Suppose the variables `handedness` and `gender` are independent, i.e. knowing someone's `gender` provides no useful information about their `handedness` and vice-versa. Then we can compute whether a randomly selected person is right-handed and female[21] using the Multiplication Rule:

$$\begin{aligned} P(\text{right-handed and female}) &= P(\text{right-handed}) \times P(\text{female}) \\ &= 0.91 \times 0.50 = 0.455 \end{aligned}$$

⊙ **Exercise 2.32** Three people are selected at random.[22]

(a) What is the probability that the first person is male and right-handed?

(b) What is the probability that the first two people are male and right-handed?.

(c) What is the probability that the third person is female and left-handed?

(d) What is the probability that the first two people are male and right-handed and the third person is female and left-handed?

Sometimes we wonder if one outcome provides useful information about another outcome. The question we are asking is, are the occurrences of the two events independent? We say that two events A and B are independent if they satisfy Equation (2.29).

● **Example 2.33** If we shuffle up a deck of cards and draw one, is the event that the card is a heart independent of the event that the card is an ace?

The probability the card is a heart is 1/4 and the probability that it is an ace is 1/13. The probability the card is the ace of hearts is 1/52. We check whether Equation 2.29 is satisfied:

$$P(\heartsuit) \times P(\text{ace}) = \frac{1}{4} \times \frac{1}{13} = \frac{1}{52} = P(\heartsuit \text{ and ace})$$

Because the equation holds, the event that the card is a heart and the event that the card is an ace are independent events.

2.2 Conditional probability (special topic)

Are students more likely to use marijuana when their parents used drugs? The `drug_use` data set contains a sample of 445 cases with two variables, `student` and `parents`, and is summarized in Table 2.11.[23] The `student` variable is either `uses` or `not`, where a student is labeled as `uses` if she has recently used marijuana. The `parents` variable takes the value `used` if at least one of the parents used drugs, including alcohol.

● **Example 2.34** If at least one parent used drugs, what is the chance their child (`student`) uses?

We will estimate this probability using the data. Of the 210 cases in this data set where `parents` = `used`, 125 represent cases where `student` = `uses`:

$$P(\texttt{student} = \texttt{uses} \text{ given } \texttt{parents} = \texttt{used}) = \frac{125}{210} = 0.60$$

[21] The actual proportion of the U.S. population that is `female` is about 50%, and so we use 0.5 for the probability of sampling a woman. However, this probability does differ in other countries.

[22] Brief answers are provided. (a) This can be written in probability notation as P(a randomly selected person is male and right-handed) = 0.455. (b) 0.207. (c) 0.045. (d) 0.0093.

[23] Ellis GJ and Stone LH. 1979. Marijuana Use in College: An Evaluation of a Modeling Explanation. Youth and Society 10:323-334.

		parents		
		used	not	Total
student	uses	125	94	219
	not	85	141	226
	Total	210	235	445

Table 2.11: Contingency table summarizing the drug_use data set.

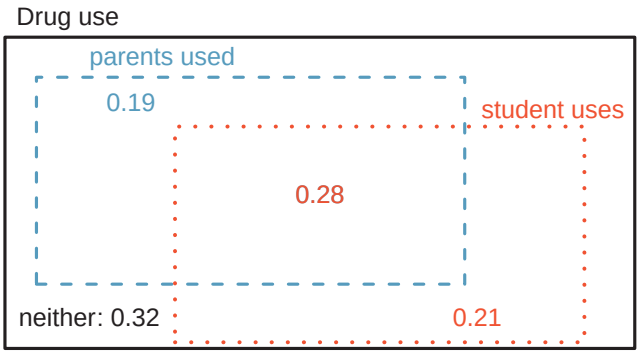

Figure 2.12: A Venn diagram using boxes for the drug_use data set.

Example 2.35 A student is randomly selected from the study and she does not use drugs. What is the probability that at least one of her parents used?

If the student does not use drugs, then she is one of the 226 students in the second row. Of these 226 students, 85 had at least one parent who used drugs:

$$P(\texttt{parents} = \texttt{used} \text{ given } \texttt{student} = \texttt{not}) = \frac{85}{226} = 0.376$$

2.2.1 Marginal and joint probabilities

Table 2.13 includes row and column totals for each variable separately in the drug_use data set. These totals represent **marginal probabilities** for the sample, which are the probabilities based on a single variable without conditioning on any other variables. For instance, a probability based solely on the student variable is a marginal probability:

$$P(\texttt{student} = \texttt{uses}) = \frac{219}{445} = 0.492$$

A probability of outcomes for two or more variables or processes is called a **joint probability**:

$$P(\texttt{student} = \texttt{uses} \text{ and } \texttt{parents} = \texttt{not}) = \frac{94}{445} = 0.21$$

It is common to substitute a comma for "and" in a joint probability, although either is acceptable.

	parents: used	parents: not	Total
student: uses	0.28	0.21	0.49
student: not	0.19	0.32	0.51
Total	0.47	0.53	1.00

Table 2.13: Probability table summarizing parental and student drug use.

Joint outcome	Probability
parents = used, student = uses	0.28
parents = used, student = not	0.19
parents = not, student = uses	0.21
parents = not, student = not	0.32
Total	1.00

Table 2.14: A joint probability distribution for the drug_use data set.

Marginal and joint probabilities
If a probability is based on a single variable, it is a *marginal probability*. The probability of outcomes for two or more variables or processes is called a *joint probability*.

We use **table proportions** to summarize joint probabilities for the drug_use sample. These proportions are computed by dividing each count in Table 2.11 by 445 to obtain the proportions in Table 2.13. The joint probability distribution of the parents and student variables is shown in Table 2.14.

⊙ **Exercise 2.36** Verify Table 2.14 represents a probability distribution: events are disjoint, all probabilities are non-negative, and the probabilities sum to 1.[24]

We can compute marginal probabilities using joint probabilities in simple cases. For example, the probability a random student from the study uses drugs is found by summing the outcomes from Table 2.14 where student = uses:

$$\begin{aligned} &P(\underline{\texttt{student = uses}}) \\ &\quad = P(\texttt{parents = used}, \underline{\texttt{student = uses}}) + \\ &\qquad\quad P(\texttt{parents = not}, \underline{\texttt{student = uses}}) \\ &\quad = 0.28 + 0.21 = 0.49 \end{aligned}$$

2.2.2 Defining conditional probability

There is some connection between drug use of parents and of the student: drug use of one is associated with drug use of the other.[25] In this section, we discuss how to use information about associations between two variables to improve probability estimation.

The probability that a random student from the study uses drugs is 0.49. Could we update this probability if we knew that this student's parents used drugs? Absolutely. To

[24] Each of the four outcome combination are disjoint, all probabilities are indeed non-negative, and the sum of the probabilities is $0.28 + 0.19 + 0.21 + 0.32 = 1.00$.

[25] This is an observational study and no causal conclusions may be reached.

do so, we limit our view to only those 210 cases where parents used drugs and look at the fraction where the student uses drugs:

$$P(\texttt{student} = \texttt{uses} \text{ given } \texttt{parents} = \texttt{used}) = \frac{125}{210} = 0.60$$

We call this a **conditional probability** because we computed the probability under a condition: `parents` = `used`. There are two parts to a conditional probability, **the outcome of interest** and the **condition**. It is useful to think of the condition as information we know to be true, and this information usually can be described as a known outcome or event.

We separate the text inside our probability notation into the outcome of interest and the condition:

$$\begin{aligned} &P(\texttt{student} = \texttt{uses} \text{ given } \texttt{parents} = \texttt{used}) \\ &= P(\texttt{student} = \texttt{uses} \mid \texttt{parents} = \texttt{used}) = \frac{125}{210} = 0.60 \end{aligned} \tag{2.37}$$

$P(A|B)$
Probability of outcome A given B

The vertical bar "|" is read as *given*.

In Equation (2.37), we computed the probability a student uses based on the condition that at least one parent used as a fraction:

$$\begin{aligned} &P(\texttt{student} = \texttt{uses} \mid \texttt{parents} = \texttt{used}) \\ &\quad = \frac{\#\text{ times } \texttt{student} = \texttt{uses} \text{ and } \texttt{parents} = \texttt{used}}{\#\text{ times } \texttt{parents} = \texttt{used}} \\ &\quad = \frac{125}{210} = 0.60 \end{aligned} \tag{2.38}$$

We considered only those cases that met the condition, `parents` = `used`, and then we computed the ratio of those cases that satisfied our outcome of interest, the student uses.

Counts are not always available for data, and instead only marginal and joint probabilities may be provided. For example, disease rates are commonly listed in percentages rather than in a count format. We would like to be able to compute conditional probabilities even when no counts are available, and we use Equation (2.38) as an example demonstrating this technique.

We considered only those cases that satisfied the condition, `parents` = `used`. Of these cases, the conditional probability was the fraction who represented the outcome of interest, `student` = `uses`. Suppose we were provided only the information in Table 2.13 on the preceding page, i.e. only probability data. Then if we took a sample of 1000 people, we would anticipate about 47% or $0.47 \times 1000 = 470$ would meet our information criterion. Similarly, we would expect about 28% or $0.28 \times 1000 = 280$ to meet both the information criterion and represent our outcome of interest. Thus, the conditional probability could be computed:

$$\begin{aligned} P(\texttt{student} = \texttt{uses} \mid \texttt{parents} = \texttt{used}) &= \frac{\#\ (\texttt{student} = \texttt{uses} \text{ and } \texttt{parents} = \texttt{used})}{\#\ (\texttt{parents} = \texttt{used})} \\ &= \frac{280}{470} = \frac{0.28}{0.47} = 0.60 \end{aligned} \tag{2.39}$$

In Equation (2.39), we examine exactly the fraction of two probabilities, 0.28 and 0.47, which we can write as

$$P(\texttt{student} = \texttt{uses} \text{ and } \texttt{parents} = \texttt{used}) \quad \text{and} \quad P(\texttt{parents} = \texttt{used}).$$

The fraction of these probabilities represents our general formula for conditional probability.

Conditional Probability
The conditional probability of the outcome of interest A given condition B is computed as the following:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \tag{2.40}$$

⊙ **Exercise 2.41** (a) Write out the following statement in conditional probability notation: "*The probability a random case has* `parents` $=$ `not` *if it is known that* `student` $=$ `not`". Notice that the condition is now based on the student, not the parent. (b) Determine the probability from part (a). Table 2.13 on page 81 may be helpful.[26]

⊙ **Exercise 2.42** (a) Determine the probability that one of the parents had used drugs if it is known the student does not use drugs. (b) Using the answers from part (a) and Exercise 2.41(b), compute

$$P(\texttt{parents} = \texttt{used}|\texttt{student} = \texttt{not}) + P(\texttt{parents} = \texttt{not}|\texttt{student} = \texttt{not})$$

(c) Provide an intuitive argument to explain why the sum in (b) is 1.[27]

⊙ **Exercise 2.43** The data indicate that drug use of parents and children are associated. Does this mean the drug use of parents causes the drug use of the students?[28]

2.2.3 Smallpox in Boston, 1721

The `smallpox` data set provides a sample of 6,224 individuals from the year 1721 who were exposed to smallpox in Boston.[29] Doctors at the time believed that inoculation, which involves exposing a person to the disease in a controlled form, could reduce the likelihood of death.

Each case represents one person with two variables: `inoculated` and `result`. The variable `inoculated` takes two levels: `yes` or `no`, indicating whether the person was inoculated or not. The variable `result` has outcomes `lived` or `died`. These data are summarized in Tables 2.15 and 2.16.

⊙ **Exercise 2.44** Write out, in formal notation, the probability a randomly selected person who was not inoculated died from smallpox, and find this probability.[30]

[26] (a) $P(\texttt{parent} = \texttt{not}|\texttt{student} = \texttt{not})$. (b) Equation (2.40) for conditional probability indicates we should first find $P(\texttt{parents} = \texttt{not} \text{ and } \texttt{student} = \texttt{not}) = 0.32$ and $P(\texttt{student} = \texttt{not}) = 0.51$. Then the ratio represents the conditional probability: $0.32/0.51 = 0.63$.

[27] (a) This probability is $\frac{P(\texttt{parents} = \texttt{used and student} = \texttt{not})}{P(\texttt{student} = \texttt{not})} = \frac{0.19}{0.51} = 0.37$. (b) The total equals 1. (c) Under the condition the student does not use drugs, the parents must either use drugs or not. The complement still appears to work *when conditioning on the same information.*

[28] No. This was an observational study. Two potential confounding variables include `income` and `region`. Can you think of others?

[29] Fenner F. 1988. *Smallpox and Its Eradication (History of International Public Health, No. 6)*. Geneva: World Health Organization. ISBN 92-4-156110-6.

[30] $P(\texttt{result} = \texttt{died} \mid \texttt{inoculated} = \texttt{no}) = \frac{P(\texttt{result} = \texttt{died and inoculated} = \texttt{no})}{P(\texttt{inoculated} = \texttt{no})} = \frac{0.1356}{0.9608} = 0.1411$.

		inoculated		
		yes	no	Total
result	lived	238	5136	5374
	died	6	844	850
	Total	244	5980	6224

Table 2.15: Contingency table for the smallpox data set.

		inoculated		
		yes	no	Total
result	lived	0.0382	0.8252	0.8634
	died	0.0010	0.1356	0.1366
	Total	0.0392	0.9608	1.0000

Table 2.16: Table proportions for the smallpox data, computed by dividing each count by the table total, 6224.

⊙ **Exercise 2.45** Determine the probability that an inoculated person died from smallpox. How does this result compare with the result of Exercise 2.44?[31]

⊙ **Exercise 2.46** The people of Boston self-selected whether or not to be inoculated. (a) Is this study observational or was this an experiment? (b) Can we infer any causal connection using these data? (c) What are some potential confounding variables that might influence whether someone lived or died and also affect whether that person was inoculated?[32]

2.2.4 General multiplication rule

Section 2.1.6 introduced the Multiplication Rule for independent processes. Here we provide the **General Multiplication Rule** for events that might not be independent.

General Multiplication Rule
If A and B represent two outcomes or events, then

$$P(A \text{ and } B) = P(A|B) \times P(B)$$

It is useful to think of A as the outcome of interest and B as the condition.

This General Multiplication Rule is simply a rearrangement of the definition for conditional probability in Equation (2.40) on page 83.

[31] $P(\texttt{result} = \texttt{died} \mid \texttt{inoculated} = \texttt{yes}) = \frac{P(\texttt{result} = \texttt{died and inoculated} = \texttt{yes})}{P(\texttt{inoculated} = \texttt{yes})} = \frac{0.0010}{0.0392} = 0.0255$. The death rate for individuals who were inoculated is only about 1 in 40 while the death rate is about 1 in 7 for those who were not inoculated.

[32] Brief answers: (a) Observational. (b) No, we cannot infer causation from this observational study. (c) Accessibility to the latest and best medical care. There are other valid answers for part (c).