CHAPTER 6

The Standard Deviation as a Ruler and the Normal Model

The women's heptathlon in the Olympics consists of seven track and field events: the 200-m and 800-m runs, 100-m high hurdles, shot put, javelin, high jump, and long jump. To determine who should get the gold medal, somehow the performances in all seven events have to be combined into one score. How can performances in such different events be compared? They don't even have the same units; the races are recorded in minutes and seconds and the throwing and jumping events in meters. In the 2004 Olympics, Austra Skujyté of Lithuania put the shot 16.4 meters, about 3 meters farther than the average of all contestants. Carolina Klüft won the long jump with a 6.78-m jump, about a meter better than the average. Which performance deserves more points? Even though both events are measured in meters, it's not clear how to compare them. The solution to the problem of how to compare scores turns out to be a useful method for comparing all sorts of values whether they have the same units or not.

The Standard Deviation as a Ruler

Grading on a Curve
If you score 79% on an exam, what grade should you get? One teaching philosophy looks only at the raw percentage, 79, and bases the grade on that alone. Another looks at your *relative* performance and bases the grade on how you did compared with the rest of the class. Teachers and students still debate which method is better.

The trick in comparing very different-looking values is to use standard deviations. The standard deviation tells us how the whole collection of values varies, so it's a natural ruler for comparing an individual value to the group. Over and over during this course, we will ask questions such as "How far is this value from the mean?" or "How different are these two statistics?" The answer in every case will be to measure the distance or difference in standard deviations.

The concept of the standard deviation as a ruler is not special to this course. You'll find statistical distances measured in standard deviations throughout Statistics, up to the most advanced levels.[1] This approach is one of the basic tools of statistical thinking.

[1] Other measures of spread could be used as well, but the standard deviation is the most common measure, and it is almost always used as the ruler.

In order to compare the two events, let's start with a picture. This time we'll use stem-and-leaf displays so we can see the individual distances.

Long Jump Stem	Long Jump Leaf	Shot Put Stem	Shot Put Leaf
67	8	16	4
66		15	
65	1	15	
64	2	14	56778
63	0566	14	24
62	11235	13	5789
61	0569	13	012234
60	2223	12	55
59	0278	12	0144
58	4	11	59
57	0	11	23

FIGURE 6.1

Stem-and-leaf displays for both the long jump and the shot put in the 2004 Olympic Heptathlon. Carolina Klüft (green scores) won the long jump, and Austra Skujyté (red scores) won the shot put. Which heptathlete did better for both events combined?

The two winning performances on the top of each stem-and-leaf display appear to be about the same distance from the center of the pack. But look again carefully. What do we mean by the *same distance*? The two displays have different scales. Each line in the stem-and-leaf for the shot put represents half a meter, but for the long jump each line is only a tenth of a meter. It's only because our eyes naturally adjust the scales and use the standard deviation as the ruler that we see each as being about the same distance from the center of the data. How can we make this hunch more precise? Let's see how many standard deviations each performance is from the mean.

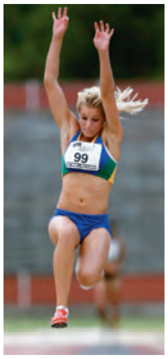
	Event: Long Jump	Event: Shot Put
Mean (all contestants)	6.16 m	13.29 m
SD	0.23 m	1.24 m
n	26	28
Klüft	6.78 m	14.77 m
Skujyté	6.30 m	16.40 m

Klüft's 6.78-m long jump is 0.62 meters longer than the mean jump of 6.16 m. How many *standard deviations* better than the mean is that? The standard deviation for this event was 0.23 m, so her jump was $(6.78 - 6.16)/0.23 = 0.62/0.23 = 2.70$ *standard deviations better* than the mean. Skujyté's winning shot put was $16.40 - 13.29 = 3.11$ meters longer than the mean shot put distance, and that's $3.11/1.24 = 2.51$ standard deviations better than the mean. That's a great performance but not quite as impressive as Klüft's long jump, which was farther above the mean, as measured in *standard deviations.*

Standardizing with *z*-Scores

NOTATION ALERT:

There goes another letter. We always use the letter z to denote values that have been standardized with the mean and standard deviation.



To compare these athletes' performances, we determined how many standard deviations from the event's mean each was.

Expressing the distance in standard deviations ***standardizes*** the performances. To standardize a value, we simply subtract the mean performance in that event and then divide this difference by the standard deviation. We can write the calculation as

$$z = \frac{y - \bar{y}}{s}.$$

These values are called **standardized values,** and are commonly denoted with the letter z. Usually, we just call them ***z*-scores.**

Standardized values have *no units*. z-scores measure the distance of each data value from the mean in standard deviations. A z-score of 2 tells us that a data value is 2 standard deviations above the mean. It doesn't matter whether the original variable was measured in inches, dollars, or seconds. Data values below the mean have negative z-scores, so a z-score of -1.6 means that the data value was 1.6 standard deviations below the mean. Of course, regardless of the direction, the farther a data value is from the mean, the more unusual it is, so a z-score of -1.3

is more extraordinary than a z-score of 1.2. Looking at the z-scores, we can see that even though both were winning scores, Klüft's long jump with a z-score of 2.70 is slightly more impressive than Skujyté's shot put with a z-score of 2.51.

FOR EXAMPLE **Standardizing skiing times**

The men's combined skiing event in the winter Olympics consists of two races: a downhill and a slalom. Times for the two events are added together, and the skier with the lowest total time wins. In the 2006 Winter Olympics, the mean slalom time was 94.2714 seconds with a standard deviation of 5.2844 seconds. The mean downhill time was 101.807 seconds with a standard deviation of 1.8356 seconds. Ted Ligety of the United States, who won the gold medal with a combined time of 189.35 seconds, skied the slalom in 87.93 seconds and the downhill in 101.42 seconds.

Question: On which race did he do better compared with the competition?

For the slalom, Ligety's z-score is found by subtracting the mean time from his time and then dividing by the standard deviation:

$$z_{Slalom} = \frac{87.93 - 94.2714}{5.2844} = -1.2$$

Similarly, his z-score for the downhill is:

$$z_{Downhill} = \frac{101.42 - 101.807}{1.8356} = -0.21$$

The z-scores show that Ligety's time in the slalom is farther below the mean than his time in the downhill. His performance in the slalom was more remarkable.

By using the standard deviation as a ruler to measure statistical distance from the mean, we can compare values that are measured on different variables, with different scales, with different units, or for different individuals. To determine the winner of the heptathlon, the judges must combine performances on seven very different events. Because they want the score to be absolute, and *not* dependent on the particular athletes in each Olympics, they use predetermined tables, but they could combine scores by standardizing each, and then adding the z-scores together to reach a total score. The only trick is that they'd have to switch the sign of the z-score for running events, because unlike throwing and jumping, it's better to have a running time below the mean (with a negative z-score).

To combine the scores Skujyté and Klüft earned in the long jump and the shot put, we standardize both events as shown in the table. That gives Klüft her 2.70 z-score in the long jump and a 1.19 in the shot put, for a total of 3.89. Skujyté's shot put gave her a 2.51, but her long jump was only 0.61 SDs above the mean, so her total is 3.12.

		Event: Long Jump	Event: Shot Put
	Mean	6.16 m	13.29 m
	SD	0.23 m	1.24 m
Klüft	Performance	6.78 m	14.77 m
	z-score	$\frac{6.78 - 6.16}{0.23} = 2.70$	$\frac{14.77 - 13.29}{1.24} = 1.19$
	Total z-score	$2.70 + 1.19 = \mathbf{3.89}$	
Skujyté	Performance	6.30 m	16.40 m
	z-score	$\frac{6.30 - 6.16}{0.23} = 0.61$	$\frac{16.40 - 13.29}{1.24} = 2.51$
	Total z-score	$0.61 + 2.51 = \mathbf{3.12}$	

Is this the result we wanted? Yes. Each won one event, but Klüft's shot put was second best, while Skujyté's long jump was seventh. The z-scores measure how far each result is from the event mean in standard deviation units. And because they are both in standard deviation units, we can combine them. Not coincidentally, Klüft went on to win the gold medal for the entire seven-event heptathlon, while Skujyté got the silver.

FOR EXAMPLE Combining z-scores

In the 2006 winter Olympics men's combined event, Ivica Kostelić of Croatia skied the slalom in 89.44 seconds and the downhill in 100.44 seconds. He thus beat Ted Ligety in the downhill, but not in the slalom. Maybe he should have won the gold medal.

Question: Considered in terms of standardized scores, which skier did better?

Kostelić's z-scores are:

$$z_{Slalom} = \frac{89.44 - 94.2714}{5.2844} = -0.91 \quad \text{and} \quad z_{Downhill} = \frac{100.44 - 101.807}{1.8356} = -0.74$$

The sum of his z-scores is approximately –1.65. Ligety's z-score sum is only about –1.41. Because the standard deviation of the downhill times is so much smaller, Kostelić's better performance there means that he would have won the event if standardized scores were used.

When we standardize data to get a z-score, we do two things. First, we shift the data by subtracting the mean. Then, we rescale the values by dividing by their standard deviation. We often shift and rescale data. What happens to a grade distribution if *everyone* gets a five-point bonus? Everyone's grade goes up, but does the shape change? (*Hint:* Has anyone's distance from the mean changed?) If we switch from feet to meters, what happens to the distribution of heights of students in your class? Even though your intuition probably tells you the answers to these questions, we need to look at exactly how shifting and rescaling work.

JUST CHECKING

1. Your Statistics 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"?

Shifting Data

Since the 1960s, the Centers for Disease Control's National Center for Health Statistics has been collecting health and nutritional information on people of all ages and backgrounds. A recent survey, the National Health and Nutrition Examination Survey (NHANES) 2001–2002,[2] measured a wide variety of variables, including body measurements, cardiovascular fitness, blood chemistry, and demographic information on more than 11,000 individuals.

[2] www.cdc.gov/nchs/nhanes.htm

WHO	80 male participants of the NHANES survey between the ages of 19 and 24 who measured between 68 and 70 inches tall
WHAT	Their weights
UNIT	Kilograms
WHEN	2001–2002
WHERE	United States
WHY	To study nutrition, and health issues and trends
HOW	National survey

Included in this group were 80 men between 19 and 24 years old of average height (between 5′8″ and 5′10″ tall). Here are a histogram and boxplot of their weights:

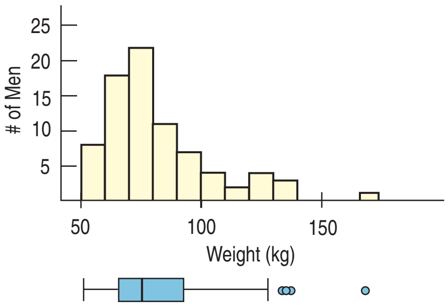


FIGURE 6.2

Histogram and boxplot for the men's weights. The shape is skewed to the right with several high outliers.

Their mean weight is 82.36 kg. For this age and height group, the National Institutes of Health recommends a maximum healthy weight of 74 kg, but we can see that some of the men are heavier than the recommended weight. To compare their weights to the recommended maximum, we could subtract 74 kg from each of their weights. What would that do to the center, shape, and spread of the histogram? Here's the picture:

A S ***Activity:* Changing the Baseline.** What happens when we shift data? Do measures of center and spread change?

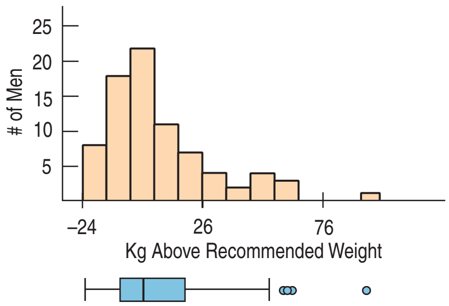


FIGURE 6.3

Subtracting 74 kilograms shifts the entire histogram down but leaves the spread and the shape exactly the same.

Doctors' height and weight charts sometimes give ideal weights for various heights that include 2-inch heels. If the mean height of adult women is 66 inches including 2-inch heels, what is the mean height of women without shoes? Each woman is shorter by 2 inches when barefoot, so the mean is decreased by 2 inches, to 64 inches.

On average, they weigh 82.36 kg, so on average they're 8.36 kg overweight. And, after subtracting 74 from each weight, the mean of the new distribution is $82.36 - 74 = 8.36$ kg. In fact, when we **shift** the data by adding (or subtracting) a constant to each value, all measures of position (center, percentiles, min, max) will increase (or decrease) by the same constant.

What about the spread? What does adding or subtracting a constant value do to the spread of the distribution? Look at the two histograms again. Adding or subtracting a constant changes each data value equally, so the entire distribution just shifts. Its shape doesn't change and neither does the spread. None of the measures of spread we've discussed—not the range, not the IQR, not the standard deviation—changes.

Adding (or subtracting) a constant to every data value adds (or subtracts) the same constant to measures of position, but leaves measures of spread unchanged.

Rescaling Data

Not everyone thinks naturally in metric units. Suppose we want to look at the weights in pounds instead. We'd have to **rescale** the data. Because there are about 2.2 pounds in every kilogram, we'd convert the weights by multiplying each value by 2.2. Multiplying or dividing each value by a constant changes the measurement units. Here are histograms of the two weight distributions, plotted on the same scale, so you can see the effect of multiplying:

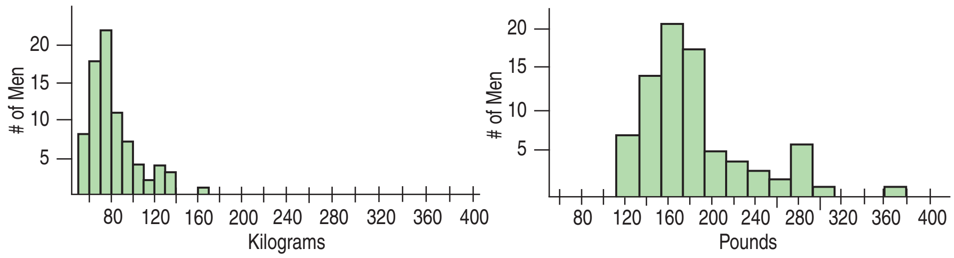


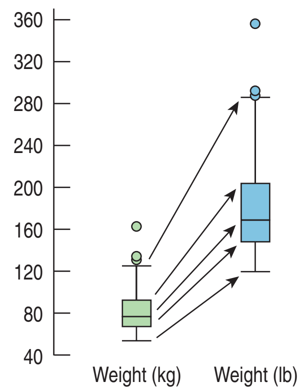
FIGURE 6.4
Men's weights in both kilograms and pounds. How do the distributions and numerical summaries change?

A S ***Simulation:*** **Changing the Units.** Change the center and spread values for a distribution and watch the summaries change (or not, as the case may be).

What happens to the shape of the distribution? Although the histograms don't look exactly alike, we see that the shape really hasn't changed: Both are unimodal and skewed to the right.

What happens to the mean? Not too surprisingly, it gets multiplied by 2.2 as well. The men weigh 82.36 kg on average, which is 181.19 pounds. As the boxplots and 5-number summaries show, all measures of position act the same way. They all get multiplied by this same constant.

What happens to the spread? Take a look at the boxplots. The spread in pounds (on the right) is larger. How much larger? If you guessed 2.2 times, you've figured out how measures of spread get rescaled.



	Weight (kg)	Weight (lb)
Min	54.3	119.46
Q1	67.3	148.06
Median	76.85	169.07
Q3	92.3	203.06
Max	161.5	355.30
IQR	25	55
SD	22.27	48.99

FIGURE 6.5
The boxplots (drawn on the same scale) show the weights measured in kilograms (on the left) and pounds (on the right). Because 1 kg is 2.2 lb, all the points in the right box are 2.2 times larger than the corresponding points in the left box. So each measure of position and spread is 2.2 times as large when measured in pounds rather than kilograms.

When we multiply (or divide) all the data values by any constant, all measures of position (such as the mean, median, and percentiles) and measures of spread (such as the range, the IQR, and the standard deviation) are multiplied (or divided) by that same constant.

FOR EXAMPLE Rescaling the slalom

Recap: The times in the men's combined event at the winter Olympics are reported in minutes and seconds. Previously, we converted these to seconds and found the mean and standard deviation of the slalom times to be 94.2714 seconds and 5.2844 seconds, respectively.

Question: Suppose instead that we had reported the times in minutes—that is, that each individual time was divided by 60. What would the resulting mean and standard deviation be?

Dividing all the times by 60 would divide both the mean and the standard deviation by 60:

$$\text{Mean} = 94.2714/60 = 1.5712 \text{ minutes}; \quad SD = 5.2844/60 = 0.0881 \text{ minutes}.$$

JUST CHECKING

2. In 1995 the Educational Testing Service (ETS) adjusted the scores of SAT tests. Before ETS recentered the SAT Verbal test, the mean of all test scores was 450.

- **a)** How would adding 50 points to each score affect the mean?
- **b)** The standard deviation was 100 points. What would the standard deviation be after adding 50 points?
- **c)** Suppose we drew boxplots of test takers' scores a year before and a year after the recentering. How would the boxplots of the two years differ?

3. A company manufactures wheels for in-line skates. The diameters of the wheels have a mean of 3 inches and a standard deviation of 0.1 inches. Because so many of their customers use the metric system, the company decided to report their production statistics in millimeters (1 inch = 25.4 mm). They report that the standard deviation is now 2.54 mm. A corporate executive is worried about this increase in variation. Should he be concerned? Explain.

Back to *z*-scores

A S ***Activity:*** **Standardizing.** What if we both shift and rescale? The result is so nice that we give it a name.

Standardizing data into *z*-scores is just shifting them by the mean and rescaling them by the standard deviation. Now we can see how standardizing affects the distribution. When we subtract the mean of the data from every data value, we shift the mean to zero. As we have seen, such a shift doesn't change the standard deviation.

When we *divide* each of these shifted values by *s*, however, the standard deviation should be divided by *s* as well. Since the standard deviation was *s* to start with, the new standard deviation becomes 1.

How, then, does standardizing affect the distribution of a variable? Let's consider the three aspects of a distribution: the shape, center, and spread.

z-scores have mean 0 and standard deviation 1.

- *Standardizing into z-scores does not change the **shape** of the distribution of a variable.*
- *Standardizing into z-scores changes the **center** by making the mean 0.*
- *Standardizing into z-scores changes the **spread** by making the standard deviation 1.*

STEP-BY-STEP EXAMPLE — Working with Standardized Variables

Many colleges and universities require applicants to submit scores on standardized tests such as the SAT Writing, Math, and Critical Reading (Verbal) tests. The college your little sister wants to apply to says that while there is no minimum score required, the middle 50% of their students have combined SAT scores between 1530 and 1850. You'd feel confident if you knew her score was in their top 25%, but unfortunately she took the ACT test, an alternative standardized test.

Question: How high does her ACT need to be to make it into the top quarter of equivalent SAT scores?

To answer that question you'll have to standardize all the scores, so you'll need to know the mean and standard deviations of scores for some group on both tests. The college doesn't report the mean or standard deviation for their applicants on either test, so we'll use the group of all test takers nationally. For college-bound seniors, the average combined SAT score is about 1500 and the standard deviation is about 250 points. For the same group, the ACT average is 20.8 with a standard deviation of 4.8.

THINK	**Plan** State what you want to find out. **Variables** Identify the variables and report the W's (if known). Check the appropriate conditions.	I want to know what ACT score corresponds to the upper-quartile SAT score. I know the mean and standard deviation for both the SAT and ACT scores based on all test takers, but I have no individual data values. ✔ **Quantitative Data Condition:** Scores for both tests are quantitative but have no meaningful units other than points.
SHOW	**Mechanics** Standardize the variables. The y-value we seek is z standard deviations above the mean.	The middle 50% of SAT scores at this college fall between 1530 and 1850 points. To be in the top quarter, my sister would have to have a score of at least 1850. That's a z-score of $z = \frac{(1850 - 1500)}{250} = 1.40$ So an SAT score of 1850 is 1.40 standard deviations above the mean of all test takers. For the ACT, 1.40 standard deviations above the mean is $20.8 + 1.40(4.8) = 27.52$.
TELL	**Conclusion** Interpret your results in context.	To be in the top quarter of applicants in terms of combined SAT score, she'd need to have an ACT score of at least 27.52.

When Is a *z*-score BIG?

A z-score gives us an indication of how unusual a value is because it tells us how far it is from the mean. If the data value sits right at the mean, it's not very far at all and its z-score is 0. A z-score of 1 tells us that the data value is 1 standard deviation above the mean, while a z-score of -1 tells us that the value is 1 standard deviation below the mean. How far from 0 does a z-score have to be to be interesting or unusual? There is no universal standard, but the larger the score is (negative or positive), the more unusual it is. We know that 50% of the data lie between the quartiles. For symmetric data, the standard deviation is usually a bit smaller than the IQR, and it's not uncommon for at least half of the data to have z-scores between -1 and 1. But no matter what the shape of the distribution, a z-score of 3 (plus or minus) or more is rare, and a z-score of 6 or 7 shouts out for attention.

To say more about how big we expect a z-score to be, we need to *model* the data's distribution. 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

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 simplify our lives.

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

—George Box, famous statistician

reality exactly. But it can still be useful. Like a physical model, it's something we can look at and manipulate in order 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 models. **Normal models** are appropriate for distributions whose shapes are unimodal and roughly symmetric. For these distributions, they provide a measure of how extreme a *z*-score is. Fortunately, there is a Normal model for every possible combination of mean and standard deviation. We write $N(\mu, \sigma)$ to represent a Normal model with a mean of μ and a standard deviation of σ. Why the Greek? Well, *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 model. Such numbers are called **parameters** of the model.

NOTATION ALERT:

$N(\mu, \sigma)$ always denotes a Normal model. 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.

We don't want to confuse the parameters with summaries of the data such as $\bar{y}$ and *s*, so we use special symbols. In Statistics, we almost always use Greek letters for parameters. By contrast, summaries of data are called **statistics** and are usually written with Latin letters.

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

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

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

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.

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.

When we use the Normal model, we assume that the distribution of the data is, well, Normal. Practically speaking, there's no way to check whether this **Normality Assumption** is true. In fact, it almost certainly is not true. Real data don't behave like mathematical models. Models are idealized; real data are real. The good news, however, is that to use a Normal model, it's sufficient to check the following condition:

A S *Activity:* **Working with Normal Models.** Learn more about the Normal model and see what data drawn at random from a Normal model might look like.

> **Nearly Normal Condition.** The shape of the data's distribution is unimodal and symmetric. Check this by making a histogram (or a Normal probability plot, which we'll explain later).

Don't model data with a Normal model without checking whether the condition is satisfied.

All models make **assumptions.** Whenever we model—and we'll do that often—we'll be careful to point out the assumptions that we're making. And, what's even more important, we'll check the associated **conditions** in the data to make sure that those assumptions are reasonable.

[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.

The 68–95–99.7 Rule

One in a Million

These magic 68, 95, 99.7 values come from the Normal model. As a model, it can give us corresponding values for any *z*-score. For example, it tells us that fewer than 1 out of a million values have *z*-scores smaller than −5.0 or larger than +5.0. So if someone tells you you're "one in a million," they must really admire your *z*-score.

TI-*nspire*

The 68–95–99.7 Rule. See it work for yourself.

Normal models give us an idea of how extreme a value is by telling us how likely it is to find one that far from the mean. We'll soon show how to find these numbers precisely—but one simple rule is usually all we need.

It turns out that in a Normal model, about 68% of the values fall within 1 standard deviation of the mean, about 95% of the values fall within 2 standard deviations of the mean, and about 99.7%—almost all—of the values fall within 3 standard deviations of the mean. These facts are summarized in a rule that we call (let's see . . .) the **68–95–99.7 Rule.**[4]

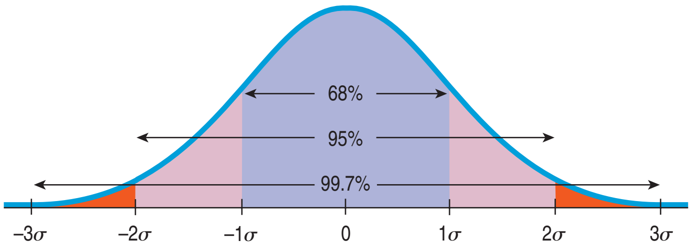


Figure 6.6

Reaching out one, two, and three standard deviations on a Normal model gives the 68–95–99.7 Rule, seen as proportions of the area under the curve.

FOR EXAMPLE Using the 68–95–99.7 Rule

Question: In the 2006 Winter Olympics men's combined event, Jean-Baptiste Grange of France skied the slalom in 88.46 seconds—about 1 standard deviation faster than the mean. If a Normal model is useful in describing slalom times, about how many of the 35 skiers finishing the event would you expect skied the slalom *faster* than Jean-Baptiste?

From the 68–95–99.7 Rule, we expect 68% of the skiers to be within one standard deviation of the mean. Of the remaining 32%, we expect half on the high end and half on the low end. 16% of 35 is 5.6, so, conservatively, we'd expect about 5 skiers to do better than Jean-Baptiste.

JUST CHECKING

4. 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?

5. Suppose it takes you 20 minutes, on average, to drive to school, 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 school 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 predictions? Explain.

[4] This rule is also called the "Empirical Rule" because it originally came from observation. The rule was first published by Abraham de Moivre in 1733, 75 years before the Normal model was discovered. Maybe it should be called "de Moivre's Rule," but that wouldn't help us remember the important numbers, 68, 95, and 99.7.

The First Three Rules for Working with Normal Models

A S ***Activity:* Working with Normal Models.** Well, actually playing with them. This interactive tool lets you do what this chapter's figures can't do, move them!

1. Make a picture.
2. Make a picture.
3. Make a picture.

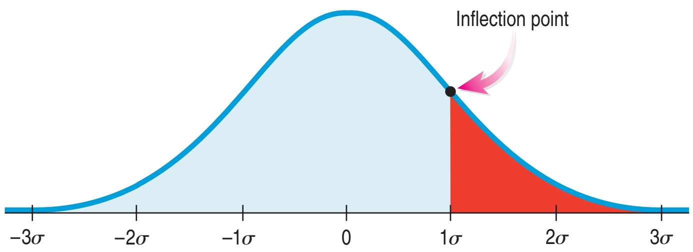
Although we're thinking about models, not histograms of data, the three rules don't change. To help you think clearly, a simple hand-drawn sketch is all you need. Even experienced statisticians sketch pictures to help them think about Normal models. You should too.

Of course, when we have data, we'll also need to make a histogram to check the **Nearly Normal Condition** to be sure we can use the Normal model to model the data's distribution. Other times, we may be told that a Normal model is appropriate based on prior knowledge of the situation or on theoretical considerations.

A S ***Activity:* Normal Models.** Normal models have several interesting properties—see them here.

How to Sketch a Normal Curve That Looks Normal To sketch a good Normal curve, you need to remember only three things:

- The Normal curve is bell-shaped and symmetric around its mean. Start at the middle, and sketch to the right and left from there.
- Even though the Normal model extends forever on either side, you need to draw it only for 3 standard deviations. After that, there's so little left that it isn't worth sketching.
- The place where the bell shape changes from curving downward to curving back up—the *inflection point*—is exactly one standard deviation away from the mean.

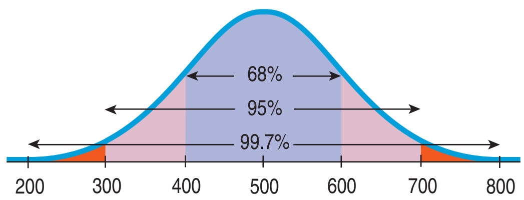


STEP-BY-STEP EXAMPLE **Working with the 68–95–99.7 Rule**

The SAT Reasoning Test has three parts: Writing, Math, and Critical Reading (Verbal). Each part has a distribution that is roughly unimodal and symmetric and is designed to have an overall mean of about 500 and a standard deviation of 100 for all test takers. In any one year, the mean and standard deviation may differ from these target values by a small amount, but they are a good overall approximation.

Question: Suppose you earned a 600 on one part of your SAT. Where do you stand among all students who took that test?

You could calculate your z-score and find out that it's $z = (600 - 500)/100 = 1.0$, but what does that tell you about your percentile? You'll need the Normal model and the 68–95–99.7 Rule to answer that question.

THINK **Plan** State what you want to know.	I want to see how my SAT score compares with the scores of all other students. To do that, I'll need to model the distribution.
Variables Identify the variable and report the W's.	Let y = my SAT score. Scores are quantitative but have no meaningful units other than points.
Be sure to check the appropriate conditions.	✓ **Nearly Normal Condition:** If I had data, I would check the histogram. I have no data, but I am told that the SAT scores are roughly unimodal and symmetric.
Specify the parameters of your model.	I will model SAT score with a *N*(500, 100) model.
SHOW **Mechanics** Make a picture of this Normal model. (A simple sketch is all you need.)	
Locate your score.	My score of 600 is 1 standard deviation above the mean. That corresponds to one of the points of the 68–95–99.7 Rule.
TELL **Conclusion** Interpret your result in context.	About 68% of those who took the test had scores that fell no more than 1 standard deviation from the mean, so 100% − 68% = 32% of all students had scores more than 1 standard deviation away. Only half of those were on the high side, so about 16% (half of 32%) of the test scores were better than mine. My score of 600 is higher than about 84% of all scores on this test.

The bounds of SAT scoring at 200 and 800 can also be explained by the 68–95–99.7 Rule. Since 200 and 800 are three standard deviations from 500, it hardly pays to extend the scoring any farther on either side. We'd get more information only on 100 − 99.7 = 0.3% of students.

The Worst-Case Scenario* Suppose we encounter an observation that's 5 standard deviations above the mean. Should we be surprised? We've just seen that when a Normal model is appropriate, such a value is exceptionally rare. After all, 99.7% of all the values should be within 3 standard deviations of the mean, so anything farther away would be unusual indeed.

But our handy 68–95–99.7 Rule applies only to Normal models, and the Normal is such a *nice* shape. What if we're dealing with a distribution that's strongly skewed (like the CEO salaries), or one that is uniform or bimodal or something really strange? A Normal model has 68% of its observations within one standard deviation of the mean, but a bimodal distribution could even be entirely empty in the middle. In that case could we still say anything at all about an observation 5 standard deviations above the mean?

Remarkably, even with really weird distributions, the worst case can't get all that bad. A Russian mathematician named Pafnuty Tchebycheff[5] answered the question by proving this theorem:

In any distribution, at least $1 - \frac{1}{k^2}$ *of the values must lie within* $\pm k$ *standard deviations of the mean.*

What does that mean?

- For $k = 1$, $1 - \frac{1}{1^2} = 0$; if the distribution is far from Normal, it's possible that none of the values are within 1 standard deviation of the mean. We should be really cautious about saying anything about 68% unless we think a Normal model is justified. (Tchebycheff's theorem really is about the worst case; it tells us nothing about the middle; only about the extremes.)
- For $k = 2$, $1 - \frac{1}{2^2} = \frac{3}{4}$; no matter how strange the shape of the distribution, at least 75% of the values must be within 2 standard deviations of the mean. Normal models may expect 95% in that 2-standard-deviation interval, but even in a worst-case scenario it can never go lower than 75%.
- For $k = 3$, $1 - \frac{1}{3^2} = \frac{8}{9}$; in any distribution, at least 89% of the values lie within 3 standard deviations of the mean.

What we see is that values beyond 3 standard deviations from the mean are uncommon, Normal model or not. Tchebycheff tells us that at least 96% of all values must be within 5 standard deviations of the mean. While we can't always apply the 68–95–99.7 Rule, we can be sure that the observation we encountered 5 standard deviations above the mean is unusual.

Finding Normal Percentiles

A S ***Activity:*** **Your Pulse z-Score.** Is your pulse rate high or low? Find its *z*-score with the Normal Model Tool.

An SAT score of 600 is easy to assess, because we can think of it as one standard deviation above the mean. If your score was 680, though, where do you stand among the rest of the people tested? Your *z*-score is 1.80, so you're somewhere between 1 and 2 standard deviations above the mean. We figured out that no more than 16% of people score better than 600. By the same logic, no more than 2.5% of people score better than 700. Can we be more specific than "between 16% and 2.5%"?

When the value doesn't fall exactly 1, 2, or 3 standard deviations from the mean, we can look it up in a table of **Normal percentiles** or use technology.[6] Either way, we first convert our data to *z*-scores before using the table. Your SAT score of 680 has a *z*-score of $(680 - 500)/100 = 1.80$.

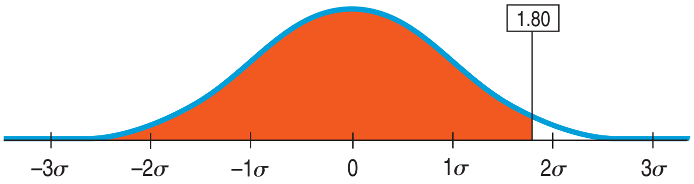
A S ***Activity:*** **The Normal Table.** Table Z just sits there, but this version of the Normal table changes so it always Makes a Picture that fits.

[5] He may have made the worst case for deviations clear, but the English spelling of his name is not. You'll find his first name spelled Pavnutii or Pavnuty and his last name spelled Chebsheff, Cebysev, and other creative versions.

[6] See Table Z in Appendix G, if you're curious. But your calculator (and any statistics computer package) does this, too—and more easily!

FIGURE 6.7

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



z	.00	.01
1.7	.9554	.9564
1.8	.9641	.9649
1.9	.9713	.9719

TI-*nspire*

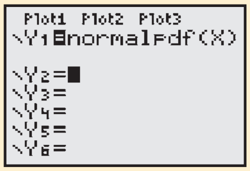
Normal percentiles. Explore the relationship between *z*-scores and areas in a Normal model.

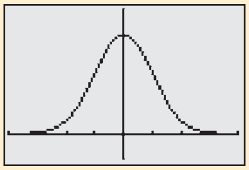
In the piece of the table shown, we find your *z*-score by looking down the left column for the first two digits, 1.8, and across the top row for the third digit, 0. The table gives the percentile as 0.9641. That means that 96.4% of the *z*-scores are less than 1.80. Only 3.6% of people, then, scored better than 680 on the SAT.

Most of the time, though, you'll do this with your calculator.

TI Tips

Finding Normal percentages



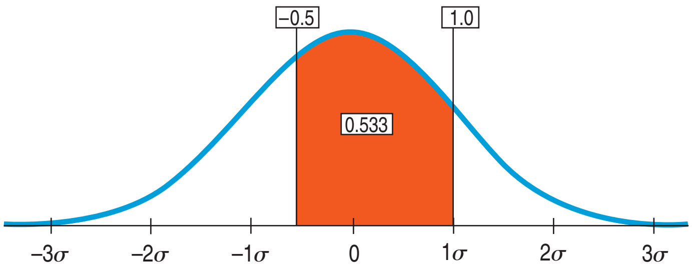


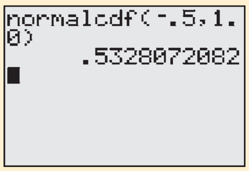
Your calculator knows the Normal model. Have a look under `2nd DISTR`. There you will see three "norm" functions, `normalpdf(`, `normalcdf(`, and `invNorm(`. Let's play with the first two.

- `normalpdf(` calculates *y*-values for graphing a Normal curve. You probably won't use this very often, if at all. If you want to try it, graph `Y1=normalpdf(X)` in a graphing `WINDOW` with `Xmin=-4`, `Xmax=4`, `Ymin=-0.1`, and `Ymax=0.5`.
- `normalcdf(` finds the proportion of area under the curve between two *z*-score cut points, by specifying `normalcdf(zLeft,zRight)`. Do make friends with this function; you will use it often!

Example 1

The Normal model shown shades the region between $z = -0.5$ and $z = 1.0$.





To find the shaded area:

Under `2nd DISTR` select `normalcdf(`; hit `ENTER`.
Specify the cut points: `normalcdf(-.5,1.0)` and hit `ENTER` again.

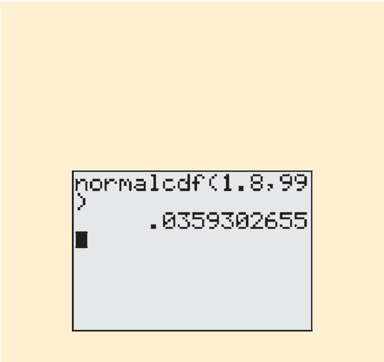
There's the area. Approximately 53% of a Normal model lies between half a standard deviation below and one standard deviation above the mean.

Example 2

In the example in the text we used Table Z to determine the fraction of SAT scores above your score of 680. Now let's do it again, this time using your `TI`.

First we need *z*-scores for the cut points:

- Since 680 is 1.8 standard deviations above the mean, your *z*-score is 1.8; that's the left cut point.



- Theoretically the standard Normal model extends rightward forever, but you can't tell the calculator to use infinity as the right cut point. Recall that for a Normal model almost all the area lies within ±3 standard deviations of the mean, so any upper cut point beyond, say, $z = 5$ does not cut off anything very important. We suggest you always use 99 (or −99) when you really want infinity as your cut point—it's easy to remember and way beyond any meaningful area.

Now you're ready. Use the command `normalcdf(1.8,99)`.

There you are! The Normal model estimates that approximately 3.6% of SAT scores are higher than 680.

STEP-BY-STEP EXAMPLE Working with Normal Models Part I

The Normal model is our first model for data. It's the first in a series of modeling situations where we step away from the data at hand to make more general statements about the world. We'll become more practiced in thinking about and learning the details of models as we progress through the book. To give you some practice in thinking about the Normal model, here are several problems that ask you to find percentiles in detail.

Question: What proportion of SAT scores fall between 450 and 600?

THINK

Plan State the problem.

I want to know the proportion of SAT scores between 450 and 600.

Variables Name the variable.

Let y = SAT score.

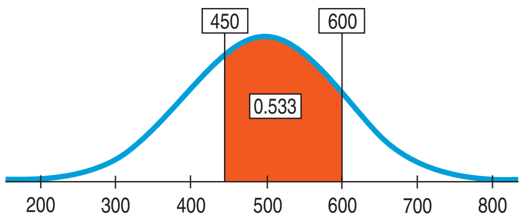
Check the appropriate conditions and specify which Normal model to use.

✔ **Nearly Normal Condition:** We are told that SAT scores are nearly Normal.

I'll model SAT scores with a N(500, 100) model, using the mean and standard deviation specified for them.

SHOW

Mechanics Make a picture of this Normal model. Locate the desired values and shade the region of interest.



Find z-scores for the cut points 450 and 600. Use technology to find the desired proportions, represented by the area under the curve. (This was Example 1 in the TI Tips—take another look.)

Standardizing the two scores, I find that

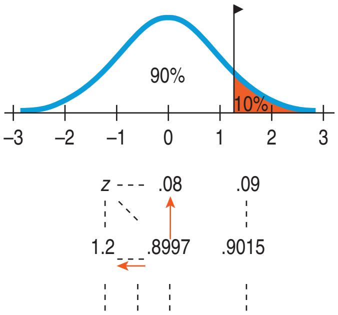
$$z = \frac{(y - \mu)}{\sigma} = \frac{(600 - 500)}{100} = 1.00$$

and

$$z = \frac{(450 - 500)}{100} = -0.50$$

(If you use a table, then you need to subtract the two areas to find the area *between* the cut points.)	So, Area $(450 < y < 600)$ = Area $(-0.5 < z < 1.0)$ $= 0.5328$ (**OR**: From Table Z, the area $(z < 1.0) = 0.8413$ and area $(z < -0.5) = 0.3085$, so the proportion of z-scores *between* them is $0.8413 - 0.3085 = 0.5328$, or 53.28%.)
TELL **Conclusion** Interpret your result in context.	The Normal model estimates that about 53.3% of SAT scores fall between 450 and 600.

From Percentiles to Scores: *z* in Reverse

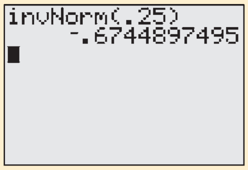


Finding areas from z-scores is the simplest way to work with the Normal model. But sometimes we start with areas and are asked to work backward to find the corresponding z-score or even the original data value. For instance, what z-score cuts off the top 10% in a Normal model?

Make a picture like the one shown, shading the rightmost 10% of the area. Notice that this is the 90th percentile. Look in Table Z for an area of 0.900. The exact area is not there, but 0.8997 is pretty close. That shows up in the table with 1.2 in the left margin and .08 in the top margin. The z-score for the 90th percentile, then, is approximately $z = 1.28$.

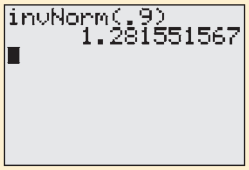
Computers and calculators will determine the cut point more precisely (and more easily).

TI Tips — Finding Normal cutpoints



To find the z-score at the 25th percentile, go to `2nd DISTR` again. This time we'll use the third of the "norm" functions, `invNorm(`.

Just specify the desired percentile with the command `invNorm(.25)` and hit `ENTER`. The calculator says that the cut point for the leftmost 25% of a Normal model is approximately $z = -0.674$.



One more example: What z-score cuts off the highest 10% of a Normal model? That's easily done—just remember to specify the *percentile*. Since we want the cut point for the *highest* 10%, we know that the other 90% must be *below* that z-score. The cut point, then, must stand at the 90th percentile, so specify `invNorm(.90)`.

Only 10% of the area in a Normal model is more than about 1.28 standard deviations above the mean.

STEP-BY-STEP EXAMPLE — Working with Normal Models Part II

Question: Suppose a college says it admits only people with SAT Verbal test scores among the top 10%. How high a score does it take to be eligible?

THINK

Plan State the problem.

How high an SAT Verbal score do I need to be in the top 10% of all test takers?

Variable Define the variable.

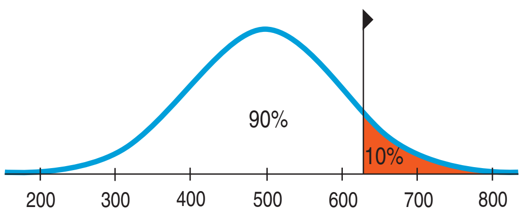
Let y = my SAT score.

Check to see if a Normal model is appropriate, and specify which Normal model to use.

✔ **Nearly Normal Condition:** I am told that SAT scores are nearly Normal. I'll model them with $N(500, 100)$.

SHOW

Mechanics Make a picture of this Normal model. Locate the desired percentile approximately by shading the rightmost 10% of the area.



The college takes the top 10%, so its cutoff score is the 90th percentile. Find the corresponding z-score using your calculator as shown in the TI Tips. (**OR**: Use Table Z as shown on p. 119.)

The cut point is $z = 1.28$.

Convert the z-score back to the original units.

A z-score of 1.28 is 1.28 standard deviations above the mean. Since the SD is 100, that's 128 SAT points. The cutoff is 128 points above the mean of 500, or 628.

TELL

Conclusion Interpret your results in the proper context.

Because the school wants SAT Verbal scores in the top 10%, the cutoff is 628. (Actually, since SAT scores are reported only in multiples of 10, I'd have to score at least a 630.)

TI-nspire

Normal models. Watch the Normal model react as you change the mean and standard deviation.

STEP-BY-STEP EXAMPLE

More Working with Normal Models

Working with Normal percentiles can be a little tricky, depending on how the problem is stated. Here are a few more worked examples of the kind you're likely to see.

A cereal manufacturer has a machine that fills the boxes. Boxes are labeled "16 ounces," 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 ounces and the Normal model 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 ounces.

Based on their experience with the packaging machine, the company believes that the amount of cereal in the boxes fits a Normal model with a standard deviation of 0.2 ounces. The manufacturer decides to set the machine to put an average of 16.3 ounces 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?

THINK

Plan State the problem.

What proportion of boxes weigh less than 16 ounces?

Variable Name the variable.

Let y = weight of cereal in a box.

Check to see if a Normal model is appropriate.

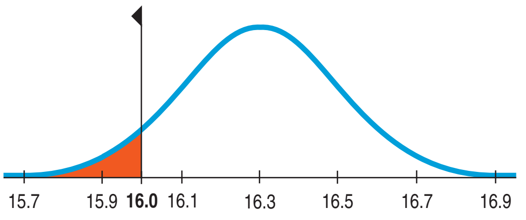
✔ **Nearly Normal Condition:** I have no data, so I cannot make a histogram, but I am told that the company believes the distribution of weights from the machine is Normal.

Specify which Normal model to use.

I'll use a $N(16.3, 0.2)$ model.

SHOW

Mechanics Make a picture of this Normal model. 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. Less than 20% for sure.

I want to know what fraction of the boxes will weigh less than 16 ounces.

Convert your cutoff value into a z-score.

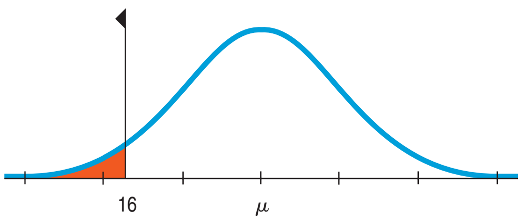
$$z = \frac{y - \mu}{\sigma} = \frac{16 - 16.3}{0.2} = -1.50$$

Find the area with your calculator (or use the Normal table).

$$\text{Area}(y < 16) = \text{Area}(z < -1.50) = 0.0668$$

TELL **Conclusion** State your conclusion, and check that it's consistent with your earlier guess. It's below 20%—seems okay.	I estimate that approximately 6.7% of the boxes will contain less than 16 ounces 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?

THINK **Plan** State the problem.	What mean weight will reduce the proportion of underweight boxes to 4%?
Variable Name the variable.	Let y = weight of cereal in a box.
Check to see if a Normal model is appropriate.	✔ **Nearly Normal Condition:** I am told that a Normal model applies.
Specify which Normal model to use. This time you are not given a value for the mean!	I don't know μ, the mean amount of cereal. The standard deviation for this machine is 0.2 ounces. The model is $N(\mu, 0.2)$.
REALITY CHECK We found out earlier that setting the machine to $\mu = 16.3$ ounces made 6.7% of the boxes too light. We'll need to raise the mean a bit to reduce this fraction.	No more than 4% of the boxes can be below 16 ounces.
SHOW **Mechanics** Make a picture of this Normal model. Center it at μ (since you don't know the mean), and shade the region below 16 ounces.	
Using your calculator (or the Normal table), 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 σ is 0.2, that's 1.75×0.2, or 0.35 ounces more than 16.	For 16 to be 1.75 standard deviations below the mean, the mean must be $16 + 1.75(0.2) = 16.35$ ounces.
TELL **Conclusion** Interpret your result in context. (This makes sense; we knew it would have to be just a bit higher than 16.3.)	The company must set the machine to average 16.35 ounces of cereal per box.

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 ounces 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?

THINK

Plan State the problem.

What standard deviation will allow the mean to be 16.2 ounces and still have only 4% of boxes underweight?

Variable Name the variable.

Let y = weight of cereal in a box.

Check conditions to be sure that a Normal model is appropriate.

✓ **Nearly Normal Condition:** The company believes that the weights are described by a Normal model.

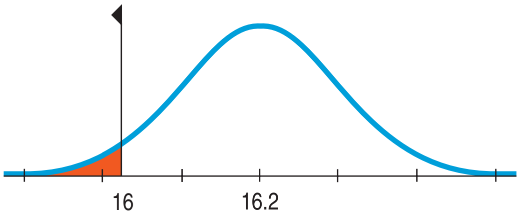
Specify which Normal model to use. This time you don't know σ.

I know the mean, but not the standard deviation, so my model is $N(16.2, \sigma)$.

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

SHOW

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



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

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

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

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

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

$$1.75\,\sigma = 0.2$$

$$\sigma = 0.114$$

TELL

Conclusion Interpret your result in context.

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 ounces. This means the machine must be more consistent (by nearly a factor of 2) in filling the boxes.

Are You Normal? Find Out with a Normal Probability Plot

In the examples we've worked through, we've assumed that the underlying data distribution was roughly unimodal and symmetric, so that using a Normal model makes sense. When you have data, you must *check* to see whether a Normal model is reasonable. How? Make a picture, of course! Drawing a histogram of the data and looking at the shape is one good way to see if a Normal model might work.

TI-*nspire*

Normal probability plots and histograms. See how a normal probability plot responds as you change the shape of a distribution.

There's a more specialized graphical display that 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.

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 model.

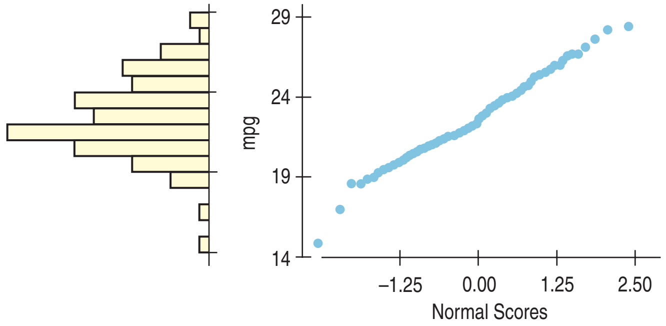


FIGURE 6.9

Histogram and Normal probability plot for gas mileage (mpg) recorded by one of the authors over the 8 years he owned a 1989 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 the men's *Weight*s from the NHANES Study 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 68–95–99.7 Rule for these data would not be very accurate.

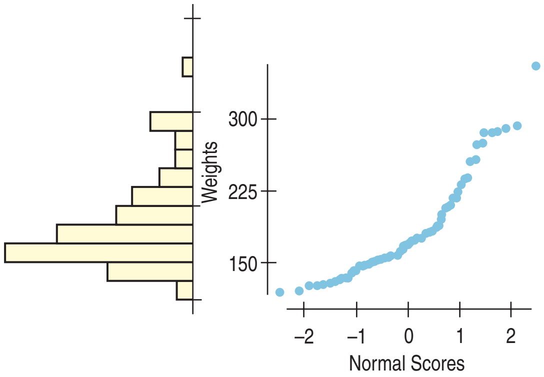


FIGURE 6.10

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

TI Tips

Creating a Normal probability plot

Let's make a Normal probability plot with the calculator. Here are the boys' agility test scores we looked at in Chapter 5; enter them in `L1`:

22, 17, 18, 29, 22, 23, 24, 23, 17, 21

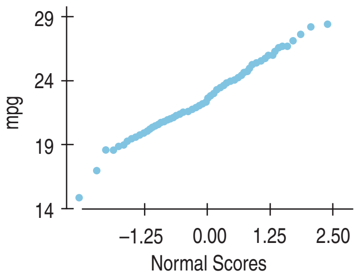
Now you can create the plot:

- Turn a `STATPLOT On`.
- Tell it to make a Normal probability plot by choosing the last of the icons.
- Specify your datalist and which axis you want the data on. (We'll use `Y` so the plot looks like the others we showed you.)
- Specify the `Mark` you want the plot to use.
- Now `ZoomStat` does the rest.

The plot doesn't look very straight. Normality is certainly questionable here.

(Not that it matters in making this decision, but that vertical line is the *y*-axis. Points to the left have negative *z*-scores and points to the right have positive *z*-scores.)

How Does a Normal Probability Plot Work?



A S ***Activity:*** **Assessing Normality.** This activity guides you through the process of checking the Nearly Normal condition using your statistics package.

Why does the Normal probability plot work like that? We looked at 100 fuel efficiency measures for the author's Nissan 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 (well, −1.1 to be exact). A Normal probability plot takes each data value and plots it against the *z*-score you'd expect that point to have if the distribution were perfectly Normal.[7]

When the values match up well, the line is straight. If one or two points are surprising from the Normal's point of view, they don't line up. When the entire distribution is skewed or different from the Normal in some other way, the values don't match up very well at all and the plot bends.

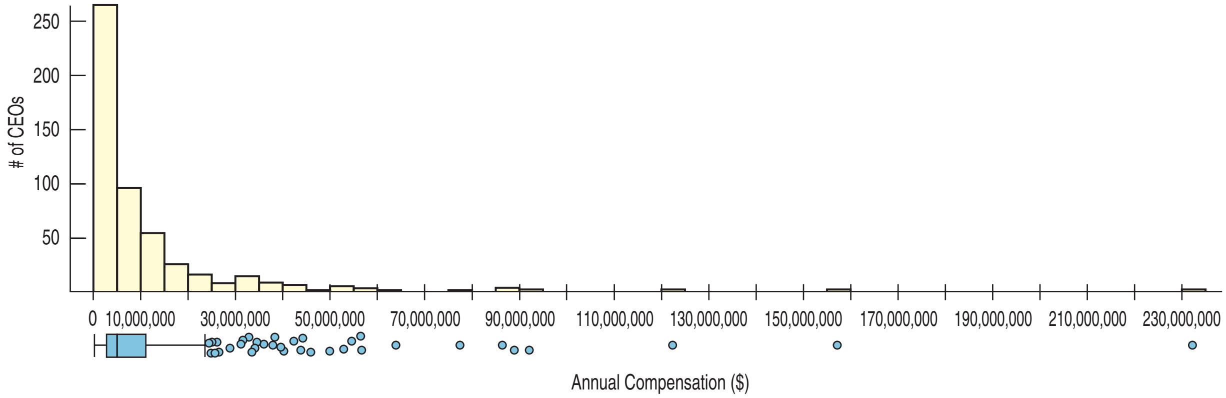
It turns out to be tricky to find the values we expect. They're called *Normal scores,* but you can't easily look them up in the tables. That's why probability plots are best made with technology and not by hand.

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.

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

WHAT CAN GO WRONG?

- **Don't use a Normal model when the distribution is not unimodal and symmetric.** Normal models are so easy and useful that it is tempting to use them even when they don't describe the data very well. That can lead to wrong conclusions. Don't use a Normal model without first checking the **Nearly Normal Condition.** Look at a picture of the data to check that it is unimodal and symmetric. A histogram, or a Normal probability plot, can help you tell whether a Normal model is appropriate.



The CEOs (p. 90) had 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 mean and standard deviation when outliers are present.** Both means and standard deviations can be distorted by outliers, and no model based on distorted values will do a good job. A z-score calculated from a distribution with outliers may be misleading. It's always a good idea to check for outliers. How? Make a picture.
- **Don't round your results in the middle of a calculation.** We *reported* the mean of the heptathletes' long jump as 6.16 meters. More precisely, it was 6.16153846153846 meters.

You should use all the precision available in the data for all the intermediate steps of a calculation. Using the more precise value for the mean (and also carrying 15 digits for the SD), the z-score calculation for Klüft's long jump comes out to

$$z = \frac{6.78 - 6.16153846153846}{0.2297597407326585} = 2.691775053755667700$$

We'd report that as 2.692, as opposed to the rounded-off value of 2.70 we got earlier from the table.

- **Don't worry about minor differences in results.** Because various calculators and programs may carry different precision in calculations, your answers may differ slightly from those we show in the text and in the Step-By-Steps, or even from the values given in the answers in the back of the book. Those differences aren't anything to worry about. They're not the main story Statistics tries to tell.

CONNECTIONS

Changing the center and spread of a variable is equivalent to changing its *units.* Indeed, the only part of the data's context changed by standardizing is the units. All other aspects of the context do not depend on the choice or modification of measurement units. This fact points out an important distinction between the numbers the data provide for calculation and the meaning of the variables and the relationships among them. Standardizing can make the numbers easier to work with, but it does not alter the meaning.

Another way to look at this is to note that standardizing may change the center and spread values, but it does not affect the *shape* of a distribution. A histogram or boxplot of standardized values looks just the same as the histogram or boxplot of the original values except, perhaps, for the numbers on the axes.

When we summarized *shape, center,* and *spread* for histograms, we compared them to unimodal, symmetric shapes. You couldn't ask for a nicer example than the Normal model. And if the shape *is* like a Normal, we'll use the the mean and standard deviation to standardize the values.

WHAT HAVE WE LEARNED?

We've learned that the story data can tell may be easier to understand after shifting or rescaling the data.

- Shifting data by adding or subtracting the same amount from each value affects measures of center and position but not measures of spread.
- Rescaling data by multiplying or dividing every value by a constant, changes all the summary statistics—center, position, and spread.

We've learned the power of standardizing data.

- Standardizing uses the standard deviation as a ruler to measure distance from the mean, creating *z*-scores.
- Using these *z*-scores, we can compare apples and oranges—values from different distributions or values based on different units.
- And a *z*-score can identify unusual or surprising values among data.

We've learned that the 68–95–99.7 Rule can be a useful rule of thumb for understanding distributions.

- For data that are unimodal and symmetric, about 68% fall within 1 SD of the mean, 95% fall within 2 SDs of the mean, and 99.7% fall within 3 SDs of the mean (see p. 130).

Again we've seen the importance of *Thinking* about whether a method will work.

- **Normality Assumption:** We sometimes work with Normal tables (Table Z). Those tables are based on the Normal model.
- Data can't be exactly Normal, so we check the **Nearly Normal Condition** by making a histogram (is it unimodal, symmetric, and free of outliers?) or a Normal probability plot (is it straight enough?). (See p. 125.)

Terms

Standardizing 105. We standardize to eliminate units. Standardized values can be compared and combined even if the original variables had different units and magnitudes.

Standardized value 105. A value found by subtracting the mean and dividing by the standard deviation.

Shifting 107. Adding a constant to each data value adds the same constant to the mean, the median, and the quartiles, but does not change the standard deviation or IQR.

Rescaling 108. Multiplying each data value by a constant multiplies both the measures of position (mean, median, and quartiles) and the measures of spread (standard deviation and IQR) by that constant.

Normal model 112. A useful family of models for unimodal, symmetric distributions.

Parameter 112. A numerically valued attribute of a model. For example, the values of μ and σ in a $N(\mu,\sigma)$ model are parameters.

Statistic 112. A value calculated from data to summarize aspects of the data. For example, the mean, $\bar{y}$ and standard deviation, s, are statistics.

***z*-score** 105. A z-score tells how many standard deviations a value is from the mean; z-scores have a mean of 0 and a standard deviation of 1. When working with data, use the statistics $\bar{y}$ and s:

$$z = \frac{y - \bar{y}}{s}.$$

112. When working with models, use the parameters μ and σ:

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

Standard Normal model 112. A Normal model, $N(\mu,\sigma)$ with mean $\mu = 0$ and standard deviation $\sigma = 1$. Also called the **standard Normal distribution.**

Nearly Normal Condition 112. A distribution is nearly Normal if it is unimodal and symmetric. We can check by looking at a histogram or a Normal probability plot.

68–95–99.7 Rule 113. In a Normal model, about 68% of values fall within 1 standard deviation of the mean, about 95% fall within 2 standard deviations of the mean, and about 99.7% fall within 3 standard deviations of the mean.

Normal percentile 116. The Normal percentile corresponding to a z-score gives the percentage of values in a standard Normal distribution found at that z-score or below.

Normal probability plot 124. 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.**

Skills

THINK

- Understand how adding (subtracting) a constant or multiplying (dividing) by a constant changes the center and/or spread of a variable.
- Recognize when standardization can be used to compare values.
- Understand that standardizing uses the standard deviation as a ruler.
- Recognize when a Normal model is appropriate.

SHOW

- Know how to calculate the z-score of an observation.
- Know how to compare values of two different variables using their z-scores.
- Be able to use Normal models and the 68–95–99.7 Rule to estimate the percentage of observations falling within 1, 2, or 3 standard deviations of the mean.
- Know how to find the percentage of observations falling below any value in a Normal model using a Normal table or appropriate technology.
- Know how to check whether a variable satisfies the **Nearly Normal Condition** by making a Normal probability plot or a histogram.

TELL

- Know what z-scores mean.
- Be able to explain how extraordinary a standardized value may be by using a Normal model.