 in. interval. The 73rd person is about midway through this interval, so the median height is approximately 69.5 in.

I don't know where the mode is, but it is more likely to be in the interval that contains the greatest number of data values, which is the 69–70 in. interval. I can assume that all three measures of central tendency have about the same value.



I drew a histogram to show the height distribution.

I drew a vertical line on my histogram to represent the mean.

Then I drew a frequency polygon by connecting the midpoints at the top of each bar of the graph.

The data is almost symmetrical about the mean and tapers off in a gradual way on both sides. The frequency polygon resembles a bell shape.

The data has a normal distribution.

Heights within one standard deviation of the mean:
$69.521 - 2.987...$ or 66.5 in.
$69.521 + 2.987...$ or 72.5 in.

Heights within two standard deviations of the mean:
$69.521 - 2(2.987...)$ or 63.5 in.
$69.521 + 2(2.987...)$ or 75.5 in.

I determined the range of heights within one standard deviation of the mean by adding and subtracting the standard deviation from the mean.

I also determined the range of heights within two standard deviations of the mean.

Range	Height Range
$\bar{x} - 1\sigma$ to $\bar{x} + 1\sigma$	about 66.5 in. to 72.5 in.
$\bar{x} - 2\sigma$ to $\bar{x} + 2\sigma$	about 63.5 in. to 75.5 in.

I summarized my results in a table and compared them to my histogram. The range for one standard deviation appears to include most of the data. The range for two standard deviations appears to include almost all of the data.

Number of males within one standard deviation of the mean from 67 in. to 73 in.:
116 + 128 + 147 + 129 + 115 + 63, or 698

I estimated the percent of the heights that were within one and two standard deviations of the mean.

To determine the number of heights in each range, I rounded up to find the lower and upper boundaries in each range. Then I summed the number of people from the table that were in these ranges.

Number of males within two standard deviations of the mean from 64 in. to 76 in.:
30 + 52 + 64 + 116 + 128 + 147 + 129 + 115 + 63 + 53 + 29 + 20, or 944

Range	Height Range	Percent of Data
$\bar{x} - 1\sigma$ to $\bar{x} + 1\sigma$	66.5 in. to 72.5 in.	$\frac{698}{1000}$ or 69.8%
$\bar{x} - 2\sigma$ to $\bar{x} + 2\sigma$	63.5 in. to 75.5 in.	$\frac{946}{1000}$ or 94.6%

About 70% of the heights are within one standard deviation of the mean.

About 95% of the heights are within two standard deviations of the mean.

Heights within one standard deviation of the mean are most common.

A high percent of heights are within two standard deviations of the mean.

I predict that about 70% of my male customers will need snowboards for heights from 66.5 to 72.5 in.

Your Turn

What percent of all the heights is within three standard deviations of the mean?

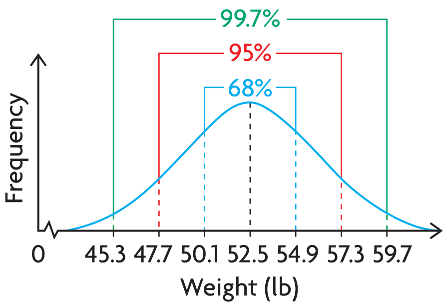
EXAMPLE 2 Analyzing a normal distribution

Jim raises Siberian husky sled dogs at his kennel. He knows, from the data he has collected over the years, that the weights of adult male dogs are normally distributed, with a mean of 52.5 lb and a standard deviation of 2.4 lb. Jim used this information to sketch a normal curve, with

- 68% of the data within one standard deviation of the mean
- 95% of the data within two standard deviations of the mean
- 99.7% of the data within three standard deviations of the mean

The Canadian Championship Dog Derby, held in Yellowknife, Northwest Territories, is one of the oldest sled-dog races in North America. Top mushers gather to challenge their dogs in the fast-paced, three-day event.

Weights of Adult Male Huskies



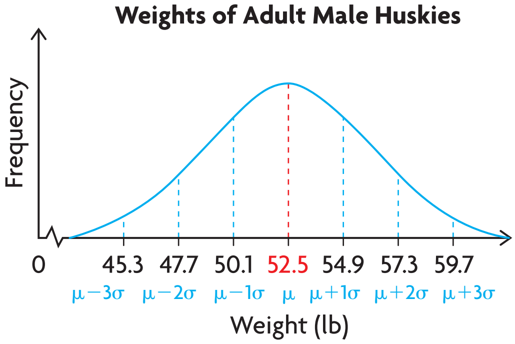
What percent of adult male dogs at Jim's kennel would you expect to have a weight between 47.7 lb and 54.9 lb?

Communication *Tip*

In statistics, when an entire population is involved, use the symbol μ (read as "mu") for the mean of the population.

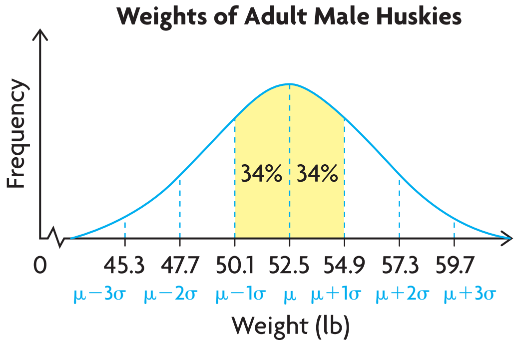
Ian's Solution

Weights of Adult Male Huskies

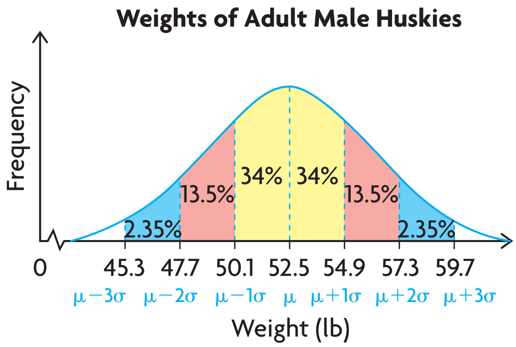


I sketched the graph and labelled the mean, μ, below the horizontal axis. Since the standard deviation is 2.4 lb, I can label the scale to the right of the mean as $\mu + 1\sigma$, $\mu + 2\sigma$, and $\mu + 3\sigma$.

I labelled the scale to the left of the mean as $\mu - 1\sigma$, $\mu - 2\sigma$, and $\mu - 3\sigma$.



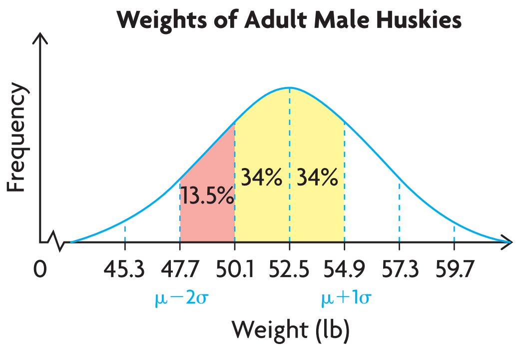
I reasoned that the area under the curve is symmetrical around μ, so if 68% of the dogs have weights within one standard deviation, then 34% must have weights between $\mu - 1\sigma$ and μ, and 34% must have weights between μ and $\mu + 1\sigma$.



I knew that 95% of the weights lie within two standard deviations of the mean. Since 68% of the weights lie within one standard deviation of the mean, 27% of the weights must lie between one and two standard deviations, or 13.5% for each side of the graph.

Using the same reasoning, I figured out the percent of data that would lie between two and three standard deviations from the mean for each side of the graph:

$$\frac{99.7\% - 95\%}{2} = 2.35\%$$



For the percent that fits between 47.7 lb and 54.9 lb, I determined the location of each weight.

$\mu + 1\sigma = 54.9$

$\mu - 2\sigma = 47.7$

I used my diagram to determine the sum of the percent of data between these locations.

The percent of dogs with weights between 47.7 lb and 54.9 lb, x, can be represented as

$$x = 13.5\% + 34\% + 34\%$$
$$x = 81.5\%$$

Approximately 81.5% of adult male dogs should have a weight between 47.7 lb and 54.9 lb.

Your Turn

a) What percent of adult male dogs at Jim's kennel would you expect to have a weight between 50.1 lb and 59.7 lb?

b) What percent of adult male dogs at Jim's kennel would you expect to have a weight less than 45.3 lb?

EXAMPLE 3 Comparing normally distributed data

Two baseball teams flew to the North American Indigenous Games. The members of each team had carry-on luggage for their sports equipment. The masses of the carry-on luggage were normally distributed, with the characteristics shown to the right.

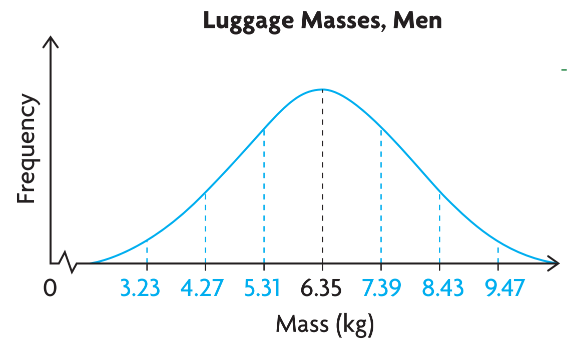
Team	μ (kg)	σ (kg)
Men	6.35	1.04
Women	6.35	0.59

a) Sketch a graph to show the distribution of the masses of the luggage for each team.

b) The women's team won the championship. Each member received a medal and a souvenir baseball, with a combined mass of 1.18 kg, which they packed in their carry-on luggage. Sketch a graph that shows how the distribution of the masses of their carry-on luggage changed for the flight home.

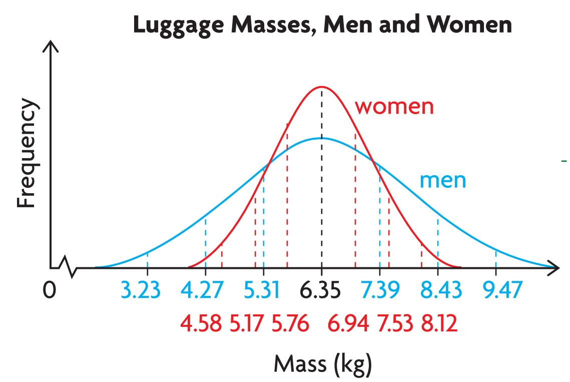
Samara's Solution

a)



I sketched the normal distribution of the masses of the luggage for the men's team. I marked the values for μ, $\mu + 1\sigma$, $\mu + 2\sigma$, $\mu + 3\sigma$, $\mu - 1\sigma$, $\mu - 2\sigma$, and $\mu - 3\sigma$.

I knew that the area under the normal curve represents 100% of the data, so I could think of the area as equal to 1 unit.

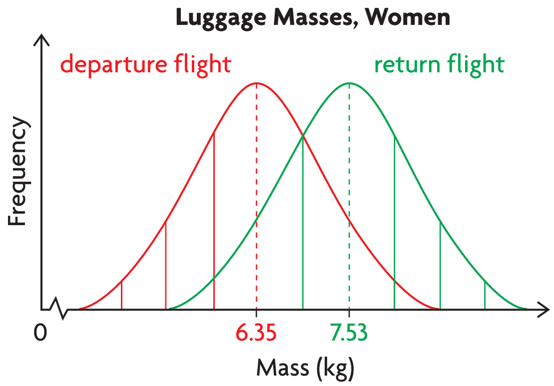


On the same graph, I sketched the normal curve for masses of the luggage for the women's team. I knew that this curve must be narrower than the curve for the men's team, since the standard deviation is lower.

I also knew that the area under this curve represents 100% of the data for the women's luggage, so the area under the red curve is also equal to 1 unit. For the area under both curves to be the same, the normal curve for the women's team must be taller.

b) Data for masses of luggage for women's team on flight home:

$\mu = 6.35 + 1.18$ or 7.53 kg
$\sigma = 0.59$ kg



Since each member of the women's team will add 1.18 kg to the mass of her carry-on luggage, the mean mass will increase by 1.18 kg. Although the mass of each piece of luggage will change, the distribution of the masses will stay the same, and the standard deviation will still be 0.59 kg.

I sketched a graph to show the new masses of the luggage for the women's team. It made sense that the new graph would simply move 1.18 to the right of the old graph. The shape stayed the same because the standard deviations of the two graphs are the same.

Your Turn

Suppose that the women had gone shopping and had also added their purchases to their carry-on luggage. How would you sketch a graph to show the distribution of the masses of their luggage for the trip home? Explain.

EXAMPLE 4 Analyzing data to solve a problem

Shirley wants to buy a new cellphone. She researches the cellphone she is considering and finds the following data on its longevity, in years.

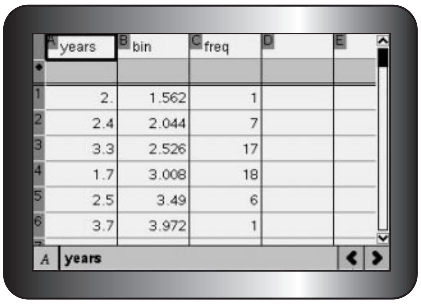
2.0	2.4	3.3	1.7	2.5	3.7	2.0	2.3	2.9	2.2
2.3	2.7	2.5	2.7	1.9	2.4	2.6	2.7	2.8	2.5
1.7	1.1	3.1	3.2	3.1	2.9	2.9	3.0	2.1	2.6
2.6	2.2	2.7	1.8	2.4	2.5	2.4	2.3	2.5	2.6
3.2	2.1	3.4	2.2	2.7	1.9	2.9	2.6	2.7	2.8

a) Does the data approximate a normal distribution?
b) If Shirley purchases this cellphone, what is the likelihood that it will last for more than three years?

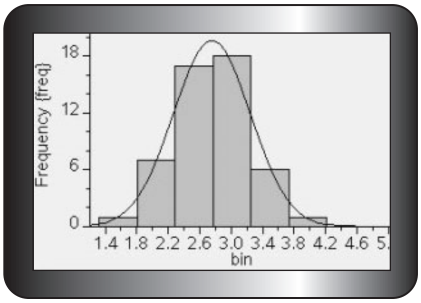
Shirley's Solution

a) $\mu = 2.526$
$\sigma = 0.482$
median $= 2.55$

Using my calculator, I determined the mean, the standard deviation, and the median. The median is close to the mean, which indicates that the data may be normally distributed.



I created a frequency table to generate a histogram.



I created a histogram of the data, with an interval width of σ. My histogram looked almost symmetrical, so I decided to check the data to see how closely it approximates a normal distribution.

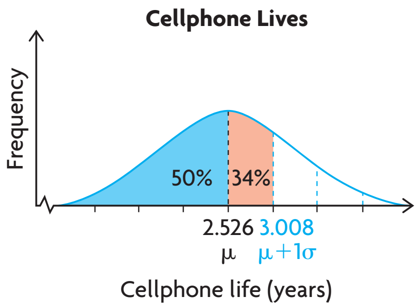
I generated a normal distribution curve on top of the histogram.

-1σ to 1σ	-2σ to 2σ	-3σ to 3σ
$\frac{35}{50} = 70\%$	$\frac{48}{50} = 96\%$	100%

The data approximates a normal distribution.

I determined the percent of the data within one, two, and three standard deviations of the mean. I think the percents are reasonably close to those for a normal distribution (68%, 95%, and 99.7%).

b)



I sketched a normal curve with the mean at 2.526 and the mean + 1 standard deviation at 3.008. I could use this location on the graph to determine the percent of values greater than three years.

I knew that the left-half area of the curve contains 50% of the data, and the area between 2.526 and 3.008 contains approximately 34% of the data.

$\mu + 1\sigma = 2.526 + 0.482$, or 3.008

$100\% - (50\% + 34\%) = 16\%$

About 16% of the cellphones lasted more than three years.

The area under the curve to the right of 3.008 is the white section. I subtracted the area of the coloured sections from 100%.

Your Turn

If Shirley purchases this cellphone, what is the likelihood that it will last at least 18 months?

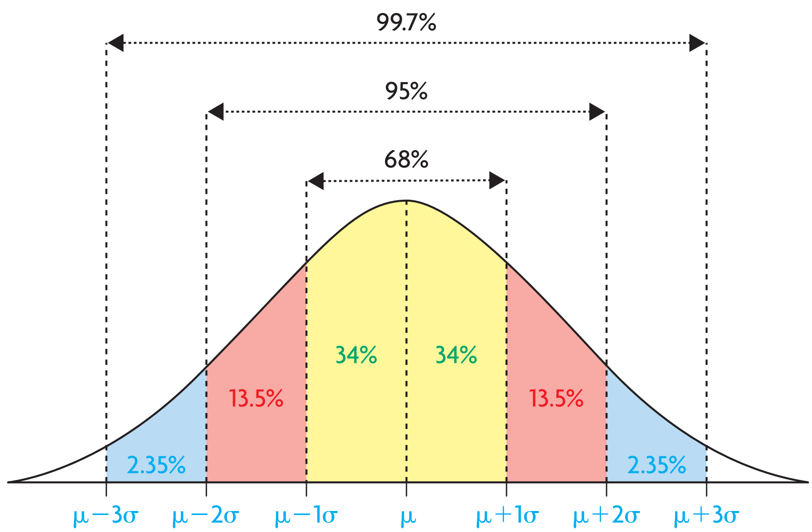
In Summary

Key Ideas

- Graphing a set of grouped data can help you determine whether the shape of the frequency polygon can be approximated by a normal curve.
- You can make reasonable estimates about data that approximates a normal distribution, because data that is normally distributed has special characteristics.
- Normal curves can vary in two main ways: the mean determines the location of the centre of the curve on the horizontal axis, and the standard deviation determines the width and height of the curve.

Need to Know

- The properties of a normal distribution can be summarized as follows:
  - The graph is symmetrical. The mean, median, and mode are equal (or close) and fall at the line of symmetry.
  - The normal curve is shaped like a bell, peaking in the middle, sloping down toward the sides, and approaching zero at the extremes.
  - About 68% of the data is within one standard deviation of the mean.
  - About 95% of the data is within two standard deviations of the mean.
  - About 99.7% of the data is within three standard deviations of the mean.
  - The area under the curve can be considered as 1 unit, since it represents 100% of the data.



- Generally, measurements of living things (such as mass, height, and length) have a normal distribution.

CHECK Your Understanding

1. The ages of members of a seniors curling club are normally distributed, with a mean of 63 years and a standard deviation of 4 years. What percent of the curlers is in each of the following age groups?
   a) between 55 and 63 years old
   b) between 67 and 75 years old
   c) older than 75 years old

2. A teacher is analyzing the class results for three biology tests. Each set of marks is normally distributed.

Test	Mean (μ)	Standard Deviation (σ)
1	77	3.9
2	83	3.9
3	77	7.4

   a) Sketch normal curves for tests 1 and 2 on one graph. Sketch normal curves for tests 1 and 3 on a different graph.
   b) Examine your graphs. How do tests 1 and 3 compare? How do tests 1 and 2 compare?
   c) Determine Oliver's marks on each test, given the information shown at the right.

Test	Oliver's Mark
1	$\mu + 2\sigma$
2	$\mu - 1\sigma$
3	$\mu + 3\sigma$

3. Is the data in each set normally distributed? Explain.

a)

Interval	10–19	20–29	30–39	40–49	50–59	60–69
Frequency	3	5	17	20	11	4

b)

Interval	2–5	6–9	10–13	14–17	18–21	22–25
Frequency	2	8	8	3	4	5

c)

Interval	10–24	25–39	40–54	55–69	70–84	85–99
Frequency	2	7	16	10	4	1

PRACTISING

4. Tiegan is organizing her movie collection. She decides to record the length of each movie, in minutes.

91	129	95	96	96	90	101	87	100	90
86	78	105	99	81	106	101	122	91	102
89	125	162	155	89	89	180	94	84	99
73	100	99	100	117	135	100	89	87	110
125	103	94	99	98	102	96	88	154	144

   a) Determine the mean and standard deviation for the set of data.
   b) Create a frequency table, using σ as the interval width.
   c) Are the lengths of Tiegan's movies normally distributed? Explain.

The Indian monsoon, or rainy season, usually begins in June or July, depending on location, and ends late in September.

5. The data in each of the following sets has been ordered from least to greatest. For each set,

i) calculate the mean, median, and standard deviation;

ii) create a frequency polygon; and

iii) explain why the distribution is or is not approximately normal.

a) daily maximum temperatures (°C) in monsoon season in India:
41.5, 42.4, 42.6, 42.7, 42.9, 43.0, 43.6, 44.0, 44.5, 44.6, 44.6, 44.8, 45.0, 45.3, 45.5, 45.5, 45.6, 45.7, 45.8, 46.1, 46.3, 46.4, 46.5, 46.6, 46.8, 47.0, 47.2, 47.6, 47.6, 47.9

b) class marks on a pop quiz out of 15:
2, 4, 5, 6, 6, 6, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 9, 9, 9, 10, 10, 11, 11, 11, 12, 12, 13, 13, 15

6. A manufacturer offers a warranty on its coffee makers. The coffee makers have a mean lifespan of 4.5 years, with a standard deviation of 0.7 years. For how long should the coffee makers be covered by the warranty, if the manufacturer wants to repair no more than 2.5% of the coffee makers sold?

Rolling 3 Dice	
Sum	**Frequency**
3	1
4	3
5	6
6	10
7	15
8	21
9	25
10	27
11	27
12	25
13	21
14	15
15	10
16	6
17	3
18	1

7. Hila found the data at the left that shows the number of ways that each sum can be obtained when rolling three dice.

a) Determine the mean and the standard deviation.

b) Draw a frequency polygon to show the data.

c) Does the data have a normal distribution? Explain.

8. The company payroll of Sweetwater Communications has a mean monthly salary of $5400, with a standard deviation of $800.

a) Sketch a normal curve to represent the salaries for the company.

b) Sketch a curve to show the effects of Proposal 1: Each employee receives a raise of $270 per month.

c) Sketch a curve to show the effects of Proposal 2: Each employee receives a 5% raise on the original salary.

9. The results for the first round of the 2009 Masters golf tournament are given below.

65	68	70	71	72	73	74	76
66	69	70	71	72	73	74	76
66	69	70	72	73	73	75	77
67	69	71	72	73	73	75	77
67	69	71	72	73	73	75	77
68	69	71	72	73	73	75	78
68	69	71	72	73	73	75	78
68	70	71	72	73	73	75	78
68	70	71	72	73	73	75	79
68	70	71	72	73	74	75	79
68	70	71	72	73	74	75	79
68	70	71	72	73	74	76	80

a) Are the golf scores normally distributed?
b) Explain how the measures of central tendency support your decision in part a).

10. A school of 130 bottlenose dolphins is living in a protected environment. The life expectancy of the dolphins is normally distributed, with a mean of 39 years and a standard deviation of 3.5 years. How many of these dolphins can be expected to live more than 46 years?

11. Julie is an engineer who designs roller coasters. She wants to design a roller coaster that 95% of the population can ride. The average adult in North America has a mass of 71.8 kg, with a standard deviation of 13.6 kg.
a) What range of masses should Julie consider in her design?
b) If Julie wanted to design a roller coaster that 99.7% of the population could ride, what range of masses should she consider?
c) What assumption is being made, which could cause problems if it is not valid?

12. A new video game is being tested with a sample of students. The scores on the first attempt for each player are recorded in the table.
a) Graph the data. Does the data appear to have a normal distribution?
b) Determine the mean and standard deviation of the data. Do these values validate your answer to part a)?

Scores	Freq.
less than 18 000	2
18 000–27 000	5
27 000–36 000	14
36 000–45 000	36
45 000–54 000	77
54 000–63 000	128
63 000–72 000	163
72 000–81 000	163
81 000–90 000	127
90 000–99 000	80
99 000–108 000	33
108 000–117 000	14
117 000–126 000	6
greater than 126 000	2

13. In a dog obedience class, the masses of the 60 dogs enrolled were normally distributed, with a mean of 11.2 kg and a standard deviation of 2.8 kg. How many dogs would you expect to fall within each range of masses?
a) between 8.4 kg and 14.0 kg
b) between 5.6 kg and 16.8 kg
c) between 2.8 kg and 19.6 kg
d) less than 11.2 kg

14. The mass of an Appaloosa horse is generally in the range of 431 kg to 533 kg. Assuming that the data is normally distributed, determine the mean and standard deviation for the mass of an Appaloosa. Justify your answers.

The Appaloosa Horse Club of Canada Museum is located in Claresholm, Alberta.

Closing

15. Explain why a selection of 10 students from a class can have marks that are not normally distributed, even when the marks of the whole class are normally distributed.

Extending

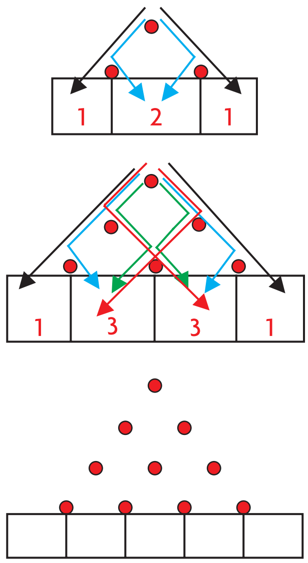
16. Newfoundland dogs have masses that are normally distributed. The mean mass of a male dog is 63.5 kg, with a standard deviation of 1.51 kg. The mean mass of a female dog is 49.9 kg, with a standard deviation of 1.51 kg. Esteban claims that he used to have two adult Newfoundland dogs: a male that was 78.9 kg and a female that was 29.9 kg. Using your knowledge of normal distribution, do you think he is being truthful? Explain.

Applying Problem-Solving Strategies

Predicting Possible Pathways

A Galton board is a triangular array of pegs that is used for statistical experiments. Balls are dropped, one at a time, onto the top peg and fall either right or left. Then, as they hit a peg in the next row, they fall either right or left again, until they finally pass through a slot at the bottom, where they can be counted. Each ball is equally likely to fall either way.

The Puzzle



A. For an array with 2 rows and 3 pegs, there is one way for a ball to fall into each end slot, and two ways for a ball to fall into the middle slot. For an array with 3 rows and 6 pegs, there is one way for a ball to fall into each end slot, and three ways for a ball to fall into the two middle slots.

Determine the number of ways for a ball to fall through an array with 4 rows and 10 pegs.

B. Examine each array in step A. Look for a pattern to determine how many pegs there would be in a 5-row array and a 6-row array.

C. Look for a pattern in the number of ways for a ball to fall into each slot in the arrays in part A. Use this pattern to determine the number of ways for a ball to fall into each slot in a 5-row array and a 6-row array.

The Strategy

D. Describe a strategy you could use to determine the number of ways for a ball to fall into each slot in an array of any size.

E. Use your strategy to determine the number of ways for a ball to fall into each slot in an array with 10 rows.

F. Draw a histogram or a frequency polygon to illustrate the results for a 10-row array. Comment on the distribution.