

# Random Variables



## What Is an Actuary?

Actuaries are the daring people who put a price on risk, estimating the likelihood and costs of rare events, so they can be insured. That takes financial, statistical, and business skills. It also makes them invaluable to many businesses. Actuaries are rather rare themselves; only about 19,000 work in North America. Perhaps because of this, they are well paid. If you're enjoying this course, you may want to look into a career as an actuary. Contact the Society of Actuaries or the Casualty Actuarial Society (who, despite what you may think, did not pay for this blurb).

Insurance companies make bets. They bet that you're going to live a long life. You bet that you're going to die sooner. Both you and the insurance company want the company to stay in business, so it's important to find a "fair price" for your bet. Of course, the right price for *you* depends on many factors, and nobody can predict exactly how long you'll live. But when the company averages over enough customers, it can make reasonably accurate estimates of the amount it can expect to collect on a policy before it has to pay its benefit.

Here's a simple example. An insurance company offers a "death and disability" policy that pays \$10,000 when you die or \$5000 if you are permanently disabled. It charges a premium of only \$50 a year for this benefit. Is the company likely to make a profit selling such a plan? To answer this question, the company needs to know the *probability* that its clients will die or be disabled in any year. From actuarial information like this, the company can calculate the expected value of this policy.

## Expected Value: Center

### NOTATION ALERT:

The most common letters for random variables are  $X$ ,  $Y$ , and  $Z$ . But be cautious: If you see any capital letter, it just might denote a random variable.

We'll want to build a probability model in order to answer the questions about the insurance company's risk. First we need to define a few terms. The amount the company pays out on an individual policy is called a **random variable** because its numeric value is based on the outcome of a random event. We use a capital letter, like  $X$ , to denote a random variable. We'll denote a particular value that it can have by the corresponding lowercase letter, in this case  $x$ . For the insurance company,  $x$  can be \$10,000 (if you die that year), \$5000 (if you are disabled), or \$0 (if neither occurs). Because we can list all the outcomes, we might formally call this random variable a **discrete** random variable. Otherwise, we'd call it a **continuous** random variable. The collection of all the possible values and the probabilities that they occur is called the **probability model** for the random variable.

**AS** **Activity: Random Variables.** Learn more about random variables from this animated tour.

Suppose, for example, that the death rate in any year is 1 out of every 1000 people, and that another 2 out of 1000 suffer some kind of disability. Then we can display the probability model for this insurance policy in a table like this:

Policyholder Outcome	Payout $x$	Probability $P(X = x)$
Death	10,000	$\frac{1}{1000}$
Disability	5000	$\frac{2}{1000}$
Neither	0	$\frac{997}{1000}$

To see what the insurance company can expect, imagine that it insures exactly 1000 people. Further imagine that, in perfect accordance with the probabilities, 1 of the policyholders dies, 2 are disabled, and the remaining 997 survive the year unscathed. The company would pay \$10,000 to one client and \$5000 to each of 2 clients. That's a total of \$20,000, or an average of  $20000/1000 = \$20$  per policy. Since it is charging people \$50 for the policy, the company expects to make a profit of \$30 per customer. Not bad!

#### NOTATION ALERT:

The expected value (or mean) of a random variable is written  $E(X)$  or  $\mu$ .

We can't predict what *will* happen during any given year, but we can say what we *expect* to happen. To do this, we (or, rather, the insurance company) need the probability model. The expected value of a policy is a parameter of this model. In fact, it's the mean. We'll signify this with the notation  $\mu$  (for population mean) or  $E(X)$  for expected value. This isn't an average of some data values, so we won't estimate it. Instead, we assume that the probabilities are known and simply calculate the expected value from them.

How did we come up with \$20 as the expected value of a policy payout? Here's the calculation. As we've seen, it often simplifies probability calculations to think about some (convenient) number of outcomes. For example, we could imagine that we have exactly 1000 clients. Of those, exactly 1 died and 2 were disabled, corresponding to what the probabilities would say.

$$\mu = E(X) = \frac{10,000(1) + 5000(2) + 0(997)}{1000}$$

So our expected payout comes to \$20,000, or \$20 per policy.

Instead of writing the expected value as one big fraction, we can rewrite it as separate terms with a common denominator of 1000.

$$\begin{aligned} \mu &= E(X) \\ &= \$10,000\left(\frac{1}{1000}\right) + \$5000\left(\frac{2}{1000}\right) + \$0\left(\frac{997}{1000}\right) \\ &= \$20. \end{aligned}$$

How convenient! See the probabilities? For each policy, there's a  $1/1000$  chance that we'll have to pay \$10,000 for a death and a  $2/1000$  chance that we'll have to pay \$5000 for a disability. Of course, there's a  $997/1000$  chance that we won't have to pay anything.

Take a good look at the expression now. It's easy to calculate the **expected value** of a (discrete) random variable—just multiply each possible value by the probability that it occurs, and find the sum:

$$\mu = E(X) = \sum xP(x).$$

Be sure that every possible outcome is included in the sum. And verify that you have a valid probability model to start with—the probabilities should each be between 0 and 1 and should sum to one.

## FOR EXAMPLE

## Love and expected values

On Valentine's Day the *Quiet Nook* restaurant offers a *Lucky Lovers Special* that could save couples money on their romantic dinners. When the waiter brings the check, he'll also bring the four aces from a deck of cards. He'll shuffle them and lay them out face down on the table. The couple will then get to turn one card over. If it's a black ace, they'll owe the full amount, but if it's the ace of hearts, the waiter will give them a \$20 Lucky Lovers discount. If they first turn over the ace of diamonds (hey—at least it's red!), they'll then get to turn over one of the remaining cards, earning a \$10 discount for finding the ace of hearts this time.

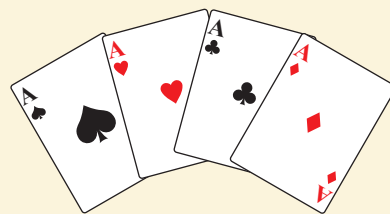
**Question:** Based on a probability model for the size of the Lucky Lovers discounts the restaurant will award, what's the expected discount for a couple?

Let  $X$  = the Lucky Lovers discount. The probabilities of the three outcomes are:

$$P(X = 20) = P(A \heartsuit) = \frac{1}{4}$$

$$\begin{aligned} P(X = 10) &= P(A \spadesuit, \text{ then } A \heartsuit) = P(A \spadesuit) \cdot P(A \heartsuit | A \spadesuit) \\ &= \frac{1}{4} \cdot \frac{1}{3} = \frac{1}{12} \end{aligned}$$

$$P(X = 0) = P(X \neq 20 \text{ or } 10) = 1 - \left( \frac{1}{4} + \frac{1}{12} \right) = \frac{2}{3}$$



My probability model is:

Outcome	A $\heartsuit$	A $\spadesuit$ , then A $\heartsuit$	Black Ace
$x$	20	10	0
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{2}{3}$

$$E(X) = 20 \cdot \frac{1}{4} + 10 \cdot \frac{1}{12} + 0 \cdot \frac{2}{3} = \frac{70}{12} \approx 5.83$$

Couples dining at the Quiet Nook can expect an average discount of \$5.83.



## JUST CHECKING

- One of the authors took his minivan in for repair recently because the air conditioner was cutting out intermittently. The mechanic identified the problem as dirt in a control unit. He said that in about 75% of such cases, drawing down and then recharging the coolant a couple of times cleans up the problem—and costs only \$60. If that fails, then the control unit must be replaced at an additional cost of \$100 for parts and \$40 for labor.
  - Define the random variable and construct the probability model.
  - What is the expected value of the cost of this repair?
  - What does that mean in this context?

Oh—in case you were wondering—the \$60 fix worked!

## First Center, Now Spread . . .

Of course, this expected value (or mean) is not what actually happens to any *particular* policyholder. No individual policy actually costs the company \$20. We are dealing with random events, so some policyholders receive big payouts, others nothing. Because the insurance company must anticipate this variability, it needs to know the *standard deviation* of the random variable.

For data, we calculated the **standard deviation** by first computing the deviation from the mean and squaring it. We do that with (discrete) random variables as well. First, we find the deviation of each payout from the mean (expected value):

Policyholder Outcome	Payout $x$	Probability $P(X = x)$	Deviation $(x - \mu)$
Death	10,000	$\frac{1}{1000}$	$(10,000 - 20) = 9980$
Disability	5000	$\frac{2}{1000}$	$(5000 - 20) = 4980$
Neither	0	$\frac{997}{1000}$	$(0 - 20) = -20$

Next, we square each deviation. **The variance** is the expected value of those squared deviations, so we multiply each by the appropriate probability and sum those products. That gives us the variance of  $X$ . Here's what it looks like:

$$\text{Var}(X) = 9980^2 \left( \frac{1}{1000} \right) + 4980^2 \left( \frac{2}{1000} \right) + (-20)^2 \left( \frac{997}{1000} \right) = 149,600.$$

Finally, we take the square root to get the standard deviation:

$$SD(X) = \sqrt{149,600} \approx \$386.78.$$

The insurance company can expect an average payout of \$20 per policy, with a standard deviation of \$386.78.

Think about that. The company charges \$50 for each policy and expects to pay out \$20 per policy. Sounds like an easy way to make \$30. In fact, most of the time (probability 997/1000) the company pockets the entire \$50. But would you consider selling your roommate such a policy? The problem is that occasionally the company loses big. With probability 1/1000, it will pay out \$10,000, and with probability 2/1000, it will pay out \$5000. That may be more risk than you're willing to take on. The standard deviation of \$386.78 gives an indication that it's no sure thing. That's a pretty big spread (and risk) for an average profit of \$30.

Here are the formulas for what we just did. Because these are parameters of our probability model, the variance and standard deviation can also be written as  $\sigma^2$  and  $\sigma$ . You should recognize both kinds of notation.

$$\begin{aligned} \sigma^2 &= \text{Var}(X) = \sum (x - \mu)^2 P(x) \\ \sigma &= SD(X) = \sqrt{\text{Var}(X)} \end{aligned}$$

## FOR EXAMPLE

## Finding the standard deviation

**Recap:** Here's the probability model for the Lucky Lovers restaurant discount.

Outcome	A♥	A♠, then A♥	Black Ace
$x$	20	10	0
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{2}{3}$

We found that couples can expect an average discount of  $\mu = \$5.83$ .

**Question:** What's the standard deviation of the discounts?

First find the variance:  $\text{Var}(X) = \sum (x - \mu)^2 \cdot P(x)$

$$\begin{aligned}
 &= (20 - 5.83)^2 \cdot \frac{1}{4} + (10 - 5.83)^2 \cdot \frac{1}{12} + (0 - 5.83)^2 \cdot \frac{2}{3} \\
 &\approx 74.306.
 \end{aligned}$$

So,  $SD(X) = \sqrt{74.306} \approx \$8.62$

Couples can expect the Lucky Lovers discounts to average \$5.83, with a standard deviation of \$8.62.

## STEP-BY-STEP EXAMPLE

## Expected Values and Standard Deviations for Discrete Random Variables

As the head of inventory for Knowway computer company, you were thrilled that you had managed to ship 2 computers to your biggest client the day the order arrived. You are horrified, though, to find out that someone had restocked refurbished computers in with the new computers in your storeroom. The shipped computers were selected randomly from the 15 computers in stock, but 4 of those were actually refurbished.

If your client gets 2 new computers, things are fine. If the client gets one refurbished computer, it will be sent back at your expense—\$100—and you can replace it. However, if both computers are refurbished, the client will cancel the order this month and you'll lose a total of \$1000.

**Question:** What's the expected value and the standard deviation of the company's loss?



**Plan** State the problem.

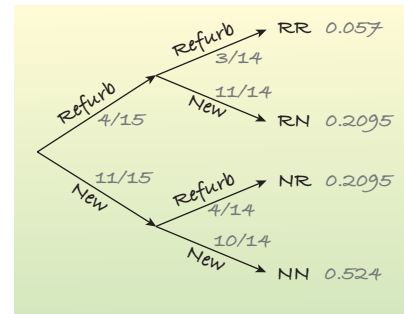
I want to find the company's expected loss for shipping refurbished computers and the standard deviation.

**Variable** Define the random variable.

Let  $X =$  amount of loss.

**Plot** Make a picture. This is another job for tree diagrams.

If you prefer calculation to drawing, find  $P(\text{NN})$  and  $P(\text{RR})$ , then use the Complement Rule to find  $P(\text{NR or RN})$ .



**Model** List the possible values of the random variable, and determine the probability model.

Outcome	x	P(X = x)
Two refurb	1000	$P(\text{RR}) = 0.057$
One refurb	100	$P(\text{NR} \cup \text{RN}) = 0.2095 + 0.2095 = 0.419$
New/new	0	$P(\text{NN}) = 0.524$



**Mechanics** Find the expected value.

Find the variance.

Find the standard deviation.

$$E(X) = 0(0.524) + 100(0.419) + 1000(0.057) = \$98.90$$

$$\begin{aligned} \text{Var}(X) &= (0 - 98.90)^2(0.524) \\ &\quad + (100 - 98.90)^2(0.419) \\ &\quad + (1000 - 98.90)^2(0.057) \\ &= 51,408.79 \end{aligned}$$

$$SD(X) = \sqrt{51,408.79} = \$226.735$$



**Conclusion** Interpret your results in context.



Both numbers seem reasonable. The expected value of \$98.90 is between the extremes of \$0 and \$1000, and there's great variability in the outcome values.

I expect this mistake to cost the firm \$98.90, with a standard deviation of \$226.74. The large standard deviation reflects the fact that there's a pretty large range of possible losses.



**TI Tips**

L1	L2	L3	Z
0	.52381		----
100	41905		
1000			
----			
L2(3) = 4/15 * 3/14			

**Finding the mean and SD of a random variable**

You can easily calculate means and standard deviations for a random variable with your TI. Let's do the Knowway computer example.

- Enter the values of the variable in a list, say, **L1**: 0, 100, 1000.
- Enter the probability model in another list, say, **L2**. Notice that you can enter the probabilities as fractions. For example, multiplying along the top branches

```
1-Var Stats L1,L2
Σ
```

```
1-Var Stats
x̄=99.04761905
Σx=99.04761905
Σx²=61333.3333
Sx=
σx=226.986569
↓n=1
```

of the tree gives the probability of a \$1000 loss to be  $\frac{4}{15} \cdot \frac{3}{14}$ . When you enter that, the TI will automatically calculate the probability as a decimal!

- Under the **STAT CALC** menu, ask for **1-Var Stats L1,L2**.

Now you see the mean and standard deviation (along with some other things). Don't fret that the calculator's mean and standard deviation aren't precisely the same as the ones we found. Such minor differences can arise whenever we round off probabilities to do the work by hand.

Beware: Although the calculator knows enough to call the standard deviation  $\sigma$ , it uses  $\bar{x}$  where it should say  $\mu$ . Make sure you don't make that mistake!

## More About Means and Variances

Our insurance company expected to pay out an average of \$20 per policy, with a standard deviation of about \$387. If we take the \$50 premium into account, we see the company makes a profit of  $50 - 20 = \$30$  per policy. Suppose the company lowers the premium by \$5 to \$45. It's pretty clear that the expected profit also drops an average of \$5 per policy, to  $45 - 20 = \$25$ .

What about the standard deviation? We know that adding or subtracting a constant from data shifts the mean but doesn't change the variance or standard deviation. The same is true of random variables.<sup>1</sup>

$$E(X \pm c) = E(X) \pm c \quad \text{Var}(X \pm c) = \text{Var}(X).$$

### FOR EXAMPLE

#### Adding a constant

**Recap:** We've determined that couples dining at the *Quiet Nook* can expect Lucky Lovers discounts averaging \$5.83 with a standard deviation of \$8.62. Suppose that for several weeks the restaurant has also been distributing coupons worth \$5 off any one meal (one discount per table).

**Question:** If every couple dining there on Valentine's Day brings a coupon, what will be the mean and standard deviation of the total discounts they'll receive?

Let  $D$  = total discount (Lucky Lovers plus the coupon); then  $D = X + 5$ .

$$\begin{aligned} E(D) &= E(X + 5) = E(X) + 5 = 5.83 + 5 = \$10.83 \\ \text{Var}(D) &= \text{Var}(X + 5) = \text{Var}(X) = 8.62^2 \\ \text{SD}(D) &= \sqrt{\text{Var}(X)} = \$8.62 \end{aligned}$$

Couples with the coupon can expect total discounts averaging \$10.83. The standard deviation is still \$8.62.

Back to insurance . . . What if the company decides to double all the payouts—that is, pay \$20,000 for death and \$10,000 for disability? This would double the average payout per policy and also increase the variability in payouts. We have seen that multiplying or dividing all data values by a constant changes both the mean and the standard deviation by the same factor. Variance, being the square of standard deviation, changes by the square of the constant. The same is true of random variables. In general, multiplying each value of a random variable by a

<sup>1</sup> The rules in this section are true for both discrete and continuous random variables.

constant multiplies the mean by that constant and the variance by the *square* of the constant.

$$E(aX) = aE(X) \quad \text{Var}(aX) = a^2\text{Var}(X)$$

### FOR EXAMPLE

#### Double the love

**Recap:** On Valentine's Day at the *Quiet Nook*, couples may get a Lucky Lovers discount averaging \$5.83 with a standard deviation of \$8.62. When two couples dine together on a single check, the restaurant doubles the discount offer—\$40 for the ace of hearts on the first card and \$20 on the second.

**Question:** What are the mean and standard deviation of discounts for such foursomes?

$$E(2X) = 2E(X) = 2(5.83) = \$11.66$$

$$\text{Var}(2x) = 2^2\text{Var}(x) = 2^2 \cdot 8.62^2 = 297.2176$$

$$\text{SD}(2X) = \sqrt{297.2176} = \$17.24$$

If the restaurant doubles the discount offer, two couples dining together can expect to save an average of \$11.66 with a standard deviation of \$17.24.

This insurance company sells policies to more than just one person. How can we figure means and variances for a collection of customers? For example, how can the company find the total expected value (and standard deviation) of policies taken over all policyholders? Consider a simple case: just two customers, Mr. Ecks and Ms. Wye. With an expected payout of \$20 on each policy, we might predict a total of  $\$20 + \$20 = \$40$  to be paid out on the two policies. Nothing surprising there. **The expected value of the sum is the sum of the expected values.**

$$E(X + Y) = E(X) + E(Y).$$

The variability is another matter. Is the risk of insuring two people the same as the risk of insuring one person for twice as much? We wouldn't expect both clients to die or become disabled in the same year. Because we've spread the risk, the standard deviation should be smaller. Indeed, this is the fundamental principle behind insurance. By spreading the risk among many policies, a company can keep the standard deviation quite small and predict costs more accurately.

But how much smaller is the standard deviation of the sum? It turns out that, if the random variables are independent, there is a simple Addition Rule for variances: **The variance of the sum of two independent random variables is the sum of their individual variances.**

For Mr. Ecks and Ms. Wye, the insurance company can expect their outcomes to be independent, so (using  $X$  for Mr. Ecks's payout and  $Y$  for Ms. Wye's)

$$\begin{aligned} \text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y) \\ &= 149,600 + 149,600 \\ &= 299,200. \end{aligned}$$

If they had insured only Mr. Ecks for twice as much, there would only be one outcome rather than two *independent* outcomes, so the variance would have been

$$\text{Var}(2X) = 2^2\text{Var}(X) = 4 \times 149,600 = 598,400, \text{ or}$$

twice as big as with two independent policies.

Of course, variances are in squared units. The company would prefer to know standard deviations, which are in dollars. The standard deviation of the payout for two independent policies is  $\sqrt{299,200} = \$546.99$ . But the standard deviation



of the payout for a single policy of twice the size is  $\sqrt{598,400} = \$773.56$ , or about 40% more.

If the company has two customers, then, it will have an expected annual total payout of \$40 with a standard deviation of about \$547.

## FOR EXAMPLE

### Adding the discounts

**Recap:** The Valentine's Day Lucky Lovers discount for couples averages \$5.83 with a standard deviation of \$8.62. We've seen that if the restaurant doubles the discount offer for two couples dining together on a single check, they can expect to save \$11.66 with a standard deviation of \$17.24. Some couples decide instead to get separate checks and pool their two discounts.

**Question:** You and your amour go to this restaurant with another couple and agree to share any benefit from this promotion. Does it matter whether you pay separately or together?

Let  $X_1$  and  $X_2$  represent the two separate discounts, and  $T$  the total; then  $T = X_1 + X_2$ .

$$E(T) = E(X_1 + X_2) = E(X_1) + E(X_2) = 5.83 + 5.83 = \$11.66,$$

so the expected saving is the same either way.

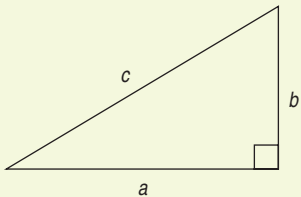
The cards are reshuffled for each couple's turn, so the discounts couples receive are independent. It's okay to add the variances:

$$\begin{aligned} \text{Var}(T) &= \text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) = 8.62^2 + 8.62^2 = 148.6088 \\ \text{SD}(T) &= \sqrt{148.6088} = \$12.19 \end{aligned}$$

When two couples get separate checks, there's less variation in their total discount. The standard deviation is \$12.19, compared to \$17.24 for couples who play for the double discount on a single check. It does, therefore, matter whether they pay separately or together.

### Pythagorean Theorem of Statistics

We often use the standard deviation to measure variability, but when we add independent random variables, we use their variances. Think of the Pythagorean Theorem. In a right triangle (only), the square of the length of the hypotenuse is the sum of the squares of the lengths of the other two sides:



$$c^2 = a^2 + b^2.$$

For independent random variables (only), the square of the standard deviation of their sum is the sum of the squares of their standard deviations:

$$\text{SD}^2(X + Y) = \text{SD}^2(X) + \text{SD}^2(Y).$$

It's simpler to write this with variances:

For independent random variables,  $X$  and  $Y$ ,  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ .

In general,

- ▶ The mean of the sum of two random variables is the sum of the means.
- ▶ The mean of the difference of two random variables is the difference of the means.
- ▶ If the random variables are independent, the variance of their sum or difference is always the sum of the variances.

$$E(X \pm Y) = E(X) \pm E(Y) \quad \text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$$

Wait a minute! Is that third part correct? Do we always *add* variances? Yes. Think about the two insurance policies. Suppose we want to know the mean and standard deviation of the *difference* in payouts to the two clients. Since each policy has an expected payout of \$20, the expected difference is  $20 - 20 = \$0$ . If we also subtract variances, we get \$0, too, and that surely doesn't make sense. Note that if the outcomes for the two clients are independent, the difference in payouts could range from  $\$10,000 - \$0 = \$10,000$  to  $\$0 - \$10,000 = -\$10,000$ , a spread of \$20,000. The variability in differences increases as much as the variability in sums. If the company has two customers, the difference in payouts has a mean of \$0 and a standard deviation of about \$547 (again).

## FOR EXAMPLE

## Working with differences

**Recap:** The Lucky Lovers discount at the *Quiet Nook* averages \$5.83 with a standard deviation of \$8.62. Just up the street, the *Wise Fool* restaurant has a competing Lottery of Love promotion. There a couple can select a specially prepared chocolate from a large bowl and unwrap it to learn the size of their discount. The restaurant's manager says the discounts vary with an average of \$10.00 and a standard deviation of \$15.00.

**Question:** How much more can you expect to save at the *Wise Fool*? With what standard deviation?

Let  $W$  = discount at the *Wise Fool*,  $X$  = the discount at the *Quiet Nook*, and  $D$  = the difference:  $D = W - X$ . These are different promotions at separate restaurants, so the outcomes are independent.

$$\begin{aligned} E(W - X) &= E(W) - E(X) = 10.00 - 5.83 = \$4.17 \\ SD(W - X) &= \sqrt{\text{Var}(W - X)} \\ &= \sqrt{\text{Var}(W) + \text{Var}(X)} \\ &= \sqrt{15^2 + 8.62^2} \\ &\approx \$17.30 \end{aligned}$$

Discounts at the *Wise Fool* will average \$4.17 more than at the *Quiet Nook*, with a standard deviation of \$17.30.

**For random variables, does  $X + X + X = 3X$ ?** Maybe, but be careful. As we've just seen, insuring one person for \$30,000 is not the same risk as insuring three people for \$10,000 each. When each instance represents a different outcome for the same random variable, it's easy to fall into the trap of writing all of them with the same symbol. Don't make this common mistake. Make sure you write each instance as a *different* random variable. Just because each random variable describes a similar situation doesn't mean that each random outcome will be the same.

These are *random* variables, not the variables you saw in Algebra. Being random, they take on different values each time they're evaluated. So what you really mean is  $X_1 + X_2 + X_3$ . Written this way, it's clear that the sum shouldn't necessarily equal 3 times *anything*.

## FOR EXAMPLE

## Summing a series of outcomes

**Recap:** The *Quiet Nook*'s Lucky Lovers promotion offers couples discounts averaging \$5.83 with a standard deviation of \$8.62. The restaurant owner is planning to serve 40 couples on Valentine's Day.

**Question:** What's the expected total of the discounts the owner will give? With what standard deviation?

Let  $X_1, X_2, X_3, \dots, X_{40}$  represent the discounts to the 40 couples, and  $T$  the total of all the discounts. Then:

$$\begin{aligned} T &= X_1 + X_2 + X_3 + \dots + X_{40} \\ E(T) &= E(X_1 + X_2 + X_3 + \dots + X_{40}) \\ &= E(X_1) + E(X_2) + E(X_3) + \dots + E(X_{40}) \\ &= 5.83 + 5.83 + 5.83 + \dots + 5.83 \\ &= \$233.20 \end{aligned}$$

Reshuffling cards between couples makes the discounts independent, so:

$$\begin{aligned} SD(T) &= \sqrt{\text{Var}(X_1 + X_2 + X_3 + \dots + X_{40})} \\ &= \sqrt{\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) + \dots + \text{Var}(X_{40})} \\ &= \sqrt{8.62^2 + 8.62^2 + 8.62^2 + \dots + 8.62^2} \\ &\approx \$54.52 \end{aligned}$$

The restaurant owner can expect the 40 couples to win discounts totaling \$233.20, with a standard deviation of \$54.52.



## JUST CHECKING

2. Suppose the time it takes a customer to get and pay for seats at the ticket window of a baseball park is a random variable with a mean of 100 seconds and a standard deviation of 50 seconds. When you get there, you find only two people in line in front of you.
- How long do you expect to wait for your turn to get tickets?
  - What's the standard deviation of your wait time?
  - What assumption did you make about the two customers in finding the standard deviation?

### STEP-BY-STEP EXAMPLE

### Hitting the Road: Means and Variances

You're planning to spend next year wandering through the mountains of Kyrgyzstan. You plan to sell your used SUV so you can purchase an off-road Honda motor scooter when you get there. Used SUVs of the year and mileage of yours are selling for a mean of \$6940 with a standard deviation of \$250. Your research shows that scooters in Kyrgyzstan are going for about 65,000 Kyrgyzstan som with a standard deviation of 500 som. One U.S. dollar is worth about 38.5 Kyrgyzstan som (38 som and 50 tlyn).

**Question:** How much cash can you expect to pocket after you sell your SUV and buy the scooter?



**Plan** State the problem.

**Variables** Define the random variables.

Write an appropriate equation.  
Think about the assumptions.

I want to model how much money I'd have (in som) after selling my SUV and buying the scooter.

Let  $A$  = sale price of my SUV (in dollars),  
 $B$  = price of a scooter (in som), and  
 $D$  = profit (in som)

$$D = 38.5A - B$$

✓ **Independence Assumption:** The prices are independent.



**Mechanics** Find the expected value, using the appropriate rules.

$$\begin{aligned} E(D) &= E(38.5A - B) \\ &= 38.5E(A) - E(B) \\ &= 38.5(6,940) - (65,000) \\ E(D) &= 202,190 \text{ som} \end{aligned}$$

Find the variance, using the appropriate rules. Be sure to check the assumptions first!

Since sale and purchase prices are independent,

$$\begin{aligned} \text{Var}(D) &= \text{Var}(38.5A - B) \\ &= \text{Var}(38.5A) + \text{Var}(B) \\ &= (38.5)^2 \text{Var}(A) + \text{Var}(B) \\ &= 1482.25(250)^2 + (500)^2 \\ \text{Var}(D) &= 92,890,625 \end{aligned}$$

Find the standard deviation.

$$SD(D) = \sqrt{92,890,625} = 9637.98 \text{ som}$$



**Conclusion** Interpret your results in context. (Here that means talking about dollars.)

I can expect to clear about 202,190 som (\$5252) with a standard deviation of 9638 som (\$250).



Given the initial cost estimates, the mean and standard deviation seem reasonable.

## Continuous Random Variables

**A S**

**Activity: Numeric**

**Outcomes.** You've seen how to simulate discrete random outcomes. There's a tool for simulating continuous outcomes, too.

**A S**

**Activity: Means of Random Variables.** Experiment with continuous random variables to learn how their expected values behave.

A company manufactures small stereo systems. At the end of the production line, the stereos are packaged and prepared for shipping. Stage 1 of this process is called “packing.” Workers must collect all the system components (a main unit, two speakers, a power cord, an antenna, and some wires), put each in plastic bags, and then place everything inside a protective styrofoam form. The packed form then moves on to Stage 2, called “boxing.” There, workers place the form and a packet of instructions in a cardboard box, close it, then seal and label the box for shipping.

The company says that times required for the packing stage can be described by a Normal model with a mean of 9 minutes and standard deviation of 1.5 minutes. The times for the boxing stage can also be modeled as Normal, with a mean of 6 minutes and standard deviation of 1 minute.

This is a common way to model events. Do our rules for random variables apply here? What's different? We no longer have a list of discrete outcomes, with their associated probabilities. Instead, we have **continuous random variables that can take on any value**. Now any single value won't have a probability. We saw this back in Chapter 6 when we first saw the Normal model (although we didn't talk then about “random variables” or “probability”). We know that the probability that  $z = 1.5$  doesn't make sense, but we *can* talk about the probability that  $z$  lies *between* 0.5 and 1.5. For a Normal random variable, the probability that it falls within an interval is just the area under the Normal curve over that interval.

Some continuous random variables have Normal models; others may be skewed, uniform, or bimodal. Regardless of shape, all continuous random variables have means (which we also call *expected values*) and variances. In this book we won't worry about how to calculate them, but we can still work with models for continuous random variables when we're given these parameters.

The good news is that nearly everything we've said about how discrete random variables behave is true of continuous random variables, as well. **When two independent continuous random variables have Normal models, so does their sum or difference.** This simple fact is a special property of Normal models and is very important. It allows us to apply our knowledge of Normal probabilities to questions about the sum or difference of independent random variables.

## STEP-BY-STEP EXAMPLE

## Packaging Stereos



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

Recall that times required to pack the stereos can be described by a Normal model with a mean of 9 minutes and standard deviation of 1.5 minutes. The times for the boxing stage can also be modeled as Normal, with a mean of 6 minutes and standard deviation of 1 minute.

Questions:

1. What is the probability that packing two consecutive systems takes over 20 minutes?
2. What percentage of the stereo systems take longer to pack than to box?

**Question 1:** What is the probability that packing two consecutive systems takes over 20 minutes?



**Plan** State the problem.

**Variables** Define your random variables.

Write an appropriate equation.

Think about the assumptions. Sums of independent Normal random variables follow a Normal model. Such simplicity isn't true in general.

I want to estimate the probability that packing two consecutive systems takes over 20 minutes.

Let  $P_1$  = time for packing the first system

$P_2$  = time for packing the second

$T$  = total time to pack two systems

$$T = P_1 + P_2$$

✓ **Normal Model Assumption:** We are told that both random variables follow Normal models.

✓ **Independence Assumption:** We can reasonably assume that the two packing times are independent.



**Mechanics** Find the expected value.

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

Find the standard deviation.

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

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

Since the times are independent,

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

$$\text{Var}(T) = 4.50$$

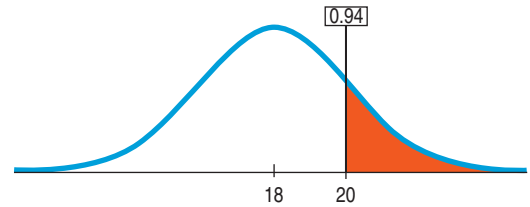
$$\text{SD}(T) = \sqrt{4.50} \approx 2.12 \text{ minutes}$$

I'll model  $T$  with  $N(18, 2.12)$ .

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

Find the z-score for 20 minutes.

Use technology or Table Z to find the probability.



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

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



**Conclusion** Interpret your result in context.

There's a little more than a 17% chance that it will take a total of over 20 minutes to pack two consecutive stereo systems.

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



**Plan** State the question.

I want to estimate the percentage of the stereo systems that take longer to pack than to box.

**Variables** Define your random variables.

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

Write an appropriate equation.

$$D = P - B$$

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

The probability that it takes longer to pack than to box a system is the probability that the difference  $P - B$  is greater than zero.

Don't forget to think about the assumptions.

✓ **Normal Model Assumption:** We are told that both random variables follow Normal models.

✓ **Independence Assumption:** We can assume that the times it takes to pack and to box a system are independent.



**Mechanics** Find the expected value.

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

For the difference of independent random variables, variances add.

Find the standard deviation.

State what model you will use.

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

Find the z-score for 0 minutes, then use Table Z or technology to find the probability.

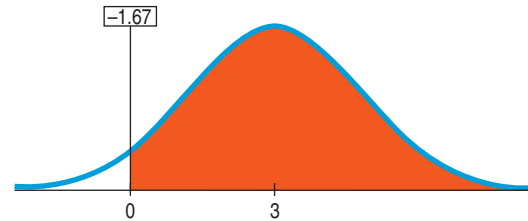
Since the times are independent,

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

$$\text{Var}(D) = 3.25$$

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

I'll model  $D$  with  $N(3, 1.80)$



$$\begin{aligned}z &= \frac{0 - 3}{1.80} = -1.67 \\ P(D > 0) &= P(z > -1.67) = 0.9525\end{aligned}$$



**Conclusion** Interpret your result in context.

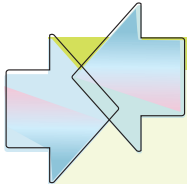
About 95% of all the stereo systems will require more time for packing than for boxing.

## WHAT CAN GO WRONG?

- ▶ **Probability models are still just models.** Models can be useful, but they are not reality. Think about the assumptions behind your models. Are your dice really perfectly fair? (They are probably pretty close.) But when you hear that the probability of a nuclear accident is  $1/10,000,000$  per year, is that likely to be a precise value? Question probabilities as you would data.
- ▶ **If the model is wrong, so is everything else.** Before you try to find the mean or standard deviation of a random variable, check to make sure the probability model is reasonable. As a start, the probabilities in your model should add up to 1. If not, you may have calculated a probability incorrectly or left out a value of the random variable. For instance, in the insurance example, the description mentions only death and disability. Good health is by far the most likely outcome, not to mention the best for both you and the insurance company (who gets to keep your money). Don't overlook that.
- ▶ **Don't assume everything's Normal.** Just because a random variable is continuous or you happen to know a mean and standard deviation doesn't mean that a Normal model will be useful. You must *Think* about whether the **Normality Assumption** is justified. Using a Normal model when it really does not apply will lead to wrong answers and misleading conclusions.

To find the expected value of the sum or difference of random variables, we simply add or subtract means. Center is easy; spread is trickier. Watch out for some common traps.

- ▶ **Watch out for variables that aren't independent.** You can add expected values of *any* two random variables, but you can only add variances of independent random variables. Suppose a survey includes questions about the number of hours of sleep people get each night and also the number of hours they are awake each day. From their answers, we find the mean and standard deviation of hours asleep and hours awake. The expected total must be 24 hours; after all, people are either asleep or awake.<sup>2</sup> The means still add just fine. Since all the totals are exactly 24 hours, however, the standard deviation of the total will be 0. We can't add variances here because the number of hours you're awake depends on the number of hours you're asleep. Be sure to check for independence before adding variances.
- ▶ **Don't forget: Variances of independent random variables add. Standard deviations don't.**
- ▶ **Don't forget: Variances of independent random variables add, even when you're looking at the difference between them.**
- ▶ **Don't write independent instances of a random variable with notation that looks like they are the same variables.** Make sure you write each instance as a different random variable. Just because each random variable describes a similar situation doesn't mean that each random outcome will be the same. These are *random* variables, not the variables you saw in Algebra. Write  $X_1 + X_2 + X_3$  rather than  $X + X + X$ .



## CONNECTIONS

We've seen means, variances, and standard deviations of data. We know that they estimate parameters of models for these data. Now we're looking at the probability models directly. We have only parameters because there are no data to summarize.

It should be no surprise that expected values and standard deviations adjust to shifts and changes of units in the same way as the corresponding data summaries. The fact that we can add variances of independent random quantities is fundamental and will explain why a number of statistical methods work the way they do.

## WHAT HAVE WE LEARNED?



We've learned to work with random variables. We can use the probability model for a discrete random variable to find its expected value and its standard deviation.

We've learned that the mean of the sum or difference of two random variables, discrete or continuous, is just the sum or difference of their means. And we've learned the Pythagorean Theorem of Statistics: *For independent random variables*, the variance of their sum or difference is always the *sum* of their variances.

Finally, we've learned that Normal models are once again special. Sums or differences of Normally distributed random variables also follow Normal models.

<sup>2</sup> Although some students do manage to attain a state of consciousness somewhere between sleeping and wakefulness during Statistics class.



## Terms

Random variable	366. A random variable assumes any of several different numeric values as a result of some random event. Random variables are denoted by a capital letter such as $X$ .
Discrete random variable	366. A random variable that can take one of a finite number <sup>3</sup> of distinct outcomes is called a discrete random variable.
Continuous random variable	366, 367. A random variable that can take any numeric value within a range of values is called a continuous random variable. The range may be infinite or bounded at either or both ends.
Probability model	366. The probability model is a function that associates a probability $P$ with each value of a discrete random variable $X$ , denoted $P(X = x)$ , or with any interval of values of a continuous random variable.
Expected value	367. The expected value of a random variable is its theoretical long-run average value, the center of its model. Denoted $\mu$ or $E(X)$ , it is found (if the random variable is discrete) by summing the products of variable values and probabilities: $\mu = E(X) = \sum xP(x).$
Variance	369. The variance of a random variable is the expected value of the squared deviation from the mean. For discrete random variables, it can be calculated as: $\sigma^2 = \text{Var}(X) = \sum (x - \mu)^2 P(x).$
Standard deviation	369. The standard deviation of a random variable describes the spread in the model, and is the square root of the variance: $\sigma = \text{SD}(X) = \sqrt{\text{Var}(X)}.$
Changing a random variable by a constant:	372. $E(X \pm c) = E(X) \pm c$ $\text{Var}(X \pm c) = \text{Var}(X)$ 373. $E(aX) = aE(X)$ $\text{Var}(aX) = a^2\text{Var}(X)$
Adding or subtracting random variables:	373. $E(X \pm Y) = E(X) \pm E(Y)$ 374. If $X$ and $Y$ are independent, $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$ . 374. (The Pythagorean Theorem of Statistics)

## Skills

THINK

- ▶ Be able to recognize random variables.
- ▶ Understand that random variables must be independent in order to determine the variability of their sum or difference by adding variances.

SHOW

- ▶ Be able to find the probability model for a discrete random variable.
- ▶ Know how to find the mean (expected value) and the variance of a random variable.
- ▶ Always use the proper notation for these population parameters:  $\mu$  or  $E(X)$  for the mean, and  $\sigma$ ,  $\text{SD}(X)$ ,  $\sigma^2$ , or  $\text{Var}(X)$  when discussing variability.
- ▶ Know how to determine the new mean and standard deviation after adding a constant, multiplying by a constant, or adding or subtracting two independent random variables.

TELL

- ▶ Be able to interpret the meaning of the expected value and standard deviation of a random variable in the proper context.

<sup>3</sup> Technically, there could be an infinite number of outcomes, as long as they're *countable*. Essentially that means we can imagine listing them all in order, like the counting numbers 1, 2, 3, 4, 5, . . .