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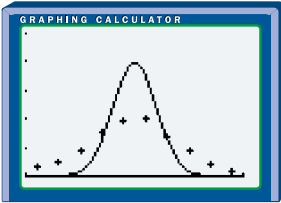
3. **a)** Using the techniques of section 2.1, create a table like the one below for each data set. Calculate the sample mean amount $\bar{x}$ of molybdenum for each supplier.

Supplier A				
Molybdenum (parts per 10 000)	Frequency (f)	Mid-Interval Value (x)	f × x	f × x²
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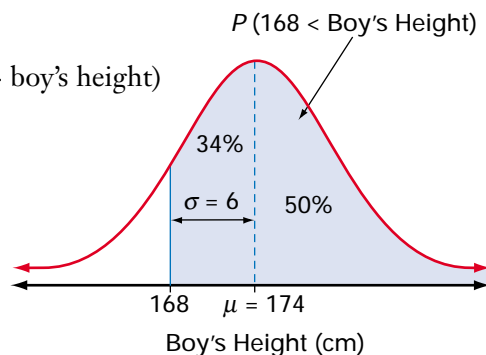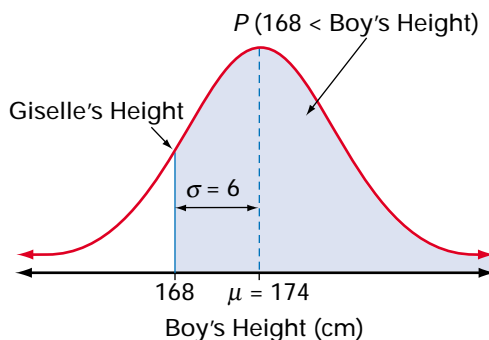 168 cm to 180 cm will contain 68% of the data. Because the normal distribution is symmetric, the bottom half of this range, 168 cm to 174 cm (the mean value), contains $\dfrac{68\%}{2} = 34\%$ of the data.

Therefore,

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

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.
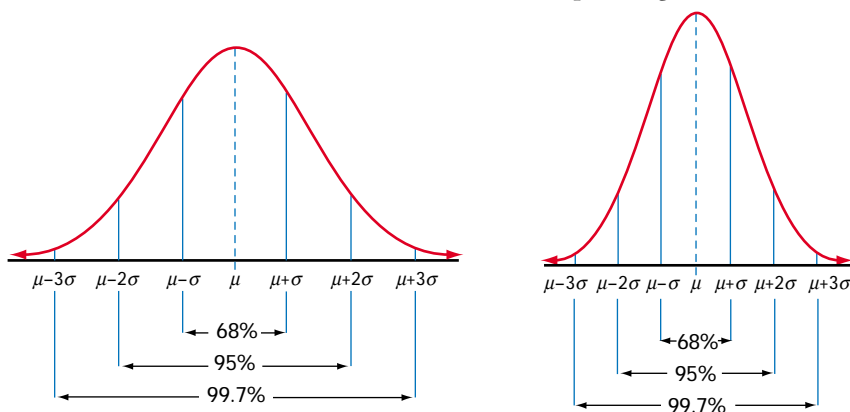
Notice that, for any normal distribution X,

- 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.

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 $\left(\dfrac{x-\mu}{\sigma}\right)$ 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 $\dfrac{1}{\sigma\sqrt{2\pi}}$ 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.

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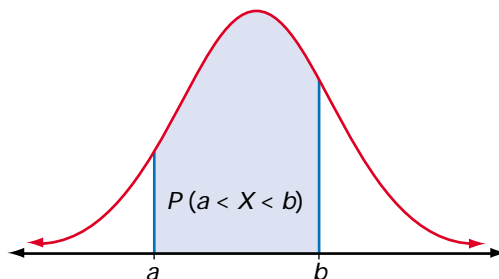
No simple formula exists for the areas under normal distribution curves. Instead, these areas have to be calculated using a more accurate version of "counting grid squares." However, you can simplify such calculations considerably by using z-scores.

Recall from section 2.6 that the z-score for a value of the random variable is $z = \dfrac{x-\mu}{\sigma}$. 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

$$P(X < 49.5) = P(Z < -1.67)$$
$$\doteq 0.0475$$
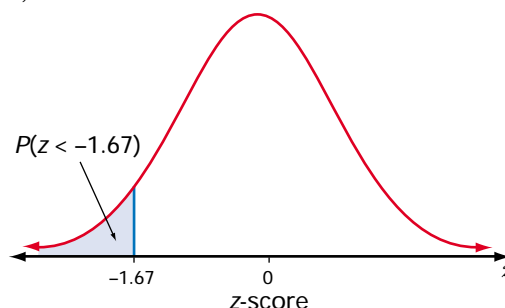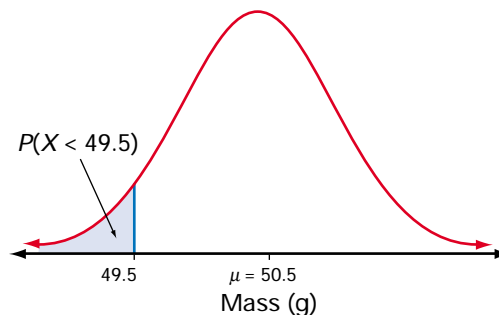
So, 4.8% of packages contain less than 49.5 g of powder.

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.
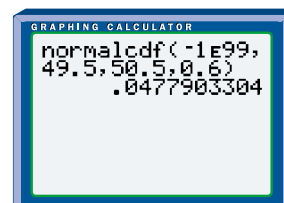
As described in section 2.6, a z-score is the number of standard deviations a value lies above the mean. Negative z-scores represent values that lie below the mean. You can use z-scores to rank any set of data, using the standard deviation as a unit of measure. Thus, in Example 2, the mass 49.5 g has 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

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

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.


```
GRAPHING CALCULATOR
invNorm(.75,540,
70)
      587.2142825
```
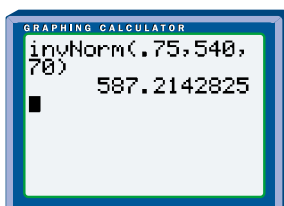
So, a score of 587 or above qualifies as third quartile (top 25%). Sharma will be accepted to the program, since her score of 655 easily exceeds 587.

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, $z = \dfrac{x - \mu}{\sigma}$, 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.

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?