The Normal Distribution

CHAPTER

9

The NYSE

The New York Stock Exchange (NYSE) was founded in 1792 by 24 stockbrokers who signed an agreement under a buttonwood tree on Wall Street in New York. The first offices were in a rented room at 40 Wall Street. In the 1830s traders who were not part of the Exchange did business in the street. They were called "curbstone brokers." It was the curbstone brokers who first made markets in gold and oil stocks and, after the Civil War, in small industrial companies such as the emerging steel, textile, and chemical industries.

By 1903 the New York Stock Exchange was established at its current home at 18 Broad Street. The curbstone brokers finally moved indoors in 1921 to a building on Greenwich street in lower Manhattan. In 1953 the curb market changed its name to the American Stock Exchange. In 1993 the American Stock Exchange pioneered the market for derivatives by introducing the first exchange-traded fund, Standard & Poor's Depositary Receipts (SPDRs).

The NYSE Euronext holding company was created in 2007 as a combination of the NYSE Group, Inc., and Euronext N.V. And in 2008, NYSE Euronext merged with the American Stock Exchange. The combined exchange is the world's largest and most liquid exchange group.

9.1 The Standard Deviation as a Ruler

WHO	Months
WHAT	CAPE10 values for the NYSE
WHEN	1880 through mid 2010
WHY	Investment guidance

Investors have always sought ways to help them decide when to buy and when to sell. Such measures have become increasingly sophisticated. But all rely on identifying when the stock market is in an unusual state—either unusually undervalued (buy!) or unusually overvalued (sell!). One such measure is the Cyclically Adjusted Price/Earnings Ratio (CAPE10) developed by Yale professor Robert Shiller. The CAPE10 is based on the standard Price/Earnings (P/E) ratio of stocks, but designed to smooth out short-term fluctuations by "cyclically adjusting" them. The CAPE10 has been as low as 4.78, in 1920, and as high as 44.20, in late 1999. The long-term average CAPE10 (since year 1881) is 16.34.

Investors who follow the CAPE10 use the metric to signal times to buy and sell. One mutual fund strategy buys only when the CAPE10 is 33% lower than the long-term average and sells (or "goes into cash") when the CAPE10 is 50% higher than the long-term average. Between January 1, 1971, and October 23, 2009, this strategy would have outperformed such standard measures as the Wiltshire 5000 in both average return and volatility, but it is important to note that the strategy would have been completely in cash from just before the stock market crash of 1987 all the way to March of 2009! Shiller popularized the strategy in his book *Irrational Exuberance*. Figure 9.1 shows a time series plot of the CAPE10 values for the New York Stock Exchange from 1880 until the middle of 2010. Generally, the CAPE10 hovers around 15. But occasionally, it can take a large excursion. One such time was in 1999 and 2000, when the CAPE10 exceeded 40. But was this just a random peak or were these values really extraordinary?

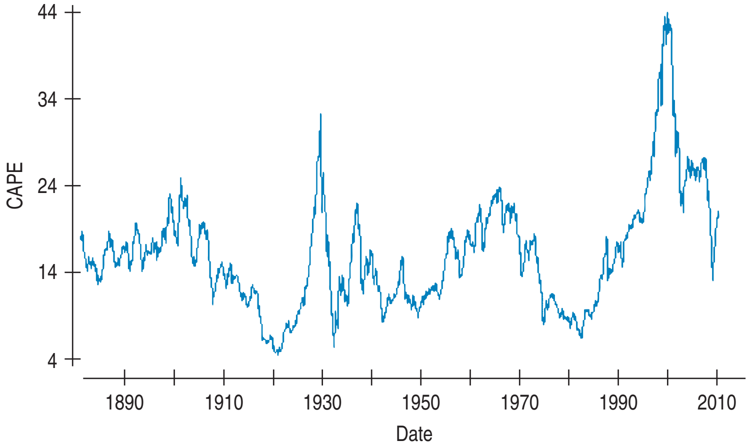


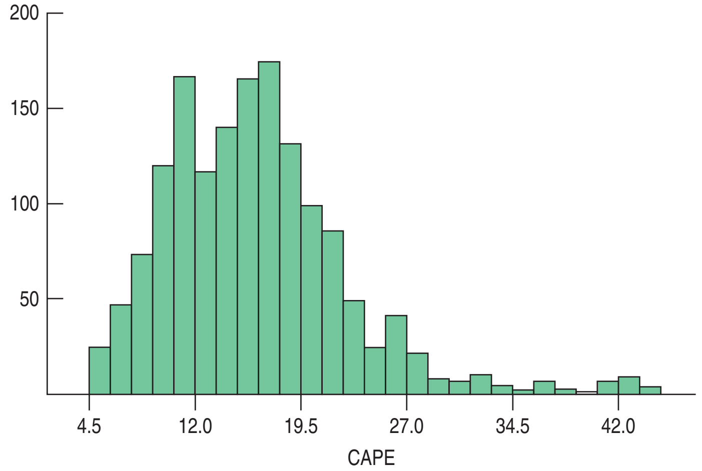
Figure 9.1 CAPE10 values for the NYSE from 1880 to 2010.

We can look at the overall distribution of CAPE10 values. Figure 9.2 shows a histogram of the same values. Now we don't see patterns over time, but we may be able to make a better judgment of whether values are extraordinary.

Overall, the main body of the distribution looks unimodal and reasonably symmetric. But then there's a tail of values that trails off to the high end. How can we assess how extraordinary they are?

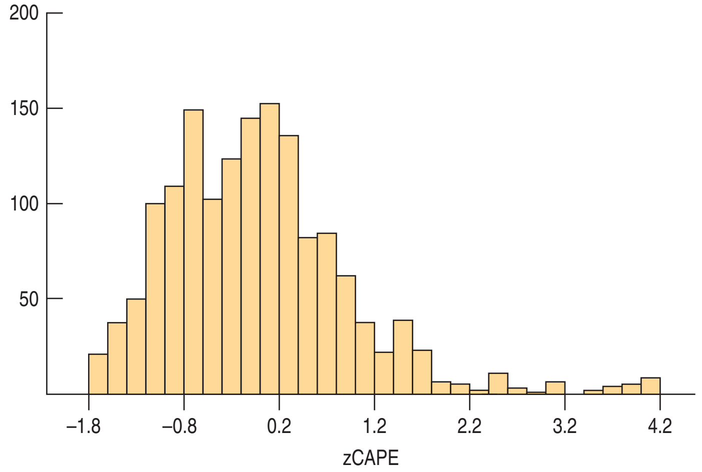
Investors follow a wide variety of measures that record various aspects of stocks, bonds, and other investments. They are usually particularly interested in identifying times when these measures are extraordinary because those often represent times of increased risk or opportunity. But these are quantitative values, not categories.

How can we characterize the usual behavior of a random variable that can take on any value in a range of values? The distributions of Chapter 8 won't provide the tools we need, but many of the basic concepts still work. The random variables we need are *continuous*.



| **Figure 9.2** The distribution of the CAPE10 values shown in Figure 9.1.

We saw in Chapter 5 that *z*-scores provide a standard way to compare values. In a sense, we use the standard deviation as a ruler, asking how many standard deviations a value is from the mean. That's what a *z*-score reports; the number of standard deviations away from the mean. We can convert the CAPE10 values to *z*-scores by subtracting their mean (16.3559) and dividing by their standard deviation (6.58). Figure 9.3 shows the resulting distribution.



| **Figure 9.3** The CAPE10 values as *z*-scores.

It's easy to see that the *z*-scores have the same distribution as the original values, but now we can also see that the largest of them is above 4. How extraordinary is it for a value to be four standard deviations away from the mean? Fortunately, there's a fact about unimodal, symmetric distributions that can guide us.[1]

The 68–95–99.7 Rule

In a unimodal, symmetric distribution, about 68% of the values fall within 1 standard deviation of the mean, about 95% fall within 2 standard deviations of the mean, and about 99.7%—almost all—fall within 3 standard deviations of the

[1] All of the CAPE10 values in the right tail occurred after 1993. Until that time the distribution of CAPE10 values was quite symmetric and clearly unimodal.

mean. Calling this rule the **68–95–99.7 Rule** provides a mnemonic for these three values.[2]

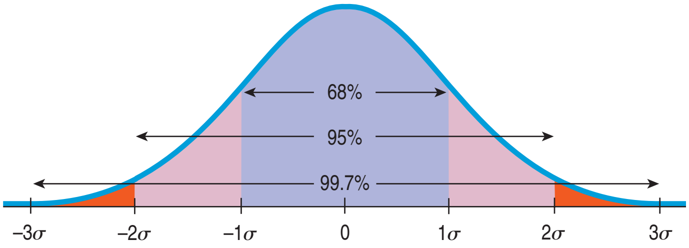


Figure 9.4 The 68–95–99.7 Rule tells us how much of most unimodal, symmetric models is found within one, two, or three standard deviations of the mean.

For Example An extraordinary day for the Dow?

On May 6, 2010, the Dow Jones Industrial Average (DJIA) lost 404.7 points. Although that wasn't the most ever lost in a day, it was a large amount for that period. During the previous year, the mean change in the DJIA was −9.767 with a standard deviation of 98.325 points. A histogram of day-to-day changes in the DJIA looks like this:

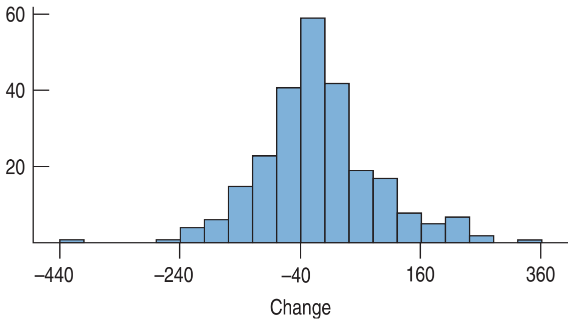


Figure 9.5 Day-to-day changes in the Dow Jones Industrial Average for the year ending June 2010.

Question: Use the 68–95–99.7 Rule to characterize how extraordinary the May 6 changes were. Is the rule appropriate?

Answer: The histogram is unimodal and symmetric, so the 68–95–99.7 rule is an appropriate model. The *z*-score corresponding to the May 6 change is

$$\frac{-404.7 - (-9.767)}{98.325} = -4.017$$

A *z*-score bigger than 3 in magnitude will occur with a probability of less than 0.015. A *z*-score of 4 is even less likely. This was a truly extraordinary event.

9.2 The Normal Distribution

"All models are wrong—but some are useful."

—George Box, famous statistician

The 68–95–99.7 Rule is useful in describing how unusual a *z*-score is. But often in business we want a more precise answer than one of these three values. To say more about how big we expect a *z*-score to be, we need to *model* the data's distribution.

[2] This rule is also called the "Empirical Rule" because it originally was observed without any proof. It was first published by Abraham de Moivre in 1733, 75 years before the underlying reason for it—which we're about to see—was known.

A model will let us say much more precisely how often we'd be likely to see z-scores of different sizes. Of course, like all models of the real world, the model will be wrong—wrong in the sense that it can't match reality exactly. But it can still be useful. Like a physical model, it's something we can look at and manipulate to learn more about the real world.

Models help our understanding in many ways. Just as a model of an airplane in a wind tunnel can give insights even though it doesn't show every rivet,[3] models of data give us summaries that we can learn from and use, even though they don't fit each data value exactly. It's important to remember that they're only *models* of reality and not reality itself. But without models, what we can learn about the world at large is limited to only what we can say about the data we have at hand.

There is no universal standard for z-scores, but there is a model that shows up over and over in Statistics. You may have heard of "bell-shaped curves." Statisticians call them Normal distributions. **Normal distributions** are appropriate models for distributions whose shapes are unimodal and roughly symmetric. There is a Normal distribution for every possible combination of mean and standard deviation. We write $N(\mu, \sigma)$ to represent a Normal distribution with a mean of μ and a standard deviation of σ. We use Greek symbols here because *this* mean and standard deviation are not numerical summaries of data. They are part of the model. They don't come from the data. Rather, they are numbers that we choose to help specify the distribution. Such numbers are called **parameters**.

If we model data with a Normal distribution and standardize them using the corresponding μ and σ, we still call the standardized values **z-scores,** and we write

$$z = \frac{y - \mu}{\sigma}.$$

If we standardize the data first (using its mean and standard deviation) it will have mean 0 and standard deviation 1. Then, to model it with a Normal, we'll need only the model $N(0,1)$. The Normal distribution with mean 0 and standard deviation 1 is called the **standard Normal distribution** (or the **standard Normal model**).

But be careful. You shouldn't use a Normal model for just any data set. Remember that standardizing won't change the shape of the distribution. If the distribution is not unimodal and symmetric to begin with, standardizing won't make it Normal.

Is Normal Normal?

Don't be misled. The name "Normal" doesn't mean that these are the *usual* shapes for histograms. The name follows a tradition of positive thinking in Mathematics and Statistics in which functions, equations, and relationships that are easy to work with or have other nice properties are called "normal," "common," "regular," "natural," or similar terms. It's as if by calling them ordinary, we could make them actually occur more often and make our lives simpler.

Notation Alert!

$N(\mu, \sigma)$ always denotes a Normal. The μ, pronounced "mew," is the Greek letter for "m," and always represents the mean in a model. The σ, sigma, is the lowercase Greek letter for "s," and always represents the standard deviation in a model.

Is the Standard Normal a Standard?

Yes. We call it the "Standard Normal" because it models standardized values. It is also a "standard" because this is the particular Normal model that we almost always use.

Just Checking

1 Your Accounting teacher has announced that the lower of your two tests will be dropped. You got a 90 on test 1 and an 80 on test 2. You're all set to drop the 80 until she announces that she grades "on a curve." She standardized the scores in order to decide which is the lower one. If the mean on the first test was 88 with a standard deviation of 4 and the mean on the second was 75 with a standard deviation of 5,

a) Which one will be dropped?

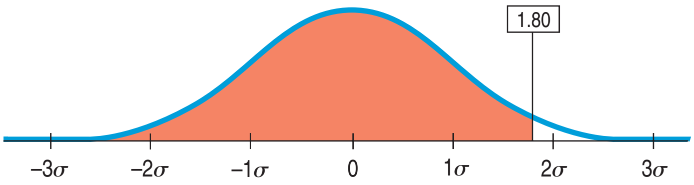
b) Does this seem "fair"?

Finding Normal Percentiles

Finding the probability that a proportion is at least 1 SD above the mean is easy. We know that 68% of the values lie within 1 SD of the mean, so 32% lie farther away. Since the Normal distribution is symmetric, half of those 32% (or 16%) are

[3] In fact, the model is useful *because* it doesn't have every rivet. It is because models offer a simpler view of reality that they are so useful as we try to understand reality.

more than 1 SD above the mean. But what if we want to know the percentage of observations that fall more than 1.8 SD above the mean? We already know that no more than 16% of observations have *z*-scores above 1. By similar reasoning, no more than 2.5% of the observations have a *z*-score above 2. Can we be more precise with our answer than "between 16% and 2.5%"?



z	.00	.01
1.7	0.9554	0.9564
1.8	0.9641	0.9649
1.9	0.9713	0.9719

Figure 9.6 A table of Normal percentiles (Table Z in Appendix D) lets us find the percentage of individuals in a standard Normal distribution falling below any specified z-score value.

When the value doesn't fall exactly 0, 1, 2, or 3 standard deviations from the mean, we can look it up in a table of **Normal percentiles**.[4] Tables use the standard Normal distribution, so we'll have to convert our data to *z*-scores before using the table. If our data value was 1.8 standard deviations above the mean, we would standardize it to a *z*-score of 1.80, and then find the value associated with a *z*-score of 1.80. If we use a table, as shown in Figure 9.6, we find the *z*-score by looking down the left column for the first two digits (1.8) and across the top row for the third digit, 0. The table gives the percentile as 0.9641. That means that 96.4% of the *z*-scores are less than 1.80. Since the total area is always 1, and 1 − 0.9641 = 0.0359 we know that only 3.6% of all observations from a Normal distribution have *z*-scores higher than 1.80. We can also find the probabilities associated with *z*-scores using technology such as calculators, statistical software, and various websites.

Finding Normal Percentiles

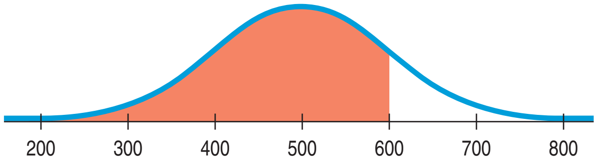
These days, finding percentiles from a Normal table is rarely necessary. Most of the time, we can use a calculator, a computer, or a website.

For Example GMAT scores and the Normal model

The Graduate Management Admission Test (GMAT) has scores from 200 to 800. Scores are supposed to follow a distribution that is roughly unimodal and symmetric and is designed to have an overall mean of 500 and a standard deviation of 100. In any one year, the mean and standard deviation may differ from these target values by a small amount, but we can use these values as good overall approximations.

Question: Suppose you earned a 600 on your GMAT test. From that information and the 68–95–99.7 Rule, where do you stand among all students who took the GMAT?

Answer: Because we're told that the distribution is unimodal and symmetric, we can approximate the distribution with a Normal model. We are also told the scores have a mean of 500 and an SD of 100. So, we'll use a *N*(500,100). It's good practice at this point to draw the distribution. Find the score whose percentile you want to know and locate it on the picture. When you finish the calculation, you should check to make sure that it's a reasonable percentile from the picture.



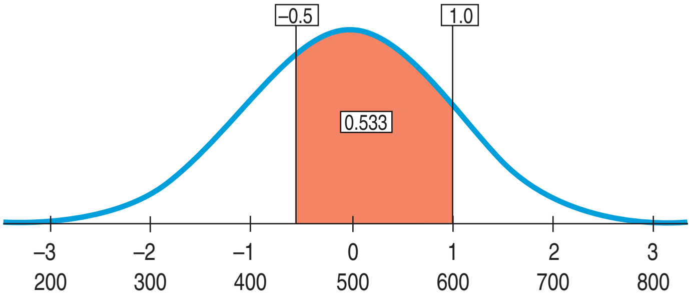
A score of 600 is 1 SD above the mean. That corresponds to one of the points in the 68–95–99.7% Rule. About 32% (100% − 68%) of those who took the test were more than one standard deviation from the mean, but only half of those were on the high side. So about 16% (half of 32%) of the test scores were better than 600.

[4] See Table Z in Appendix D. Many calculators and statistics computer packages do this as well.

For Example — More GMAT scores

Question: Assuming the GMAT scores are nearly Normal with $N(500,100)$, what proportion of GMAT scores falls between 450 and 600?

Answer: *The first step is to find the z-scores associated with each value.* Standardizing the scores we are given, we find that for 600, $z = (600 - 500)/100 = 1.0$ and for 450, $z = (450 - 500)/100 = -0.50$. We can label the axis below the picture either in the original values or the z-scores or even use both scales as the following picture shows.



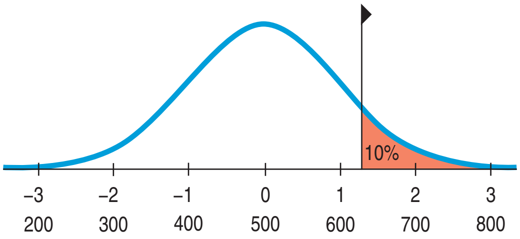
From Table Z, we find the area $z \le 1.0 = 0.8413$, which means that 84.13% of scores fall below 1.0, and the area $z \le -0.50 = 0.3085$, which means that 30.85% of the values fall below -0.5, so the proportion of z-scores *between* them is $84.13\% - 30.85\% = 53.28\%$. So, the Normal model estimates that about 53.3% of GMAT scores fall between 450 and 600.

Finding areas from z-scores is the simplest way to work with the Normal distribution. But sometimes we start with areas and are asked to work backward to find the corresponding z-score or even the original data value. For instance, what z-score represents the first quartile, Q1, in a Normal distribution? In our first set of examples, we knew the z-score and used the table or technology to find the percentile. Now we want to find the cut point for the 25th percentile. Make a picture, shading the leftmost 25% of the area. Look in Table Z for an area of 0.2500. The exact area is not there, but 0.2514 is the closest number. That shows up in the table with -0.6 in the left margin and .07 in the top margin. The z-score for Q1, then, is approximately $z = -0.67$. Computers and calculators can determine the cut point more precisely (and more easily).[5]

For Example — An exclusive MBA program

Question: Suppose an MBA program says it admits only people with GMAT scores among the top 10%. How high a GMAT score does it take to be eligible?

Answer: The program takes the top 10%, so their cutoff score is the 90th percentile. Draw an approximate picture like this one.



	0.07	*0.08*	*0.09*
1.0	0.8577	0.8599	0.8621
1.1	0.8790	0.8810	0.8830
1.2	0.8980	0.8997	0.9015
1.3	0.9147	0.9162	0.9177
1.4	0.9292	0.9306	0.9319

(continued)

[5] We'll often use those more precise values in our examples. If you're finding the values from the table you may not get *exactly* the same number to all decimal places as your classmate who's using a computer package.

From our picture we can see that the z-value is between 1 and 1.5 (if we've judged 10% of the area correctly), and so the cutoff score is between 600 and 650 or so. Using technology, you may be able to select the 10% area and find the z-value directly. Using a table, such as Table Z, locate 0.90 (or as close to it as you can; here 0.8997 is closer than 0.9015) in the *interior* of the table and find the corresponding z-score (see table above). Here the 1.2 is in the left margin, and the .08 is in the margin above the entry. Putting them together gives 1.28. Now, convert the z-score back to the original units. From Table Z, the cut point is $z = 1.28$. A z-score of 1.28 is 1.28 standard deviations above the mean. Since the standard deviation is 100, that's 128 GMAT points. The cutoff is 128 points above the mean of 500, or 628. Because the program wants GMAT scores in the top 10%, the cutoff is 628. (Actually since GMAT scores are reported only in multiples of 10, you'd have to score at least a 630.)

Guided Example Cereal Company

A cereal manufacturer has a machine that fills the boxes. Boxes are labeled "16 oz," so the company wants to have that much cereal in each box. But since no packaging process is perfect, there will be minor variations. If the machine is set at exactly 16 oz and the Normal distribution applies (or at least the distribution is roughly symmetric), then about half of the boxes will be underweight, making consumers unhappy and exposing the company to bad publicity and possible lawsuits. To prevent underweight boxes, the manufacturer has to set the mean a little higher than 16.0 oz. Based on their experience with the packaging machine, the company believes that the amount of cereal in the boxes fits a Normal distribution with a standard deviation of 0.2 oz. The manufacturer decides to set the machine to put an average of 16.3 oz in each box. Let's use that model to answer a series of questions about these cereal boxes.

Question 1: What fraction of the boxes will be underweight?

PLAN

Setup State the variable and the objective.

The variable is weight of cereal in a box. We want to determine what fraction of the boxes risk being underweight.

Model Check to see if a Normal distribution is appropriate.

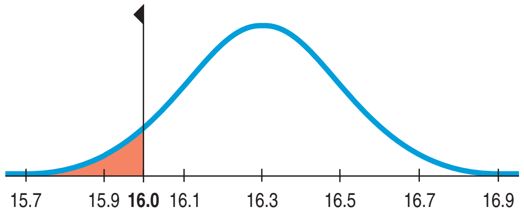
We have no data, so we cannot make a histogram. But we are told that the company believes the distribution of weights from the machine is Normal.

Specify which Normal distribution to use.

We use an N(16.3, 0.2) model.

DO

Mechanics Make a graph of this Normal distribution. Locate the value you're interested in on the picture, label it, and shade the appropriate region.



REALITY CHECK Estimate from the picture the percentage of boxes that are underweight. (This will be useful later to check that your answer makes sense.)

(It looks like a low percentage—maybe less than 10%.)

We want to know what fraction of the boxes will weigh less than 16 oz.

Convert your cutoff value into a *z*-score.

$$z = \frac{y - \mu}{\sigma} = \frac{16 - 16.3}{0.2} = -1.50.$$

Look up the area in the Normal table, or use technology.

Area ($y < 16$) = Area ($Z < -1.50$) = 0.0668.

REPORT

Conclusion State your conclusion in the context of the problem.

We estimate that approximately 6.7% of the boxes will contain less than 16 oz of cereal.

Question 2: The company's lawyers say that 6.7% is too high. They insist that no more than 4% of the boxes can be underweight. So the company needs to set the machine to put a little more cereal in each box. What mean setting do they need?

PLAN

Setup State the variable and the objective.

The variable is weight of cereal in a box. We want to determine a setting for the machine.

Model Check to see if a Normal model is appropriate.

We have no data, so we cannot make a histogram. But we are told that a Normal model applies.

Specify which Normal distribution to use. This time you are not given a value for the mean!

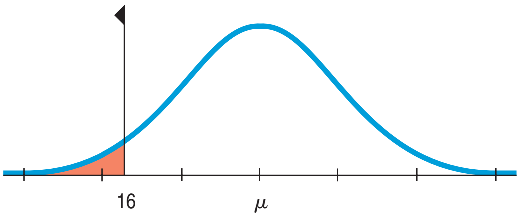
We don't know μ, the mean amount of cereal. The standard deviation for this machine is 0.2 oz. The model, then, is $N(\mu, 0.2)$.

REALITY CHECK We found out earlier that setting the machine to $\mu = 16.3$ oz made 6.7% of the boxes too light. We'll need to raise the mean a bit to reduce this fraction.

We are told that no more than 4% of the boxes can be below 16 oz.

DO

Mechanics Make a graph of this Normal distribution. Center it at μ (since you don't know the mean) and shade the region below 16 oz.



Using the Normal table, a calculator, or software, find the *z*-score that cuts off the lowest 4%.

The *z*-score that has 0.04 area to the left of it is $z = -1.75$.

Use this information to find μ. It's located 1.75 standard deviations to the right of 16.

Since 16 must be 1.75 standard deviations below the mean, we need to set the mean at $16 + 1.75 \cdot 0.2 = 16.35$.

REPORT

Conclusion State your conclusion in the context of the problem.

The company must set the machine to average 16.35 oz of cereal per box.

(continued)

Question 3: The company president vetoes that plan, saying the company should give away less free cereal, not more. Her goal is to set the machine no higher than 16.2 oz and still have only 4% underweight boxes. The only way to accomplish this is to reduce the standard deviation. What standard deviation must the company achieve, and what does that mean about the machine?

PLAN

Setup State the variable and the objective.

The variable is weight of cereal in a box. We want to determine the necessary standard deviation to have only 4% of boxes underweight.

Model Check that a Normal model is appropriate.

The company believes that the weights are described by a Normal distribution.

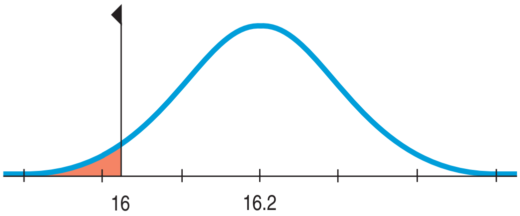
Specify which Normal distribution to use. This time you don't know σ.

Now we know the mean, but we don't know the standard deviation. The model is therefore $N(16.2, \sigma)$.

REALITY CHECK We know the new standard deviation must be less than 0.2 oz.

DO

Mechanics Make a graph of this Normal distribution. Center it at 16.2, and shade the area you're interested in. We want 4% of the area to the left of 16 oz.



Find the z-score that cuts off the lowest 4%.

We already know that the z-score with 4% below it is $z = -1.75$.

Solve for σ. (Note that we need 16 to be 1.75 σ's below 16.2, so 1.75σ must be 0.2 oz. You could just start with that equation.)

$$z = \frac{y - \mu}{\sigma}$$

$$-1.75 = \frac{16 - 16.2}{\sigma}$$

$$1.75\sigma = 0.2$$

$$\sigma = 0.114.$$

REPORT

Conclusion State your conclusion in the context of the problem.

As we expected, the standard deviation is lower than before—actually, quite a bit lower.

The company must get the machine to box cereal with a standard deviation of only 0.114 oz. This means the machine must be more consistent (by nearly a factor of 2) in filling the boxes.

9.3 Normal Probability Plots

A specialized graphical display can help you to decide whether the Normal model is appropriate: the **Normal probability plot.** If the distribution of the data is roughly Normal, the plot is roughly a diagonal straight line. Deviations from a straight line indicate that the distribution is not Normal. This plot is usually able to show deviations from Normality more clearly than the corresponding histogram, but it's usually easier to understand *how* a distribution fails to be Normal by looking at its histogram. Normal probability plots are difficult to make by hand, but are provided by most statistics software.

Some data on a car's fuel efficiency provide an example of data that are nearly Normal. The overall pattern of the Normal probability plot is straight. The two trailing low values correspond to the values in the histogram that trail off the low end. They're not quite in line with the rest of the data set. The Normal probability plot shows us that they're a bit lower than we'd expect of the lowest two values in a Normal distribution.

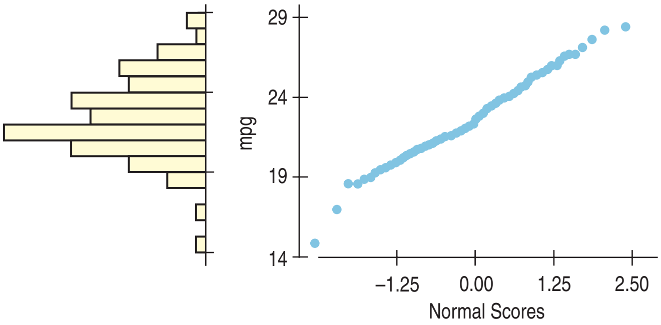


Figure 9.7 Histogram and Normal probability plot for gas mileage (mpg) recorded for a Nissan Maxima. The vertical axes are the same, so each dot on the probability plot would fall into the bar on the histogram immediately to its left.

By contrast, the Normal probability plot of a sample of men's *Weights* in Figure 9.8 from a study of lifestyle and health is far from straight. The weights are skewed to the high end, and the plot is curved. We'd conclude from these pictures that approximations using the Normal model for these data would not be very accurate.

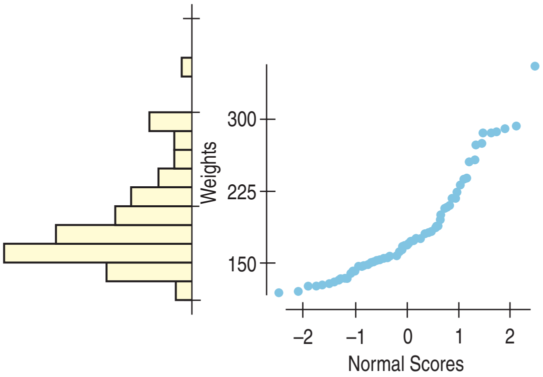
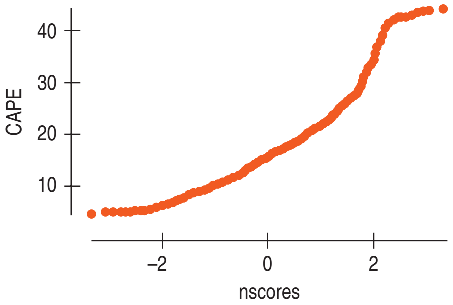


Figure 9.8 Histogram and Normal probability plot for men's weights. Note how a skewed distribution corresponds to a bent probability plot.

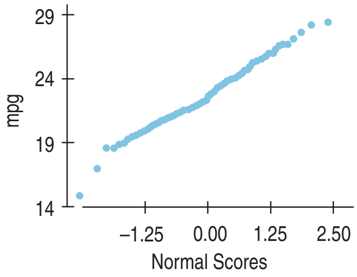
For Example Using a normal probability plot

A normal probability plot of the CAPE10 prices from page 247 looks like this:



Question: What does this plot say about the distribution of the CAPE10 scores?

Answer: The bent shape of the probability plot—and in particular, the sharp bend on the right—indicates a deviation from Normality—in this case the CAPE scores in the right tail do not stretch out as far as we'd expect for a Normal model.



How does a normal probability plot work?

Why does the Normal probability plot work like that? We looked at 100 fuel efficiency measures for a car. The smallest of these has a *z*-score of −3.16. The Normal model can tell us what value to expect for the smallest *z*-score in a batch of 100 if a Normal model were appropriate. That turns out to be −2.58. So our first data value is smaller than we would expect from the Normal.

We can continue this and ask a similar question for each value. For example, the 14th-smallest fuel efficiency has a *z*-score of almost exactly −1, and that's just what we should expect (−1.1 to be exact). We can continue in this way, comparing each observed value with the value we'd expect from a Normal model. The easiest way to make the comparison, of course, is to graph it.[6] If our observed values look like a sample from a Normal model, then the probability plot stretches out in a straight line from lower left to upper right. But if our values deviate from what we'd expect, the plot will bend or have jumps in it. The values we'd expect from a Normal model are called Normal scores, or sometimes nscores. You can't easily look them up in the table, so probability plots are best made with technology and not by hand.

The best advice on using Normal probability plots is to see whether they are straight. If so, then your data look like data from a Normal model. If not, make a histogram to understand how they differ from the model.

9.4 The Distribution of Sums of Normals

Another reason normal models show up so often is that they have some special properties. An important one is that the sum or difference of two independent Normal random variables is also Normal.

A company manufactures small stereo systems. At the end of the production line, the stereos are packaged and prepared for shipping. Stage 1 of this process is called "packing." Workers must collect all the system components (a main unit, two speakers, a power cord, an antenna, and some wires), put each in plastic bags, and then place everything inside a protective form. The packed form then moves on to

[6] Sometimes the Normal probability plot switches the two axes, putting the data on the *x*-axis and the *z*-scores on the *y*-axis.

Stage 2, called "boxing," in which workers place the form and a packet of instructions in a cardboard box and then close, seal, and label the box for shipping.

The company says that times required for the packing stage are unimodal and symmetric and can be described by a Normal distribution with a mean of 9 minutes and standard deviation of 1.5 minutes. (See Figure 9.9.) The times for the boxing stage can also be modeled as Normal, with a mean of 6 minutes and standard deviation of 1 minute.

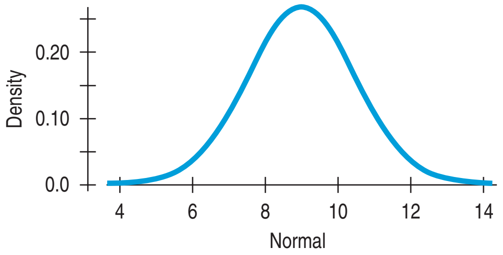
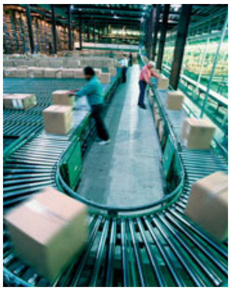


Figure 9.9 The Normal model for the packing stage with a mean of 9 minutes and standard deviation of 1.5 minutes.

The company is interested in the total time that it takes to get a system through both packing and boxing, so they want to model the sum of the two random variables. Fortunately, the special property that adding independent Normals yields another Normal allows us to apply our knowledge of Normal probabilities to questions about the sum or difference of independent random variables. To use this property of Normals, we'll need to check two assumptions: that the variables are Independent and that they can be modeled by the Normal distribution.

Guided Example Packaging Stereos



Consider the company that manufactures and ships small stereo systems that we discussed previously.

If the time required to pack the stereos can be described by a Normal distribution, with a mean of 9 minutes and standard deviation of 1.5 minutes, and the times for the boxing stage can also be modeled as Normal, with a mean of 6 minutes and standard deviation of 1 minute, what is the probability that packing an order of two systems takes over 20 minutes? What percentage of the stereo systems takes longer to pack than to box?

Question 1: What is the probability that packing an order of two systems takes more than 20 minutes?

PLAN

Setup State the problem.

We want to estimate the probability that packing an order of two systems takes more than 20 minutes.

Variables Define your random variables.

Let P_1 = time for packing the first system

P_2 = time for packing the second system

T = total time to pack two systems

Write an appropriate equation for the variables you need.

$T = P_1 + P_2$

(continued)

Think about the model assumptions.

✓ **Normal Model Assumption.** We are told that packing times are well modeled by a Normal model, and we know that the sum of two Normal random variables is also Normal.

✓ **Independence Assumption.** There is no reason to think that the packing time for one system would affect the packing time for the next, so we can reasonably assume the two are independent.

DO

Mechanics Find the expected value. (Expected values always add.)

$$\begin{aligned} E(T) &= E(P_1 + P_2) \\ &= E(P_1) + E(P_2) \\ &= 9 + 9 = 18 \text{ minutes} \end{aligned}$$

Find the variance.

For sums of independent random variables, variances add. (In general, we don't need the variables to be Normal for this to be true—just independent.)

Since the times are independent,

$$\begin{aligned} Var(T) &= Var(P_1 + P_2) \\ &= Var(P_1) + Var(P_2) \\ &= 1.5^2 + 1.5^2 \\ Var(T) &= 4.50 \end{aligned}$$

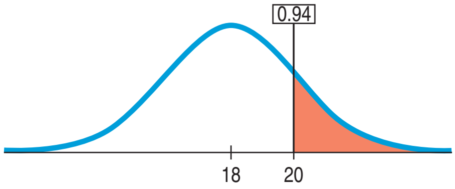
Find the standard deviation.

$$SD(T) = \sqrt{4.50} \approx 2.12 \text{ minutes}$$

Now we use the fact that both random variables follow Normal distributions to say that their sum is also Normal.

We can model the time, T, with a $N(18, 2.12)$ model.

Sketch a picture of the Normal distribution for the total time, shading the region representing over 20 minutes.



Find the z-score for 20 minutes.

$$z = \frac{20 - 18}{2.12} = 0.94$$

Use technology or a table to find the probability.

$$P(T > 20) = P(z > 0.94) = 0.1736$$

REPORT

Conclusion Interpret your result in context.

MEMO

Re: Computer Systems Packing

Using past history to build a model, we find slightly more than a 17% chance that it will take more than 20 minutes to pack an order of two stereo systems.

Question 2: What percentage of stereo systems take longer to pack than to box?

PLAN

Setup State the question.

We want to estimate the percentage of the stereo systems that takes longer to pack than to box.

Variables Define your random variables.

Let P = time for packing a system
B = time for boxing a system
D = difference in times to pack and box a system

Write an appropriate equation.

$D = P - B$

What are we trying to find? Notice that we can tell which of two quantities is greater by subtracting and asking whether the difference is positive or negative.

A system that takes longer to pack than to box will have $P > B$, and so D will be positive. We want to find $P(D > 0)$.

Remember to think about the assumptions.

✓ **Normal Model Assumption.** We are told that both random variables are well modeled by Normal distributions, and we know that the difference of two Normal random variables is also Normal.

✓ **Independence Assumption.** There is no reason to think that the packing time for a system will affect its boxing time, so we can reasonably assume the two are independent.

DO

Mechanics Find the expected value.

$$\begin{aligned} E(D) &= E(P - B) \\ &= E(P) - E(B) \\ &= 9 - 6 = 3 \text{ minutes} \end{aligned}$$

For the difference of independent random variables, the variance is the sum of the individual variances.

Since the times are independent,

$$\begin{aligned} Var(D) &= Var(P - B) \\ &= Var(P) + Var(B) \\ &= 1.5^2 + 1^2 \end{aligned}$$

$Var(D) = 3.25$

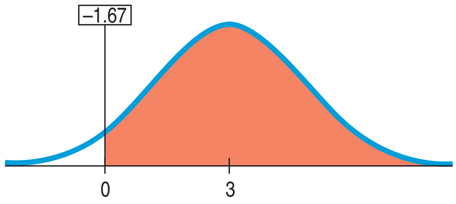
Find the standard deviation.

$SD(D) = \sqrt{3.25} \approx 1.80$ minutes

State what model you will use.

We can model D with $N\,(3, 1.80)$.

Sketch a picture of the Normal distribution for the difference in times and shade the region representing a difference greater than zero.



Find the *z*-score. Then use a table or technology to find the probability.

$$z = \frac{0 - 3}{1.80} = -1.67$$

$$P(D > 0) = P(z > -1.67) = 0.9525$$

REPORT

Conclusion Interpret your result in context.

MEMO

Re: Computer Systems Packing

In our second analysis, we found that just over 95% of all the stereo systems will require more time for packing than for boxing.

9.5 The Normal Approximation for the Binomial

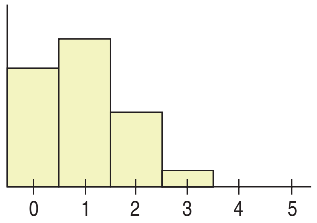
The Normal distribution can approximate discrete events when the number of possible events is large. In particular, it is a good model for sums of independent random variables of which a Binomial random variable is a special case. Here's an example of how the Normal can be used to calculate binomial probabilities. Suppose that the Tennessee Red Cross anticipates the need for at least 1850 units of O-negative blood this year. It estimates that it will collect blood from 32,000 donors. How likely is the Tennessee Red Cross to meet its need? We learned how to calculate such probabilities in Chapter 8. We could use the binomial model with $n = 32{,}000$ and $p = 0.06$. The probability of getting *exactly* 1850 units of O-negative blood from 32,000 donors is $\binom{32000}{1850} \times 0.06^{1850} \times 0.94^{30150}$. No calculator on earth can calculate 32000 choose 1850 (it has more than 100,000 digits).[7] And that's just the beginning. The problem said *at least* 1850, so we would have to calculate it again for 1851, for 1852, and all the way up to 32,000. When we're dealing with a large number of trials like this, making direct calculations of the probabilities becomes tedious (or outright impossible).

The Binomial model has mean $np = 1920$ and standard deviation $\sqrt{npq} \approx 42.48$. We can approximate its distribution with a Normal distribution using the same mean and standard deviation. Remarkably enough, that turns out to be a very good approximation. Using that mean and standard deviation, we can find the *probability:*

$$P(X \geq 1850) = P\left(z \geq \frac{1850 - 1920}{42.48}\right) \approx P(z \;\geq -1.65) \approx 0.95$$

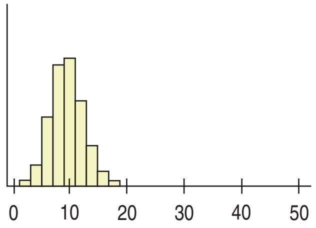
There seems to be about a 95% chance that this Red Cross chapter will have enough O-negative blood.

We can't always use a Normal distribution to make estimates of Binomial probabilities. The success of the approximation depends on the sample size. Suppose we are searching for a prize in cereal boxes, where the probability of finding a prize is 20%. If we buy five boxes, the actual Binomial probabilities that we get 0, 1, 2, 3, 4, or 5 prizes are 33%, 41%, 20%, 5%, 1%, and 0.03%, respectively. The histogram just below shows that this probability model is skewed. We shouldn't try to estimate these probabilities by using a Normal model.



But if we open 50 boxes of this cereal and count the number of prizes we find, we'll get the histogram below. It is centered at $np = 50(0.2) = 10$ prizes, as expected, and it appears to be fairly symmetric around that center.

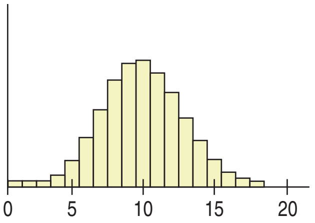
[7] If your calculator *can* find Binom(32000, 0.06), then apparently it's smart enough to use an approximation.



The third histogram (shown just below) shows the same distribution, still centered at the expected value of 10 prizes. It looks close to Normal for sure. With this larger sample size, it appears that a Normal distribution might be a useful approximation.

*The continuity correction

When we use a continuous model to model a set of discrete events, we may need to make an adjustment called the **continuity correction**. We approximated the Binomial distribution (50, 0.2) with a Normal distribution. But what does the Normal distribution say about the probability that $X = 10$? Every specific value in the Normal probability model has probability 0. That's not the answer we want.



Because X is really discrete, it takes on the exact values 0, 1, 2, . . . , 50, each with positive probability. The histogram holds the secret to the correction. Look at the bin corresponding to $X = 10$ in the histogram. It goes from 9.5 to 10.5. What we really want is to find the area under the normal curve *between* 9.5 and 10.5. So when we use the Normal distribution to approximate discrete events, we go halfway to the next value on the left and/or the right. We approximate $P(X = 10)$ by finding $P(9.5 \le X \le 10.5)$. For a Binomial (50, 0.2), $\mu = 10$ and $\sigma = 2.83$.

$$\text{So } P(9.5 \le X \le 10.5) \approx P\left(\frac{9.5 - 10}{2.83} \le z \le \frac{10.5 - 10}{2.83}\right)$$
$$= P(-0.177 \le z \le 0.177)$$
$$= 0.1405$$

By comparison, the *exact* Binomial probability is 0.1398.

A Normal distribution is a close enough approximation to the Binomial only for a large enough number of trials. And what we mean by "large enough" depends on the probability of success. We'd need a larger sample if the probability of success were very low (or very high). It turns out that a Normal distribution works pretty well if we expect to see at least 10 successes and 10 failures. That is, we check the Success/Failure Condition.

Success/Failure Condition: A Binomial model is approximately Normal if we expect at least 10 successes and 10 failures:

$$np \ge 10 \text{ and } nq \ge 10.$$

Why 10? Well, actually it's 9, as revealed in the following Math Box.

Math Box

Why Check *np* > 10?

It's easy to see where the magic number 10 comes from. You just need to remember how Normal models work. The problem is that a Normal model extends infinitely in both directions. But a Binomial model must have between 0 and n successes, so if we use a Normal to approximate a Binomial, we have to cut off its tails. That's not very important if the center of the Normal model is so far from 0 and n that the lost tails have only a negligible area. More than three standard deviations should do it because a Normal model has little probability past that.

So the mean needs to be at least 3 standard deviations from 0 and at least 3 standard deviations from n. Let's look at the 0 end.

We require: $\mu - 3\sigma > 0$
Or, in other words: $\mu > 3\sigma$
For a Binomial that's: $np > 3\sqrt{npq}$
Squaring yields: $n^2p^2 > 9npq$
Now simplify: $np > 9q$
Since $q \leq 1$, we require: $np > 9$

For simplicity we usually demand that np (and nq for the other tail) be at least 10 to use the Normal approximation which gives the Success/Failure Condition.[8]

For Example

Using the Normal distribution

Some LCD panels have stuck or "dead" pixels that have defective transistors and are permanently unlit. If a panel has too many dead pixels, it must be rejected. A manufacturer knows that, when the production line is working correctly, the probability of rejecting a panel is .07.

Questions:

a) How many screens do they expect to reject in a day's production run of 500 screens? What is the standard deviation?

b) If they reject 40 screens today, is that a large enough number that they should be concerned that something may have gone wrong with the production line?

c) In the past week of 5 days of production, they've rejected 200 screens—an average of 40 per day. Should that raise concerns?

Answers:

a) $\mu = 0.07 \times 500 = 35$ is the expected number of rejects

$\sigma = \sqrt{npq} = \sqrt{500 \times 0.07 \times 0.93} = 5.7$

b) $P(X \geq 45) = P\left(z \geq \dfrac{40 - 35}{5.7}\right) = P(z \geq 0.877) \approx 0.29$, not an extraordinarily large number of rejects

c) Using the Normal approximation:

$\mu = 0.07 \times 2500 = 1.75$

$\sigma = \sqrt{2500 \times 0.07 \times 0.93} = 12.757$

$$P(X \geq 200) = P\left(z \geq \frac{200 - 175}{12.757}\right) = P(z \geq 1.96) \approx 0.025$$

[8] Looking at the final step, we see that we need $np > 9$ in the worst case, when q (or p) is near 1, making the Binomial model quite skewed. When q and p are near 0.5—for example, between 0.4 and 0.6—the Binomial model is nearly symmetric, and $np > 5$ ought to be safe enough. Although we'll always check for 10 expected successes and failures, keep in mind that for values of p near 0.5, we can be somewhat more forgiving.

9.6 Other Continuous Random Variables

The Normal distribution differs from the probability distributions we saw in Chapter 8 because it doesn't specify probabilities for individual values, but rather, for intervals of values. When a random variable can take on any value in an interval, we can't model it using a discrete probability model and must use a continuous probability model instead. For any continuous random variable, the distribution of its probability can be shown with a curve. That curve is called the **probability density function (pdf),** usually denoted as $f(x)$. Technically, the curve we've been using to work with the Normal distribution is known as the Normal probability density function.

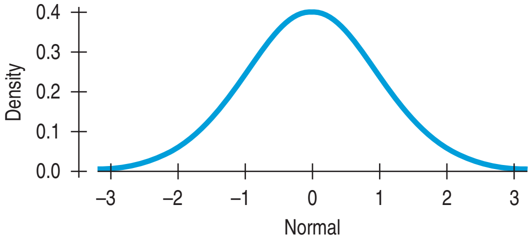


Figure 9.10 The standard Normal density function (a normal with mean 0 and standard deviation 1). The probability of finding a *z*-score in any interval is the area over that interval under the curve. For example, the probability that the *z*-score falls between −1 and 1 is about 68%, which can be seen approximately from the density function or found more precisely from a table or technology.

Density functions must satisfy two requirements. They must stay nonnegative for every possible value, and the total area under the curve must be exactly 1.0. This last requirement corresponds to the Probability Assignment Rule of Chapter 7, which said that the total probability (equal to 1.0) must be assigned somewhere.

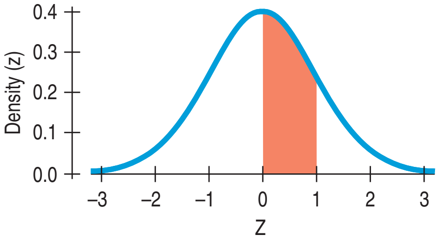
Any density function can give the probability that the random variable lies in an interval. But remember, the probability that X lies in the interval from a to b is the *area* under the density function, $f(x)$, between the values a and b and not the value $f(a)$ or $f(b)$. In general, finding that area requires calculus or numerical analysis, and is beyond the scope of this text. But for the models we'll discuss, the probabilities are found either from tables (the Normal) or simple computations (Uniform).

There are many (in fact, there are an infinite number of) possible continuous distributions, but we'll explore only three of the most commonly used to model business phenomena. In addition to the Normal distribution, we'll look at the Uniform distribution and the Exponential distribution.

How can *every* value have probability 0?

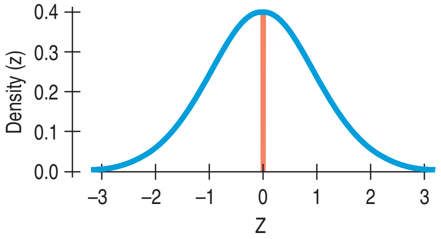
At first it may seem illogical that each value of a continuous random variable has probability 0. Let's look at the standard Normal random variable, Z. We could find (from a table, website, or computer program) that the probability that Z lies between 0 and 1 is 0.3413.

(continued)

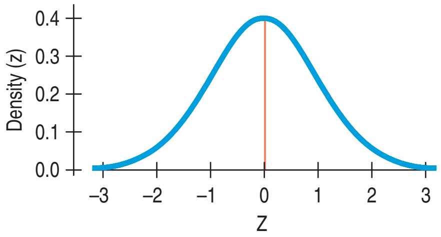


That's the area under the Normal pdf (in red) between the values 0 and 1.

So, what's the probability that Z is between 0 and 1/10?



That area is only 0.0398. What is the chance then that Z will fall between 0 and 1/100? There's not much area—the probability is only 0.0040. If we kept going, the probability would keep getting smaller. The probability that Z is between 0 and 1/100,000 is less than 0.0001.



So, what's the probability that Z is *exactly* 0? Well, there's *no* area under the curve right at $x = 0$, so the probability is 0. It's only intervals that have positive probability, but that's OK. In real life we never mean exactly 0.0000000000 or any other value. If you say "exactly 164 pounds," you might really mean between 163.5 and 164.5 pounds or even between 163.99 and 164.01 pounds, but realistically not 164.000000000 . . . pounds.

The Uniform Distribution

We've already seen the discrete version of the uniform probability model. A continuous uniform shares the principle that all events should be equally likely, but with a continuous distribution we can't talk about the probability of a particular value because each value has probability zero. Instead, for a continuous random variable X, we say that the probability that X lies in any interval depends only on the length of that interval. Not surprisingly the density function of a continuous uniform random variable looks flat (see Figure 9.11).

The density function of a continuous uniform random variable defined on the interval a to b can be defined by the formula (see Figure 9.11)

$$f(x) = \begin{cases} \dfrac{1}{b - a} & \text{if} \quad a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

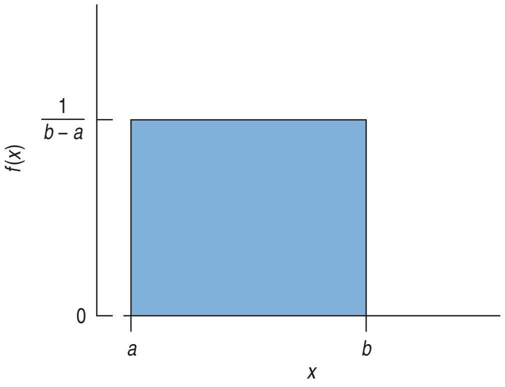


Figure 9.11 The density function of a continuous uniform random variable on the interval from *a* to *b*.

From Figure 9.11, it's easy to see that the probability that *X* lies in any interval between *a* and *b* is the same as any other interval of the same length. In fact, the probability is just the ratio of the length of the interval to the total length: $b - a$. In other words:

For values c and d ($c \le d$) *both within the interval* $[a, b]$:

$$P(c \le X \le d) = \frac{(d - c)}{(b - a)}$$

As an example, suppose you arrive at a bus stop and want to model how long you'll wait for the next bus. The sign says that busses arrive about every 20 minutes, but no other information is given. You might assume that the arrival is equally likely to be anywhere in the next 20 minutes, and so the density function would be

$$f(x) = \begin{cases} \frac{1}{20} & \textit{if} \quad 0 \le x \le 20 \\ 0 & \textit{otherwise} \end{cases}$$

and would look as shown in Figure 9.12.

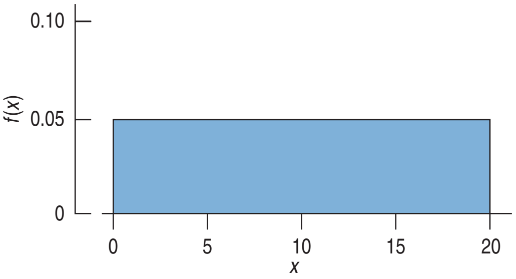


Figure 9.12 The density function of a continuous uniform random variable on the interval [0,20]. Notice that the mean (the balancing point) of the distribution is at 10 minutes and that the area of the box is 1.

Just as the mean of a data distribution is the balancing point of a histogram, the mean of any continuous random variable is the balancing point of the density function. Looking at Figure 9.12, we can see that the balancing point is halfway between the end points at 10 minutes. In general, the expected value is:

$$E(X) = \frac{a + b}{2}$$

for a uniform distribution on the interval (a, b). With $a = 0$ and $b = 20$, the expected value would be 10 minutes.

The variance and standard deviation are less intuitive:

$$Var(X) = \frac{(b-a)^2}{12}; SD(X) = \sqrt{\frac{(b-a)^2}{12}}.$$

Using these formulas, our bus wait will have an expected value of 10 minutes with a standard deviation of $\sqrt{\frac{(20-0)^2}{12}} = 5.77$ minutes.

Just Checking

2 As a group, the Dutch are among the tallest people in the world. The average Dutch man is 184 cm tall—just over 6 feet (and the average Dutch woman is 170.8 cm tall—just over 5′7″). If a Normal model is appropriate and the standard deviation for men is about 8 cm, what percentage of all Dutch men will be over 2 meters (6′6″) tall?

3 Suppose it takes you 20 minutes, on average, to drive to work, with a standard deviation of 2 minutes. Suppose a Normal model is appropriate for the distributions of driving times.

a) How often will you arrive at work in less than 22 minutes?

b) How often will it take you more than 24 minutes?

c) Do you think the distribution of your driving times is unimodal and symmetric?

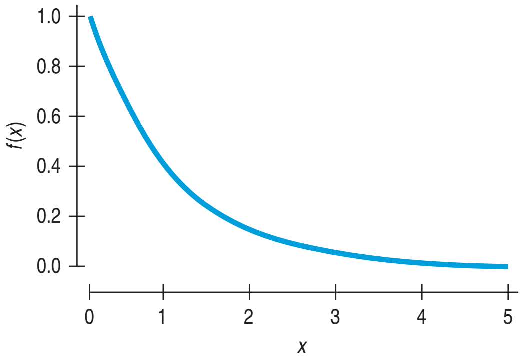
d) What does this say about the accuracy of your prediction? Explain.

The Exponential Model

We saw in Chapter 8 that the Poisson distribution is a good model for the arrival of, or occurrence, of events. We found, for example, the probability that x visits to our website will occur within the next minute. The exponential distribution with parameter λ can be used to model the time *between* those events. Its density function has the form:

$$f(x) = \lambda e^{-\lambda x} \quad for\ x \geq 0\ and\ \lambda > 0$$

The use of the parameter λ again is not coincidental. It highlights the relationship between the exponential and the Poisson.



| Figure 9.13 The exponential density function with $\lambda = 1$.

If a discrete random variable can be modeled by a Poisson model with rate λ, then the times between events can be modeled by an exponential model with the same parameter λ. The mean of the exponential is $1/\lambda$. The inverse relationship between the two means makes intuitive sense. If λ increases and we expect *more* hits per minute, then the expected time between hits should go down. The standard deviation of an exponential random variable is $1/\lambda$.

Like any continuous random variable, probabilities of an exponential random variable can be found only through the density function. Fortunately, the area under the exponential density between any two values, s and t ($s \leq t$), has a particularly easy form:

$$P(s \leq X \leq t) = e^{-\lambda s} - e^{-\lambda t}.$$

In particular, by setting s to be 0, we can find the probability that the waiting time will be less than t from

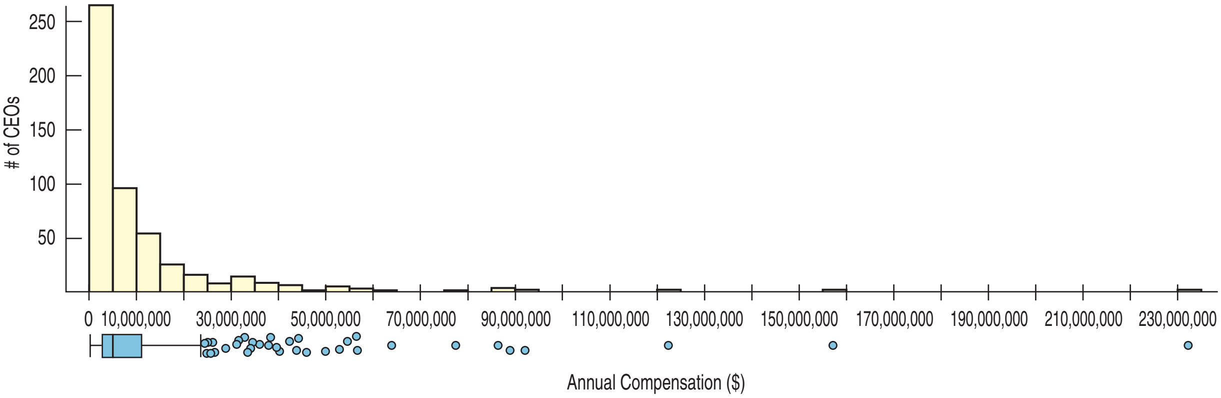
$$P(X \leq t) = P(0 \leq X \leq t) = e^{-\lambda 0} - e^{-\lambda t} = 1 - e^{-\lambda t}.$$

The function $P(X \leq t) = F(t)$ is called the **cumulative distribution function (cdf)** of the random variable X. If arrivals of hits to our website can be well modeled by a Poisson with $\lambda = 4$/minute, then the probability that we'll have to wait less than 20 seconds (1/3 of a minute) is $F(1/3) = P(0 \leq X \leq 1/3) = 1 - e^{-4/3} = 0.736$. That seems about right. Arrivals are coming about every 15 seconds on average, so we shouldn't be surprised that nearly 75% of the time we won't have to wait more than 20 seconds for the next hit.

What Can Go Wrong?

- **Probability models are still just models.** Models can be useful, but they are not reality. Think about the assumptions behind your models. Question probabilities as you would data.
- **Don't assume everything's Normal.** Just because a random variable is continuous or you happen to know a mean and standard deviation doesn't mean that a Normal model will be useful. You must think about whether the **Normality Assumption** is justified. Using a Normal model when it really does not apply will lead to wrong answers and misleading conclusions.

  A sample of CEOs has a mean total compensation of \$10,307,311.87 with a standard deviation of \$17,964,615.16. Using the Normal model rule, we should expect about 68% of the CEOs to have compensations between −\$7,657,303.29 and \$28,271,927.03. In fact, more than 90% of the CEOs have annual compensations in this range. What went wrong? The distribution is skewed, not symmetric. Using the 68–95–99.7 Rule for data like these will lead to silly results.



- **Don't use the Normal approximation with small *n*.** To use a Normal approximation in place of a Binomial model, there must be at least 10 expected successes and 10 expected failures.

Ethics in Action

Although e-government services are available online, many Americans, especially those who are older, prefer to deal with government agencies in person. For this reason, the U.S. Social Security Administration (SSA) has local offices distributed across the country. Pat Mennoza is the office manager for one of the larger SSA offices in Phoenix. Since the initiation of the SSA website, his staff has been severely reduced. Yet, because of the number of retirees in the area, his office is one of the busiest. Although there have been no formal complaints, Pat expects that customer waiting times have increased. He decides to keep track of customer wait times for a one-month period in the hopes of making a case for hiring additional staff. He finds that the average wait time is 5 minutes with a standard deviation of 6 minutes. He reasons that 50% of customers who visit his office wait longer than 5 minutes for service. The target wait time is 10 minutes or less. Applying the Normal probability model, Pat finds that more than 20% of customers will have to wait longer than 10 minutes! He has uncovered what he suspected. His next step is to request additional staff based on his findings.

ETHICAL ISSUE *Waiting times are generally skewed and therefore not usually modeled using the Normal distribution. Pat should have checked the data to see if a Normal model was appropriate. Using the Normal for data that are highly skewed to the right will inflate the probability a customer will have to wait longer than 10 minutes. Related to Item A, ASA Ethical Guidelines.*

ETHICAL SOLUTION *Check reasonableness of applying the Normal probability model.*

What Have We Learned?

Learning Objectives

- Recognize normally distributed data by making a histogram and checking whether it is unimodal, symmetric, and bell-shaped, or by making a normal probability plot using technology and checking whether the plot is roughly a straight line.
  - The Normal model is a distribution that will be important for much of the rest of this course.
  - Before using a Normal model, we should check that our data are plausibly from a normally distributed population.
  - A Normal probability plot provides evidence that the data are Normally distributed if it is linear.
- Understand how to use the Normal model to judge whether a value is extreme.
  - Standardize values to make z-scores and obtain a standard scale. Then refer to a standard Normal distribution.
  - Use the 68–95–99.7 Rule as a rule-of-thumb to judge whether a value is extreme.
- Know how to refer to tables or technology to find the probability of a value randomly selected from a Normal model falling in any interval.
  - Know how to perform calculations about Normally distributed values and probabilities.
- Recognize when independent random Normal quantities are being added or subtracted.
  - The sum or difference will also follow a Normal model
  - The *variance* of the sum or difference will be the sum of the individual variances.
  - The mean of the sum or difference will be the sum or difference, respectively, of the means.
- Recognize when other continuous probability distributions are appropriate models.

Terms

68–95–99.7 Rule (or Empirical Rule) In a Normal model, 68% of values fall within one standard deviation of the mean, 95% fall within two standard deviations of the mean, and 99.7% fall within three standard deviations of the mean. This is also approximately true for most unimodal, symmetric distributions.

Cumulative distribution function (cdf) A function for a continuous probability model that gives the probability of all values below a given value.

Exponential Distribution A continuous distribution appropriate for modeling the times between events whose occurrences follow a Poisson model.

Normal Distribution A unimodal, symmetric, “bell-shaped” distribution that appears throughout Statistics.

Normal probability plot A display to help assess whether a distribution of data is approximately Normal. If the plot is nearly straight, the data satisfy the Nearly Normal Condition.

Probability Density Function (pdf) A function for any continuous probability model that gives the probability of a random value falling between any two values as the area under the pdf between those two values.

Standard Normal model or Standard Normal distribution A Normal model, $N(\mu, \sigma)$ with mean $\mu = 0$ and standard deviation $\sigma = 1$.

Uniform Distribution A continuous distribution that assigns a probability to any range of values (between 0 and 1) proportional to the difference between the values.

The CAPE10 index is based on the Price/Earnings (P/E) ratios of stocks. We can examine the P/E ratios without applying the smoothing techniques used to find the CAPE10. The file **CAPE10** holds the data, giving dates, CAPE10 values, and P/E values.

Examine the P/E values. Would you judge that a Normal model would be appropriate for those values from the 1880s through the 1980s? Explain (and show the plots you made.)

Now consider the more recent P/E values in this context. Do you think they have been extreme? Explain.

Technology Help: Making Normal Probability Plots

The best way to tell whether your data can be modeled well by a Normal model is to make a picture or two. We've already talked about making histograms. Normal probability plots are almost never made by hand because the values of the Normal scores are tricky to find. But most statistics software make Normal plots, though various packages call the same plot by different names and array the information differently.

EXCEL

XLSTAT

Excel offers a "Normal probability plot" as part of the Regression command in the Data Analysis extension, but (as of this writing) it is not a correct Normal probability plot and should not be used.

JMP

To make a "Normal Quantile Plot" in JMP,

- Make a histogram using **Distributions** from the **Analyze** menu.
- Click on the drop-down menu next to the variable name.
- Choose **Normal Quantile Plot** from the drop-down menu.
- JMP opens the plot next to the histogram.

Comments

JMP places the ordered data on the vertical axis and the Normal scores on the horizontal axis. The vertical axis aligns with the histogram's axis, a useful feature.

MINITAB

To make a "Normal Probability Plot" in MINITAB,

- Choose **Probability Plot** from the **Graph** menu.
- Select "Single" for the type of plot. Click **OK.**
- Enter the name of the variable in the "Graph variables" box. Click **OK.**

Comments

MINITAB places the ordered data on the horizontal axis and the Normal scores on the vertical axis.

SPSS

To make a Normal "P-P plot" in SPSS,

- Choose **P-P** from the **Graphs** menu.
- Select the variable to be displayed in the source list.
- Click the arrow button to move the variable into the target list.
- Click the **OK** button.

Comments

SPSS places the ordered data on the horizontal axis and the Normal scores on the vertical axis. You may safely ignore the options in the P-P dialog.

Exercises

SECTION 9.1

1. An incoming MBA student took placement exams in economics and mathematics. In economics, she scored 82 and in math 86. The overall results on the economics exam had a mean of 72 and a standard deviation of 8, while the mean math score was 68, with a standard deviation of 12. On which exam did she do better compared with the other students?

2. The first Statistics exam had a mean of 65 and a standard deviation of 10 points; the second had a mean of 80 and a standard deviation of 5 points. Derrick scored an 80 on both tests. Julie scored a 70 on the first test and a 90 on the second. They both totaled 160 points on the two exams, but Julie claims that her total is better. Explain.

3. Your company's Human Resources department administers a test of "Executive Aptitude." They report test grades as z-scores, and you got a score of 2.20. What does this mean?

4. After examining a child at his 2-year checkup, the boy's pediatrician said that the z-score for his height relative to American 2-year-olds was -1.88. Write a sentence to explain to the parents what that means.

5. Your company will admit to the executive training program only people who score in the top 3% on the executive aptitude test discussed in Exercise 1.

a) With your z-score of 2.20, did you make the cut?

b) What do you need to assume about test scores to find your answer in part a?

6. The pediatrician in Exercise 4 explains to the parents that the most extreme 5% of cases often require special treatment or attention.

a) Does this child fall into that group?
b) What do you need to assume about the heights of 2-year-olds to find your answer to part a?

SECTION 9.2

7. The Environmental Protection Agency (EPA) fuel economy estimates for automobiles suggest a mean of 24.8 mpg and a standard deviation of 6.2 mpg for highway driving. Assume that a Normal model can be applied.

a) Draw the model for auto fuel economy. Clearly label it, showing what the 68–95–99.7 Rule predicts about miles per gallon.
b) In what interval would you expect the central 68% of autos to be found?
c) About what percent of autos should get more than 31 mpg?
d) About what percent of cars should get between 31 and 37.2 mpg?
e) Describe the gas mileage of the worst 2.5% of all cars.

8. Some IQ tests are standardized to a Normal model with a mean of 100 and a standard deviation of 16.

a) Draw the model for these IQ scores. Clearly label it, showing what the 68–95–99.7 Rule predicts about the scores.
b) In what interval would you expect the central 95% of IQ scores to be found?
c) About what percent of people should have IQ scores above 116?
d) About what percent of people should have IQ scores between 68 and 84?
e) About what percent of people should have IQ scores above 132?

9. What percent of a standard Normal model is found in each region? Be sure to draw a picture first.

a) $z > 1.5$
b) $z < 2.25$
c) $-1 < z < 1.15$
d) $|z| > 0.5$

10. What percent of a standard Normal model is found in each region? Draw a picture first.

a) $z > -2.05$
b) $z < -0.33$
c) $1.2 < z < 1.8$
d) $|z| < 1.28$

11. In a standard Normal model, what value(s) of z cut(s) off the region described? Don't forget to draw a picture.

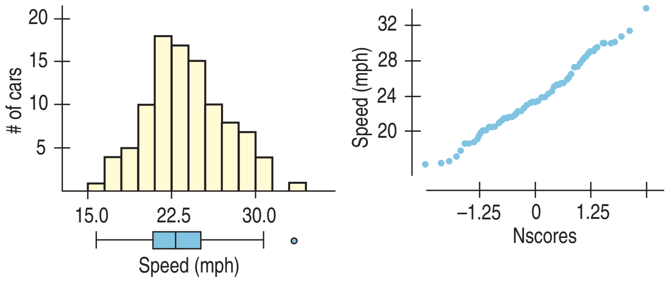
a) the highest 20%
b) the highest 75%
c) the lowest 3%
d) the middle 90%

12. In a standard Normal model, what value(s) of z cut(s) off the region described? Remember to draw a picture first.

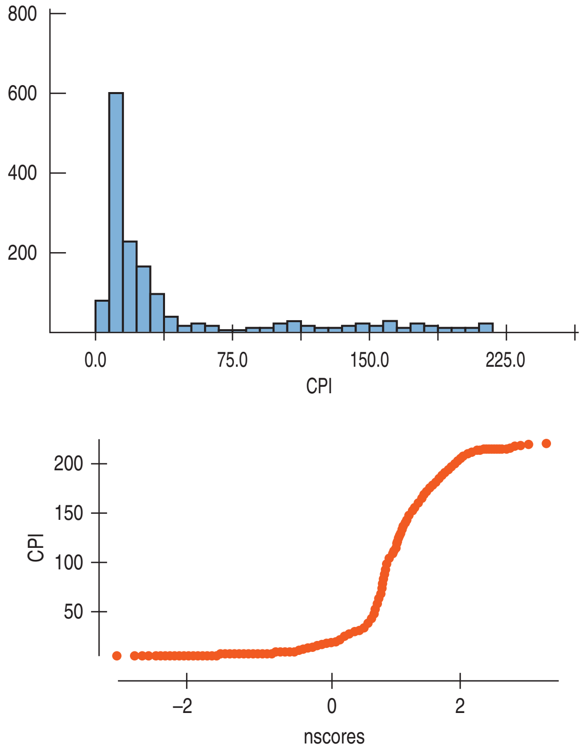
a) the lowest 12%
b) the highest 30%
c) the highest 7%
d) the middle 50%

SECTION 9.3

13. Speeds of cars were measured as they passed one point on a road to study whether traffic speed controls were needed. Here's a histogram and normal probability plot of the measured speeds. Is a Normal model appropriate for these data? Explain.



14. Has the Consumer Price Index (CPI) fluctuated around its mean according to a Normal model? Here are some displays. Is a Normal model appropriate for these data? Explain.



SECTION 9.4

15. For a new type of tire, a NASCAR team found the average distance a set of tires would run during a race is 168 miles, with a standard deviation of 14 miles. Assume that tire mileage is independent and follows a Normal model.

a) If the team plans to change tires twice during a 500-mile race, what is the expected value and standard deviation of miles remaining after two changes?
b) What is the probability they won't have to change tires a third time before the end of a 500-mile race?

16. In the 4 × 100 medley relay event, four swimmers swim 100 yards, each using a different stroke. A college team preparing for the conference championship looks at the times their swimmers have posted and creates a model based on the following assumptions:

- The swimmers' performances are independent.
- Each swimmer's times follow a Normal model.
- The means and standard deviations of the times (in seconds) are as shown here.

Swimmer	Mean	SD
1 (backstroke)	50.72	0.24
2 (breaststroke)	55.51	0.22
3 (butterfly)	49.43	0.25
4 (freestyle)	44.91	0.21

a) What are the mean and standard deviation for the relay team's total time in this event?
b) The team's best time so far this season was 3:19.48. (That's 199.48 seconds.) What is the probability that they will beat that time in the next event?

SECTION 9.5

17. Because many passengers who make reservations do not show up, airlines often overbook flights (sell more tickets than there are seats). A Boeing 767-400ER holds 245 passengers. If the airline believes the rate of passenger no-shows is 5% and sells 255 tickets, is it likely they will not have enough seats and someone will get bumped?

a) Use the Normal model to approximate the Binomial to determine the probability of at least 246 passengers showing up.
b) Should the airline change the number of tickets they sell for this flight? Explain.

18. Shortly after the introduction of the Belgian euro coin, newspapers around the world published articles claiming the coin is biased. The stories were based on reports that someone had spun the coin 250 times and gotten 140 heads—that's 56% heads.

a) Use the Normal model to approximate the Binomial to determine the probability of spinning a fair coin 250 times and getting at least 140 heads.
b) Do you think this is evidence that spinning a Belgian euro is unfair? Would you be willing to use it at the beginning of a sports event? Explain.

SECTION 9.6

19. A cable provider wants to contact customers in a particular telephone exchange to see how satisfied they are with the new digital TV service the company has provided. All numbers are in the 452 exchange, so there are 10,000 possible numbers from 452-0000 to 452-9999. If they select the numbers with equal probability:

a) What distribution would they use to model the selection?
b) The new business "incubator" was assigned the 200 numbers between 452-2500 and 452-2699, but these businesses don't subscribe to digital TV. What is the probability that the randomly selected number will be for an incubator business?
c) Numbers above 9000 were only released for domestic use last year, so they went to newly constructed residences. What is the probability that a randomly selected number will be one of these?

20. In an effort to check the quality of their cell phones, a manufacturing manager decides to take a random sample of 10 cell phones from yesterday's production run, which produced cell phones with serial numbers ranging (according to when they were produced) from 43005000 to 43005999. If each of the 1000 phones is equally likely to be selected:

a) What distribution would they use to model the selection?
b) What is the probability that a randomly selected cell phone will be one of the last 100 to be produced?
c) What is the probability that the first cell phone selected is either from the last 200 to be produced or from the first 50 to be produced?

CHAPTER EXERCISES

For Exercises 21–28, use the 68–95–99.7 Rule to approximate the probabilities rather than using technology to find the values more precisely. Answers given for probabilities or percentages from Exercise 29 and on assume that a calculator or software has been used. Answers found from using *Z*-tables may vary slightly.

T **21. Mutual fund returns.** In the last quarter of 2007, a group of 64 mutual funds had a mean return of 2.4% with a standard deviation of 5.6%. If a Normal model can be used to model them, what percent of the funds would you expect to be in each region?

Be sure to draw a picture first.

a) Returns of 8.0% or more
b) Returns of 2.4% or less
c) Returns between −8.8% and 13.6%
d) Returns of more than 19.2%

22. **Human resource testing.** Although controversial and the subject of some recent law suits (e.g., *Satchell et al. vs. FedEx Express*), some human resource departments administer standard IQ tests to all employees. The Stanford-Binet test scores are well modeled by a Normal model with mean 100 and standard deviation 16. If the applicant pool is well modeled by this distribution, a randomly selected applicant would have what probability of scoring in the following regions?

a) 100 or below
b) Above 148
c) Between 84 and 116
d) Above 132

23. **Mutual funds, again.** From the 64 mutual funds in Exercise 21 with quarterly returns that are well modeled by a Normal model with a mean of 2.4% and a standard deviation of 5.6%, find the cutoff return value(s) that would separate the

a) highest 50%.
b) highest 16%.
c) lowest 2.5%.
d) middle 68%.

24. **Human resource testing, again.** For the IQ test administered by human resources and discussed in Exercise 22, what cutoff value would separate the

a) lowest 0.15% of all applicants?
b) lowest 16%?
c) middle 95%?
d) highest 2.5%?

25. **Currency exchange rates.** The daily exchange rates for the five-year period 2003 to 2008 between the euro (EUR) and the British pound (GBP) are well modeled by a Normal distribution with mean 1.459 euros (to pounds) and standard deviation 0.033 euros. Given this model, what is the probability that on a randomly selected day during this period, the pound was worth

a) less than 1.459 euros?
b) more than 1.492 euros?
c) less than 1.393 euros?
d) Which would be more unusual, a day on which the pound was worth less than 1.410 euros or more than 1.542 euros?

26. **Stock prices.** For the 900 trading days from January 2003 through July 2006, the daily closing price of IBM stock (in $) is well modeled by a Normal model with mean $85.60 and standard deviation $6.20. According to this model, what is the probability that on a randomly selected day in this period the stock price closed

a) above $91.80?
b) below $98.00?
c) between $73.20 and $98.00?
d) Which would be more unusual, a day on which the stock price closed above $93 or below $70?

27. **Currency exchange rates, again.** For the model of the EUR/GBP exchange rate discussed in Exercise 25, what would the cutoff rates be that would separate the

a) highest 16% of EUR/GBP rates?
b) lowest 50%?
c) middle 95%?
d) lowest 2.5%?

28. **Stock prices, again.** According to the model in Exercise 26, what cutoff value of price would separate the

a) lowest 16% of the days?
b) highest 0.15%?
c) middle 68%?
d) highest 50%?

29. **Mutual fund probabilities.** According to the Normal model $N(0.024, 0.056)$ describing mutual fund returns in the 4th quarter of 2007 in Exercise 21, what percent of this group of funds would you expect to have return

a) over 6.8%?
b) between 0% and 7.6%?
c) more than 1%?
d) less than 0%?

30. **Normal IQs.** Based on the Normal model $N(100, 16)$ describing IQ scores from Exercise 22, what percent of applicants would you expect to have scores

a) over 80?
b) under 90?
c) between 112 and 132?
d) over 125?

31. **Mutual funds, once more.** Based on the model $N(0.024, 0.056)$ for quarterly returns from Exercise 21, what are the cutoff values for the

a) highest 10% of these funds?
b) lowest 20%?
c) middle 40%?
d) highest 80%?

32. **More IQs.** In the Normal model $N(100, 16)$ for IQ scores from Exercise 22, what cutoff value bounds the

a) highest 5% of all IQs?
b) lowest 30% of the IQs?
c) middle 80% of the IQs?
d) lowest 90% of all IQs?

33. Mutual funds, finis. Consider the Normal model $N(0.024, 0.056)$ for returns of mutual funds in Exercise 21 one last time.

a) What value represents the 40th percentile of these returns?
b) What value represents the 99th percentile?
c) What's the IQR of the quarterly returns for this group of funds?

34. IQs, finis. Consider the IQ model $N(100, 16)$ one last time.

a) What IQ represents the 15th percentile?
b) What IQ represents the 98th percentile?
c) What's the IQR of the IQs?

35. Parameters. Every Normal model is defined by its parameters, the mean and the standard deviation. For each model described here, find the missing parameter. As always, start by drawing a picture.

a) $\mu = 20$, 45% above 30; $\sigma = ?$
b) $\mu = 88$, 2% below 50; $\sigma = ?$
c) $\sigma = 5$, 80% below 100; $\mu = ?$
d) $\sigma = 15.6$, 10% above 17.2; $\mu = ?$

36. Parameters, again. Every Normal model is defined by its parameters, the mean and the standard deviation. For each model described here, find the missing parameter. Don't forget to draw a picture.

a) $\mu = 1250$, 35% below 1200; $\sigma = ?$
b) $\mu = 0.64$, 12% above 0.70; $\sigma = ?$
c) $\sigma = 0.5$, 90% above 10.0; $\mu = ?$
d) $\sigma = 220$, 3% below 202; $\mu = ?$

37. SAT or ACT? Each year thousands of high school students take either the SAT or ACT, standardized tests used in the college admissions process. Combined SAT scores can go as high as 1600, while the maximum ACT composite score is 36. Since the two exams use very different scales, comparisons of performance are difficult. (A convenient rule of thumb is $SAT = 40 \times ACT + 150$; that is, multiply an ACT score by 40 and add 150 points to estimate the equivalent SAT score.) Assume that one year the combined SAT can be modeled by $N(1000, 200)$ and the ACT can be modeled by $N(27, 3)$. If an applicant to a university has taken the SAT and scored 1260 and another student has taken the ACT and scored 33, compare these students scores using z-values. Which one has a higher relative score? Explain.

38. Economics. Anna, a business major, took final exams in both Microeconomics and Macroeconomics and scored 83 on both. Her roommate Megan, also taking both courses, scored 77 on the Micro exam and 95 on the Macro exam. Overall, student scores on the Micro exam had a mean of 81 and a standard deviation of 5, and the Macro scores had a mean of 74 and a standard deviation of 15. Which student's overall performance was better? Explain.

39. Claims. Two companies make batteries for cell phone manufacturers. One company claims a mean life span of 2 years, while the other company claims a mean life span of 2.5 years (assuming average use of minutes/month for the cell phone).

a) Explain why you would also like to know the standard deviations of the battery life spans before deciding which brand to buy.
b) Suppose those standard deviations are 1.5 months for the first company and 9 months for the second company. Does this change your opinion of the batteries? Explain.

T **40. Car speeds.** The police department of a major city needs to update its budget. For this purpose, they need to understand the variation in their fines collected from motorists for speeding. As a sample, they recorded the speeds of cars driving past a location with a 20 mph speed limit, a place that in the past has been known for producing fines. The mean of 100 readings was 23.84 mph, with a standard deviation of 3.56 mph. (The police actually recorded every car for a two-month period. These are 100 representative readings.)

a) How many standard deviations from the mean would a car going the speed limit be?
b) Which would be more unusual, a car traveling 34 mph or one going 10 mph?

41. CEOs. A business publication recently released a study on the total number of years of experience in industry among CEOs. The mean is provided in the article, but not the standard deviation. Is the standard deviation most likely to be 6 months, 6 years, or 16 years? Explain which standard deviation is correct and why.

42. Stocks. A newsletter for investors recently reported that the average stock price for a blue chip stock over the past 12 months was \$72. No standard deviation was given. Is the standard deviation more likely to be \$6, \$16, or \$60? Explain.

43. Web visitors. A website manager has noticed that during the evening hours, about 3 people per minute check out from their shopping cart and make an online purchase. She believes that each purchase is independent of the others.

a) What model might you suggest to model the number of purchases per minute?
b) What model would you use to model the time between events?
c) What is the mean time between purchases?
d) What is the probability that the time to the next purchase will be between 1 and 2 minutes?

44. Monitoring quality. A cell phone manufacturer samples cell phones from the assembly to test. She noticed that the number of faulty cell phones in a production run of cell phones is usually small and that the quality of one day's run seems to have no bearing on the next day.

a) What model might you use to model the number of faulty cell phones produced in one day?
She wants to model the time between the events of producing a faulty phone. The mean number of defective cell phones is 2 per day.
b) What model would you use to model the time between events?
c) What would the probability be that the time to the next failure is 1 day or less?
d) What is the mean time between failures?

45. Lefties. A lecture hall has 200 seats with folding arm tablets, 30 of which are designed for left-handers. The typical size of classes that meet there is 188, and we can assume that about 13% of students are left-handed. Use a Normal approximation to find the probability that a right-handed student in one of these classes is forced to use a lefty arm tablet.

46. Seatbelts. Police estimate that 80% of drivers wear their seatbelts. They set up a safety roadblock, stopping cars to check for seatbelt use. If they stop 120 cars, what's the probability they find at least 20 drivers not wearing their seatbelt? Use a Normal approximation.

47. Rickets. Vitamin D is essential for strong, healthy bones. Although the bone disease rickets was largely eliminated in England during the 1950s, some people there are concerned that this generation of children is at increased risk because they are more likely to watch TV or play computer games than spend time outdoors. Recent research indicated that about 20% of British children are deficient in vitamin D. A company that sells vitamin D supplements tests 320 elementary school children in one area of the country. Use a Normal approximation to find the probability that no more than 50 of them have vitamin D deficiency.

48. Tennis. A tennis player has taken a special course to improve her serving. She thinks that individual serves are independent of each other. She has been able to make a successful first serve 70% of the time. Use a Normal approximation to find the probability she'll make at least 65 of her first serves out of the 80 she serves in her next match if her success percentage has not changed.

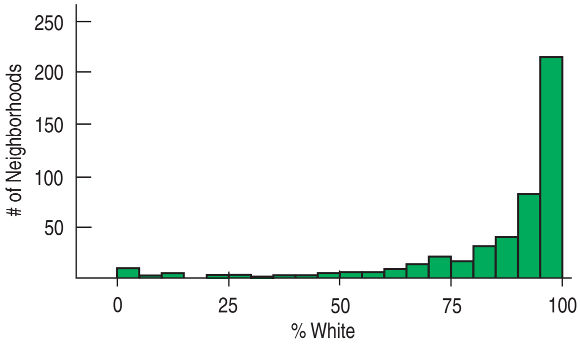
49. Low job satisfaction. Suppose that job satisfaction scores can be modeled with $N(100, 12)$. Human resource departments of corporations are generally concerned if the job satisfaction drops below a certain score. What score would you consider to be unusually low? Explain.

50. Low return. Exercise 21 proposes modeling quarterly returns of a group of mutual funds with $N(0.024, 0.056)$. The manager of this group of funds would like to flag any fund whose return is unusually low for a quarter. What level of return would you consider to be unusually low? Explain.

51. Management survey. A survey of 200 middle managers showed a distribution of the number of hours of exercise they participated in per week with a mean of 3.66 hours and a standard deviation of 4.93 hours.

a) According to the Normal model, what percent of managers will exercise fewer than one standard deviation below the mean number of hours?
b) For these data, what does that mean? Explain.
c) Explain the problem in using the Normal model for these data.

52. Customer database. A large philanthropic organization keeps records on the people who have contributed to their cause. In addition to keeping records of past giving, the organization buys demographic data on neighborhoods from the U.S. Census Bureau. Eighteen of these variables concern the ethnicity of the neighborhood of the donor. Here is a histogram and summary statistics for the percentage of whites in the neighborhoods of 500 donors.



Count	500
Mean	83.59
Median	93
StdDev	22.26
IQR	17
Q1	80
Q3	97

a) Which is a better summary of the percentage of white residents in the neighborhoods, the mean or the median? Explain.
b) Which is a better summary of the spread, the IQR or the standard deviation? Explain.
c) From a Normal model, about what percentage of neighborhoods should have a percent white residents within one standard deviation of the mean?
d) What percentage of neighborhoods actually have a percent white within one standard deviation of the mean?
e) Explain the problem in using the Normal model for these data.

53. Drug company. Manufacturing and selling drugs that claim to reduce an individual's cholesterol level is big business. A company would like to market their drug to women if their cholesterol is in the top 15%. Assume the cholesterol levels of adult American women can be described by a Normal model with a mean of 188 mg/dL and a standard deviation of 24.

a) Draw and label the Normal model.
b) What percent of adult women do you expect to have cholesterol levels over 200 mg/dL?
c) What percent of adult women do you expect to have cholesterol levels between 150 and 170 mg/dL?
d) Estimate the interquartile range of the cholesterol levels.
e) Above what value are the highest 15% of women's cholesterol levels?

54. Tire company. A tire manufacturer believes that the tread life of its snow tires can be described by a Normal model with a mean of 32,000 miles and a standard deviation of 2500 miles.

a) If you buy a set of these tires, would it be reasonable for you to hope that they'll last 40,000 miles? Explain.
b) Approximately what fraction of these tires can be expected to last less than 30,000 miles?
c) Approximately what fraction of these tires can be expected to last between 30,000 and 35,000 miles?
d) Estimate the IQR for these data.
e) In planning a marketing strategy, a local tire dealer wants to offer a refund to any customer whose tires fail to last a certain number of miles. However, the dealer does not want to take too big a risk. If the dealer is willing to give refunds to no more than 1 of every 25 customers, for what mileage can he guarantee these tires to last?

Just Checking Answers

1 a) On the first test, the mean is 88 and the SD is 4, so $z = (90 - 88)/4 = 0.5$. On the second test, the mean is 75 and the SD is 5, so $z = (80 - 75)/5 = 1.0$. The first test has the lower z-score, so it is the one that will be dropped.

b) The second test is 1 standard deviation above the mean, farther away than the first test, so it's the better score relative to the class.

2 The mean is 184 centimeters, with a standard deviation of 8 centimeters. 2 meters is 200 centimeters, which is 2 standard deviations above the mean. We expect 5% of the men to be more than 2 standard deviations below or above the mean, so half of those, 2.5%, are likely to be above 2 meters.

3 a) We know that 68% of the time we'll be within 1 standard deviation (2 min) of 20. So 32% of the time we'll arrive in less than 18 or more than 22 minutes. Half of those times (16%) will be greater than 22 minutes, so 84% will be less than 22 minutes.

b) 24 minutes is 2 standard deviations above the mean. Because of the 95% rule, we know 2.5% of the times will be more than 24 minutes.

c) Traffic incidents may occasionally increase the time it takes to get to school, so the driving times may be skewed to the right, and there may be outliers.

d) If so, the Normal model would not be appropriate and the percentages we predict would not be accurate.