

6.1 Discrete and Continuous Random Variables

A probability model describes the possible outcomes of a chance process and the likelihood that those outcomes will occur. For example, suppose we toss a fair coin 3 times. The sample space for this chance process is

HHH HHT HTH THH HTT THT TTH TTT

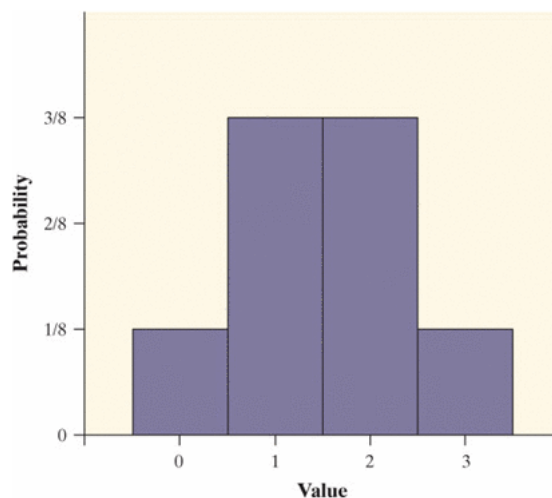
Since there are 8 equally likely outcomes, the probability is $1/8$ for each possible outcome. Define the variable X = the number of heads obtained. The value of X will vary from one set of tosses to another but will always be one of the numbers 0, 1, 2, or 3. How likely is X to take each of those values? It will be easier to answer this question if we group the possible outcomes by the number of heads obtained:

$X = 0$: TTT
 $X = 1$: HTT THT TTH
 $X = 2$: HHT HTH THH
 $X = 3$: HHH

We can summarize the **probability distribution** of X as follows:

Value:	0	1	2	3
Probability:	$1/8$	$3/8$	$3/8$	$1/8$

In Graphical Form:



We can use the probability distribution to answer questions about the variable X . What's the probability that we get at least one head in three tosses of the coin? In symbols, we want to find $P(X \geq 1)$. We could add probabilities to get the answer:

$$\begin{aligned} P(X \geq 1) &= P(X = 1) + P(X = 2) + P(X = 3) \\ &= 1/8 + 3/8 + 3/8 = 7/8 \end{aligned}$$

Or

$$\begin{aligned} P(X \geq 1) &= 1 - P(X < 1) = 1 - P(X = 0) \\ &= 1 - 1/8 = 7/8 \end{aligned}$$

A numerical variable that describes the outcomes of a chance process (like X in the coin-tossing scenario) is called a **random variable**. The probability model for a random variable is its probability distribution.

Random Variable - A **random variable** takes numerical values that describe the outcomes of some chance process. The **probability distribution** of a random variable gives its possible values and their probabilities.

There are two main types of random variables, corresponding to two types of probability distributions: *discrete* and *continuous*.

6.1.1 Discrete Random Variables

We have learned several rules of probability but only one way of assigning probabilities to events: assign a probability to every individual outcome, then add these probabilities to find the probability of any event. This idea works well if we can find a way to list all possible outcomes. We will call random variables having probability assigned in this way **discrete random variables**. The probability distribution for a discrete random variable must have outcome probabilities that are between 0 and 1 and that add up to 1.

Discrete Random Variables and Their Probability Distributions

A **discrete random variable** X takes a fixed set of possible values with gaps between. The probability distribution of a discrete random variable X lists the values x_i , and their probabilities P_i :

Value:	x_1	x_2	x_3	\dots
Probability:	p_1	p_2	p_3	\dots

The probabilities P_i must satisfy two requirements:

1. Every probability P_i is a number between 0 and 1.
2. The sum of the probabilities is 1: $P_1 + P_2 + P_3 + \dots = 1$.

To find the probability of any event, add the probabilities P_i of the particular values x_i that make up the event.

Example – Apgar Scores: Babies' Health at Birth
Discrete random variables

In 1952, Dr. Virginia Apgar suggested five criteria for measuring a baby's health at birth: skin color, heart rate, muscle tone, breathing, and response when stimulated. She developed a 0-1-2 scale to rate a newborn on each of the five criteria. A baby's Apgar score is the sum of the ratings on each of the five scales, which gives a whole-number from 0 to 10. Apgar scores are still used today to evaluate the health of newborns.

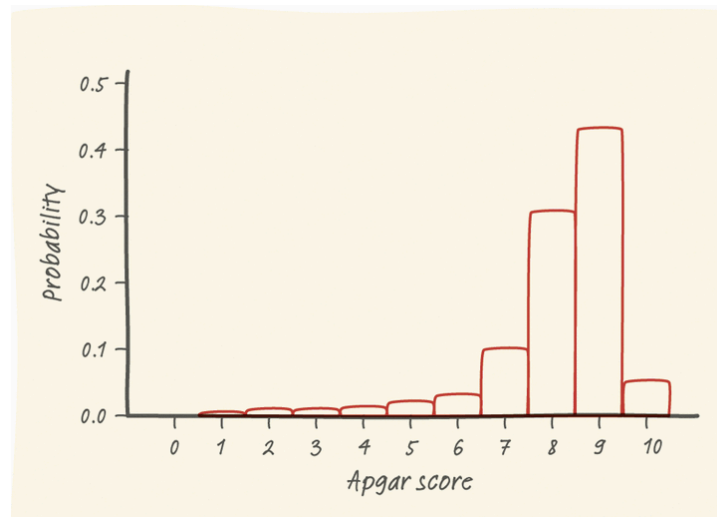
What Apgar scores are typical? To find out, researchers recorded the Apgar scores of over 2 million newborn babies in a single year. Imagine selecting one of these newborns at random. (That's our chance process.) Define the random variable X = Apgar score of a randomly selected baby one minute after birth. The table below gives the probability distribution for X .

Value:	0	1	2	3	4	5	6	7	8	9	10
Probability:	0.001	0.006	0.007	0.008	0.012	0.020	0.038	0.099	0.319	0.437	0.053

PROBLEM:

(a) Show that the probability distribution for X is legitimate.

(b) Make a histogram of the probability distribution. *Describe* what you see.



(c) *Doctors decided that Apgar scores of 7 or higher indicate a healthy baby. What's the probability that a randomly selected baby is healthy?*

Note that the probability of randomly selecting a newborn whose Apgar score is greater than or equal to 7 is not the same as the probability that the baby's Apgar score is strictly greater than 7. The latter probability is

$$\begin{aligned} P(X \geq 7) &= P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10) \\ &= 0.098 + 0.319 + 0.437 + 0.053 = 0.907 \end{aligned}$$

CHECK YOUR UNDERSTANDING

North Carolina State University posts the grade distributions for its courses online.³ Students in Statistics 101 in a recent semester received 26% A's, 42% B's, 20% C's, 10% D's, and 2% F's. Choose a Statistics 101 student at random. The student's grade on a four-point scale (with A = 4) is a discrete random variable X with this probability distribution:

Value of X :	0	1	2	3	4
Probability:	0.02	0.10	0.20	0.42	0.26

1. Say in words what the meaning of $P(X \geq 3)$ is. What is this probability?
2. Write the event "the student got a grade worse than C" in terms of values of the random variable X . What is the probability of this event?
3. Sketch a graph of the probability distribution. Describe what you see.

6.1.2 Mean (Expected Value) of a Discrete Random Variable

When we analyzed distributions of quantitative data in **Chapter 1**, we made it a point to discuss their shape, center, and spread, and to identify any outliers. We'll follow the same strategy with probability distributions of random variables. You can use what you learned earlier to describe the shape of a probability distribution histogram. We've already seen examples of symmetric (number of heads in three coin tosses) and left-skewed (Apgar score of a randomly chosen baby) probability distributions. What about center and spread?

The mean \bar{X} of a set of observations is their ordinary average. The **mean μ_X a discrete random variable X** is also an average of the possible values of X , but with an important change to take into account the fact that not all outcomes may be equally likely. A simple example will show what we need to do.

Example – Winning (and Losing) at Roulette

Finding the mean of a discrete random variable

On an American roulette wheel, there are 38 slots numbered 1 through 36, plus 0 and 00. Half of the slots from 1 to 36 are red; the other half are black. Both the 0 and 00 slots are green. Suppose that a player places a simple \$1 bet on red. If the ball lands in a red slot, the player gets the original dollar back, plus an additional dollar for winning the bet. If the ball lands in a different-colored slot, the player loses the dollar bet to the casino.

Let's define the random variable X = net gain from a single \$1 bet on red. The possible values of X are $-\$1$ and $\$1$. (The player either gains a dollar or loses a dollar.) What are the corresponding probabilities? The chance that the ball lands in a red slot is $18/38$. The chance that the ball lands in a different-colored slot is $20/38$. Here is the probability distribution of X :

Value:	$-\$1$	$\$1$
Probability:	$20/38$	$18/38$

What is the player's average gain? The ordinary average of the two possible outcomes $-\$1$ and $\$1$ is $\$0$. But $\$0$ isn't the average winnings because the player is less likely to win $\$1$ than to lose $\$1$. In the long run, the player gains a dollar 18 times in every 38 games played and loses a dollar on the remaining 20 of 38 bets. The player's long-run average gain for this simple bet is

$$\mu_X = (-\$1)\left(\frac{20}{38}\right) + (\$1)\left(\frac{18}{38}\right) = -\$0.05$$

You see that in the long run the player loses (and the casino gains) five cents per bet.

If someone actually played several games of roulette, we would call the mean amount the person gained \bar{X} . The mean in the previous example is a different quantity—it is the long-run average gain we'd expect if someone played roulette a very large number of times. For this reason, the mean of a random variable is often referred to as its expected value. Just as probabilities are an idealized description of long-run proportions, the mean of a discrete random variable describes the long-run average outcome. There are two ways of denoting the mean of a random variable X . We can use the notation μ_X or we can write $E(X)$, as in the "expected value of X ". In the roulette example, $\mu_X = E(X) = -\$0.05$.

The mean of any discrete random variable is found just as in the roulette example. It is an average of the possible outcomes, but a weighted average in which each outcome is weighted by its probability. Here (finally!) is the definition.

Mean (expected value) of a discrete random variable – Suppose that X is a discrete random variable whose probability is

Value:	x_1	x_2	x_3	\dots
Probability:	p_1	p_2	p_3	\dots

To find the **mean (expected value)** of X , multiply each possible value by its probability, then add all the products:

$$\begin{aligned}\mu_X = E(X) &= x_1p_1 + x_2p_2 + x_3p_3 + \dots \\ &= \sum x_i p_i\end{aligned}$$

Example – Apgar Scores: What’s Typical?
Mean and expected value as an average

In 1952, Dr. Virginia Apgar suggested five criteria for measuring a baby’s health at birth: skin color, heart rate, muscle tone, breathing, and response when stimulated. She developed a 0-1-2 scale to rate a newborn on each of the five criteria. A baby’s Apgar score is the sum of the ratings on each of the five scales, which gives a whole-number from 0 to 10. Apgar scores are still used today to evaluate the health of newborns.

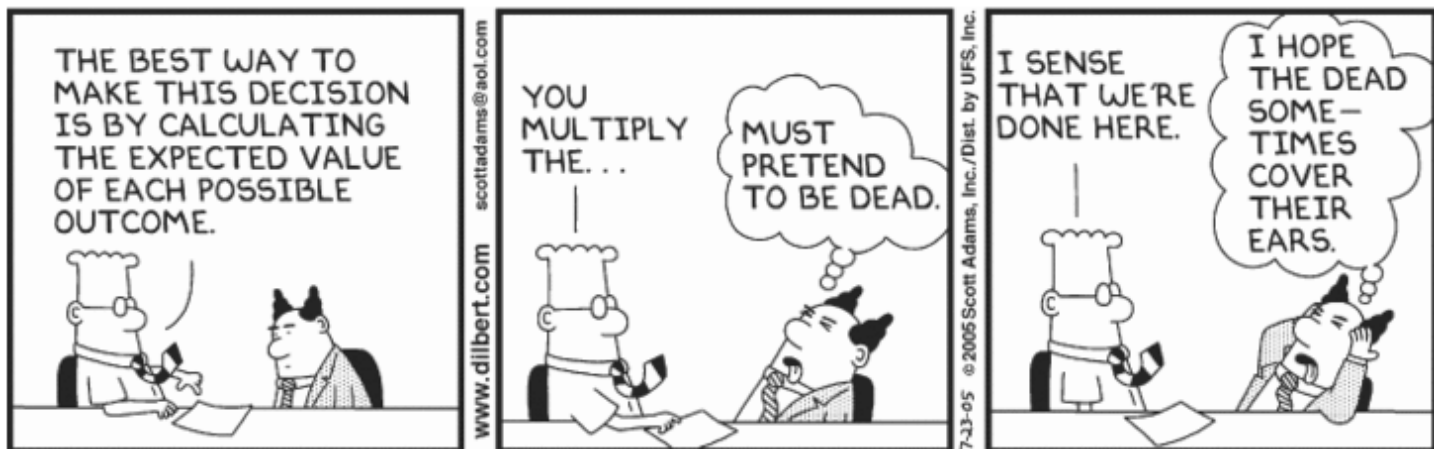
What Apgar scores are typical? To find out, researchers recorded the Apgar scores of over 2 million newborn babies in a single year. Imagine selecting one of these newborns at random. (That’s our chance process.) Define the random variable X = Apgar score of a randomly selected baby one minute after birth. The table below gives the probability distribution for X .

Value x_i:	0	1	2	3	4	5	6	7	8	9	10
Probability p_i:	0.001	0.006	0.007	0.008	0.012	0.020	0.038	0.099	0.319	0.437	0.053

PROBLEM: Compute the mean of the random variable X and interpret this value in context.

Notice that the mean Apgar score, 8.128, is not a possible value of the random variable X . It's also not an integer. If you think of the mean as a long-run average over many repetitions, these facts shouldn't bother you.

AP EXAM TIP If the mean of a random variable has a non-integer value, but you report it as an integer, your answer will be marked as incorrect



6.1.3 Standard Deviation (and Variance) of a Discrete Random Variable

With the mean as our measure of center for a discrete random variable, it shouldn't surprise you that we'll use the standard deviation as our measure of spread. In chapter 1, we first defined the sample variance s_x^2 as the "average squared deviation from the mean" and then took the square root of the variance to get the sample standard deviation σ_x . The definition of the **variance of a random variable** σ_X^2 is similar to the definition of the variance for a set of quantitative data. That is, the variance is an average of the squared deviation $(x_i - \mu_X)^2$ of the values of the variable X from its mean μ_X .

As with the mean, the average we use is a weighted average in which each outcome is weighted by its probability to take account of outcomes that are not equally likely. To get the **standard deviation of a random variable**, we take the square root of the variance.

DEFINITION: Variance and standard deviation of a discrete random variable

Suppose that X is a discrete random variable whose probability distribution is

Value:	x_1	x_2	x_3	\dots
Probability:	p_1	p_2	p_3	\dots

and that μ_X is the mean of X . The **variance of X** is

$$\begin{aligned}\text{Var}(X) &= \sigma_X^2 = (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + (x_3 - \mu_X)^2 p_3 + \dots \\ &= \sum (x_i - \mu_X)^2 p_i\end{aligned}$$

The standard deviation of X , σ_X , is the square root of the variance.

The standard deviation of a random variable X is a measure of how much the values of the variable tend to vary, on average, from the mean μ_X .

Example – Apgar Scores: How Variable Are They?

Calculating measure of spread

In 1952, Dr. Virginia Apgar suggested five criteria for measuring a baby's health at birth: skin color, heart rate, muscle tone, breathing, and response when stimulated. She developed a 0-1-2 scale to rate a newborn on each of the five criteria. A baby's Apgar score is the sum of the ratings on each of the five scales, which gives a whole-number from 0 to 10. Apgar scores are still used today to evaluate the health of newborns.

What Apgar scores are typical? To find out, researchers recorded the Apgar scores of over 2 million newborn babies in a single year. Imagine selecting one of these newborns at random. (That's our chance process.) Define the random variable X = Apgar score of a randomly selected baby one minute after birth. The table below gives the probability distribution for X .

Value x_i :	0	1	2	3	4	5	6	7	8	9	10
Probability p_i :	0.001	0.006	0.007	0.008	0.012	0.020	0.038	0.099	0.319	0.437	0.053

The mean Apgar score of a randomly chosen newborn to be $\mu_X = 8.128$

PROBLEM: Compute and interpret the standard deviation of the random variable X .

The formula for the variance of X is $\sigma_X^2 = \sum (x_i - \mu_X)^2 p_i$. Plugging in values gives

$$\begin{aligned}\sigma_X^2 &= (0 - 8.128)^2(0.001) + (1 - 8.128)^2(0.006) \\ &\quad + (2 - 8.128)^2(0.007) + \cdots + (10 - 8.128)^2(0.053) \\ \sigma_X^2 &= 2.066\end{aligned}$$

The standard deviation of X is $\sigma_X = \sqrt{2.066} = 1.437$. On average, a randomly selected baby's Apgar score will differ from the mean (8.128) by about 1.4 units.

Learn how to use the calculator to analyze random variables: **Analyzing random variables on the calculator**

CHECK YOUR UNDERSTANDING

A large auto dealership keeps track of sales made during each hour of the day. Let X = the number of cars sold during the first hour of business on a randomly selected Friday. Based on previous records, the probability distribution of X is as follows:

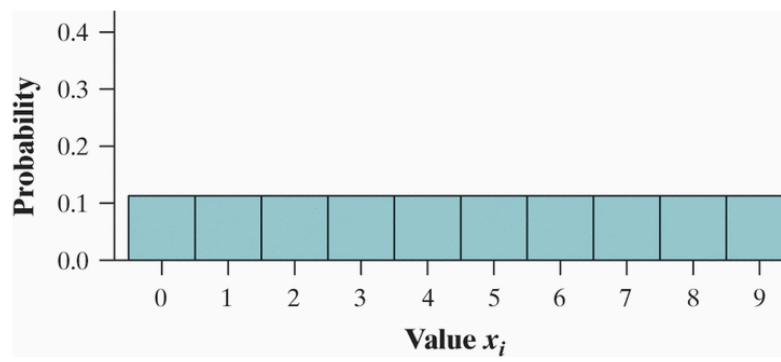
Cars sold:	0	1	2	3
Probability:	0.3	0.4	0.2	0.1

1. Compute and interpret the mean of X .

2. Compute and interpret the standard deviation of X .

6.1.4 Continuous Random Variables

When we use the table of random digits to select a digit between 0 and 9, the result is a discrete random variable (call it X). The probability model assigns probability $1/10$ to each of the 10 possible values of X . The figure below shows the probability distribution for this random variable.



Suppose we want to choose a number at random between 0 and 1, allowing *any* number between 0 and 1 as the outcome. Calculator and computer random number generators will do this. The sample space of this chance process is an entire interval of numbers:

$S = \{\text{all numbers between 0 and 1}\}$

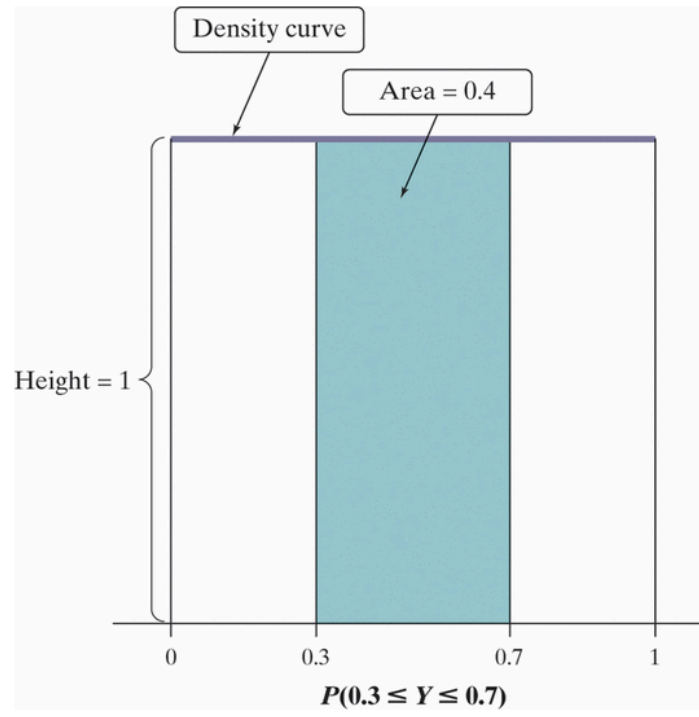
Call the outcome of the random number generator Y for short. How can we find probabilities of events like $P(0.3 \leq Y \leq 0.7)$? As in the case of selecting a random digit, we would like all possible outcomes to be equally likely. But we cannot assign probabilities to each individual value of Y and then add them, because there are infinitely many possible values.

In situations like this, we use a different way of assigning probabilities directly to events—as *areas under a density curve*. Recall from **Chapter 2** that any density curve has area exactly 1 underneath it, corresponding to total probability 1.

Example – Random Numbers

Density curves and probability distributions

The random number generator will spread its output uniformly across the entire interval from 0 to 1 as we allow it to generate a long sequence of random numbers. The results of many trials are represented by the density curve of a *uniform distribution*. This density curve below appears in purple. It has height 1 over the interval from 0 to 1. The area under the density curve is 1, and the probability of any event is the area under the density curve and above the event in question.



As the figure above shows, the probability that the random number generator produces a number Y between 0.3 and 0.7 is:

$$P(0.3 \leq Y \leq 0.7) = 0.4$$

That's because the area of the shaded rectangle is

$$\text{length} \times \text{width} = 0.4 \times 1 = 0.4$$

The figure above shows the probability distribution of the random variable Y = random number between 0 and 1. We call Y a **continuous random variable** because its values are not isolated numbers but rather an entire interval of numbers.

Continuous Random Variable X - takes all values in an interval of numbers. The probability distribution of X is described by a density curve. The probability of any event is the area under the density curve and above the values of X that make up the event.

In many cases, discrete random variables arise from counting something—for instance, the number of siblings that a randomly selected student has. Continuous random variables often arise from measuring something—for instance, height, SAT score, or blood pressure of a randomly selected student.

The probability distribution for a continuous random variable assigns probabilities to intervals of outcomes rather than to individual outcomes. In fact, *all continuous probability models assign probability 0 to every individual outcome*. Only intervals of values have positive probability. To see that this is true, consider a specific outcome from the random number generator of the previous example, such as $P(Y = 0.7)$. The probability of this event is the area under the density curve that's above the point 0.7 on the horizontal axis. But this vertical line segment has no width, so the area is 0. For that reason,

$$P(0.3 \leq Y \leq 0.7) = P(0.3 \leq Y < 0.7) = P(0.3 < Y \leq 0.7) = 0.4$$

We can use any density curve to assign probabilities. The density curves that are most familiar to us are the Normal curves. Normal distributions can be probability distributions as well as descriptions of data. There is a close connection between a Normal distribution as an idealized description for data and a Normal probability model. The following example shows what we mean.

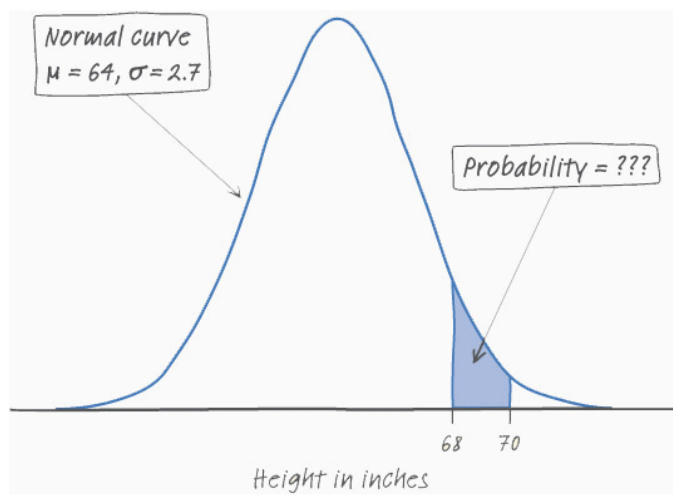
Example – Young Women's Heights Normal Probability distributions

The heights of young women closely follow the Normal distribution with mean $\mu = 64$ inches and standard deviation $\sigma = 2.7$ inches. This is a distribution for a large set of data. Now choose one young woman at random.

Call her height Y . If we repeat the random choice very many times, the distribution of values of Y is the same Normal distribution that describes the heights of all young women. Find the probability that the chosen woman is between 68 and 70 inches tall.

STATE: What's the probability that a randomly chosen young woman has height between 68 and 70 inches?

PLAN: The height Y of the woman we choose has the $N(64, 2.7)$ distribution. We want to find $P(68 \leq Y \leq 70)$. This is the area under the Normal curve below. We'll standardize the heights and then use a z-table to find the shaded area.



DO: The standardized scores for the two heights are

$$z = \frac{68 - 64}{2.7} = 1.48 \quad \text{and} \quad z = \frac{70 - 64}{2.7} = 2.22$$

If we let Z represent the random variable that follows a standard Normal distribution, then the desired probability is $P(1.48 \leq Z \leq 2.22)$. From the z -table, we find that $P(Z \leq 2.22) = 0.9868$ and $P(Z \leq 1.48) = 0.9306$. So we have

$$\begin{aligned} P(1.48 \leq Z \leq 2.22) &= P(Z \leq 2.22) - P(Z \leq 1.48) \\ &= 0.9868 - 0.9306 = 0.0562 \end{aligned}$$

CONCLUDE: There's about a 5.6% chance that a randomly chosen young woman has a height between 68 and 70 inches.

We can check that our answer is correct using the *Normal curve* applet or the `normalcdf` command on the TI-83/84 (`normCdf` on the TI-89).

Using the *Normal curve* applet or `normalcdf` (68, 70, 64, 2.7) yields $P(68 \leq Y \leq 70) = 0.0561$. As usual, there is a small roundoff error from using a z -table. You can also check that $P(1.48 \leq Z \leq 2.22) = 0.0562$ using `normalcdf` (1.48, 2.22).

The calculation in the preceding example is the same as those we did in chapter 2. Only the language of probability is new.

What about the mean and standard deviation for continuous random variables? The probability distribution of a continuous random variable X is described by a density curve. Chapter 2 showed how to find the mean of the distribution: it is the point at which the area under the density curve would balance if it were made out of solid material. The mean lies at the center of symmetric density curves such as the Normal curves. We can locate the standard deviation of a Normal distribution from its inflection points. Exact calculation of the mean and standard deviation for most continuous random variables requires advanced mathematics.

AP EXAM TIP When you solve problems involving random variables, start by defining the random variable of interest. For example, let X = the Apgar score of a randomly selected baby or let Y = the height of a randomly selected young woman. Then state the probability you're trying to find in terms of the random variable: $P(68 \leq Y \leq 70)$ or $P(X \geq 7)$.