GLOBAL EDITION

STATISTICS

THE ART AND SCIENCE OF LEARNING FROM DATA

FIFTH EDITION

Alan Agresti • Christine A. Franklin • Bernhard Klingenberg

Statistics

The Art and Science of Learning from Data

Fifth Edition
Global Edition

Alan Agresti
University of Florida

Christine Franklin
University of Georgia

Bernhard Klingenberg
Williams College and New College of Florida

Pearson

Events are usually denoted by letters from the beginning of the alphabet, such as A and B, or by a letter or string of letters that describes the event. For a student taking the three-question pop quiz, some possible events are

$$A = \text{student answers all three questions correctly} = \{CCC\}$$
$$B = \text{student passes (at least two correct)} = \{CCI, CIC, ICC, CCC\}.$$

Finding Probabilities of Events

Each outcome in a sample space has a probability. So does each event. To find such probabilities, we list the sample space and specify plausible assumptions about its outcomes. Sometimes, for instance, we can assume that the outcomes are equally likely. In any case, the probabilities for the outcomes in a sample space must follow two basic rules:

- The probability of each individual outcome is between 0 and 1.
- The sum of the probabilities of the individual outcomes in a sample space equals 1.

Example 4

Probability for a Sample Space With Equally Likely Outcomes ◀

Treating Colds

Picture the Scenario

The University of Wisconsin is conducting a randomized experiment[1] to compare an herbal remedy (echinacea) to a placebo for treating the common cold. The response variables are the cold's severity and its duration. Suppose a particular clinic in Madison, Wisconsin, has four volunteers, of whom two are men (Jamal and Ken) and two are women (Linda and Mei). Two of these volunteers will be randomly chosen to receive the herbal remedy, and the other two will receive the placebo.

Questions to Explore

a. Identify the possible samples to receive the herbal remedy. For each possible sample, what is the probability that it is the one chosen?

b. What's the probability of the event that the sample chosen to receive the herbal remedy consists of one man and one woman?

Think It Through

a. To find the sample space, here it is best to think about all possible two-person groups that can be formed from the four volunteers. These are the 6 possible samples: {(Jamal, Ken), (Jamal, Linda), (Jamal, Mei), (Ken, Linda), (Ken, Mei), (Linda, Mei)}. Randomly choosing 2 out of the 4 participants means that each of these 6 samples in the sample space is equally likely to be chosen. Since there are six possible samples, each one has probability 1/6. These probabilities fall between 0 and 1, and their sum equals 1, as is necessary for probabilities for a sample space.

b. The event in which the sample chosen has one man and one woman consists of the outcomes {(Jamal, Linda), (Jamal, Mei), (Ken, Linda), (Ken, Mei)}. These are the possible pairings of one man with one woman. Each outcome has probability 1/6, so the probability of this event is $4(1/6) = 4/6 = 2/3$.

[1] See www.fammed.wisc.edu/research/news0503d.html.

Insight

When each outcome is equally likely, the probability of a single outcome is simply 1/(number of possible outcomes), such as 1/6 in part a. The probability of an event is then (number of outcomes in the event)/(number of possible outcomes), such as 4/6 in part b.

▶ ***Try Exercise 5.19***

This example shows that to find the probability for an event, we can (1) find the probability for each outcome in the sample space and (2) add the probabilities of the outcomes that the event comprises.

SUMMARY: Probability of an Event

- The probability of an event A, denoted by P(A), is obtained by adding the probabilities of the individual outcomes in the event.
- When all the possible outcomes are equally likely,

$$P(A) = \frac{\text{number of outcomes in event A}}{\text{number of outcomes in the sample space}}.$$

In Example 4, to find the probability of choosing one man and one woman, we first determined the probability of each of the six possible outcomes. Because the probability is the same for each, 1/6, and because the event contains four of those outcomes, the probability is 1/6 added four times. The answer equals 4/6, the number of outcomes in the event divided by the number of outcomes in the sample space.

In Practice Equally Likely Outcomes Are Unusual

Except for simplistic situations such as random sampling or flipping balanced coins or rolling fair dice, different outcomes are not usually equally likely. Then, probabilities are often estimated, using sample proportions from simulations or from large samples of data.

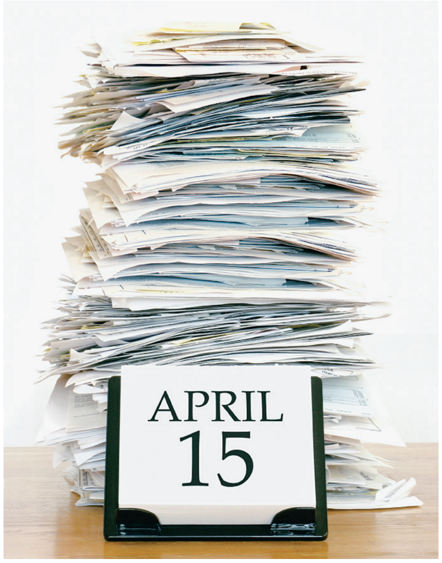
Example 5

Probabilities for a Sample Space With Outcomes That Are Not Equally Likely

Tax Audit

Picture the Scenario

April 15 is tax day in the United States—the deadline for filing federal income tax forms. The main factor in the amount of tax owed is a taxpayer's income level. Each year, the IRS audits a sample of tax forms to verify their accuracy. Table 5.2 is a contingency table that cross-tabulates the nearly 150 million long-form federal returns received in 2017 by the taxpayer's income level and whether the tax form was audited.



Recall

Section 3.1 introduced **contingency tables** to summarize the relationship between two categorical variables. ◀

Did You Know?

Most people find it easier to interpret probabilities, especially very small ones, by using their reciprocal. Thus, "there is a one in 1.9 million chance of being struck by lightning in the United States" is more accessible to most people than "the probability is 0.00000053," which is 1/1,900,000. (*Source:* http://www.lightningsafety.noaa.gov/odds.htm) ◀

Table 5.2 Contingency Table Cross-Tabulating Tax Forms by Income Level and Whether Audited

There were 149.9 million returns filed. The frequencies in the table are reported in thousands. For example, 839 represents approximately 839,000 tax forms that reported income under $200,000 and were audited.

	Audited		
Income Level	**Yes**	**No**	**Total**
Under $200,000	839	141,686	**142,525**
$200,000–$1,000,000	72	6,803	**6,875**
More than $1,000,000	23	496	**519**
Total	**934**	**148,986**	**149,919**

Source: https://www.irs.gov/pub/irs-news/fy_2017_enforcement_and_services_results_final.pdf, U.S. Department of Treasury

Questions to Explore

a. Consider, for each return, the income level and whether it was audited. If we draw a return at random, what are the possible outcomes that define the sample space?

b. What is the probability of each outcome in the sample space? Are they equally likely?

c. For a randomly selected return in 2017, what is the probability of (i) an audit, (ii) the return showing an income of more than $1,000,000?

Think It Through

a. The sample space is the set of all possible outcomes for a tax return. For each return, we recorded the income level (3 possibilities) and whether it was audited (2 possibilities), resulting in a total of 6 possible outcomes: (Under $200,000, Yes), (Under $200,000, No), ($200,000–$1,000,000, Yes), and so forth.

b. The probabilities for the 6 outcomes in the sample space are not equally likely. For instance, it is far more likely for a return to be not audited. Those outcomes that include a "No" for auditing are therefore far more likely than those that include a "Yes." To find the probability for each of the 6 outcomes, we divide the number of returns that fall in the corresponding cell of Table 5.2 by the total number of returns; e.g., the probability for the outcome (Under $200,000, Yes) equals 839/149,919 = 0.005596, about 0.56% or a chance of 1 in 179.

c. (i) The event of being audited consists of the three outcomes (Under $200,000, Yes), ($200,000 – $1,000,000, Yes) and (Over $1,000,000, Yes), with estimated probabilities of 839/149,919, 72/149,919, and 23/149,919, for a total of (839 + 72 + 23)/149,919 = 934/149,919 = 0.0062, or about a 1 in 161 chance.

(ii) The event of a return showing an income of more than $1,000,000 consists of the outcomes (Over $1,000,000, No) and (Over $1,000,000, Yes), with probability (23 + 496)/149,919 = 519/149,919 = 0.0035, or about a 1 in 289 chance.

Insight

The event of a randomly selected return being audited consists of three of the possible six outcomes. If we wrongly assumed that the six possible outcomes are equally likely, then the probability for this event would be calculated as 3/6 = 0.5, vastly different from 0.0062.

► ***Try Exercise 5.23, parts a and b***

Basic Rules for Finding Probabilities About a Pair of Events

Some events are expressed as the outcomes that (a) are *not* in some other event, or (b) are in one event *and* in another event, or (c) are in one event *or* in another event. We'll next learn how to calculate probabilities for these three cases.

The Complement of an Event For an event A, all outcomes of the sample space that are *not* in A are called the **complement** of A.

In Words

A^C reads as "**A-complement.**" The C in the superscript denotes the term complement. You can think of A^C as meaning "not A."

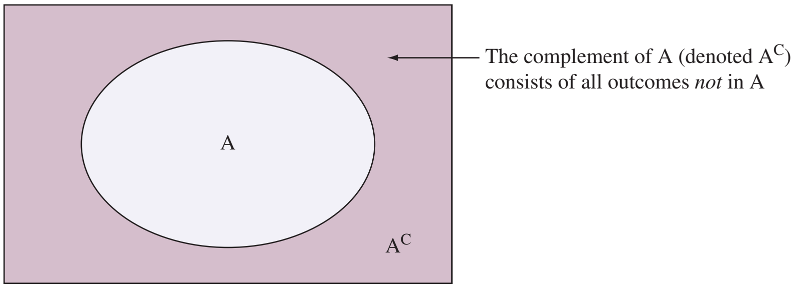
Complement of an Event

The **complement** of an event A consists of all outcomes in the sample space that are *not* in A. It is denoted by A^C. The probabilities of A and of A^C add to 1, so

$$P(A^C) = 1 - P(A).$$

In Example 5, for instance, the event of having an income *less than* \$200,000 is the complement of the event of having an income of \$200,000 or more. Because the probability that a randomly selected taxpayer had an income of under \$200,000 is 142,525/149,919 $= 0.951$, the probability of income at least \$200,000 is $1 - 0.951 = 0.049$.

Figure 5.4 illustrates the complement of an event. The box represents the entire sample space. The event A is the oval in the box. The complement of A, which is shaded, is everything else in the box that is not in A. Together, A and A^C cover the sample space. Because an event and its complement contain all possible outcomes, their total probability is 1, and the probability of either one of them is 1 minus the probability of the other. A diagram like Figure 5.4 that uses areas inside a box to represent events is called a **Venn diagram.**



▲ **Figure 5.4 Venn Diagram Illustrating an Event A and Its Complement A^C. Question** Can you sketch a Venn diagram of two events A and B such that they share some common outcomes, but some outcomes are only in A or only in B?

To find the probability of an event, it's sometimes easier to find the probability of its complement and then subtract that probability from 1. An example is when we need to find the probability that *at least one* of several events will occur. It's usually easier to first find the probability of its complement, that *none* of these events will occur and then subtract it from 1.

Complement of an Event ◀

Example 6

Women on a Jury

Picture the Scenario

A jury of 12 people is chosen for a trial. The defense attorney claims it must have been chosen in a biased manner because 50% of the city's adult residents are female, yet the jury contains no women.

Questions to Explore

If the jury were randomly chosen from the population, what is the probability that the jury would have (a) no females, (b) at least one female?

Think It Through

Let's use 12 letters, with F for female and M for male, to represent a possible jury selection. For instance, MFMMMMMMMMMM denotes the jury in which only the second person selected is female. The number of possible outcomes is $2 \times 2 \times 2 \times \ldots \times 2$, that is, 2 multiplied 12 times, which is $2^{12} = 4{,}096$. This is a case in which listing the entire sample space is not practical. Since the population is 50% male and 50% female, these 4,096 possible outcomes are equally likely.

a. Only 1 of the 4,096 possible outcomes corresponds to a no-female jury, namely, MMMMMMMMMMMM. Based on what we learned about equally likely outcomes, the probability of this outcome is 1/4,096, or 0.00024. So, if a jury is truly chosen by random sampling, the event of having no women in it is extremely unlikely (0.02%).

b. As noted previously, it would be tedious to list all possible outcomes in which at least one female is on the jury. But this is not necessary. The event that the jury contains *at least one* female is the complement of the event that it contains *no* females. Thus,

$$P(\text{at least one female}) = 1 - P(\text{no females}) = 1 - 0.00024 = 0.99976.$$

Insight

You might instead let the sample space be the possible values for the *number* of females on the jury, namely 0, 1, 2, . . . 12. But these outcomes are not equally likely. For instance, only one of the 4,096 possible samples has 0 females, but 12 of them have 1 female: The female could be the first person chosen, or the second (as in MFMMMMMMMMMM), or the third, and so on. Chapter 6 will show a formula (binomial) that gives probabilities for this alternative sample space and will allow for cases when the outcomes are not equally likely.

▶ ***Try Exercise 5.17***

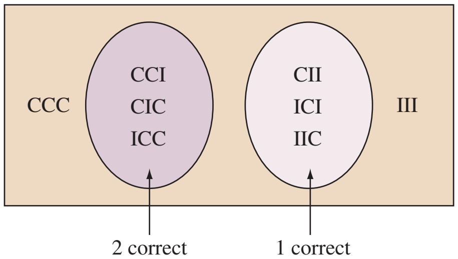
Disjoint Events Events that do not share any outcomes in common are said to be **disjoint**.

> **Disjoint Events**
>
> Two events, A and B, are **disjoint** if they do not have any common outcomes.

Did You Know?

Disjoint events are also referred to as **mutually exclusive** events. We will use only the term "disjoint." ◀

Example 3 discussed a pop quiz with three questions. The event that the student answers exactly one question correctly is {CII, ICI, IIC}. The event that the student answers exactly two questions correctly is {CCI, CIC, ICC}. These two events have no outcomes in common, so they are disjoint. In a Venn diagram, they have no overlap (Figure 5.5). By contrast, neither is disjoint from the event that the student answers the first question correctly, which is {CCC, CCI, CIC, CII}, because this event has outcomes in common with each of the other two events.



▲ **Figure 5.5 Venn Diagram Illustrating Disjoint Events.** The event of a student answering exactly one question correctly is disjoint from the event of answering exactly two questions correctly. **Question** Identify on this figure the event that the student answers the first question correctly. Is this event disjoint from either of the two events identified in the Venn diagram?

Consider an event A and its complement, A^C. They share no common outcomes, so they are disjoint events.

Intersection and Union of Events Some events are composed from other events. For instance, for two events A and B, the event that *both* occur is also an event. Called the **intersection** of A and B, it consists of the outcomes that are in both A and B. By contrast, the event that the outcome is in A *or* B or both is the **union** of A and B. It is a larger set, containing the intersection as well as outcomes that are in A but not in B and outcomes that are in B but not in A. Figure 5.6 shows Venn diagrams illustrating the intersection and union of two events.

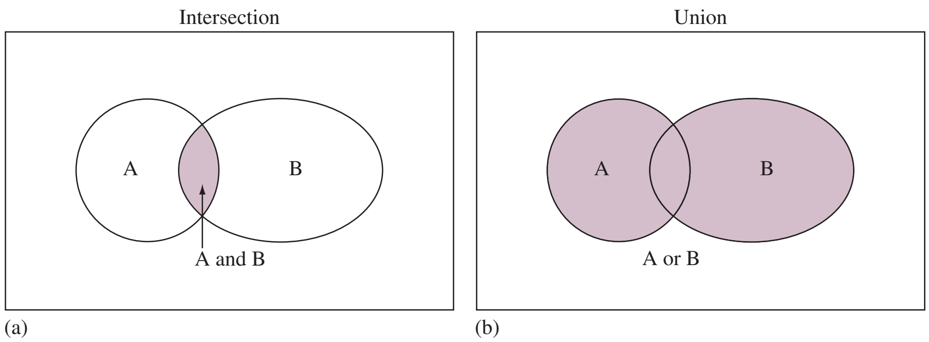
In Words

Intersection means A *and* B (the "overlap" of the events). **Union** means A *or* B or both.

Intersection and Union of Two Events

The **intersection** of A and B consists of outcomes that are in both A *and* B.

The **union** of A and B consists of outcomes that are in A *or* B or both. In probability, "A *or* B" denotes that A occurs or B occurs or both occur.



▲ **Figure 5.6 The Intersection and the Union of Two Events.** Intersection means A occurs *and* B occurs, denoted "A and B." The intersection consists of the shaded "overlap" part in Figure 5.6 (a). Union means A occurs *or* B occurs *or* both occur, denoted "A or B." It consists of all the shaded parts in Figure 5.6 (b). **Question** How could you find P(A or B) if you know P(A), P(B), and P(A and B)?