EDD RETRIEVAL RECEIPT

Order: 563983
For: EDD
Copied: 09/14/2020
Shipped: 09/14/2020
Deliver To: RECAPPUL02
Patron E-Mail: rgomila@princeton.edu
Oth Patron Info:
Def PickUp Loc: EDD-ReCAP EDD
Delivery Meth: EDD

Item BarCode: CU25528394
Item Title: Quantitative social science : an introduction \/ Kosuke Imai
Item Author: Imai, Kosuke, author.
Item Call Number: H62 .I5365 2017g
Item Vol/Part:

Article Title: Probability
Article Author: Kosuke Imai
Art Vol/Part: ,
Beg Page: 242 End Page: 310 Total Pages: 0
Other Info:
Notes:

TOTAL COUNT: 1

Chapter 6

Probability

Probability is the very guide of life.
—Cicero, *De Natura*

Until now, we have studied how to identify patterns in data. While some patterns are indisputably clear, in many cases we must figure out ways to distinguish systematic patterns from noise. Noise, also known as random error, is the irrelevant variation that occurs in every real-world data set. Quantifying the degree of statistical uncertainty of empirical findings is the topic for the *next* chapter, but this requires an understanding of probability. Probability is a set of mathematical tools that measure and model randomness in the world. As such, this chapter introduces the derivation of the fundamental rules of probability, with the use of mathematical notation. In the social sciences, we use probability to model the randomly determined nature of various real-world events, and even human behavior and beliefs. Randomness does not necessarily imply complete unpredictability. Rather, our task is to identify systematic patterns from noisy data.

6.1 Probability

We use *probability* as a measure of uncertainty. Probability is based on a set of three simple axioms, from which a countless number of useful theorems have been derived. In this section, we show how to define, interpret, and compute probability.

6.1.1 FREQUENTIST VERSUS BAYESIAN

In everyday life, we often hear statements such as "the probability of winning a coin toss is 50%" and "the probability of Obama winning the 2008 US presidential election is 80%." What do we mean by "probability"? There are at least two different interpretations. One interpretation, which is called the *frequentist* interpretation, states that probability represents the *limit* of relative frequency, defined as the ratio between the number of times the event occurs and the number of trials, in repeated trials under the same conditions. For example, the above statement about coin tosses can be interpreted as follows: if a coin toss is repeatedly conducted under the same conditions,

Figure 6.1. Reverend Thomas Bayes (1701–1761).

the fraction of times a coin lands on heads approaches 0.5 as the number of coin tosses increases. Here, the mathematical term, "limit," represents the value to which a sequence of relative frequencies converges as the number of (hypothetically) repeated experiments approaches infinity.

The frequentist interpretation of probability faces several difficulties. First, it is unclear what we mean by "the same conditions." In the case of coin flips, such conditions may include initial angle and velocity as well as air pressure and temperature. However, if all conditions are identical, then the laws of physics imply that a coin flip will always yield the same outcome. Second, in practice, we can never conduct experiments like coin flips under the exact same conditions infinitely many times. This means that probability may be unable to describe the randomness of many events in the real world. In fact, coin flips may be among the easiest experiments to repeat under nearly identical conditions. Many other events covered in this book happen in dynamic social environments that are constantly changing.

How should we think about the probability of Obama winning the 2008 US presidential election from the frequentist perspective? Since the 2008 US presidential election occurs only once, it is strange to consider a hypothetical scenario in which this particular election occurs repeatedly under the same conditions. In addition, since Obama either wins the election or not, the probability of his victory should be either 0 or 1. Here, what is random is the election forecast (due to sampling variability etc.) not the actual election outcome.

An alternative framework is the *Bayesian* interpretation of probability, named after an 18th century English mathematician and minister, Thomas Bayes (see figure 6.1). According to this paradigm, probability is a measure of one's subjective belief about the likelihood of an event occurring. A probability of 0 means that an individual thinks an event is impossible, whereas a probability of 1 implies that the individual

thinks the event is sure to happen. Any probability value between 0 and 1 indicates the degree to which one feels uncertain about the occurrence of the event. In contrast to the frequentist perspective, the Bayesian framework makes it easy to interpret the statement, "the probability of Obama winning the 2008 US presidential election is x%," because x simply reflects the speaker's subjective belief about the likelihood of Obama's victory.

Critics of Bayesian interpretation argue that if scientists have identical sets of empirical evidence, they should arrive at the same conclusion rather than reporting different probabilities of the same event. Such subjectivity may hinder scientific progress because under the Bayesian framework, probability simply becomes a tool to describe one's belief system. In contrast, Bayesians contend that human beings, including scientists, are inherently subjective, so they should explicitly recognize the role of their subjective beliefs in scientific research.

Regardless of the ongoing controversy about its interpretation, probability was established as a mathematical theory by Soviet mathematician Andrey Kolmogorov in the early 20th century. Since both frequentists and Bayesians use this mathematical theory, the disagreement is about interpretation and is not mathematical.

There are two dominant ways to interpret probability. According to the **frequentist framework**, probability represents the limit of the relative frequency with which an event of interest occurs when the number of experiments repeatedly conducted under the same conditions approaches infinity. The **Bayesian framework**, in contrast, interprets probability as one's subjective belief about the likelihood of event occurrence.

6.1.2 DEFINITION AND AXIOMS

We define probability using the following three concepts: *experiment*, *sample space*, and *event*.

The definition of probability requires the following concepts:

1. **experiment**: an action or a set of actions that produce stochastic events of interest
2. **sample space**: a set of all possible outcomes of the experiment, typically denoted by Ω
3. **event**: a subset of the sample space

We can briefly illustrate each concept using the aforementioned two examples. Flipping a coin or holding an election would be the experiment, while the sample space would be given by {lands on heads, lands on tails} or {Obama wins, McCain wins, a third-party candidate wins}. The mathematical term *set* refers to a collection of distinct objects. An event represents *any* subset of sample space, and hence it may include multiple outcomes. In fact, the entire sample space that contains all outcomes is also an event. Moreover, an event is said to *occur* if the set that defines the event

includes an actual outcome of the experiment. In the election example, events include {Obama wins, McCain wins}, which contains two outcomes and can be understood in English as "either Obama or McCain wins." Since Obama won the election, this event did occur in 2008.

As another example, consider a voter's decision in the 2008 US presidential election as an experiment. The idea is that a voter's decision can be modeled as a stochastic, rather than deterministic, event. By considering all four possible outcomes, we can define the sample space of this experiment as Ω = {abstain, vote for Obama, vote for McCain, vote for a third-party candidate}. Within this sample space, we may consider the occurrence of various events including {vote for Obama, vote for McCain, vote for a third-party candidate} (i.e., "do not abstain") and {abstain, vote for McCain, vote for a third-party candidate} (i.e., "do not vote for Obama").

We now discuss how to compute probability, starting with the simplest case in which all outcomes are equally likely to occur. In this case, the probability of event A occurring, denoted by $P(A)$, can be computed as the ratio of the number of elements in the corresponding set A to that in the entire sample space Ω:

$$P(A) = \frac{\text{number of elements in } A}{\text{number of elements in } \Omega}. \tag{6.1}$$

To illustrate this, consider an experiment of tossing a fair coin 3 times. In this experiment, if we denote {lands on heads} and {lands on tails} as H and T, respectively, then the sample space is equal to the set of 8 outcomes, $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. We can then compute the probability of, for example, landing on heads at least twice by counting the number of elements in the relevant set, $A = \{HHH, HHT, HTH, THH\}$. In this case, therefore, using the formula above we obtain $P(A) = 4/8 = 0.5$.

Having defined probability, we next consider its basic rules or *axioms*. Modern probability theory rests on the following three simple axioms. Remarkably, from these axioms, the entire theory of probability, including all the existing rules and theorems, can be derived.

The **probability axioms** are given by the following three rules:

1. The probability of any event A is nonnegative:

$$P(A) \geq 0.$$

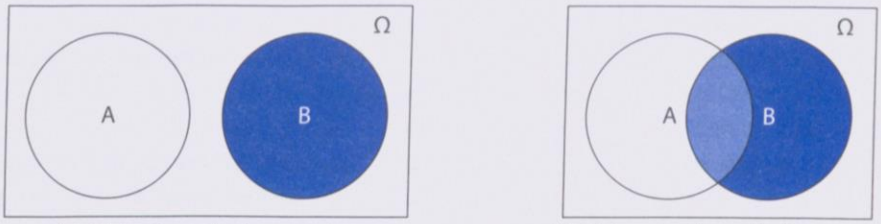
3. The probability that one of the outcomes in the sample space occurs is 1:

$$P(\Omega) = 1.$$

3. (*Addition rule*) If events A and B are mutually exclusive, then

$$P(A \text{ or } B) = P(A) + P(B). \tag{6.2}$$

The first two axioms together imply that probability ranges from 0 to 1. To understand the last axiom, recall the previous example in which the 2008 US presidential



(a) Mutually exclusive events (b) Not mutually exclusive events

Figure 6.2. Venn Diagram. Two events, A and B, can be mutually exclusive, having two disjoint sets of outcomes (left plot) or not mutually exclusive, sharing some outcomes (right plot). The rectangular box represents the sample space Ω. Source: Adapted from example by Uwe Ziegenhagen, http://texample.net.

election is considered as an experiment. "Mutually exclusive" in the last axiom means that two events, A and B, do not share an outcome. As illustrated by the *Venn diagram* (named after John Venn, an English philosopher) in figure 6.2a, mutually exclusive events imply two disjoint sets, meaning that they do not share any element. Consider two events: $A =$ Obama wins and $B =$ McCain wins. Clearly, these two events are mutually exclusive in that both Obama and McCain cannot win at the same time. Hence, we can apply the addition rule to conclude that $P(\{\text{Obama wins}\} \text{ or } \{\text{McCain wins}\}) = P(\text{Obama wins}) + P(\text{McCain wins})$.

Now, consider two events that are not mutually exclusive because they share an outcome: $A =$ Obama loses and $B =$ McCain loses. In this case, the addition rule does not apply because both A and B contain the same outcome: a third-party candidate wins. For events that are not mutually exclusive, we can apply the following general addition rule.

For any given events A and B, the **addition rule** is given by

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B). \tag{6.3}$$

Applying this to the current example, we have $P(\{\text{Obama loses}\}\text{or } \{\text{McCain loses}\}) = P(\text{Obama loses}) + P(\text{McCain loses}) - P(\{\text{Obama loses}\} \text{ and } \{\text{McCain loses}\})$.

This result can be immediately seen from the *Venn diagram* shown in figure 6.2b. In the diagram, we observe that the event, $\{A \text{ or } B\}$, can be decomposed into three mutually exclusive events, $\{A \text{ and } B^c\}$ (white region), $\{B \text{ and } A^c\}$ (dark blue region), and $\{A \text{ and } B\}$ (overlapped light blue region). The superscript c represents the *complement* of a set, which consists of all elements in the sample space except those in the set. For example, A^c represents the collection of all outcomes in the sample space that do not belong to A. The notation $\{A \text{ and } B^c\}$ translates to "all outcomes of A that do not belong to B." Since any outcome in the sample space belongs either to A or A^c, in general, we have

$$P(A^c) = 1 - P(A). \tag{6.4}$$

The equation directly follows from the probability axioms since events A and A^c are mutually exclusive and together they constitute the entire sample space.

Using the third probability axiom, given in equation (6.2), we have

$$P(A \text{ or } B) = P(A \text{ and } B^c) + P(B \text{ and } A^c) + P(A \text{ and } B). \tag{6.5}$$

When A and B are mutually exclusive, $P(A \text{ and } B^c)$ and $P(B \text{ and } A^c)$ reduce to $P(A)$ and $P(B)$, respectively (see figure 6.2a). In addition, we have $P(A \text{ and } B) = 0$ in this mutually exclusive case.

Finally, notice that event A can be decomposed as two mutually exclusive events, $\{A \text{ and } B\}$ (overlapped light blue region) and $\{A \text{ and } B^c\}$ (nonoverlapped white region). This is called the *law of total probability*.

For any given events A and B, the **law of total probability** is given by

$$P(A) = P(A \text{ and } B) + P(A \text{ and } B^c). \tag{6.6}$$

According to the law of total probability, we can write $P(A \text{ and } B^c) = P(A) - P(A \text{ and } B)$ by subtracting $P(A \text{ and } B)$ from both sides of equation (6.6). Similarly, the law of total probability can be applied to event B, yielding $P(B \text{ and } A^c) = P(B) - P(A \text{ and } B)$. Substituting these results into equation (6.5) and simplifying the expression leads to the general addition rule given in equation (6.3). We emphasize that this result is obtained by using the probability axioms alone. In addition, readers are encouraged to confirm the results shown in equations (6.3)–(6.6) using the Venn diagram of figure 6.2.

6.1.3 PERMUTATIONS

When each outcome is equally likely, in order to compute the probability of event A, we need to count the number of elements in event A as well as the total number of elements in the sample space Ω (see equation (6.1)). We introduce a useful counting technique, called *permutations*. Permutations refer to the number of ways in which objects can be arranged. For example, consider three unique objects A, B, and C. There are 6 unique ways to arrange them: $\{ABC, ACB, BAC, BCA, CAB, CBA\}$.

How can we compute the number of permutations without enumerating every arrangement, especially when the number of objects is large? It turns out that there is an easy way to do this. Let's consider the above example of arranging three objects, A, B, and C. First, there are three ways to choose the first object: A, B, or C. Once the first object is selected, there are two ways to choose the second object. Finally, the third object remains, leaving us only one way to choose this last object. We can conceptualize this process as a tree shown in figure 6.3, where the total number of leaves equals the number of permutations. Thus, to compute the number of leaves, we only need to sequentially multiply the number of branches at each level, i.e., $3 \times 2 \times 1$.

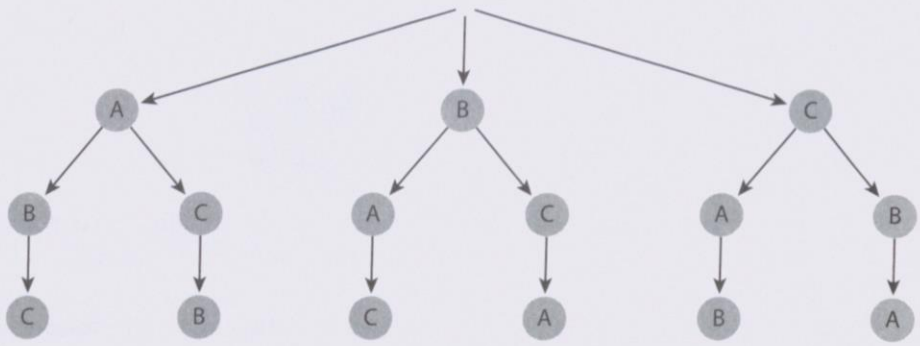


Figure 6.3. A Tree Diagram for Permutations. There are 6 ways to arrange 3 unique objects. Source: Adapted from example by Madit, http://texample.net.

Generalizing this idea, we can compute the number of permutations of k objects out of a set of n unique objects, denoted by ${}_nP_k$ where $k \leq n$, using the following formula.

The number of **permutations** when arranging k objects out of n unique objects is given by

$$ {}_nP_k = n \times (n-1) \times \cdots \times (n-k+2) \times (n-k+1) = \frac{n!}{(n-k)!}. \tag{6.7} $$

In this equation, ! represents the *factorial* operator. When n is a nonnegative integer, $n! = n \times (n-1) \times \cdots \times 2 \times 1$. Note that 0! is defined as 1.

In the previous example, $n = 3$ and $k = 3$. Therefore,

$$ {}_3P_3 = \frac{3!}{0!} = \frac{3 \times 2 \times 1}{1} = 6. $$

As another example, compute the number of ways in which you can arrange 4 cards out of 13 unique cards. This can be computed by setting $n = 13$ and $k = 4$ in equation (6.7):

$$ {}_{13}P_4 = \frac{13!}{(13-4)!} = 13 \times 12 \times 11 \times 10 = 17160. $$

The *birthday problem* is a well-known counterintuitive example of permutations. The problem asks how many people one needs in order for the probability that at least two people have the same birthday to exceed 0.5, assuming that each birthday is equally likely. What is surprising about this problem is that the answer is only 23 people, which is much lower than what most people guess. To solve this problem using permutations, first notice the following relationship:

$$ \begin{aligned} &P(\text{at least two people have the same birthday}) \\ &\quad = 1 - P(\text{nobody has the same birthday}). \end{aligned} \tag{6.8} $$

This equality holds because the event {nobody has the same birthday} is the complement of the event {at least two people have the same birthday} (see equation (6.4)). This means that we only need to compute the probability that nobody has the same birthday.

Let k be the number of people. To compute the probability that nobody has the same birthday, we count the number of ways in which k people can have different birthdays. Since each birthday is assumed to be equally likely, we can use permutations to count the number of ways in which k unique birthdays can be arranged out of 365 days. This is given by ${}_{365}P_k = 365!/(365-k)!$. Applying equation (6.1), we then divide this number by the total number of elements in the sample space. The latter is equal to the total number of ways to select k possibly nonunique birthdays out of 365 days. The first person could have any of 365 days as his/her birthday, and so could any other person. Hence, the denominator is equal to $365 \times 365 \times \cdots \times 365 = 365^k$. Therefore, we have

$$\begin{aligned} &P(\text{nobody has the same birthday}) \\ &\quad = \frac{\#\text{ of ways in which } k \text{ unique birthdays can be arranged}}{\#\text{ of ways in which } k \text{ possibly nonunique birthdays can be arranged}} \\ &\quad = \frac{{}_{365}P_k}{365^k} = \frac{365!}{365^k(365-k)!}. \end{aligned} \tag{6.9}$$

Together with equation (6.8), the solution to the birthday problem is $1 - 365!/\{365^k(365-k)!\}$.

Computing equation (6.9) is not easy even for a moderate value of k because both the denominator and numerator can take extremely large values. In such cases, it is often convenient to use the natural *logarithmic transformation* (see section 3.4.1). For the natural logarithm, $e^A = B$ implies $A = \log B$. In addition, the basic rules of logarithms we use here are

$$\log AB = \log A + \log B, \quad \log \frac{A}{B} = \log A - \log B, \quad \text{and} \quad \log A^B = B \log A.$$

Applying these rules, we have

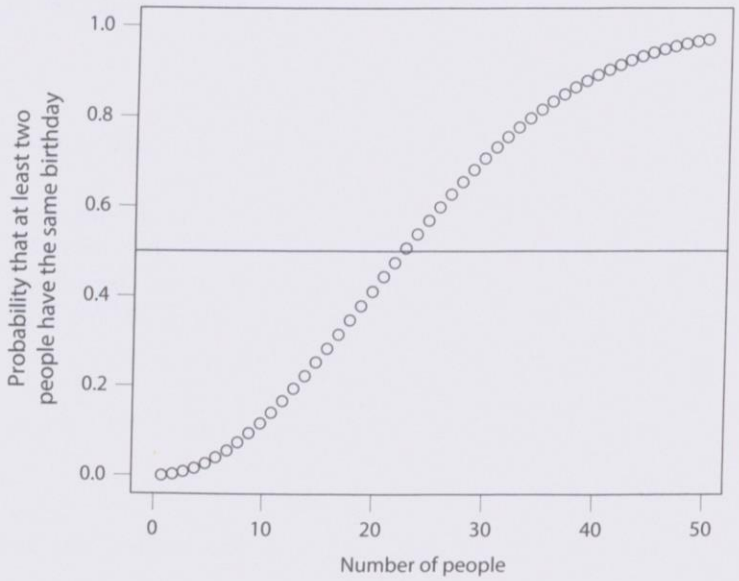
$$\log P(\text{nobody has the same birthday}) = \log 365! - k \log 365 - \log(365-k)!.$$

After computing this probability on a logarithmic scale, we then take the exponential transformation of it to obtain the desired probability. In R, we use the `lfactorial()` function to compute the logarithm of a factorial instead of the `factorial()` function, which computes a factorial without the logarithmic transformation. We now create a new function called `birthday`, which computes the probability that at least two people have the same birthday given k. The function is written so that it takes a vector of k values. We plot the results.

```
birthday <- function(k) {
    logdenom <- k * log(365) + lfactorial(365 - k) # log denominator
    lognumer <- lfactorial(365) # log numerator
    ## P(at least two have the same bday) = 1 - P(nobody has the same bday)
```

```
    pr <- 1 - exp(lognumer - logdenom) # transform back
    return(pr)
}
k <- 1:50
bday <- birthday(k)  # call the function
names(bday) <- k  # add labels
plot(k, bday, xlab = "Number of people", xlim = c(0, 50), ylim = c(0, 1),
     ylab = "Probability that at least two\n people have the same birthday")
abline(h = 0.5) # horizontal 0.5 line
bday[20:25]
```

```
##        20        21        22        23        24        25
## 0.4114384 0.4436883 0.4756953 0.5072972 0.5383443 0.5686997
```



We observe that when the number of people equals 23, the probability of at least two people having the same birthday exceeds 0.5. When the number of people is more than 50, this probability is close to 1.

6.1.4 SAMPLING WITH AND WITHOUT REPLACEMENT

While we derived an exact analytical solution to the birthday problem above, we can also produce an approximate solution using a *Monte Carlo simulation method*. The name originates from the Monte Carlo Casino in Monaco, but we may also simply call it a *simulation* method. The Monte Carlo simulation method refers to a general class of stochastic (as opposed to deterministic) methods that can be used to approximately solve analytical problems by randomly generating quantities of interest.

For the birthday problem, we sample k possibly nonunique birthdays out of 365 days and check whether or not the sampled k birthdays are all different. We use *sampling with replacement* because for each of k draws, every one of 365 days is equally likely to be sampled *regardless of* which dates were sampled before. In other words, the fact that one person is born on a certain day of the year should not exclude someone else from being born on the same day. After repeating this sampling procedure many times, we compute the fraction of simulation trials where at least two birthdays are the same, and this fraction serves as an estimate of the corresponding probability. This simulation procedure is intuitive because it emulates the *data-generating process*, or the actual process in which the data are generated, as described in the birthday problem.

In R, we can use the `sample()` function to implement sampling with or without replacement by setting the `replace` argument to either `TRUE` or `FALSE`. While unused in the birthday problem, *sampling without replacement* means that once an element is sampled, it will not be available for subsequent draws. For example, in the discussion of sample surveys in section 3.4.1, we introduced *simple random sampling* (SRS) as a method to randomly choose a representative sample of respondents from a population. SRS is an example of sampling without replacement because we typically do not interview the same individual multiple times. For sampling with replacement, the basic syntax is `sample(x, size = k, replace = TRUE)`, where `x` is a vector of elements to sample from, and `size` is the number of elements to choose. In addition, we can feed a vector of probability weights into the `prob` argument if unequal probabilities should be used to sample each element.

```
k <- 23 # number of people
sims <- 1000 # number of simulations
event <- 0 # counter
for (i in 1:sims) {
    days <- sample(1:365, k, replace = TRUE)
    days.unique <- unique(days) # unique birthdays
    ## if there are duplicates, the number of unique birthdays
    ## will be less than the number of birthdays, which is "k"
    if (length(days.unique) < k) {
        event <- event + 1
    }
}
## fraction of trials where at least two bdays are the same
answer <- event / sims
answer

## [1] 0.509
```

While our simulation estimate is close to the analytical solution, which is 0.507, they are not identical. This difference is called the *Monte Carlo error*, but is the inevitable consequence of the simulation approach. The size of the Monte Carlo error depends on the nature of the problem and it differs from one simulation to another. It is difficult to eliminate such an error but it is possible to reduce it. To obtain a more accurate estimate, we increase the number of simulations. In the above code, we set the number

of simulations to 1000. Next, we run the same code with the number of simulations set to one million and obtain an estimate of 0.508, which is closer to the true answer.

The **Monte Carlo simulation method** refers to a general class of repeated random sampling procedures used to approximately solve analytical problems. Commonly used procedures include **sampling with replacement**, in which the same unit can be repeatedly sampled, and **sampling without replacement**, in which each unit can be sampled at most once.

6.1.5 COMBINATIONS

We introduce another useful counting method called *combinations*. Combinations are similar to permutations, but the former ignores ordering while the latter does not. That is, combinations are ways to choose k distinct elements out of n elements without regard to their order. This means that when choosing 2 elements, two *different* permutations, AB and BA, represent one *identical* combination. Since the order in which the elements are arranged does not matter, the number of combinations is never greater than the number of permutations. For example, if we choose 2 distinct elements out of 3 elements, A, B, and C, the number of permutations is ${}_3P_2 = 6$ (AB, BA, AC, CA, BC, CB), whereas the number of combinations is 3 (AB, AC, BC).

In fact, to compute combinations, we first calculate permutations ${}_nP_k$ and then divide by $k!$. This is because given k sampled elements, there are $k!$ ways to arrange them in a different order, and yet all these arrangements are counted as a single combination. In the above example, for every set of two sampled elements (e.g., A and B), we have $2!(= 2 \times 1 = 2)$ ways of arranging them (i.e., AB and BA) but these two permutations count as one combination. Here, we obtain the number of combinations through the division of ${}_3P_2$ by $2!$. In general, the formula for combinations is given as follows.

The number of **combinations** when choosing k distinct elements from n elements is denoted by either ${}_nC_k$ or $\binom{n}{k}$ and is computed as

$$ {}_nC_k = \binom{n}{k} = \frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!}. \tag{6.10} $$

Suppose that we are creating a committee of 5 out of 20 people (10 men and 10 women). Assume that each person is equally likely to be assigned to the committee. What is the probability that at least 2 women are on the committee? To compute this probability, we first note the following equality:

$$ \begin{aligned} &P(\text{at least 2 women are on the committee}) \\ &\quad = 1 - P(\text{no woman is on the committee}) \\ &\qquad - P(\text{exactly 1 woman is on the committee}). \end{aligned} $$

To compute the two probabilities on the right-hand side of this equation, we count the total number of ways we can assign 5 people to the committee out of 20 people regardless of their gender. This is given by ${}_{20}C_5 = 15{,}504$. Similarly, the number of

To the Members of the California State Assembly:

I am returning Assembly Bill 1176 without my signature.

For some time now I have lamented the fact that major issues are overlooked while many unnecessary bills come to me for consideration. Water reform, prison reform, and health care are major issues my Administration has brought to the table, but the Legislature just kicks the can down the alley.

Yet another legislative year has come and gone with out the major reforms Californians overwhelmingly deserve. In light of this, and after careful consideration, I believe it is unnecessary to sign this measure at this time.

Sincerely,

Arnold Schwarzenegger

Figure 6.4. California Governor Arnold Schwarzenegger's Veto Message in 2009.

ways in which we can have no woman on the committee is given by ${}_{10}C_0 \times {}_{10}C_5 = 252$ because there is ${}_{10}C_0$ way to choose no woman and there are ${}_{10}C_5$ ways to choose 5 out of 10 men. Thus, the probability of having no woman is 0.016. The number of ways in which we can have exactly 1 woman on the committee is ${}_{10}C_1 \times {}_{10}C_4 = 2100$, giving a probability of 0.135. Altogether, the probability of having at least 2 women on the committee equals $0.84 = 1 - 0.016 - 0.135$.

As a more complex example of combinations, we discuss an incident that occurred in 2009 when California Governor Arnold Schwarzenegger wrote a message to the state assembly regarding his veto of Assembly Bill 1176.[1] This message is displayed in figure 6.4. When the message was released, many observed that the first letters of each line in the main text, starting with "F" and ending with "u," constitute a sentence of profanity. Asked whether this was intentional, Schwarzenegger's spokesman replied, "My goodness. What a coincidence. I suppose when you do so many vetoes, something like this is bound to happen." Below, we consider the probability of this acrostic happening by chance.

For the sake of simplicity, suppose that the Governor gave his veto message to his secretary who then typed it in her computer but hit the return key at random. That is, the 85 words ("For" to "time") were divided by (random) line breaks into 7 lines, each with at least one word. We further assume that there are no broken words, every way of breaking the lines was equally likely, and the total number of lines is fixed at seven. Under this scenario, what is the probability of the coincidence happening?

To compute this probability using equation (6.1), we first consider the number of ways in which the 85 words can be divided into 7 lines. Note that to end up with 7 lines, 6 line breaks must be inserted. A line break may be inserted before the second word, before the third word, ..., or before the 85th word. There are thus 84 places into which 6 line breaks must be inserted. How many ways can we insert line breaks into 6 out of these 84 places? To compute this number, we use combinations rather than permutations because the order in which 6 line breaks are inserted does not matter. (Of course, the words in the acrostic must be ordered in a particular way to generate the profanity.) Therefore, the application of combinations leads to ${}_{84}C_6 = 84!/(6!78!)$ equally likely partitions. To compute combinations in R, we use the

[1] This section is based on Philip B. Stark (2009) "Null and vetoed: Chance coincidence?" *Chance*, vol. 23, no. 4, pp. 43–46.

`choose()` function. When the number is large, we may use the `lchoose()` function so that combinations are calculated on the logarithmic scale.

```
choose(84, 6)
## [1] 406481544
```

Therefore, there are more than 400 million ways to insert 6 line breaks. However, there are only 12 ways to produce this particular acrostic. The break to produce "u" at the beginning of the second line can be in only one place ("unnecessary"). The break to produce "c" at the beginning of the third line can happen in any of 3 places ("come," "consideration," "care"). The break for the "k" can be in only one place ("kicks"). The break for the "y" can be in any of two places ("Yet," "year"). The break for the "o" can be in any of two places ("overwhelmingly," "of"). The break for the "u" can be in only one place ("unnecessary"). These scenarios lead to $12 = 1 \times 3 \times 1 \times 2 \times 2 \times 1$. Hence, the probability that this randomization scheme would produce the acrostic is $12/{}_{84}C_6$, or about one in 34 million. The analysis suggests that according to this probabilistic model, the "coincidence" is a highly unlikely event.

6.2 Conditional Probability

We next introduce conditional probability, which concerns how the probability of an event changes after we observe other events. Conditional probability follows the rules of probability described in section 6.1. The difference is that conditional probability enables us to take into account observed evidence.

6.2.1 CONDITIONAL, MARGINAL, AND JOINT PROBABILITIES

We begin by defining the conditional probability of event A occurring, given the information that event B has occurred. This conditional probability, denoted as $P(A \mid B)$, has the following definition.

The **conditional probability** of event A occurring given that event B occurred is defined as

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}. \tag{6.11}$$

In this equation, $P(A \text{ and } B)$ is the **joint probability** of both events occurring, whereas $P(B)$ is the **marginal probability** of event B. By rearranging, we obtain the **multiplication rule**

$$P(A \text{ and } B) = P(A \mid B)P(B) = P(B \mid A)P(A). \tag{6.12}$$

Using this rule, we can derive an alternative form of the ***law of total probability*** introduced in equation (6.6):

$$P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c). \tag{6.13}$$

To see the importance of conditioning, consider two couples who are both expecting twins. One couple had an ultrasound exam, but the technician was able to determine only that one of the two was a boy. The other couple did not find out the genders of their twins until the delivery when they saw the first baby was a boy. What is the probability that both babies are boys? Is this probability different between the two couples? We begin by noting that there are four outcomes in the sample space. Denoting the baby gender by "G" for girl and "B" for boy, respectively, we can represent the sample space by $\Omega = \{GG, GB, BG, BB\}$. For example, GB means that the elder twin is a girl and the younger one is a boy.

Then, for the first couple, the probability of interest is

$$\begin{aligned} P(BB \mid \text{at least one is a boy}) &= \frac{P(BB \text{ and } \{\text{at least one is a boy}\})}{P(\text{at least one is a boy})} \\ &= \frac{P(BB \text{ and } \{BB \text{ or } BG \text{ or } GB\})}{P(BB \text{ or } BG \text{ or } GB)} \\ &= \frac{P(BB)}{P(BB \text{ or } BG \text{ or } GB)} = \frac{1/4}{3/4} = \frac{1}{3}. \end{aligned}$$

The third equality follows from the fact that event BB is a subset of event $\{\text{at least one is a boy}\}$, i.e., BB and $\{BB \text{ or } BG \text{ or } GB\} = BB$.

In contrast, for the second couple, we have

$$\begin{aligned} P(BB \mid \text{elder twin is a boy}) &= \frac{P(BB \text{ and } \{\text{the elder twin is a boy}\})}{P(\text{elder twin is a boy})} \\ &= \frac{P(BB \text{ and } \{BB \text{ or } BG\})}{P(BB \text{ or } BG)} \\ &= \frac{P(BB)}{P(BB \text{ or } BG)} = \frac{1/4}{1/2} = \frac{1}{2}. \end{aligned}$$

Therefore, this example illustrates that the information upon which we condition matters. Knowing that the first baby is a boy, as opposed to knowing that at least one is a boy, gives a different conditional probability of the same event.

Probability and conditional probability can also be used to describe the characteristics of a population. For example, if 10% of a population of voters are black, then we may write $P(\text{black}) = 0.1$. We can interpret this probability as stating that if we randomly sample a voter from this population there is a 10% chance this voter is black. Similarly, $P(\text{black} \mid \text{hispanic or black})$ represents the population proportion of blacks among minority (i.e., black and Hispanic) voters.

As an illustration, we will use a random sample of 10,000 registered voters from Florida contained in the CSV file `FLVoters.csv`. Table 6.1 shows the names and descriptions of variables in this sample list of registered voters. To begin, we load the data and remove those voters who contain a missing value using the `na.omit()` function.

Table 6.1. Florida Registered Voter List Sample.

Variable	*Description*
surname	surname
county	county ID of the voter's residence
VTD	voting district ID of the voter's residence
age	age
gender	gender: m = male and f = female
race	self-reported race

```
FLVoters <- read.csv("FLVoters.csv")
dim(FLVoters) # before removal of missing data
## [1] 10000     6
FLVoters <- na.omit(FLVoters)
dim(FLVoters) # after removal
## [1] 9113    6
```

For the sake of illustration, we will treat this sample of 9113 voters as a population of interest. To compute the *marginal probability* for each racial category, we can use the `table()` and `prop.table()` functions (see section 2.5.2) and calculate the proportion of voters who belong to each racial group in this population.

```
margin.race <- prop.table(table(FLVoters$race))
margin.race
##
##       asian       black    hispanic      native       other
## 0.019203336 0.131021617 0.130802151 0.003182267 0.034017338
##       white
## 0.681773291
```

The result shows, for example, that $P(\text{black}) = 0.13$ and $P(\text{white}) = 0.68$. Similarly, we can obtain the marginal probabilities of gender as follows.

```
margin.gender <- prop.table(table(FLVoters$gender))
margin.gender
##
##         f         m
## 0.5358279 0.4641721
```

Therefore, we have $P(\text{female}) = 0.54$ and $P(\text{male}) = 0.46$. Next, to compute the *conditional probability* of race given gender, we can look at the proportion of each racial group among female voters and among male voters, separately.

```
prop.table(table(FLVoters$race[FLVoters$gender == "f"]))
##
##       asian       black    hispanic      native       other
## 0.016997747 0.138849068 0.136391563 0.003481466 0.032357157
##       white
## 0.671922998
```

The result suggests, for example, $P(\text{black} \mid \text{female}) = 0.14$ and $P(\text{white} \mid \text{female}) = 0.67$. Lastly, the *joint probability* of race and gender can be computed by calculating the proportion of voters who belong to specific racial and gender groups.

```
joint.p <- prop.table(table(race = FLVoters$race, gender = FLVoters$gender))
joint.p
##          gender
## race                f           m
##   asian    0.009107868 0.010095468
##   black    0.074399210 0.056622408
##   hispanic 0.073082410 0.057719741
##   native   0.001865467 0.001316800
##   other    0.017337869 0.016679469
##   white    0.360035115 0.321738176
```

This joint probability table gives, for example, $P(\text{black and female}) = 0.07$ and $P(\text{white and male}) = 0.32$. From this joint probability, we can compute the marginal and conditional probability. First, to obtain the marginal probability, we apply the *law of total probability* given in equation (6.6). For example, we can compute the probability of being a black voter by

$$P(\text{black}) = P(\text{black and female}) + P(\text{black and male}).$$

Thus, summing over columns for each row results in the marginal probability of race. This operation yields results identical to those obtained above.

```
rowSums(joint.p)
##       asian       black    hispanic      native       other
## 0.019203336 0.131021617 0.130802151 0.003182267 0.034017338
##       white
## 0.681773291
```

Similarly, we can obtain the marginal probability of gender from the joint probability table by summing over racial categories. Since we have a total of six racial categories, we will extend the law of total probability given in equation (6.6) to

$$P(A) = \sum_{i=1}^{N} P(A \text{ and } B_i), \tag{6.14}$$

where $B_1, \ldots, B_N$ is a set of mutually exclusive events which together cover the entire sample space. In the current setting, for example, since racial categories are mutually exclusive, we have

$$\begin{aligned} P(\text{female}) = {} & P(\text{female and asian}) + P(\text{female and black}) \\ & + P(\text{female and hispanic}) + P(\text{female and native}) \\ & + P(\text{female and other}) + P(\text{female and white}). \end{aligned}$$

Therefore, the marginal probability of gender is obtained by summing over rows for each column of the joint probability table.

```
colSums(joint.p)
##         f         m
## 0.5358279 0.4641721
```

Finally, the *conditional probability* can be obtained as the ratio of joint probability to the marginal probability (see equation (6.11)). For example, the conditional probability of being black among female voters is calculated as

$$P(\text{black} \mid \text{female}) = \frac{P(\text{black and female})}{P(\text{female})} \approx \frac{0.074}{0.536} \approx 0.139,$$

which, as expected, is equal to what we computed earlier.

The results of this example are summarized in table 6.2. From the joint probability, both marginal and conditional probabilities can be obtained. To compute marginal probability, we sum over either rows or columns. Once marginal probability is obtained in this way, we can divide joint probability by marginal probability in order to calculate the desired conditional probability.

We can extend the definition of conditional probability to settings with more than two types of events. For events A, B, and C, the joint probability is defined as $P(A \text{ and } B \text{ and } C)$, whereas there are three marginal probabilities $P(A)$, $P(B)$, and $P(C)$. In this case, there are two types of conditional probabilities: the joint probability of two events conditional on the remaining event (e.g., $P(A \text{ and } B \mid C)$) and the

Table 6.2. An Example of a Joint Probability Table.

Racial groups	*Gender*		
	Female	*Male*	*Marginal prob.*
Asian	0.009	0.010	0.019
Black	0.074	0.057	0.131
Hispanic	0.073	0.058	0.131
Native	0.002	0.001	0.003
White	0.360	0.322	0.682
Marginal prob.	0.536	0.464	1

Note: The table is based on Florida voter registration data. The marginal probability of gender (far right column) and that of race (bottom row) can be obtained by summing the joint probabilities over columns and over rows, respectively.

conditional probability of one event given the other two (e.g., $P(A \mid B \text{ and } C)$). These conditional probabilities can be defined analogously to the two-event case as

$$P(A \text{ and } B \mid C) = \frac{P(A \text{ and } B \text{ and } C)}{P(C)}, \tag{6.15}$$

$$P(A \mid B \text{ and } C) = \frac{P(A \text{ and } B \text{ and } C)}{P(B \text{ and } C)} = \frac{P(A \text{ and } B \mid C)}{P(B \mid C)}. \tag{6.16}$$

The second equality in equation (6.16) follows from the equality $P(A \text{ and } B \text{ and } C) = P(A \text{ and } B \mid C)P(C)$, which is obtained by rearranging the terms in equation (6.15).

To illustrate the above conditional probabilities, we create a new `age.group` variable indicating four age groups: 20 and below, 21–40, 41–60, and above 60.

```
FLVoters$age.group <- NA # initialize a variable
FLVoters$age.group[FLVoters$age <= 20] <- 1
FLVoters$age.group[FLVoters$age > 20 & FLVoters$age <= 40] <- 2
FLVoters$age.group[FLVoters$age > 40 & FLVoters$age <= 60] <- 3
FLVoters$age.group[FLVoters$age > 60] <- 4
```

The joint probability of age group, race, and gender can be calculated as a three-way table. Below, this three-way table is displayed as two separate two-way tables: one two-way (race and age group) table for female voters and the other two-way table for male voters.

```
joint3 <-
    prop.table(table(race = FLVoters$race, age.group = FLVoters$age.group,
                     gender = FLVoters$gender))
```

```
joint3
## , , gender = f
##
##           age.group
## race                 1            2            3
##   asian    0.0001097333 0.0026336004 0.0041698672
##   black    0.0016460002 0.0280917371 0.0257873368
##   hispanic 0.0015362669 0.0260068035 0.0273236036
##   native   0.0001097333 0.0004389334 0.0006584001
##   other    0.0003292000 0.0062548008 0.0058158674
##   white    0.0059256008 0.0796664106 0.1260836168
##           age.group
## race                 4
##   asian    0.0021946670
##   black    0.0188741358
##   hispanic 0.0182157358
##   native   0.0006584001
##   other    0.0049380007
##   white    0.1483594864
##
## , , gender = m
##
##           age.group
## race                 1            2            3
##   asian    0.0002194667 0.0028530670 0.0051574674
##   black    0.0016460002 0.0228245364 0.0189838692
##   hispanic 0.0016460002 0.0197520026 0.0221661363
##   native   0.0000000000 0.0004389334 0.0003292000
##   other    0.0004389334 0.0069132009 0.0055964007
##   white    0.0040601339 0.0750576100 0.1184022825
##           age.group
## race                 4
##   asian    0.0018654669
##   black    0.0131680018
##   hispanic 0.0141556019
##   native   0.0005486667
##   other    0.0037309338
##   white    0.1242181499
```

For example, the proportion of black female voters who are above 60 or $P(\text{black and above 60 and female})$ is equal to 0.019. Suppose that we wish to obtain the conditional probability of being black and female given that a voter is above 60 years old or $P(\text{black and female} \mid \text{above 60})$. Using equation (6.15), we can compute this conditional probability by dividing the joint probability by the marginal probability of being above 60 or $P(\text{above 60})$. To extract a specific joint probability from the above three-way table, we specify the corresponding value for each demographic characteristic.

```
## marginal probabilities for age groups
margin.age <- prop.table(table(FLVoters$age.group))
margin.age

##
##          1          2          3          4
## 0.01766707 0.27093164 0.36047405 0.35092725

## P(black and female | above 60)
joint3["black", 4, "f"] / margin.age[4]

##          4
## 0.05378361
```

According to equation (6.16), the conditional probability of being black given that a voter is female and above 60 years old or $P(\text{black} \mid \text{female and above 60})$ can be computed by dividing the three-way joint probability $P(\text{black and above 60 and female})$ by the two-way joint probability $P(\text{above 60 and female})$. To obtain this two-way joint probability, we can create a two-way joint probability table for age group and gender.

```
## two-way joint probability table for age group and gender
joint2 <- prop.table(table(age.group = FLVoters$age.group,
                           gender = FLVoters$gender))
joint2

##           gender
## age.group           f           m
##         1 0.009656535 0.008010534
##         2 0.143092286 0.127839350
##         3 0.189838692 0.170635356
##         4 0.193240426 0.157686821

joint2[4, "f"] # P(above 60 and female)

## [1] 0.1932404

## P(black | female and above 60)
joint3["black", 4, "f"] / joint2[4, "f"]

## [1] 0.09767178
```

6.2.2 INDEPENDENCE

Having defined conditional probability, we can now formally discuss the concept of *independence*. Intuitively, the independence of two events implies that the knowledge of one event does not give us any additional information about the occurrence of the other event. That is, if events A and B are independent of each other, the conditional probability of A given B does not differ from the marginal probability of A. Similarly, the conditional probability of B given A does not depend on A:

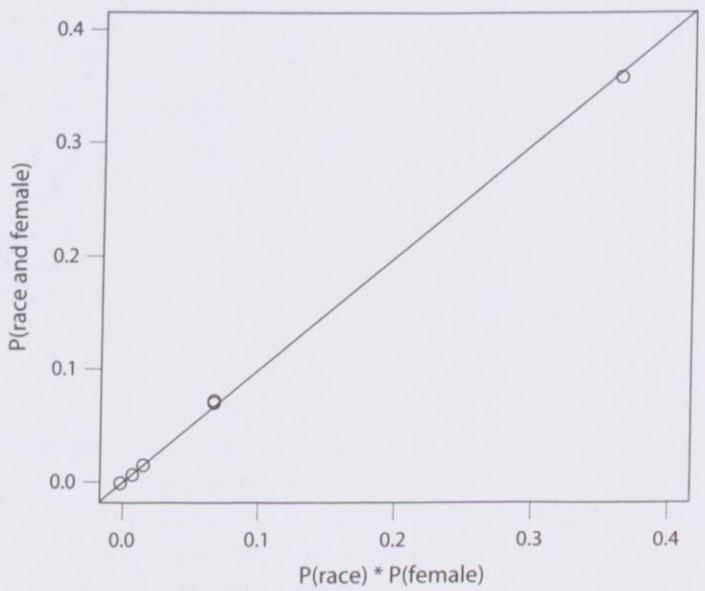
$$P(A \mid B) = P(A) \quad \text{and} \quad P(B \mid A) = P(B). \tag{6.17}$$

Together with equation (6.12), this equality implies the following formal definition of independence between events A and B.

> Events A and B are **independent** if and only if the joint probability is equal to the product of the marginal probabilities:
>
> $$P(A \text{ and } B) = P(A)P(B). \tag{6.18}$$

We investigate whether race and gender are independent of each other in the sample of Florida registered voters analyzed earlier. Although we do not expect this relationship to be exactly independent, we examine whether the proportion of female voters, for example, is greater than expected in some racial groups. Note that if independence holds, we should have, for example, $P(\text{black and female}) = P(\text{black})P(\text{female})$, $P(\text{white and male}) = P(\text{white})P(\text{male})$, and so on. We compare the products of marginal probabilities for race and female with their joint probabilities using a scatter plot. We use the `c()` function, which combines its inputs into a vector, to coerce a table format into a vector so that its elements can be used in the `plot()` function.

```
plot(c(margin.race * margin.gender["f"]), # product of marginal probs.
     c(joint.p[, "f"]), # joint probabilities
     xlim = c(0, 0.4), ylim = c(0, 0.4),
     xlab = "P(race) * P(female)", ylab = "P(race and female)")
abline(0, 1) # 45-degree line
```



The scatter plot shows that the points fall neatly along the 45-degree line, implying that $P(\text{race})P(\text{female})$ (horizontal axis) and $P(\text{race and female})$ (vertical axis) are approximately equal. This means that race and gender are approximately independent in this sample of registered voters. That is, the knowledge of a voter's gender does not help us predict her race. Similarly, one's race does not predict gender either.

The notion of independence extends to situations with more than two events. For example, if we have three events A, B, and C, the *joint independence* among these events implies that the joint probability can be written as the product of marginal probabilities:

$$P(A \text{ and } B \text{ and } C) = P(A)P(B)P(C). \tag{6.19}$$

Furthermore, we can define the independence between two events conditional on another event. The *conditional independence* of events A and B given event C implies that the joint probability of A and B given C is equal to the product of two conditional probabilities:

$$P(A \text{ and } B \mid C) = P(A \mid C)P(B \mid C). \tag{6.20}$$

Joint independence given in equation (6.19) implies pairwise independence given in equation (6.18). This result can be obtained by applying the *law of total probability*:

$$\begin{aligned} P(A \text{ and } B) &= P(A \text{ and } B \text{ and } C) + P(A \text{ and } B \text{ and } C^c) \\ &= P(A)P(B)P(C) + P(A)P(B)P(C^c) \\ &= P(A)P(B)\big(P(C) + P(C^c)\big) = P(A)P(B). \end{aligned}$$

In addition, joint independence implies conditional independence, defined in equation (6.20), but the converse is not necessarily true. This result is based on the definition of conditional probability given in equation (6.15):

$$P(A \text{ and } B \mid C) = \frac{P(A \text{ and } B \text{ and } C)}{P(C)} = \frac{P(A)P(B)P(C)}{P(C)} = P(A \mid C)P(B \mid C).$$

The last equality follows from the fact that joint independence implies pairwise independence (and hence equation (6.17) holds for A and C as well as B and C).

To examine joint independence among our sample of registered Florida voters, we compare the elements of the three-way proportion table `joint3` with the corresponding product of marginal probabilities, `margin.race`, `margin.age`, and `margin.gender`. As an illustration, we set the age group to the above 60 category and examine female voters. We also examine conditional independence between race and gender, given age. For this, we again set the age and gender groups to the above 60 and female categories, respectively. The results show that both joint (left-hand plot) and conditional (right-hand plot) independence relationships approximately hold, despite small deviations.

```
## joint independence
plot(c(joint3[, 4, "f"]), # joint probability
     margin.race * margin.age[4] * margin.gender["f"], # product of marginals
     xlim = c(0, 0.3), ylim = c(0, 0.3), main = "Joint independence",
     xlab = "P(race and above 60 and female)",
     ylab = "P(race) * P(above 60) * P(female)")
abline(0, 1)
## conditional independence given female
plot(c(joint3[, 4, "f"]) / margin.gender["f"], # joint prob. given female
     ## product of marginals
     (joint.p[, "f"] / margin.gender["f"]) *
         (joint2[4, "f"] / margin.gender["f"]),
     xlim = c(0, 0.3), ylim = c(0, 0.3), main = "Marginal independence",
     xlab = "P(race and above 60 | female)",
     ylab = "P(race | female) * P(above 60 | female)")
abline(0, 1)
```

Finally, the well-known *Monty Hall problem* illustrates how tricky conditional probability and independence can be. The problem goes as follows. You are on a game show and must choose one of three doors, where one conceals a new car and two conceal old goats. After you randomly choose one door, the host of the game show, Monty, opens a different door, which does not conceal a car. Then, Monty asks you if you would like to switch to the (unopened) third door. You will win the new car if it is behind the door of your final choice. Should you switch, or stay with your original choice? Does switching make a difference? Most people think switching makes no difference because after Monty reveals one door with a goat, the two remaining doors have a goat or a car behind them. Therefore, the chance of winning a car is 50%. However, it turns out that this seemingly sensible reasoning is incorrect.

Let's think about this problem carefully. Consider the strategy of not switching. In this case, your initial choice determines the outcome regardless of what Monty does. Therefore, the probability of winning the car is 1/3. Now, consider the strategy of switching. There are two scenarios. First, suppose that you initially choose a door with the car. The probability of this event is 1/3. Swapping the door in this scenario is a bad choice because you will not win the car. Next, suppose that the door you selected first has a goat. The probability of your initially choosing a door with a goat is 2/3. Then, since Monty opens another door with a goat, the remaining door to which you will switch contains a car. Hence, under this scenario, you will always win the car. Therefore, switching gives you a probability of winning the car that is twice as high as not switching.

We formalize this logic by applying the rules of probability covered so far. To compute the probability of winning a car given that you switch, we first apply the law of total probability in equation (6.13):

$$\begin{aligned} P(\text{car}) &= P(\text{car} \mid \text{car first})P(\text{car first}) + P(\text{car} \mid \text{goat first})P(\text{goat first}) \\ &= P(\text{goat first}) = \frac{2}{3}. \end{aligned}$$

To see why the second equality holds, notice that if you initially select the door with a car then switching makes you lose the car, i.e., $P(\text{car} \mid \text{car first}) = 0$. In contrast, if you first pick a door with a goat, then you have a 100% chance of winning a car by switching, i.e., $P(\text{car} \mid \text{goat first}) = 1$.

This rather counterintuitive problem can also be solved with *Monte Carlo simulations*. For emulating random choice in R, we use the `sample()` function. We set the `size` argument to `1` in order to randomly choose one element from a vector.

```
sims <- 1000
doors <- c("goat", "goat", "car") # order does not matter
result.switch <- result.noswitch <- rep(NA, sims)
for (i in 1:sims) {
    ## randomly choose the initial door
    first <- sample(1:3, size = 1)
    result.noswitch[i] <- doors[first]
    remain <- doors[-first] # remaining two doors
    ## Monty chooses one door with a goat
    monty <- sample((1:2)[remain == "goat"], size = 1)
    result.switch[i] <- remain[-monty]
}
mean(result.noswitch == "car")

## [1] 0.338

mean(result.switch == "car")

## [1] 0.662
```

6.2.3 BAYES' RULE

We discussed different interpretations of probability at the beginning of this chapter. One interpretation, proposed by Reverend Thomas Bayes, was that probability measures one's subjective belief in an event's occurrence. From this Bayesian perspective, it is natural to ask the question of how we should update our beliefs after observing some data. *Bayes' rule* shows how updating beliefs can be done in a mathematically coherent manner.

Bayes' rule is given by

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)}. \quad (6.21)$$

In this equation, $P(A)$ is called the **prior probability** and reflects one's initial belief about the likelihood of event A occurring. After observing the data, represented as event B, we update our belief and obtain $P(A \mid B)$, which is called the **posterior probability**.

Regardless of whether we interpret probability as subjective belief, Bayes' rule shows mathematically how the knowledge of $P(A)$ (*prior probability*), $P(B \mid A)$, and $P(B \mid A^c)$ yields that of $P(A \mid B)$ (*posterior probability*). Bayes' rule is simply the result of rewriting the definition of conditional probability given in equation (6.11) using the law of total probability shown in equation (6.13):

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}.$$

A well-known application of Bayes' rule is the interpretation of medical diagnostic tests, which can have false positives and false negatives (defined in section 4.1.3). Consider the following first-trimester screening test problem. A 35-year-old pregnant woman is told that 1 in 378 women of her age will have a baby with Down syndrome (DS). A first-trimester ultrasound screening procedure indicates that she is in a high-risk category. Of 100 cases of DS, 86 mothers will receive a high-risk result and 14 cases of DS will be missed. Also, there is a 1 in 20 chance for a normal pregnancy to be diagnosed as high risk. Given the result of the screening procedure, what is the probability that her baby has DS? What would the probability be if the result had been negative?

To solve this problem, we first specify the prior probability. Without any testing, the probability that a baby has DS, $P(\text{DS})$, is equal to 1/378 or approximately 0.003. The ultrasound screening procedure gives a high-risk result 86% of times when a baby actually has DS. This is called the *true positive rate* of the test and can be expressed as $P(\text{HR} \mid \text{DS}) = 0.86$, where HR denotes a high-risk result. However, the screening procedure also produces a *false positive rate* of 5%, which can be formally written as $P(\text{HR} \mid \text{not DS}) = 0.05$. Using this information, we can apply Bayes' rule to obtain the posterior probability that the baby has DS, given that the woman received a high-risk

result, or the *positive predictive value* of the test:

$$P(\text{DS} \mid \text{HR}) = \frac{P(\text{HR} \mid \text{DS})P(\text{DS})}{P(\text{HR} \mid \text{DS})P(\text{DS}) + P(\text{HR} \mid \text{not DS})P(\text{not DS})}$$

$$= \frac{0.86 \times \frac{1}{378}}{0.86 \times \frac{1}{378} + 0.05 \times \frac{377}{378}} \approx 0.04.$$

Similarly, if the woman received a normal pregnancy result, the posterior probability becomes

$$P(\text{DS} \mid \text{not HR}) = \frac{P(\text{not HR} \mid \text{DS})P(\text{DS})}{P(\text{not HR} \mid \text{DS})P(\text{DS}) + P(\text{not HR} \mid \text{not DS})P(\text{not DS})}$$

$$= \frac{0.14 \times \frac{1}{378}}{0.14 \times \frac{1}{378} + 0.95 \times \frac{377}{378}} \approx 0.0004.$$

We see that even when the woman receives a high-risk result, the posterior probability of having a baby with DS is small. This is because DS is a relatively rare disease, as reflected by a small prior probability. As expected, if the woman receives a normal pregnancy result, then the posterior probability becomes even smaller than the prior probability.

We can use Bayes' rule to solve the Monty Hall problem introduced in section 6.2.2. Let A represent the event that the first door has a car behind it. Define B and C similarly for the second and third doors, respectively. Since each door is equally likely to have a car behind it, the prior probabilities are $P(A) = P(B) = P(C) = 1/3$. Suppose that we choose the first door and let MC represent the event that Monty opens the third door. We want to know whether switching to the second door increases the chance of winning the car, i.e., $P(B \mid \text{MC}) > P(A \mid \text{MC})$. We apply Bayes' rule after noting that $P(\text{MC} \mid A) = 1/2$ (Monty chooses between the second and third door with equal probability), $P(\text{MC} \mid B) = 1$ (Monty has no option but to open the third door, which has a goat), and $P(\text{MC} \mid C) = 0$ (Monty cannot open the third door, which has a car):

$$P(A \mid \text{MC}) = \frac{P(\text{MC} \mid A)P(A)}{P(\text{MC} \mid A)P(A) + P(\text{MC} \mid B)P(B) + P(\text{MC} \mid C)P(C)}$$

$$= \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \times \frac{1}{3}} = \frac{1}{3},$$

$$P(B \mid \text{MC}) = \frac{P(\text{MC} \mid B)P(B)}{P(\text{MC} \mid A)P(A) + P(\text{MC} \mid B)P(B) + P(\text{MC} \mid C)P(C)}$$

$$= \frac{1 \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \times \frac{1}{3}} = \frac{2}{3}.$$

Thus, switching doors will give a probability of winning a car that is twice as great as staying with the initial choice.

6.2.4 PREDICTING RACE USING SURNAME AND RESIDENCE LOCATION

This section contains an advanced application of conditional probability and Bayes' rule in the social sciences. Readers may skip this section without affecting their ability to understand the materials in the remainder of the book.

It is often of interest to infer certain unknown attributes of individuals from their known characteristics. We consider the problem of predicting individual race using surname and residence location.[2] Accurate prediction of individual race is useful, for example, when studying turnout rates among racial groups.

The US Census Bureau releases a list of common surnames with their frequency. For example, the most common surname was "Smith" with 2,376,206 occurrences, followed by "Johnson" and "Williams" with 1,857,160 and 1,534,042, respectively. This data set is quite comprehensive, including a total of more than 150,000 surnames that occurred at least 100 times. In addition, the census provides the relative frequencies of individual race within each surname, using a six-category self-reported race measure: non-Hispanic white, non-Hispanic black, non-Hispanic Asian and Pacific Islander, Hispanic origin, non-Hispanic American Indian and Alaskan Native, and non-Hispanic of two or more races. We will combine the last two categories into a single category of non-Hispanic others, so that we have five categories in total. The aggregate information, which can be written as $P(\text{race} \mid \text{surname})$, enables us to predict race given an individual's surname.

Note that $P(\text{race})$, $P(\text{race} \mid \text{surname})$, and $P(\text{race and surname})$ are examples of general ways to represent the marginal, conditional, and joint probabilities, respectively. For example, $P(\text{race})$ represents a collection of marginal probabilities, i.e., $P(\text{white})$, $P(\text{black})$, $P(\text{asian})$, $P(\text{hispanic})$, and $P(\text{others})$. Similarly, $P(\text{race} \mid \text{surname})$ can be evaluated for any given racial group and surname, for example, $P(\text{black} \mid \text{Smith})$. To illustrate the convenience of this general notation, we apply the law of total probability in equation (6.14) to the joint probability of race and surname:

$$P(\text{surname}) = \sum_{\text{race}} P(\text{race and surname}),$$

where the summation is taken over all racial categories (i.e., white, black, asian, hispanic, and others. In terms of the notation used in equation (6.14), A represents any given surname while B_i is a racial category. This equality applies to any surname of interest, and the summation is taken over all five racial categories.

This census name list is contained in the CSV data file `names.csv`. Table 6.3 lists the names and descriptions of variables in this census surname list data set.[3]

[2] This section is in part based on Kosuke Imai and Kabir Khanna (2016) "Improving ecological inference by predicting individual ethnicity from voter registration records." *Political Analysis*, vol. 24, no. 2 (Spring), pp. 263–272.

[3] To protect anonymity, the Census Bureau does not reveal small race percentages for given surnames. For the sake of simplicity, we impute these missing values by assuming that residual values will be equally allocated to the racial categories with missing values. That is, for each last name, we subtract the sum of the percentages of all races without missing values from 100% and divide the remaining percentage equally among those races that do have missing values.

Table 6.3. US Census Bureau Surname List Data.

Variable	*Description*
`surname`	surname
`count`	number of individuals with a specific surname
`pctwhite`	percentage of non-Hispanic whites among those who have a specific surname
`pctblack`	percentage of non-Hispanic blacks among those who have a specific surname
`pctapi`	percentage of non-Hispanic Asians and Pacific Islanders among those who have a specific surname
`pcthispanic`	percentage of Hispanic origin among those who have a specific surname
`pctothers`	percentage of the other racial groups among those who have a specific surname

```
cnames <- read.csv("names.csv")
dim(cnames)
## [1] 151671      7
```

The total number of surnames contained in this data set is 151,671. For these surnames, the data set gives the probability of belonging to a particular racial group given a voter's surname, i.e., $P(\text{race} \mid \text{surname})$. We begin by using this conditional probability to classify the race of individual voters. To validate the accuracy of our prediction of individual race, we use the sample of 10,000 registered voters from Florida analyzed earlier (see table 6.1). In some Southern states including Florida, voters are asked to self-report their race when registering. This makes the Florida data an ideal validation data set. If the accuracy of a prediction method is empirically validated in Florida, we may use the method to predict individual race in other states where such information is not available.

For matching names between the voter file and census name data, we use the `match()` function. This function takes the syntax of `match(x, y)` and returns a vector of indices of vector `y`'s correspondence to each element of vector `x`. The function returns `NA` if there is no match found in `y` for an element of `x`. Here is a simple example illustrating the use of the `match()` function.

```
x <- c("blue", "red", "yellow")
y <- c("orange", "blue")
## match x with y
match(x, y) # "blue" appears in the 2nd element of y
## [1]  2 NA NA
## match y with x
```

```
match(y, x) # "blue" appears in the first element of x
## [1] NA  1
```

Going back to the problem of predicting individual racial groups, we remove voters whose surnames do not appear in the census surname list. To do so, we utilize the fact that the syntax `match(x, y)` returns `NA` if the corresponding element of `x` is not matched with any element of `y`.

```
FLVoters <- FLVoters[!is.na(match(FLVoters$surname, cnames$surname)), ]
dim(FLVoters)
## [1] 8022    7
```

The syntax `!is.na()` represents "not `NA`," where `!` indicates negation, so that only the matched elements are retained. Thus, we focus on the resulting 80% of the original sample. We first compute the proportion of voters whose race is correctly classified in each racial category. Race is considered correctly classified if the racial category with the greatest conditional probability $P(\text{race} \mid \text{surname})$ is identical to the self-reported race. These represent *true positives* of classification (see table 4.3).

We calculate the *true positive rate* for each racial group, which represents, for example, the proportion of white voters who are correctly predicted as white. To compute this, we first subset white voters from the Florida voter file and then match the surname of each voter with the same surname in the census surname data.

```
whites <- subset(FLVoters, subset = (race == "white"))
w.indx <- match(whites$surname, cnames$surname)
head(w.indx)
## [1]  8610   237  4131  2244 27852  3495
```

The outputted row index `w.indx` contains, for each observation in the `whites` data frame, the number of the row with the same surname in the `cnames` data frame. For example, the second observation in the `whites` data frame has the surname Lynch. This surname appears in the 237th row of the `cnames` data set. Accordingly, the second value in `w.indx` is 237. More specifically, for each surname belonging to a white voter in Florida, we use `apply(cnames[w.indx, vars], 1, max)` to compare the predicted probabilities across the five racial categories in the vector `vars`, and extract the highest predicted probability. We then check whether the highest predicted probability for that voter is the same as the predicted probability of their being white. If these two numbers are identical, the classification is correct. Finally, we compute the mean of the resulting binary vector to obtain the proportion of correct classifications, the true positive rate.

```
## relevant variables
vars <- c("pctwhite", "pctblack", "pctapi", "pcthispanic", "pctothers")
mean(apply(cnames[w.indx, vars], 1, max) == cnames$pctwhite[w.indx])

## [1] 0.950218
```

The result shows that 95% of white voters are correctly predicted as whites. We repeat the same analysis for black, Hispanic, and Asian voters.

```
## black
blacks <- subset(FLVoters, subset = (race == "black"))
b.indx <- match(blacks$surname, cnames$surname)
mean(apply(cnames[b.indx, vars], 1, max) == cnames$pctblack[b.indx])

## [1] 0.1604824

## Hispanic
hispanics <- subset(FLVoters, subset = (race == "hispanic"))
h.indx <- match(hispanics$surname, cnames$surname)
mean(apply(cnames[h.indx, vars], 1, max) == cnames$pcthispanic[h.indx])

## [1] 0.8465298

## Asian
asians <- subset(FLVoters, subset = (race == "asian"))
a.indx <- match(asians$surname, cnames$surname)
mean(apply(cnames[a.indx, vars], 1, max) == cnames$pctapi[a.indx])

## [1] 0.5642857
```

We find that surname alone can correctly classify 85% of Hispanic voters as Hispanic. In contrast, classification of Asian and black voters is much worse. In particular, only 16% of black voters are correctly classified as African-Americans. The high true positive rate for whites may simply arise from the fact that they far outnumber voters from other racial categories.

We next look at *false positives*. Below, we calculate the *false discovery rate* for each racial group, which, for example, represents the proportion of voters who are not white among those classified as white. We use the same indexing trick as above and compute the proportion of white voters among those classified as whites. Subtracting the resulting value from 1 yields the false discovery rate for whites.

```
indx <- match(FLVoters$surname, cnames$surname)
## white false discovery rate
1 - mean(FLVoters$race[apply(cnames[indx, vars], 1, max) ==
                       cnames$pctwhite[indx]] == "white")

## [1] 0.1973603
```

Table 6.4. Florida Census Data at the Voting District Level.

Variable	*Description*
`county`	county census ID of the voting district
`VTD`	voting district census ID (only unique within the county)
`total.pop`	total population of the voting district
`white`	proportion of non-Hispanic whites in the voting district
`black`	proportion of non-Hispanic blacks in the voting district
`api`	proportion of non-Hispanic Asians and Pacific Islanders in the voting district
`hispanic`	proportion of voters of Hispanic origin in the voting district
`others`	proportion of the other racial groups in the voting district

```
## black false discovery rate
1 - mean(FLVoters$race[apply(cnames[indx, vars], 1, max) ==
                        cnames$pctblack[indx]] == "black")
## [1] 0.3294574
## Hispanic false discovery rate
1 - mean(FLVoters$race[apply(cnames[indx, vars], 1, max) ==
                        cnames$pcthispanic[indx]] == "hispanic")
## [1] 0.2274755
## Asian false discovery rate
1 - mean(FLVoters$race[apply(cnames[indx, vars], 1, max) ==
                        cnames$pctapi[indx]] == "asian")
## [1] 0.3416667
```

The results show that the false discovery rate is the highest for Asian and black voters, while it is much lower for whites and Hispanics.

Next, we attempt to improve the above prediction by taking into account where voters live. This approach should be helpful to the extent that there exists residential segregation based on race. In the United States, voter files contain voters' addresses. Using this information, our data set also provides the voting district where each voter lives. In addition, we will utilize the Florida census data, which contains the racial composition of each voting district. The names and descriptions of variables in this census data set, `FLCensusVTD.csv`, are given in table 6.4.

How does the knowledge of residence location improve the prediction of individual race? Whereas the census name data set contains information about the conditional probability $P(\text{race} \mid \text{surname})$, the Florida census data set provides additional information about $P(\text{race} \mid \text{residence})$ (proportion of each racial category among

residents in a given voting district) and $P(\text{residence})$ (proportion of residents who live in a given voting district). We wish to combine them and compute the desired conditional probability $P(\text{race} \mid \text{surname and residence})$. Recall that these are general ways to represent marginal, conditional, and joint probabilities. Each expression can be evaluated using a specific racial group, surname, and residential location.

Computing $P(\text{race} \mid \text{surname and residence})$ requires Bayes' rule. So far, we have employed Bayes' rule for one event A conditional on an event B, but now we need to use Bayes' rule conditional on both B and another event C:

$$P(A \mid B, C) = \frac{P(B \mid A \text{ and } C)P(A \mid C)}{P(B \mid C)},$$

where every probability on the right-hand side is defined conditional on another event C (see equation (6.21)). Applying this rule yields

$$P(\text{race} \mid \text{surname and residence}) = \frac{P(\text{surname} \mid \text{race and residence})P(\text{race} \mid \text{residence})}{P(\text{surname} \mid \text{residence})}. \quad (6.22)$$

In this equation, while $P(\text{race} \mid \text{residence})$ is available from the Florida census data, the other two conditional probabilities, $P(\text{surname} \mid \text{race and residence})$ and $P(\text{surname} \mid \text{residence})$, are not directly given either in the census name data set or the Florida census data set.

To overcome this difficulty, we make an additional assumption that a voter's surname and residence location are independent of each other, given race. This *conditional independence* assumption implies that once we know a voter's race, their residence location does not give us any additional information about their surname. So long as there is no strong geographical concentration of certain surnames in Florida within a racial category, this assumption is reasonable. The assumption is violated, for example, if Hispanic Cubans tend to have distinct names and are concentrated in certain neighborhoods. Unfortunately, our data cannot tell us whether this assumption is appropriate, but we will proceed assuming it is. Applying equation (6.20), the assumption can be written as

$$\begin{aligned} P(\text{surname} \mid \text{race and residence}) &= \frac{P(\text{surname and race} \mid \text{residence})}{P(\text{race} \mid \text{residence})} \\ &= \frac{P(\text{surname} \mid \text{residence})P(\text{race} \mid \text{residence})}{P(\text{race} \mid \text{residence})} \\ &= P(\text{surname} \mid \text{race}). \end{aligned} \quad (6.23)$$

The first equality follows from the definition of conditional probability, whereas the second equality is due to the application of equation (6.20).

The assumption transforms equation (6.22) into

$$P(\text{race} \mid \text{surname and residence}) = \frac{P(\text{surname} \mid \text{race})P(\text{race} \mid \text{residence})}{P(\text{surname} \mid \text{residence})}.$$

We should keep this key version of the equation in mind as the one we will ultimately use.

Note that applying the law of total probability defined in equation (6.14) and then invoking the assumption given in equation (6.23), the denominator of equation (6.22) can be written as the following equation, which sums over all racial categories:

$$P(\text{surname} \mid \text{residence}) = \sum_{\text{race}} P(\text{surname} \mid \text{race and residence}) P(\text{race} \mid \text{residence})$$

$$= \sum_{\text{race}} P(\text{surname} \mid \text{race}) P(\text{race} \mid \text{residence}). \tag{6.24}$$

In the above equations, we use $\sum_{\text{race}}$ to indicate summation over all categories of the race variable (i.e., black, white, Asian, Hispanic, and others).

While the census surname list gives $P(\text{race} \mid \text{surname})$, the prediction of individual race based on equation (6.22) requires the computation of $P(\text{surname} \mid \text{race})$, which is included in both the numerator and the denominator (see equation (6.24)). Fortunately, we can use Bayes' rule to obtain

$$P(\text{surname} \mid \text{race}) = \frac{P(\text{race} \mid \text{surname}) P(\text{surname})}{P(\text{race})}. \tag{6.25}$$

The two terms in the numerator of equation (6.25) can be computed using the census name list. We compute $P(\text{race})$, which is not included in that data, from the Florida census data by using the law of total probability:

$$P(\text{race}) = \sum_{\text{residence}} P(\text{race} \mid \text{residence}) P(\text{residence}). \tag{6.26}$$

In this equation, $\sum_{\text{residence}}$ indicates summation over all values of the residence variable (i.e., all voting districts in the data).

To implement this prediction methodology in R, we first compute $P(\text{race})$ using equation (6.26). We do so by calculating a *weighted average* of percentages for each racial category across voting districts with the population of the voting district, which is proportional to $P(\text{residence})$, as the weight. The `weighted.mean()` function can be used to compute weighted averages, in which the `weights` argument takes a vector of weights.

```
FLCensus <- read.csv("FLCensusVTD.csv")
## compute proportions by applying weighted.mean() to each column
race.prop <-
    apply(FLCensus[, c("white", "black", "api", "hispanic", "others")],
          2, weighted.mean, weights = FLCensus$total.pop)
race.prop # race proportions in Florida

##      white      black        api   hispanic     others
## 0.60451586 0.13941679 0.02186662 0.21279972 0.02140101
```

We can now compute $P(\text{surname} \mid \text{race})$ using equation (6.25) and the census name list.

```
total.count <- sum(cnames$count)
## P(surname | race) = P(race | surname) * P(surname) / P(race)
cnames$name.white <- (cnames$pctwhite / 100) *
    (cnames$count / total.count) / race.prop["white"]
cnames$name.black <- (cnames$pctblack / 100) *
    (cnames$count / total.count) / race.prop["black"]
cnames$name.hispanic <- (cnames$pcthispanic / 100) *
    (cnames$count / total.count) / race.prop["hispanic"]
cnames$name.asian <- (cnames$pctapi / 100) *
    (cnames$count / total.count) / race.prop["api"]
cnames$name.others <- (cnames$pctothers / 100) *
    (cnames$count / total.count) / race.prop["others"]
```

Next, we compute the denominator of equation (6.22), $P(\text{surname} \mid \text{residence})$, using equation (6.24). To do this, we merge the census data into the voter file data using the `county` and `VTD` variables. In the `merge()` function, we set the `all` argument to `FALSE` so that nonmatching rows in both data sets will be dropped (see section 4.2.5). Since the census data includes $P(\text{race} \mid \text{residence})$ as a variable for each racial category, the merged data set will as well.

```
FLVoters <- merge(x = FLVoters, y = FLCensus, by = c("county", "VTD"),
                  all = FALSE)
## P(surname | residence) = sum_race P(surname | race) P(race | residence)
indx <- match(FLVoters$surname, cnames$surname)
FLVoters$name.residence <- cnames$name.white[indx] * FLVoters$white +
    cnames$name.black[indx] * FLVoters$black +
    cnames$name.hispanic[indx] * FLVoters$hispanic +
    cnames$name.asian[indx] * FLVoters$api +
    cnames$name.others[indx] * FLVoters$others
```

We have now calculated every quantity contained in our key version of equation (6.22): $P(\text{surname} \mid \text{race})$, $P(\text{race} \mid \text{residence})$, and $P(\text{surname} \mid \text{residence})$. Finally, we plug the quantities into the equation to compute the predicted probability that an individual belongs to a particular race, given his or her surname and residence.

```
## P(race | surname, residence) = P(surname | race) * P(race | residence)
##                                / P(surname | residence)
FLVoters$pre.white <- cnames$name.white[indx] * FLVoters$white /
    FLVoters$name.residence
```

```
FLVoters$pre.black <- cnames$name.black[indx] * FLVoters$black /
    FLVoters$name.residence
FLVoters$pre.hispanic <- cnames$name.hispanic[indx] * FLVoters$hispanic /
    FLVoters$name.residence
FLVoters$pre.asian <- cnames$name.asian[indx] * FLVoters$api /
    FLVoters$name.residence
FLVoters$pre.others <- 1 - FLVoters$pre.white - FLVoters$pre.black -
    FLVoters$pre.hispanic - FLVoters$pre.asian
```

We evaluate the accuracy of this prediction methodology and assess how much improvement knowledge of the voters' location of residence yields. We begin by examining true positives for each race using the same programming trick as before.

```
## relevant variables
vars1 <- c("pre.white", "pre.black", "pre.hispanic", "pre.asian",
           "pre.others")
## white
whites <- subset(FLVoters, subset = (race == "white"))
mean(apply(whites[, vars1], 1, max) == whites$pre.white)

## [1] 0.9371366

## black
blacks <- subset(FLVoters, subset = (race == "black"))
mean(apply(blacks[, vars1], 1, max) == blacks$pre.black)

## [1] 0.6474954

## Hispanic
hispanics <- subset(FLVoters, subset = (race == "hispanic"))
mean(apply(hispanics[, vars1], 1, max) == hispanics$pre.hispanic)

## [1] 0.85826

## Asian
asians <- subset(FLVoters, subset = (race == "asian"))
mean(apply(asians[, vars1], 1, max) == asians$pre.asian)

## [1] 0.6071429
```

The true positive rate for blacks has jumped from 16% to 65%. Minor improvements are also made for Hispanic and Asian voters. Since African-Americans tend to live close to one another in the United States, the location of voters' residences can be informative. For example, according to the census data, among people whose surname is "White," 27% are black. However, once we incorporate the location of their residence, the predicted probability of such individuals being black ranges from 1% to 98%. This implies that we predict some voters to be highly likely black and others highly likely nonblack.

```
## proportion of blacks among those with surname "White"
cnames$pctblack[cnames$surname == "WHITE"]

## [1] 27.38

## predicted probability of being black given residence location
summary(FLVoters$pre.black[FLVoters$surname == "WHITE"])

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.005207 0.081150 0.176300 0.264000 0.320000 0.983700
```

Finally, we compute the false positive rate for each race.

```
## white
1 - mean(FLVoters$race[apply(FLVoters[, vars1], 1, max)==
                       FLVoters$pre.white] == "white")

## [1] 0.1187425

## black
1 - mean(FLVoters$race[apply(FLVoters[, vars1], 1, max)==
                       FLVoters$pre.black] == "black")

## [1] 0.2346491

## Hispanic
1 - mean(FLVoters$race[apply(FLVoters[, vars1], 1, max) ==
                       FLVoters$pre.hispanic] == "hispanic")

## [1] 0.2153709

## Asian
1 - mean(FLVoters$race[apply(FLVoters[, vars1], 1, max) ==
                       FLVoters$pre.asian] == "asian")

## [1] 0.3461538
```

We find that the false positive rate for whites is significantly reduced. This is in large part due to the fact that many of the black voters who were incorrectly classified as whites using surname alone are now predicted to be black. In addition, the false positive rate for blacks lowered by a similar amount. This example illustrates the powerful use of conditional probability and Bayes' rule.

6.3 Random Variables and Probability Distributions

We have so far considered various events including a coin landing on heads, twins being both boys, and a voter being African-American. In this section, we introduce the concept of *random variables* and their *probability distributions*, which further widens the scope of mathematical analyses of these events.

6.3.1 RANDOM VARIABLES

A random variable assigns a number to each event. For example, two outcomes of a coin flip can be represented by a binary random variable where 1 indicates landing on heads and 0 denotes landing on tails. Another example is one's income measured in dollars. The values of random variables must represent *mutually exclusive and exhaustive* events. That is, different values cannot represent the same event and all events should be represented by some values. For example, consider a random variable that represents one's racial group using five unique integers: black $=$ 1, white $=$ 2, hispanic $=$ 3, asian $=$ 4, and others $=$ 5. According to this definition, someone who self-identifies as black and white will be assigned the value of 5 instead of taking the values of 1 and 2 at the same time.

There are two types of random variables, depending on the type of values they take. The first is a *discrete random variable*, which takes a finite (or at most countably infinite) number of distinct values. Examples include categorical or factor variables such as racial groups and number of years of education. The second type is a *continuous random variable*, which takes a value within an interval of the real line. That is, the variable can assume uncountably many values. Examples of continuous random variables include height, weight, and gross domestic product (GDP). The use of random variables, instead of events, facilitates the development of mathematical rules for probability because a random variable takes numeric values. Once we define a random variable, we can formalize a *probability model* using the distribution of the random variable.

A **random variable** assigns a numeric value to each event of the experiment. These values represent mutually exclusive and exhaustive events, together forming the entire sample space. A **discrete random variable** takes a finite or at most countably infinite number of distinct values, whereas a **continuous random variable** assumes an uncountably infinite number of values.

6.3.2 BERNOULLI AND UNIFORM DISTRIBUTIONS

We first consider the simplest example of a *discrete random variable*: a coin flip. For this experiment, we define a *binary random variable* X, which is equal to 1 if a coin lands on heads, and 0 otherwise. In general, a random variable that takes two distinct values is called a *Bernoulli random variable*. Notice that this setup applies to any experiment with two distinct events. Examples include {vote, abstain}, {win election, lose election}, and {correct classification, misclassification}. Thus, whether a voter turns out ($X = 1$) or not ($X = 0$) can be represented by a Bernoulli random variable. Generically, we consider the event $X = 1$ a success and the event $X = 0$ a failure. We use p to denote the probability of success.

The distribution of a discrete random variable can be characterized by the *probability mass function* (*PMF*). The PMF $f(x)$ of a random variable X is defined as the probability that the random variable takes a particular value x, i.e., $f(x) = P(X = x)$. That is, given the input x, which is a specific value of choice, the PMF $f(x)$ returns as

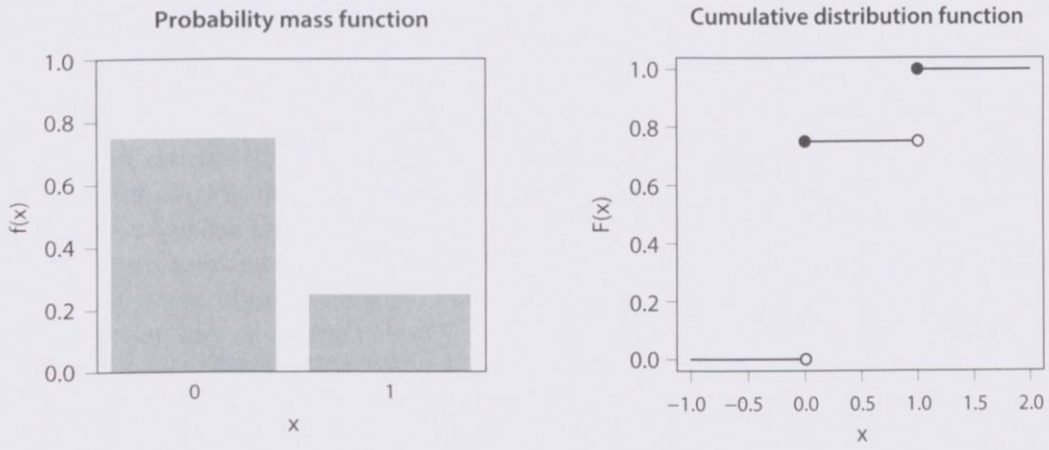


Figure 6.5. The Probability Mass and Cumulative Distribution Functions for a Bernoulli Random Variable. The probability of success is 0.25. The open and solid circles represent the exclusion and inclusion of the corresponding points, respectively.

the output the probability that a random variable X takes that value x. In the case of a Bernoulli random variable, the PMF takes the value of p when $x = 1$ and that of $1 - p$ when $x = 0$. The function is zero at all other values of x.

Another important function related to probability distribution is the *cumulative distribution function* (*CDF*). The CDF $F(x)$ represents the cumulative probability that a random variable X takes a value equal to or less than a specific value x, i.e., $F(x) = P(X \leq x)$. The CDF, therefore, represents the sum of the PMF $f(x)$ evaluated at all values up to x. Formally, the relationship between the PMF $f(x)$ and the CDF $F(x)$ for a discrete random variable can be written as

$$F(x) = P(X \leq x) = \sum_{k \leq x} f(k),$$

where k represents all values the random variable X can take that are less than or equal to x. That is, the CDF equals the sum of the PMFs. The CDF ranges from 0 to 1 for any random variable, whether continuous or discrete. It is a nondecreasing function because as x increases, more probability will be added.

The CDF $F(x)$ for a Bernoulli random variable is simple. It is zero for all negative values of x because the random variable never assumes any of those values. The CDF then takes the value of $1 - p$ when $x = 0$, which is the probability that X equals 0. The function stays flat at $1 - p$ when $0 \leq x < 1$ because none of these values will be realized. At $x = 1$, the CDF equals 1 because the random variable takes either the value of 0 or 1, and stays at this value when $x \geq 1$ because X does not take any value greater than 1. Figure 6.5 graphically displays the PMF and CDF of a Bernoulli random variable when $p = 0.25$. The open and solid circles represent the exclusion and inclusion of the corresponding points, respectively.

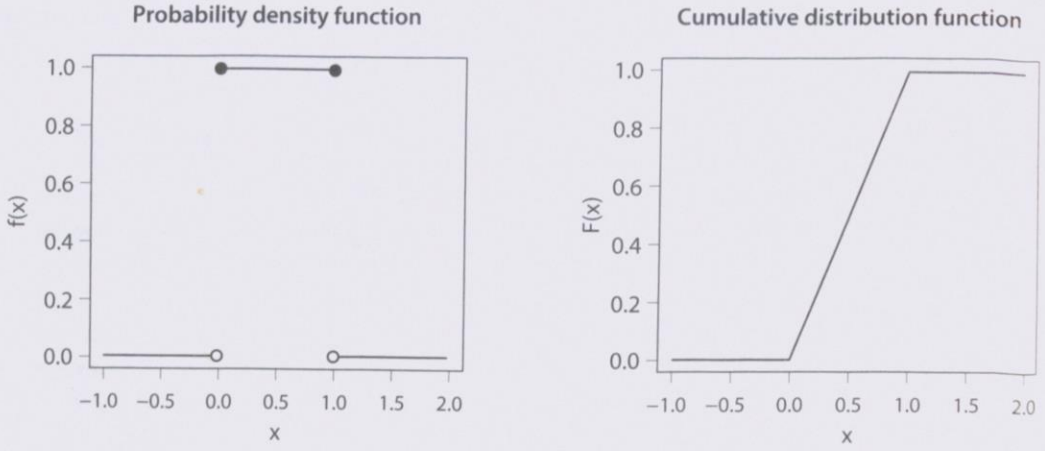


Figure 6.6. The Probability Density and Cumulative Distribution Functions for a Uniform Random Variable. The interval is set to [0, 1]. The open and solid circles represent the exclusion and inclusion of the corresponding points, respectively.

The **probability mass function** (PMF) of a **Bernoulli random variable** with success probability p is given by

$$f(x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $f(1)$ and $f(0)$ represent the probability of success and failure, respectively. The **cumulative distribution function** (CDF) is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - p & \text{if } 0 \leq x < 1, \\ 1 & \text{if } x \geq 1. \end{cases}$$

We now discuss a *uniform random variable* as a simple example of a *continuous random variable*. A uniform random variable takes every value within a given interval $[a, b]$ with equal likelihood. The PMF is not defined for a continuous random variable because this variable assumes an uncountably infinite number of values. Instead, we use the *probability density function* (*PDF*) $f(x)$ (or simply, density function), which quantifies the likelihood that a continuous random variable X will take a specific value x. We have already seen the concept of *density*, which is used to measure the height of bins in a histogram (see section 3.3.2). The value of the PDF is nonnegative and can be greater than 1. Moreover, like density in histograms, the area under the PDF must sum to 1.

Since each value within the interval is equally likely to be realized, the PDF for the uniform distribution is a flat horizontal line defined by $1/(b - a)$. In other words, the PDF does not depend on x and always equals $1/(b - a)$ within the interval. The height

is determined so that the area below the line equals 1 as required. The left-hand plot of figure 6.6 graphically displays the PDF for a uniform distribution when the interval is set to [0, 1].

We can also define the *cumulative distribution function* (CDF) for a continuous random variable. The definition of the CDF is the same as the case of discrete random variables. That is, the CDF $F(x)$ represents the probability that a random variable X takes a value less than or equal to a specific value x, i.e., $P(X \leq x)$. Graphically, the CDF corresponds to the area under the probability density function curve up to the value x (from negative infinity). Mathematically, this notion can be expressed using integration instead of summation:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt.$$

Since the entire area under the probability density curve has to sum to 1, we have $F(x) = 1$ when $x = \infty$. The CDF for the uniform distribution is shown in the right-hand plot of figure 6.6. In this case, the CDF is a straight line, as shown in the right-hand plot of the figure, because the area under the PDF increases at a constant rate.

The **probability density function** (PDF) of a **uniform random variable** with interval $[a, b]$ is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

The **cumulative probability function** (CDF) is given by

$$F(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} & \text{if } a \leq x < b, \\ 1 & \text{if } x \geq b. \end{cases}$$

We can easily compute the PDF and CDF of a uniform distribution in R. For the PDF $f(x)$, we use the `dunif()` function where the main argument is the value x at which the function is evaluated and the interval is specified using the `min` and `max` arguments. We can compute the CDF in a similar manner using the `punif()` function. The `d` in `dunif()` indicates density, whereas the `p` in `punif()` stands for probability.

```
## uniform PDF: x = 0.5, interval = [0, 1]
dunif(0.5, min = 0, max = 1)

## [1] 1

## uniform CDF: x = 1, interval = [-2, 2]
punif(1, min = -2, max = 2)

## [1] 0.75
```

The two distributions we have introduced here share a useful connection. We can use a uniform random variable to generate a Bernoulli random variable. To do this, notice that under the uniform distribution with unit interval $[0, 1]$, the CDF is given by the 45-degree line, i.e., $F(x) = x$. Therefore, the probability that this uniform random variable X takes a value less than or equal to x is equal to x when $0 \leq x \leq 1$. Thus, in order to generate a Bernoulli random variable Y with success probability p, we can first sample a uniform random variable X and then set $Y = 1$ when X is less than p (similarly, set $Y = 0$ if $X \geq p$) so that Y takes a value of 1 with probability p. To do this *Monte Carlo simulation* in R, we use the `runif()` function to generate a uniform random variable by setting the `min` and `max` arguments to `0` and `1`, respectively.

```
sims <- 1000
p <- 0.5 # success probabilities
x <- runif(sims, min = 0, max = 1) # uniform [0, 1]
head(x)

## [1] 0.292614295 0.619951024 0.004618747 0.162426728
## [5] 0.001157040 0.655518809

y <- as.integer(x <= p) # Bernoulli; turn TRUE/FALSE to 1/0
head(y)

## [1] 1 0 1 1 1 0

mean(y) # close to success probability p, proportion of 1s vs. 0s

## [1] 0.521
```

6.3.3 BINOMIAL DISTRIBUTION

The *binomial distribution* is a generalization of the Bernoulli distribution. Instead of a single coin flip, we consider an experiment in which the same coin is flipped independently and multiple times. That is, a binomial random variable can represent the number of times a coin lands on heads in multiple trials of independent coin flips.

More generally, a binomial random variable X records the number of successes in a total of n independent and identical trials with success probability p. In other words, a binomial random variable is the sum of n *independently and identically distributed* (or *i.i.d.* in short) Bernoulli random variables. Recall that a Bernoulli random variable equals either 1 or 0 with success probability p. Thus, X can take an integer value from 0 to n. Since the binomial distribution is discrete, its PMF can be interpreted as the probability of X taking a specific value x. The CDF represents the cumulative probability that a binomial random variable has x or fewer successes out of n trials. The PMF and CDF of a binomial random variable are given by the following formulas, which involve combinations (see equation (6.10)). No simple expression exists for the CDF, which is written as the sum of the PMFs.

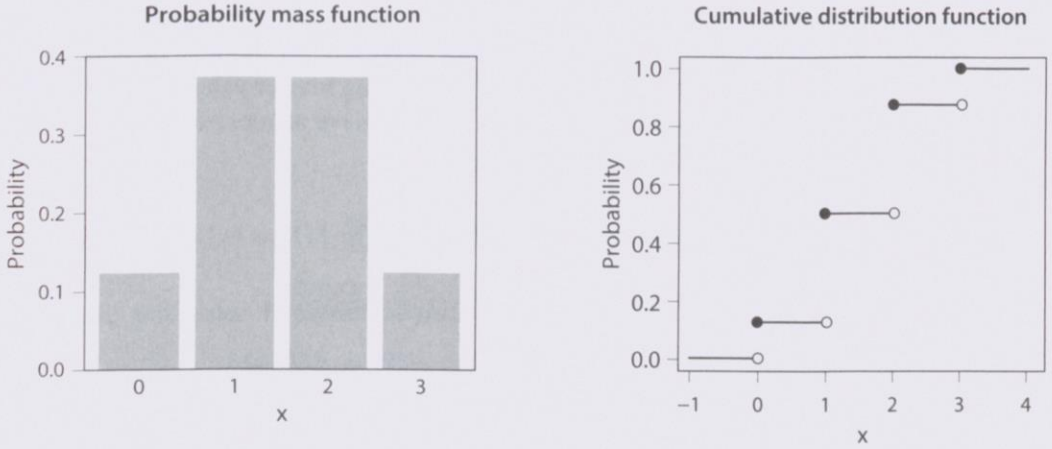


Figure 6.7. The Probability Mass and Cumulative Distribution Functions for a Binomial Random Variable. The success probability is 0.5 and the total number of trials is 3. The open and solid circles represent the exclusion and inclusion of the corresponding points, respectively. Source: Adapted from example by Paul Gaborit, http://texample.net.

The probability mass function (PMF) of a **binomial random variable** with success probability p and n trials is given by

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{6.27}$$

The cumulative distribution function (CDF) can be written as

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k},$$

for $x = 0, 1, \ldots, n$.

Figure 6.7 shows the PMF and CDF when $p = 0.5$ and $n = 3$. For example, we can compute the probability that we obtain two successes out of three trials, which is the height of the third bar in the left-hand plot of the figure:

$$f(2) = P(X = 2) = \binom{3}{2} \times 0.5^2 \times (1 - 0.5)^{3-2} = \frac{3!}{(3-2)!2!} \times 0.5^3 = 0.375.$$

Calculating the PMF of a binomial distribution is straightforward. The `dbinom()` function takes the number of successes as the main argument, and the `size` and `prob` arguments specify the number of trials and success probability, respectively.

```
## PMF when x = 2, n = 3, p = 0.5
dbinom(2, size = 3, prob = 0.5)

## [1] 0.375
```

The CDF, shown in the right-hand plot of the figure, is a *step function* where the function is flat and then jumps at each nonnegative integer value. The size of each jump equals the height of the PMF at the corresponding integer value. Using the CDF, we can compute the cumulative probability that we have at most one success out of three trials:

$$F(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = f(0) + f(1) = 0.125 + 0.375 = 0.5.$$

We can compute the CDF of a binomial distribution in R using the `pbinom()` function.

```
## CDF when x = 1, n = 3, p = 0.5
pbinom(1, size = 3, prob = 0.5)

## [1] 0.5
```

An intuitive explanation covers why the PMF of a binomial distribution looks like equation (6.27). When we flip a coin n times, each unique sequence of n outcomes is equally likely. For example, if $n = 5$, then the event that only the last two coin flips land on tails $\{HHHTT\}$ is equally as likely as the event that the flips alternate landing on heads and tails $\{HTHTHT\}$, where we use H and T to denote the events that a coin lands on heads and tails, respectively. However, for the binomial distribution only the number of heads matters. As a result, these two events represent the same outcome. We use combinations to count the number of ways we can have x successes out of n trials, which is equal to ${}_nC_x = \binom{n}{x}$. We multiply this by the probability of x successes, which is equal to p^x (because each trial is independent), and the probability of $n - x$ failures, which is given by $(1 - p)^{n-x}$ (again because of independence).

As an application of the binomial distribution, consider the probability that one's vote is pivotal in an election. Your vote is pivotal if the election is tied before you cast your ballot. Suppose that in a large population exactly 50% of voters support an incumbent while the other half support a challenger. Further, assume that whether voters turn out or not has nothing to do with their vote choice. Under this scenario, what is the probability that the election ends up with an exact tie? We compute this probability when the number of voters who turn out equals 1000, then 10,000, and then 100,000. To compute this probability, we can evaluate the PMF of the binomial distribution by setting the success probability to 50% and the size to the total number of voters who turn out. We then evaluate the PMF at exactly half of all voters who turn out. We find that the probability of a tie is quite small, even when the population of voters is evenly divided.

```
## number of voters who turn out
voters <- c(1000, 10000, 100000)
dbinom(voters / 2, size = voters, prob = 0.5)

## [1] 0.025225018 0.007978646 0.002523126
```

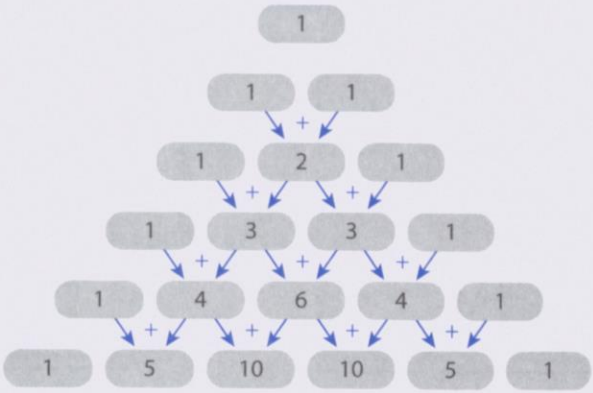


Figure 6.8. Pascal's Triangle. Binomial coefficients can be represented as Pascal's triangle, where the xth element of the nth row returns the binomial coefficient $\binom{n-1}{x-1}$. Source: Adapted from example by Paul Gaborit, http://texample.net.

Where does the name "binomial distribution" come from? The name of this distribution is based on the following *binomial theorem*.

The **binomial theorem** shows how to compute the coefficient of each term when expanding the power of a binomial, i.e., $(a + b)^n$. That is, the coefficient for the term $a^x b^{n-x}$ when expanding $(a + b)^n$ is equal to $\binom{n}{x}$.

For example, according to the binomial theorem, when $n = 4$, the coefficient for the term a^2b^2 when expanding $(a + b)^4$ is equal to $\binom{4}{2} = 6$. This result is confirmed by writing out the entire expansion:

$$(a + b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4. \tag{6.28}$$

These binomial coefficients can be organized as *Pascal's triangle*, as shown in figure 6.8. For example, the coefficients for the terms resulting from the expansion of $(a + b)^4$ in equation (6.28) are shown in the fifth row of Pascal's triangle. More generally, in Pascal's triangle, the xth element of the nth row represents the binomial coefficient $\binom{n-1}{x-1}$. In addition, as shown in the figure, each element equals the sum of the two elements just above it, leading to a straightforward sequential computation of binomial coefficients. This makes sense because, for example, $(a + b)^4$ can be written as the product of $(a + b)^3$ and $(a + b)$,

$$(a + b)^4 = (a^3 + 3a^2b + 3ab^2 + b^3)(a + b).$$

In this example, the coefficient for a^2b^2 is based on the sum of two products, i.e., $3a^2b \times b$ and $3ab^2 \times a$, and hence is equal to $6 = 3 + 3$. In general, to obtain x

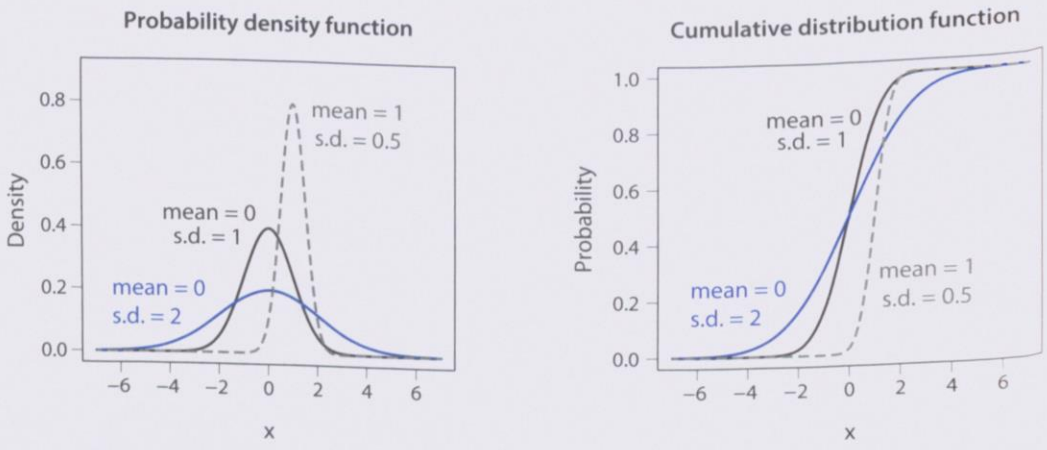


Figure 6.9. The Probability Density and Cumulative Distribution Functions of the Normal Distribution.

success combinations out of n trials, we consider two scenarios—the last trial ending in a success or ending in a failure—and add the total number of combinations under each scenario:

$$\binom{n-1}{x} + \binom{n-1}{x-1} = \frac{(n-1)!}{x!(n-x-1)!} + \frac{(n-1)!}{(x-1)!(n-x)!}$$

$$= (n-1)! \times \frac{(n-x)+x}{x!(n-x)!} = \binom{n}{x}.$$

The first (second) term corresponds to the scenario where there are x $(x-1)$ successes out of $(n-1)$ trials and the last trial ends in a failure (success).

6.3.4 NORMAL DISTRIBUTION

As another important example of a continuous random variable, we introduce the *normal distribution*. This distribution is also called the *Gaussian distribution*, named after German mathematician Carl Friedrich Gauss. As implied by its name, the normal distribution is special because, as section 6.4.2 will explore, the sum of many random variables from the same distribution tends to follow the normal distribution even when the original distribution is not normal.

A normal random variable can take any number on the real line $(-\infty, \infty)$. The normal distribution has two parameters, mean μ and standard deviation σ. If X is a normal random variable, we may write $X \sim \mathcal{N}(\mu, \sigma^2)$, where σ^2 represents the variance (the square of standard deviation). The PDF and the CDF of the normal distribution are given by the following formulas.

The probability density function (PDF) of a **normal random variable** is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},$$

for any x on the real line. The cumulative probability distribution (CDF) has no analytically tractable form and is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(t-\mu)^2\right\} dt, \tag{6.29}$$

where $X \sim \mathcal{N}(\mu, \sigma^2)$ and $\exp(\cdot)$ is the exponential function (see section 3.4.1). The CDF represents the area under the PDF from negative infinity up to x.

Figure 6.9 plots the PDF (left-hand plot) and CDF (right-hand plot) for the normal distribution, with three different sets of the mean and standard deviation. The PDF of the normal distribution is bell shaped and centered around its mean, with the standard deviation controlling the spread of the distribution. When the mean is 0 and standard deviation is 1, we have the *standard normal distribution*. The PDF is symmetric around the mean. Different means shift the PDF and CDF without changing their shape. In contrast, a larger standard deviation means more variability, yielding a flatter PDF and a more gradually increasing CDF.

The normal distribution has two important properties. First, adding a constant to (or subtracting it from) a normal random variable yields a normal random variable with appropriately shifted mean. Second, multiplying (or dividing) a normal random variable by a constant also yields another normal random variable with an appropriately scaled mean and standard deviation. Accordingly, the z-score of a normal random variable follows the standard normal distribution. We formally state these properties below.

Suppose X is a normal random variable with mean μ and standard deviation σ, i.e., $X \sim \mathcal{N}(\mu, \sigma^2)$. Let c be an arbitrary constant. Then, the following properties hold:

1. A random variable defined by $Z = X + c$ also follows a normal distribution, with $Z \sim \mathcal{N}(\mu + c, \sigma^2)$.
2. A random variable defined by $Z = cX$ also follows a normal distribution, with $Z \sim \mathcal{N}(c\mu, (c\sigma)^2)$.

These properties imply that the **z-score** of a normal random variable follows the standard normal distribution, which has zero mean and unit variance:

$$z\text{-score} = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1).$$

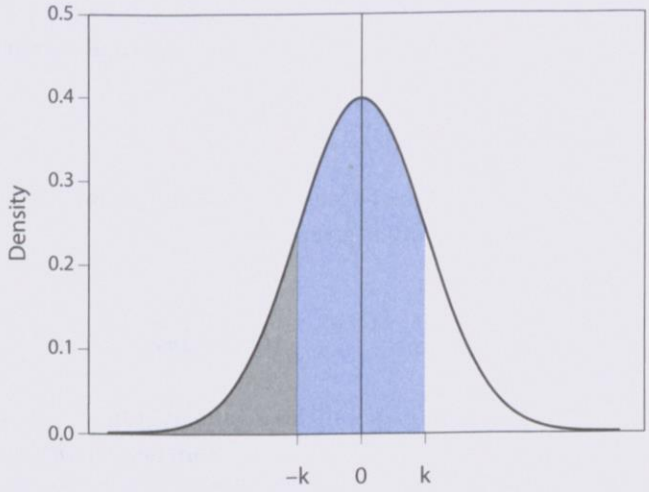


Figure 6.10. The Area under the Probability Density Function Curve of the Normal Distribution. The blue area can be computed as the difference between the cumulative distribution function (CDF) evaluated at k and $-k$ (i.e., the gray and blue areas minus the gray area).

In addition, it is important to note that if the data are distributed according to the normal distribution, about two-thirds are within 1 standard deviation from the mean and approximately 95% are within 2 standard deviations from the mean. Let us compute the probability that a normal random variable with mean μ and standard deviation σ lies within k standard deviations from the mean for a positive constant $k > 0$. To simplify the computation, consider the *z-score*, which has the standard normal distribution:

$$\begin{aligned} P(\mu - k\sigma \leq X \leq \mu + k\sigma) &= P(-k\sigma \leq X - \mu \leq k\sigma) \\ &= P\left(-k \leq \frac{X - \mu}{\sigma} \leq k\right) \\ &= P(-k \leq Z \leq k), \end{aligned}$$

where Z is a standard normal random variable. The first equality holds because we subtract μ from each term whereas the second inequality holds since we divide each term by a positive constant σ.

Thus, the desired probability equals the probability that a standard normal random variable lies between $-k$ and k. As illustrated in figure 6.10, this probability can be written as the difference in the CDF evaluated at k and $-k$:

$$P(-k \leq Z \leq k) = P(Z \leq k) - P(Z \leq -k) = F(k) - F(-k),$$

where $F(k)$ represents the sum of the blue and gray areas in the figure, whereas $F(-k)$ equals the gray area. These results can be confirmed in R with the `pnorm()` function, which evaluates the CDF at its input value. This function takes the mean (`mean`)

and standard deviation (`sd`) as two important arguments. The default is the standard normal distribution with `mean = 0` and `sd = 1`.

```
## plus minus 1 standard deviation from the mean
pnorm(1) - pnorm(-1)

## [1] 0.6826895

## plus minus 2 standard deviations from the mean
pnorm(2) - pnorm(-2)

## [1] 0.9544997
```

The result suggests that, under the standard normal distribution, approximately 2/3 are within 1 standard deviation from the mean and about 95% are within 2 standard deviations from the mean. We can also directly specify mean and standard deviation without transforming a variable into a standard normal random variable. Suppose that the original distribution has a mean of 5 and standard deviation of 2, i.e., $\mu = 5$ and $\sigma = 2$. We can compute the same probabilities as above in the following way.

```
mu <- 5
sigma <- 2
## plus minus 1 standard deviation from the mean
pnorm(mu + sigma, mean = mu, sd = sigma) - pnorm(mu - sigma, mean = mu, sd = sigma)

## [1] 0.6826895

## plus minus 2 standard deviations from the mean
pnorm(mu + 2*sigma, mean = mu, sd = sigma) - pnorm(mu - 2*sigma, mean = mu, sd = sigma)

## [1] 0.9544997
```

As an application of the normal distribution, consider the *regression towards the mean* phenomenon discussed in section 4.2.4. In that section, we presented evidence from US presidential elections demonstrating that in states where Obama received a large share of votes in 2008, he was likely to receive a *smaller* share of votes in 2012 (see section 4.2.5). Recall that our regression model used Obama's 2008 statewide vote share to predict his vote share for the same state in the 2012 election. We use the regression object `fit1`, as created in section 4.2.5.

```
## see the page reference above
## "Obama2012.z" is Obama's 2012 standardized vote share
## "Obama2008.z" is Obama's 2008 standardized vote share
fit1

##
## Call:
## lm(formula = Obama2012.z ~ -1 + Obama2008.z, data = pres)
```