

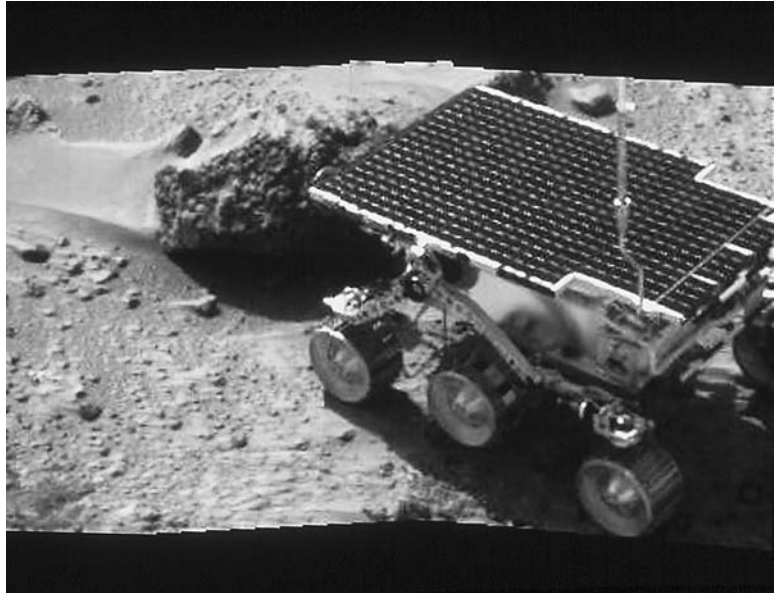
# 3

## Descriptive Study of Bivariate Data

1. Introduction
2. Summarization of Bivariate Categorical Data
3. A Designed Experiment for Making a Comparison
4. Scatter Diagram of Bivariate Measurement Data
5. The Correlation Coefficient —  
A Measure of Linear Relation
6. Prediction of One Variable from Another  
(Linear Regression)
7. Review Exercises

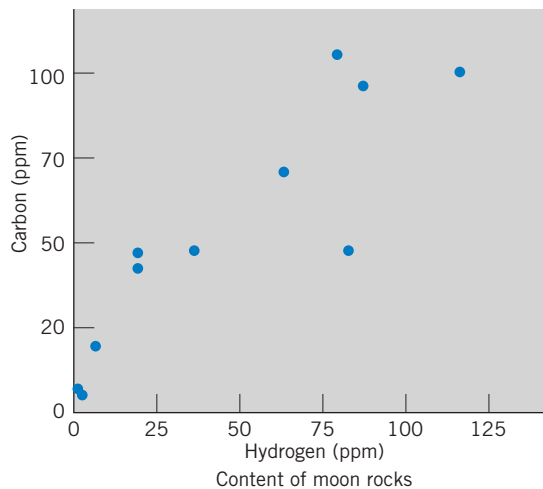
---

## Hydrogen–Carbon Association in Moon Rocks



© Photo Researchers.

In their quest for clues to the origin and composition of the planets, scientists performed chemical analyses of rock specimens collected by astronauts and unmanned space probes. The Apollo moon landings made it possible to study firsthand the geology of the moon. Eleven lunar rocks were analyzed for carbon and hydrogen content.



Rocks with large amounts of hydrogen tend to have large amounts of carbon. Other rocks tend to have small amounts of both elements, indicating a positive association between hydrogen and carbon content.

---

## 1. INTRODUCTION

In Chapter 2, we discussed the organization and summary description of data concerning a single variable. Observations on *two or more* variables are often recorded for the individual sampling units; the height and weight of individuals, or the number of goals scored by and against a team. By studying such **bivariate** or **multivariate** data, one typically wishes to discover if any relationships exist between the variables, how strong the relationships appear to be, and whether one variable of primary interest can be effectively predicted from information on the values of the other variables. To illustrate the concepts, we restrict our attention to the simplest case where only two characteristics are observed on the individual sampling units. Some examples are:

Gender and the type of occupation of college graduates.

Smoking habit and lung capacity of adult males.

Average daily carbohydrate intake and protein intake of 10-year-old children.

The age of an aircraft and the time required for repairs.

The two characteristics observed may both be qualitative traits, both numerical variables, or one of each kind. For brevity, we will only deal with situations where the characteristics observed are either both categorical or both numerical. Summarization of bivariate categorical data is discussed in Section 2. Sections 4, 5, and 6 are concerned with bivariate measurement data and treat such issues as graphical presentations, examination of relationship, and prediction of one variable from another.

## 2. SUMMARIZATION OF BIVARIATE CATEGORICAL DATA

When two traits are observed for the individual sampling units and each trait is recorded in some qualitative categories, the resulting data can be summarized in the form of a two-way frequency table. The categories for one trait are marked along the left margin, those for the other along the upper margin, and the frequency counts recorded in the cells. The total frequency for any row is given in the right-hand margin and those for any column given at the bottom margin. Both are called **marginal totals**.

Data in this summary form are commonly called **cross-classified** or **cross-tabulated data**. In statistical terminology, they are also called **contingency tables**.

### Example 1 Calculation of Relative Frequencies Aids Interpretation

Four hundred undergraduates were surveyed concerning their part-time work during the semester. The number of hours worked last week was categorized as: worked 10 or few hours, worked more than 10 hours, or did not work. The students were also categorized as underclassman or upperclassman. The cross-tabulated frequency counts are presented as Table 1.

**TABLE 1** Cross-Tabulated Frequency Counts of Work Hours

	No Job	Work 10 Hours or Less	Work More Than 10 Hours	Total
Underclassman	132	28	20	180
Upperclassman	124	44	52	220
Total	256	72	72	400

The entries of this table are self-explanatory. For instance, of the 400 students polled, 180 were underclassman. Among these, 132 did not work, 28 worked 10 hours or less, and 20 worked more than 10 hours. To gain further understanding of how the responses are distributed, calculate the relative frequencies.

**SOLUTION** For this purpose, we divide each cell frequency by the sample size 400. The relative frequencies, for instance  $44/400 = .11$ , are shown in Table 2.

**TABLE 2** Relative Frequencies for the Data of Table 1

	No Job	Work 10 Hours or Less	Work More Than 10 Hours	Total
Underclassman	.33	.07	.05	.45
Upperclassman	.31	.11	.13	.55
Total	.64	.18	.18	1.00

Depending on the specific context of a cross-tabulation, one may also wish to examine the cell frequencies relative to a marginal total. In Example 1, you may wish to compare the pattern of part-time work for underclassmen with that of the upperclassman. This is accomplished by calculating the relative frequencies separately for the two groups. For instance,  $44/220 = .200$ , as Table 3 shows.

**TABLE 3** Relative Frequencies by Class

	No Job	Work 10 Hours or Less	Work More Than 10 Hours	Total
Underclassman	.733	.156	.111	1.000
Upperclassman	.564	.200	.236	1.000

From the calculations presented in Table 3, it appears that a larger proportion of upperclassmen hold part-time jobs and that they tend to work more than 10 hours a week.

Now the pertinent question is: Can these observed differences be explained by chance or are there real differences in the pattern of part-time work between the populations of the two classes? We will pursue this aspect of statistical inference in Chapter 13.

## SIMPSON'S PARADOX

Quite surprising and misleading conclusions can occur when data from different sources are combined into a single table. We illustrate this reversal of implications with graduate school admissions data.

### Example 2 Combining Tables Can Produce Misleading Summaries

We consider graduate school admissions at a large midwestern university but, to simplify, we use only two departments as the whole school. We are interested in comparing admission rates by gender and obtain the data in Table 4 for the school.

**TABLE 4** School Admission Rates

	Admit	Not Admit	Total Applicants
Male	233	324	557
Female	88	194	282
Total	321	518	839

Does there appear to be a gender bias?

**SOLUTION** It is clear from these admissions statistics that the proportion of males admitted,  $233/557 = .418$ , is greater than the proportion of females admitted,  $88/282 = .312$ .

Does this imply some type of discrimination? Not necessarily. By checking the admission records, we were able to further categorize the cases according to department in Table 5. Table 4 is the aggregate of these two sets of data.

**TABLE 5** Admission Rates by Department

	Mechanical Engineering			History		
	Admit	Not Admit	Total	Admit	Not Admit	Total
Male	151	35	186	82	289	371
Female	16	2	18	72	192	264
Total	167	37	204	154	481	635

One of the two departments, mechanical engineering, has mostly male applicants. Even so, the proportion of males admitted,  $151/186 = .812$ , is smaller

than the proportion of females admitted,  $16/18 = .889$ . The same is true for the history department where the proportion of males admitted,  $82/371 = .221$ , is again smaller than the proportion of females admitted,  $72/264 = .273$ . When the data are studied department by department, the reverse but correct conclusion holds; females have a higher admission rate in both cases!

To obtain the correct interpretation, these data need to be presented as the full three-way table of gender-admission action-department as given above. If department is ignored as in Table 4, and the data aggregated across this variable, “department” can act as an unrecorded or lurking variable. In this example, it has reversed the direction of possible gender bias and led to the erroneous conclusion that males have a higher admission rate than females.

The reversal of the comparison, such as in Example 2, when data are combined from several groups is called **Simpson’s paradox**.

When data from several sources are aggregated into a single table, there is always the danger that unreported variables may cause a reversal of the findings. In practical applications, there is not always agreement on how much effort to expend following up on unreported variables. When comparing two medical treatments, the results often need to be adjusted for the age, gender, and sometimes current health of the subjects and other variables.

## Exercises

3.1 Nausea from air sickness affects some travelers. A drug company, wanting to establish the effectiveness of its motion sickness pill, randomly gives either its pill or a look-alike sugar pill (placebo) to 200 passengers.

	Degree of Nausea				Total
	None	Slight	Moderate	Severe	
Pill	43	36	18	3	100
Placebo	19	33	36	12	100
Total					

- Complete the marginal totals.
- Calculate the relative frequencies separately for each row.
- Comment on any apparent differences in response between the pill and the placebo.

3.2 Breakfast cereals from three leading manufacturers can be classified either above average or below average in sugar content. Data for ten cereals from each manufacturer are given below:

	Below Average	Above Average	Total
General Mills	3	7	10
Kellogg’s	4	6	10
Quaker	6	4	10
Total			

- Complete the marginal totals.
- Calculate the relative frequencies separately for each row.
- Comment on any apparent differences between the cereals produced by the three companies.

- 3.3 At the conclusion of one semester, a sample of 250 juniors was questioned about how much they had studied for each of their final exams. Students were also classified as social science, biological, or physical science majors.

Number of Hours of Study for Each Final

Major	10 or less	More than 10
Biological	30	45
Physical	15	35
Social	65	60

Compare the times studying for finals by calculating the relative frequencies.

- 3.4 A survey was conducted to study the attitudes of the faculty, academic staff, and students in regard to a proposed measure for reducing the heating and air-conditioning expenses on campus.

	Favor	Indifferent	Opposed	Total
Faculty	36	42	122	200
Academic staff	44	77	129	250
Students	106	178	116	400

Compare the attitude patterns of the three groups by computing the relative frequencies.

- 3.5 Groundwater from 19 wells was classified as low or high in alkalinity and low or high in dissolved iron. There were 9 wells with high alkalinity, 7 that were high in iron, and 5 that were high in both.
- Based on these data, complete the following two-way frequency table.
  - Calculate the relative frequencies of the cells.
  - Calculate the relative frequencies separately for each row.

Alkalinity	Iron	
	Low	High
Low		
High		

- 3.6 Interviews with 150 persons engaged in a stressful occupation revealed that 57 were alcoholics, 64 were mentally depressed, and 42 were both.
- Based on these records, complete the following two-way frequency table.
  - Calculate the relative frequencies.

	Alcoholic	Not Alcoholic	Total
Depressed			
Not depressed			
Total			

- 3.7 Cross-tabulate the “Class data” of Exercise 2.100 according to gender (M, F) and the general areas of intended major (H, S, B, P). Calculate the relative frequencies.
- 3.8 A psychologist interested in obese children gathered data on a group of children and their parents.

Parent	Child	
	Obese	Not Obese
At least one obese	12	24
Neither obese	8	36

- Calculate the marginal totals.
  - Convert the frequencies to relative frequencies.
  - Calculate the relative frequencies separately for each row.
- 3.9 Typically, there is a gender unbalance among tenured faculty, especially in the sciences. At a large university, tenured faculty members in two departments, English and Computer Science, were categorized according to gender.

	Male	Female
English	23	19
Computer Science	27	5

- Calculate relative frequencies separately for each row.
- Comment on major differences in the patterns for the two rows.



- 3.10 A large research hospital and a community hospital are located in your area. The surgery records for the two hospitals are:

	Died	Survived	Total
Research hospital	90	2110	2200
Community hospital	23	677	700

The outcome is “survived” if the patient lives at least six weeks.

- (a) Calculate the proportion of patients that survive surgery at each of the hospitals.  
 (b) Which hospital do these data suggest you should choose for surgery?

- 3.11 Refer to Exercise 3.10. Not all surgery cases, even of the same type, are equally serious. Large research hospitals tend to get the most serious surgery cases, whereas community hospitals tend to get more of the routine cases. Suppose that patients can be classified as being in either “Good” or “Poor” condition and the outcomes of surgery are as shown in table below.

- (a) Calculate the proportions that survive for each hospital and each condition.  
 (b) From these data, which hospital would you choose if you were in good condition? If you were in bad condition?  
 (c) Compare your answer with that to Exercise 3.10. Explain this reversal as an example of Simpson’s paradox and identify the lurking variable in Exercise 3.10.

Survival Rates by Condition

Good Condition				Poor Condition			
	Died	Survived	Total		Died	Survived	Total
Research hospital	15	685	700	Research hospital	75	1425	1500
Community hospital	16	584	600	Community hospital	7	93	100
Total	31	1269	1300	Total	82	1518	1600

### 3. A DESIGNED EXPERIMENT FOR MAKING A COMPARISON

We regularly encounter claims that, as a group, smokers have worse health records than nonsmokers with respect to one disease or another or that a new medical treatment is better than the former one. Properly designed experiments can often provide data that are so conclusive that a comparison is clear-cut. An example of a comparative study will illustrate the design issue and how to conduct an experiment.

During the early development of a medicated skin patch to help smokers break the habit, a test was conducted with 112 volunteers. The experimenters wanted to avoid erroneous conclusions caused by the so-called **placebo effect** when a treatment, with no therapeutic value, is administered to a subject but their symptoms improve anyway. One explanation is the positive thinking of subjects having high expectations of getting better and who believe the real treatment will work. Consequently, half the volunteers were given an unmedicated skin patch. The data will consist of a count of the number of persons who are abstinent at the end of the study.

**Purpose:** To determine the effectiveness of a medicated nicotine patch for smoking cessation based on the end-of-therapy numbers of abstinent persons in the medicated and unmedicated groups.



What is involved in comparing two approaches or methods for doing something? First the subjects, or experimental units, must be assigned to the two groups in such a manner that neither method is favored. One approach is to list the subjects' names on a piece of paper, cut the paper into strips, each with one name on it, and then draw one at a time until half the names are drawn. Ideally, we like to have groups of equal size, so if there is an odd number of subjects, draw just over one-half. These subjects are assigned to the first approach. The other subjects are assigned to the second approach. This step, called **random assignment**, helps guarantee a valid comparison. Any subject likely to respond positively has the same chance of supporting the first approach as supporting the second approach. When subjects cannot be randomly assigned, we will never know if observed differences in the number of abstinent smokers is due to the approaches or some other variables associated with the particular persons assigned to the two groups.

Subjects were randomly assigned to the medicated or unmedicated (placebo) groups. They were not told which group. As with many medical trials, this was a **double blind trial**. That is, the medical staff in contact with the patients was also kept unaware of which patients were getting the treated patch and which were not. At the end of the study, the number of persons in each group who were abstinent and who were smoking were recorded.

The data<sup>1</sup> collected from this experiment are summarized in Table 6.

**TABLE 6** Quitting Smoking

	Abstinent	Smoking	
Medicated patch	21	36	57
Unmedicated patch	11	44	55
	32	80	112

The proportion abstinent is  $21/57 = .368$  for the medicated skin patch group and only  $11/55 = .200$  for the control. The medicated patch seems to work. Later, in Chapter 10, we verify that the difference  $.368 - .200 = .168$  is greater than can be explained by chance variation.

In any application where the subjects might learn from the subjects before them, it would be a poor idea to perform all the trials for treatment 1 and then all those for treatment 2. Learning or other uncontrolled variables must not be given the opportunity to systematically affect the experiment. We could number the subjects 1 to 112 and write each of these numbers on a separate slip of paper. The 112 slips of paper should be mixed and then drawn one at a time to determine the sequence in which the trials are conducted.

Researchers continue to investigate the effectiveness of patches. One study presents evidence against the effectiveness of patches.<sup>2</sup>

<sup>1</sup>M. Fiore, S. Kenford, D. Jorenby, D. Wetter, S. Smith, and T. Baker. "Two Studies of the Clinical Effectiveness of the Nicotine Patch with Different Counseling Treatments." *Chest* 105 (1994), pp. 524–533.

<sup>2</sup>A. Albert, et al. "Nicotine replacement therapy use among a cohort of smokers," *Journal of Addictive Diseases* 24(1) (2005), pp. 101–113.

## Exercises

---

- 3.12 With reference to the quit-smoking experiment, suppose two new subjects are available. Explain how you would assign one subject to receive the placebo and one to receive the medicated patch.
- 3.13 With reference to the quit-smoking experiment:
- Suppose the placebo trials were ignored and you were only told that 21 of 57 were abstinent after using the medicated patches. Would this now appear to be stronger evidence in favor of the patches?
  - Explain why the placebo trials provide a more valid reference for results of the medicated patch trials.

## 4. SCATTER DIAGRAM OF BIVARIATE MEASUREMENT DATA

---

We now turn to a description of data sets concerning two variables, each measured on a numerical scale. For ease of reference, we will label one variable  $x$  and the other  $y$ . Thus, two numerical observations  $(x, y)$  are recorded for each sampling unit. These observations are *paired* in the sense that an  $(x, y)$  pair arises from the same sampling unit. An  $x$  observation from one pair and an  $x$  or  $y$  from another are unrelated. For  $n$  sampling units, we can write the measurement pairs as

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

The set of  $x$  measurements alone, if we disregard the  $y$  measurements, constitutes a data set for one variable. The methods of Chapter 2 can be employed for descriptive purposes, including graphical presentation of the pattern of distribution of the measurements, and calculation of the mean, standard deviation, and other quantities. Likewise, the  $y$  measurements can be studied disregarding the  $x$  measurements. However, a major purpose of collecting bivariate data is to answer such questions as:

- Are the variables related?
- What form of relationship is indicated by the data?
- Can we quantify the strength of their relationship?
- Can we predict one variable from the other?

Studying either the  $x$  measurements by themselves or the  $y$  measurements by themselves would not help answer these questions.

An important first step in studying the relationship between two variables is to graph the data. To this end, the variable  $x$  is marked along the horizontal axis and  $y$  on the vertical axis on a graph paper. The pairs  $(x, y)$  of observations are then plotted as dots on the graph. The resulting diagram is called a **scatter diagram** or **scatter plot**. By looking at the scatter diagram, a visual impression can be formed about the relation between the variables. For instance, we can observe whether the points band around a line or a curve or if they form a patternless cluster.

**Example 3** A Scatter Diagram Provides a Visual Display of a Relationship

Recorded in Table 7 are the data of

$$x = \text{Undergraduate GPA}$$

and

$$y = \text{Score in the Graduate Management Aptitude Test (GMAT)}$$

for applicants seeking admission to an MBA program.

Construct a scatter diagram.

**TABLE 7** Data of Undergraduate GPA  $x$  and GMAT Score  $y$

$x$	$y$	$x$	$y$	$x$	$y$
3.63	447	2.36	399	2.80	444
3.59	588	2.36	482	3.13	416
3.30	563	2.66	420	3.01	471
3.40	553	2.68	414	2.79	490
3.50	572	2.48	533	2.89	431
3.78	591	2.46	509	2.91	446
3.44	692	2.63	504	2.75	546
3.48	528	2.44	336	2.73	467
3.47	552	2.13	408	3.12	463
3.35	520	2.41	469	3.08	440
3.39	543	2.55	538	3.03	419
				3.00	509

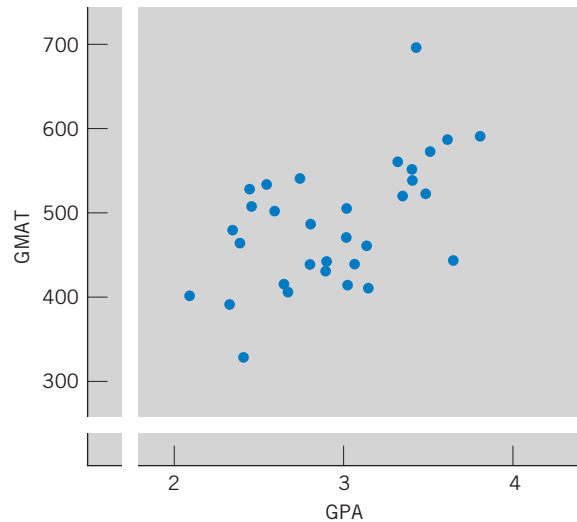


Figure 1 Scatter diagram of applicants' scores.

**SOLUTION** The scatter diagram is plotted in Figure 1. The southwest-to-northeast pattern of the points indicates a positive relation between  $x$  and  $y$ . That is, the applicants with a high GPA tend to have a high GMAT. Evidently, the relation is far from a perfect mathematical relation.

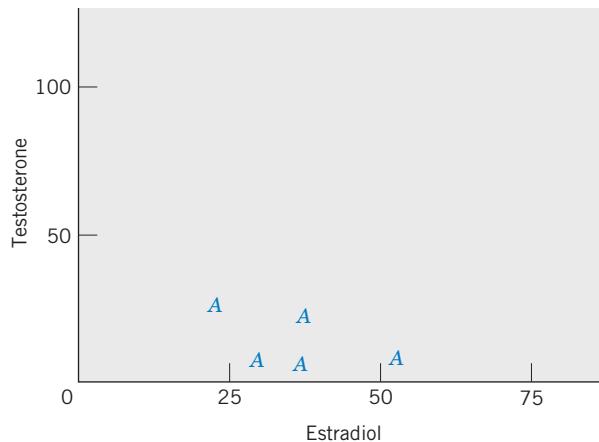
When the two measurements are made on two or more groups, visual comparisons between groups are made by plotting the points on the same scatter plot. A different symbol is used for each group. The resulting graph is called a **multiple scatter plot** where “multiple” refers to groups.

**Example 4** A Multiple Scatter Diagram for Visually Comparing Relationships

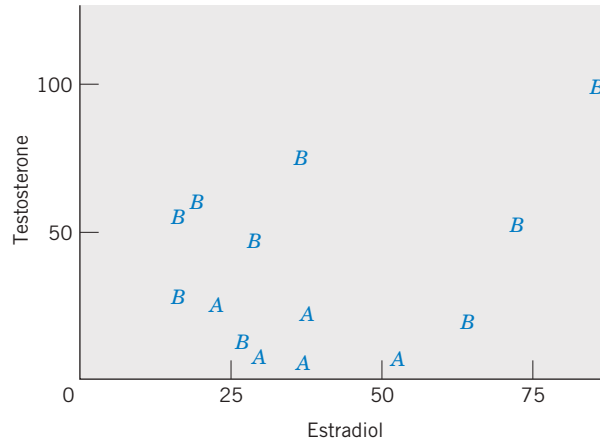
Concern was raised by environmentalists that spills of contaminants were affecting wildlife in and around an adjacent lake. Estrogenic contaminants in the environment can have grave consequences on the ability of living things to reproduce. Researchers examined the reproductive development of young male alligators hatched from eggs taken from around (1) Lake Apopka, the lake which was contaminated, and (2) Lake Woodruff, which acted as a control. The contaminants were thought to influence sex steroid concentrations. The concentrations of two steroids, estradiol and testosterone, were determined by radioimmunoassay.

		Lake Apopka								
Estradiol		38	23	53	37	30				
Testosterone		22	24	8	6	7				
		Lake Woodruff								
Estradiol		29	64	19	36	27	16	15	72	85
Testosterone		47	20	60	75	12	54	33	53	100

- (a) Make a scatter diagram of the two concentrations for the Lake Apopka alligators.
- (b) Create a multiple scatter diagram by adding to the same plot the pairs of concentrations for the Lake Woodruff male alligators. Use a different symbol for the two lakes.
- (c) Comment on any major differences between the two groups of male alligators.



(a) Scatter diagram for Lake Apopka



(b) Multiple scatter diagram

Figure 2 Scatter diagrams. *A* = Lake Apopka.  
*B* = Lake Woodruff.

**SOLUTION**

- (a) Figure 2a gives the scatter diagram for the Lake Apopka alligators.
- (b) Figure 2b is the multiple scatter diagram with the points for Lake Woodruff marked as *B*.
- (c) The most prominent feature of the data is that the male alligators from the contaminated lake have, generally, much lower levels of testosterone than those from the nearly pollution-free control lake. (The *A*'s are at the bottom third of the multiple scatter diagram.) Low testosterone levels in males have grave consequences regarding reproduction.

## 5. THE CORRELATION COEFFICIENT—A MEASURE OF LINEAR RELATION

The scatter diagram provides a visual impression of the nature of the relation between the  $x$  and  $y$  values in a bivariate data set. In a great many cases, the points appear to band around a straight line. Our visual impression of the closeness of the scatter to a linear relation can be quantified by calculating a numerical measure, called the **correlation coefficient**.

The correlation coefficient, denoted by  $r$ , is a measure of strength of the linear relation between the  $x$  and  $y$  variables. Before introducing its formula, we outline some important features of the correlation coefficient and discuss the manner in which it serves to measure the strength of a linear relation.

1. The value of  $r$  is always between  $-1$  and  $+1$ .
2. The magnitude of  $r$  indicates the strength of a linear relation, whereas its sign indicates the direction. More specifically,

- $r > 0$  if the pattern of  $(x, y)$  values is a band that runs from lower left to upper right.
- $r < 0$  if the pattern of  $(x, y)$  values is a band that runs from upper left to lower right.
- $r = +1$  if all  $(x, y)$  values lie exactly on a straight line with a positive slope (perfect positive linear relation).
- $r = -1$  if all  $(x, y)$  values lie exactly on a straight line with a negative slope (perfect negative linear relation).

A high numerical value of  $r$ , that is, a value close to  $+1$  or  $-1$ , represents a strong linear relation.

3. A value of  $r$  close to zero means that the linear association is very weak.

The correlation coefficient is close to zero when there is no visible pattern of relation; that is, the  $y$  values do not change in any direction as the  $x$  values change. A value of  $r$  near zero could also happen because the points band

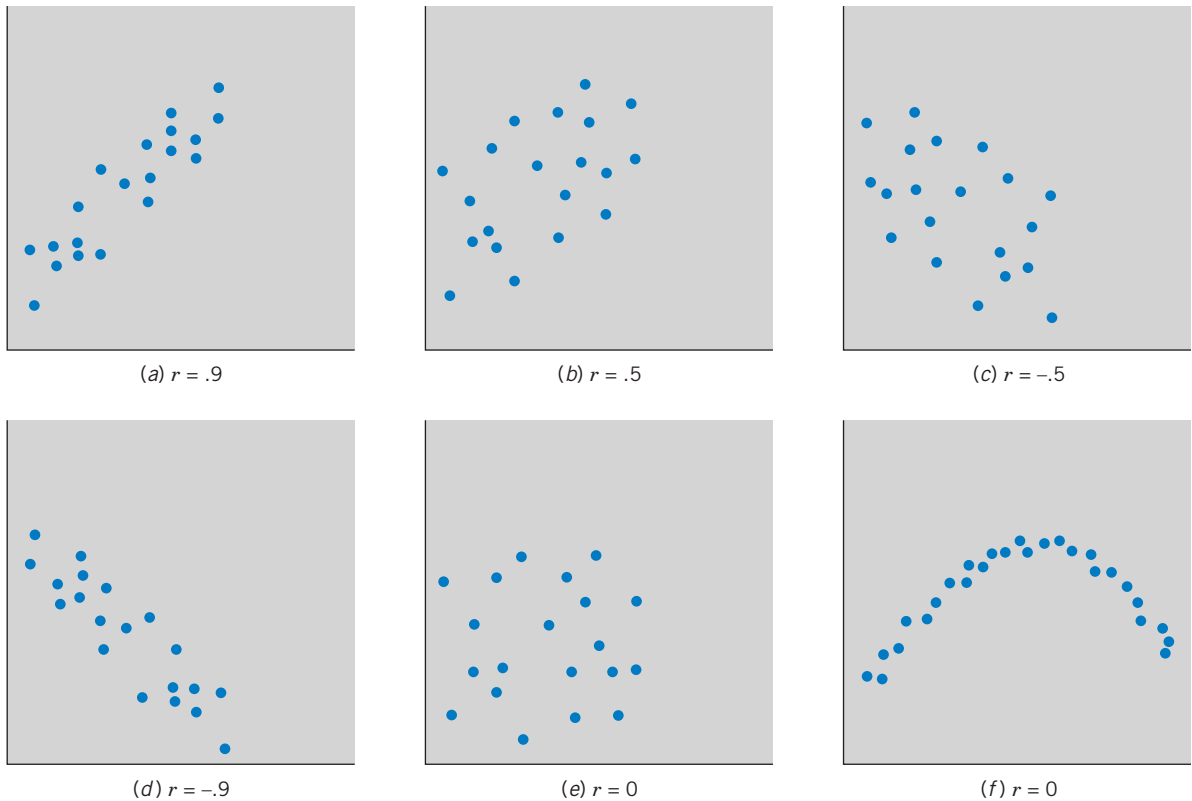


Figure 3 Correspondence between the values of  $r$  and the amount of scatter.

around a curve that is far from linear. After all,  $r$  measures linear association, and a markedly bent curve is far from linear.

Figure 3 shows the correspondence between the appearance of a scatter diagram and the value of  $r$ . Observe that (e) and (f) correspond to situations where  $r = 0$ . The zero correlation in (e) is due to the absence of any relation between  $x$  and  $y$ , whereas in (f) it is due to a relation following a curve that is far from linear.

### CALCULATION OF $r$

The sample correlation coefficient quantifies the association between two numerically valued characteristics. It is calculated from  $n$  pairs of observations on the two characteristics

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

The correlation coefficient is best interpreted in terms of the **standardized observations**, or **sample z values**

$$\frac{\text{Observation} - \text{Sample mean}}{\text{Sample standard deviation}} = \frac{x_i - \bar{x}}{s_x}$$

where  $s_x = \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1)}$  and the subscript  $x$  on  $s$  distinguishes the sample standard deviation of the  $x$  observations from the sample standard deviation  $s_y$  of the  $y$  observations.

Since the difference  $x_i - \bar{x}$  has the units of  $x$  and the sample standard deviation  $s_x$  also has the same units, the standardized observation is free of the units of measurements. The **sample correlation coefficient** is the sum of the products of the standardized  $x$  observation times the standardized  $y$  observations divided by  $n - 1$ .

#### Sample Correlation Coefficient

$$r = \frac{1}{n - 1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

When the pair  $(x_i, y_i)$  has both components above their sample means or both below their sample means, the product of standardized observations will be positive; otherwise it will be negative. Consequently, if most pairs have both components simultaneously above or simultaneously below their means,  $r$  will be positive.

An alternative formula for  $r$  is used for calculation. It is obtained by canceling the common term  $n - 1$ .



**Calculation Formula for the Sample Correlation Coefficient**

where

$$r = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}}$$

$$S_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

$$S_{xx} = \sum (x - \bar{x})^2 \quad S_{yy} = \sum (y - \bar{y})^2$$

The quantities  $S_{xx}$  and  $S_{yy}$  are the sums of squared deviations of the  $x$  observations and the  $y$  observations, respectively.  $S_{xy}$  is the sum of cross products of the  $x$  deviations with the  $y$  deviations. This formula will be examined in more detail in Chapter 11.

**Example 5** Calculation of Sample Correlation

Calculate  $r$  for the  $n = 4$  pairs of observations

(2, 5) (1, 3) (5, 6) (0, 2)

**SOLUTION** We first determine the mean  $\bar{x}$  and deviations  $x - \bar{x}$  and then  $\bar{y}$  and the deviations  $y - \bar{y}$ . See Table 8.

**TABLE 8** Calculation of  $r$

	$x$	$y$	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
	2	5	0	1	0	1	0
	1	3	-1	-1	1	1	1
	5	6	3	2	9	4	6
	0	2	-2	-2	4	4	4
Total	8	16	0	0	14	10	11
	$\bar{x} = 2$	$\bar{y} = 4$			$S_{xx}$	$S_{yy}$	$S_{xy}$

Consequently,

$$r = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}} = \frac{11}{\sqrt{14} \sqrt{10}} = .930$$

The value .930 is large and it implies a strong association where both  $x$  and  $y$  tend to be small or both tend to be large.

It is sometimes convenient, when using hand-held calculators, to evaluate  $r$  using the alternative formulas for  $S_{xx}$ ,  $S_{yy}$ , and  $S_{xy}$  (see Appendix A1.2).

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} \quad S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

This calculation is illustrated in Table 9.

**TABLE 9** Alternate Calculation of  $r$

	$x$	$y$	$x^2$	$y^2$	$xy$
	2	5	4	25	10
	1	3	1	9	3
	5	6	25	36	30
	0	2	0	4	0
Total	8	16	30	74	43
	$\Sigma x$	$\Sigma y$	$\Sigma x^2$	$\Sigma y^2$	$\Sigma xy$

$$r = \frac{43 - \frac{8 \times 16}{4}}{\sqrt{30 - \frac{8^2}{4}} \sqrt{74 - \frac{(16)^2}{4}}} = .930$$

We remind the reader that  $r$  measures the closeness of the pattern of scatter to a line. Figure 3f on page 94 presents a strong relationship between  $x$  and  $y$ , but one that is not linear. The small value of  $r$  for these data does not properly reflect the strength of the relation. Clearly,  $r$  is not an appropriate summary of a curved pattern. Another situation where the sample correlation coefficient  $r$  is not appropriate occurs when the scatter plot breaks into two clusters. Faced with separate clusters as depicted in Figure 4, it is best to try and determine the underlying cause. It may be that a part of the sample has come from one population and a part from another.



Figure 4  $r$  is not appropriate—samples from two populations.

## CORRELATION AND CAUSATION

Data analysts often jump to unjustified conclusions by mistaking an observed correlation for a cause-and-effect relationship. A high sample correlation coefficient does not necessarily signify a causal relation between two variables. A classic example concerns an observed high positive correlation between the number of storks sighted and the number of births in a European city. It is hoped no one would use this evidence to conclude that storks bring babies or, worse yet, that killing storks would control population growth.

The observation that two variables tend to simultaneously vary in a certain direction does not imply the presence of a direct relationship between them. If we record the monthly number of homicides  $x$  and the monthly number of religious meetings  $y$  for several cities of widely varying sizes, the data will probably indicate a high positive correlation. It is the fluctuation of a third variable (namely, the city population) that causes  $x$  and  $y$  to vary in the same direction, despite the fact that  $x$  and  $y$  may be unrelated or even negatively related. Picturesquely, the third variable, which in this example is actually causing the observed correlation between crime and religious meetings, is referred to as a **lurking variable**. The false correlation that it produces is called a **spurious correlation**. It is more a matter of common sense than of statistical reasoning to determine if an observed correlation has a practical interpretation or is spurious.

An observed correlation between two variables may be **spurious**. That is, it may be caused by the influence of a third variable.

When using the correlation coefficient as a measure of relationship, we must be careful to avoid the possibility that a lurking variable is affecting any of the variables under consideration.

### Example 6 Spurious Correlation Caused by Lurking Variables

Figure 5 gives a scatter diagram of the number of person in prison, under state or federal jurisdiction, and the number of cell phone subscribers in each of 10 years. Both variables are measured in millions (see Exercise 3.29).

This plot exhibits a pattern of strong positive correlation; the numerical value  $r = .987$ . Would restricting the number of cell phones result in fewer persons in prison?

**SOLUTION** The scatter diagram reveals a strong positive correlation, but common sense suggests there is no cause-and-effect relation to tie an increase in the number of cell phone subscribers to an increase in the prison population. Realistically, the two variables should not have a causal relationship.

In Figure 6 we have repeated the scatter diagram but have labeled each point according to the year. For example, 03 stands for 2003. The years increase exactly in the same order as the points from lower left to upper right in the scatter diagram. More things change over time or from

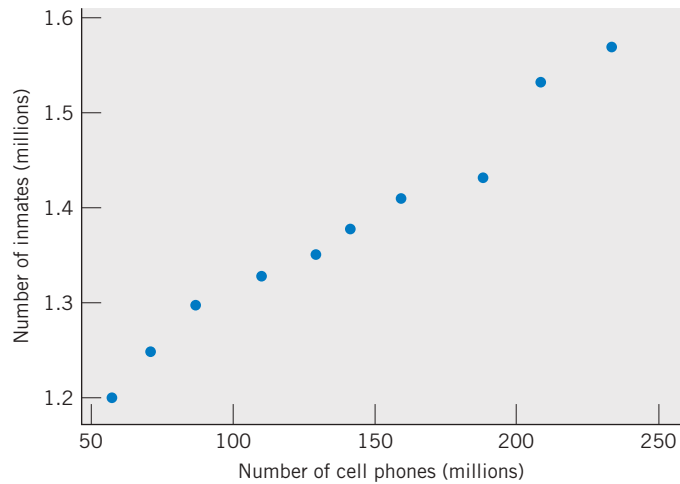


Figure 5 Scatter diagram reveals pattern of strong positive correlation.

year to year. Year is just a stand-in, or proxy, for all of them. Since the population of the United States increased over these years, population size could be one lurking variable.

Once the year notation is added to the graph, it is clear that other variables are leading to the observed correlation. A graph, with time order added, can often help discredit false claims of causal relations. If the order of the years had been scrambled, we could not discredit the suggestion of a causal relation.

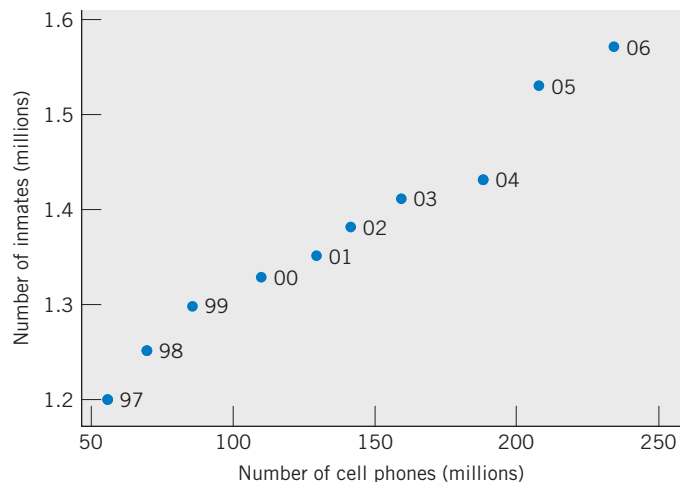


Figure 6 Scatter diagram pattern has strong relation to year.

### Lurking Variables

The Insurance Institute for Highway Safety 2007 report announced the safest and unsafest 2001–2004 car models for the period 2002 to 2004 in terms of fewest fatalities per one million registered vehicle years. The death rates, shown in parentheses, are given in terms of one million cars that are registered for the year.

Lowest Fatality Rates	Highest Fatality Rates
Chevrolet Astro (7)	Chevrolet Blazer 2 dr. (232)
Infiniti G35 (11)	Acura RSX (202)
BMW 7 Series (11)	Nissan 350Z (193)
Toyota 4Runner (13)	Kia Spectra (191)
Audi A4/S4 Quattro (14)	Pontiac Sunfire (179)
Mercedes-Benz E-Class (14)	Kia Rio (175)
Toyota Highlander (14)	
Mercedes-Benz M-Class (14)	

Although it must be acknowledged there is truth in the statement that larger cars are generally safer than small cars, there is a big lurking variable here—the driver. How often does the teenager cruise in the luxury car? There is a strong correlation between the age of the driver and the type of car driven and also between the age of the driver and driver behavior.

To its credit, the institute’s report states that although the Chevrolet Astro’s performance in frontal crash tests is abysmal, it does much better on fatalities than the Blazer. The Infiniti G35 shares many features with the Nissan 350Z. To reiterate, driver behavior is an important lurking variable.

**WARNING.** Don’t confuse the existence of a high correlation between two variables with the situation where one is the cause of the other. Recall Example 6, where the number of cell phones and the number of persons in prison have a high correlation. There is no commonsense connection—no causal relation.

A high correlation between two variables can sometimes be caused when there is a third, “lurking” variable that strongly influences both of them.

### Exercises

- 3.14 Would you expect a positive, negative, or nearly zero correlation for each of the following? Give reasons for your answers.
- (a) The physical fitness of a dog and the physical fitness of the owner.
- (b) For each person, the number of songs downloaded from the Internet last month and the number of hours listening to MP3 format music.
- (c) For a student, the number of friends listed on their personal Internet sites and the number of hours they are active on the Internet.
- 3.15 In each of the following instances, would you expect a positive, negative, or zero correlation?
- (a) Number of salespersons and total dollar sales for real estate firms.
- (b) Total payroll and percent of wins of national league baseball teams.

- (c) The amount spent on a week of TV advertising and sales of a cola.
  - (d) Age of adults and their ability to maintain a strenuous exercise program.
- 3.16 Data collected since 2000 revealed a positive correlation between the federal debt and attendance at National Football League games. Would restricting the number of persons attending games reduce the national debt? Explain your answer.
- 3.17 If the value of  $r$  is small, can we conclude that there is not a strong relationship between the two variables?
- 3.18 For the data set

$x$	1	2	7	4	6
$y$	6	5	2	4	3

- (a) Construct a scatter diagram.
  - (b) Guess the sign and value of the correlation coefficient.
  - (c) Calculate the correlation coefficient.
- 3.19 Refer to the alligator data in Table D.11 of the Data Bank. Using the data on  $x_3$  and  $x_4$  for male and female alligators from Lake Apopka:
- (a) Make a scatter diagram of the pairs of concentrations for the male alligators. Calculate the sample correlation coefficient.
  - (b) Create a multiple scatter diagram by adding, on the same plot, the pairs of concentrations for the female alligators. Use a different symbol for females. Calculate the sample correlation coefficient for this latter group.
  - (c) Comment on any major differences between the male and female alligators.
- 3.20 (a) Construct scatter diagrams of the data sets

(i) 

$x$	0	4	2	6	3
$y$	4	6	2	8	5

(ii) 

$x$	0	4	2	6	3
$y$	8	2	5	4	6

- (b) Calculate  $r$  for the data set (i).
- (c) Guess the value of  $r$  for the data set (ii) and then calculate  $r$ . (Note: The  $x$  and  $y$  values are the same for both sets, but they are paired differently in the two cases.)

3.21 Match the following values of  $r$  with the correct diagrams (Figure 7).

- (a)  $r = -.3$  (b)  $r = .1$  (c)  $r = .9$

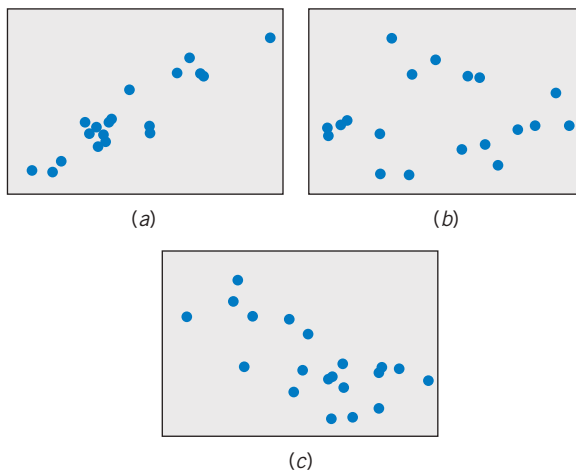


Figure 7 Scatter diagrams for Exercise 3.21.

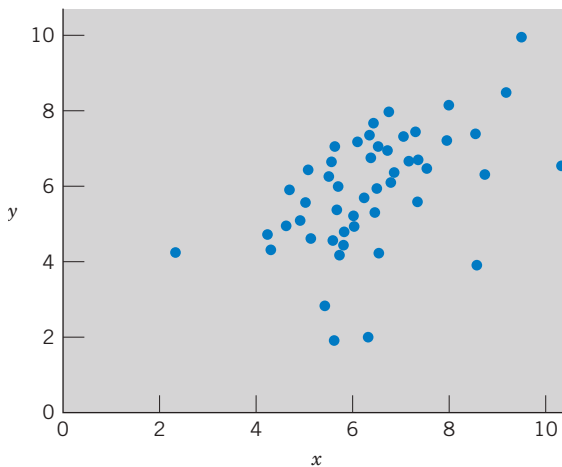


Figure 8 Scatter diagram for Exercise 3.22.

- 3.22 Is the correlation in Figure 8 about (a) .1, (b) .5, (c) .9, or (d)  $-.7$ ?
- 3.23 Calculations from a data set of  $n = 36$  pairs of  $(x, y)$  values have provided the following results.

$$\sum (x - \bar{x})^2 = 530.7 \quad \sum (y - \bar{y})^2 = 235.4$$

$$\sum (x - \bar{x})(y - \bar{y}) = -204.3$$

Obtain the correlation coefficient.

3.24 Over the years, a traffic officer noticed that cars with fuzzy dice hanging on the rear-view mirror always seemed to be speeding. Perhaps tongue in cheek, he suggested that outlawing the sale of fuzzy dice would reduce the number of cars exceeding the speed limit. Comment on lurking variables.

3.25 Heating and combustion analyses were performed in order to study the composition of moon rocks. Recorded here are the determinations of hydrogen (H) and carbon (C) in parts per million (ppm) for 11 specimens.

Hydrogen (ppm)	120	82	90	8	38	20	2.8	66	2.0	20	85
Carbon (ppm)	105	110	99	22	50	50	7.3	74	7.7	45	51

Calculate  $r$ .

3.26 A zoologist collected 20 wild lizards in the southwestern United States. After measuring their total length (mm), they were placed on a treadmill and their speed (m/sec) recorded.

Speed	1.28	1.36	1.24	2.47	1.94	2.52	2.67
Length	179	157	169	146	143	131	159

Speed	1.29	1.56	2.66	2.17	1.57	2.10	2.54
Length	142	141	130	142	116	130	140

Speed	1.63	2.11	2.57	1.72	0.76	1.02
Length	138	137	134	114	90	114

- (a) Create a scatter plot. Comment on any unusual observations.
- (b) Calculate the sample correlation coefficient.

3.27 An ongoing study of wolves is being conducted at the Yukon-Charley Rivers National Preserve. Table D.9 in the Data Bank gives the physical characteristics of wolves that were captured.

Males						
Body length (cm)	134	143	148	127	136	146
Weight (lb)	71	93	101	84	88	117

Body length (cm)	142	139	140	133	123
Weight (lb)	86	86	93	86	106

- (a) Plot length versus weight for the male wolves. From your visual inspection, estimate the value of the correlation coefficient.
- (b) Calculate the sample correlation coefficient for male wolves.
- (c) Create a multiple scatter diagram by adding the points for female wolves from Table D.9 to your plot in part (a). Do the patterns of correlation for males and females appear to be similar or different?

3.28 An ongoing study of wolves is being conducted at the Yukon-Charley Rivers National Preserve. Table D.9 in the Data Bank gives the physical characteristics of wolves that were captured.

Females								
Body length (cm)	123	129	143	124	125	122	125	122
Weight (lb)	57	84	90	71	71	77	68	73

- (a) Plot length versus weight for the female wolves. From your visual inspection, estimate the value of the correlation coefficient.
- (b) Calculate the sample correlation coefficient for female wolves.

3.29 Refer to Example 6 concerning spurious correlation. Replace number of cell phone subscribers with the number of registered motorcycles in millions.

**TABLE 10** Variables Showing Spurious Correlation

Inmates (mil)	Cell phones (mil)	Motorcycles (mil)	Year
1.20	55.3	3.8	1997
1.25	69.2	3.9	1998
1.30	86.0	4.2	1999
1.33	109.4	4.3	2000
1.35	128.4	4.9	2001
1.38	140.8	5.0	2002
1.41	158.7	5.4	2003
1.43	187.1	5.8	2004
1.53	207.9	6.0	2005
1.57	233.0	6.2	2006



- (a) Create a scatter diagram and identify the kind of association.
- (b) Comment on possible lurking variables.

Year	1960	1970	1980	1990	2000	2007
Garbage (millions of tons)	88	121	152	205	232	254
Population (millions)	179	203	227	249	282	302

3.30 **A further property of  $r$ .** Suppose all  $x$  measurements are changed to  $x' = ax + b$  and all  $y$  measurements to  $y' = cy + d$ , where  $a, b, c$ , and  $d$  are fixed numbers ( $a \neq 0, c \neq 0$ ). Then the correlation coefficient remains unchanged if  $a$  and  $c$  have the same signs; it changes sign but not numerical value if  $a$  and  $c$  are of opposite signs.

This property of  $r$  can be verified along the lines of Exercise 2.74 in Chapter 2. In particular, the deviations  $x - \bar{x}$  change to  $a(x - \bar{x})$  and the deviations  $y - \bar{y}$  change to  $c(y - \bar{y})$ . Consequently,  $\sqrt{S_{xx}}$ ,  $\sqrt{S_{yy}}$ , and  $S_{xy}$  change to  $|a|\sqrt{S_{xx}}$ ,  $|c|\sqrt{S_{yy}}$ , and  $acS_{xy}$ , respectively (recall that we must take the positive square root of a sum of squares of the deviations). Therefore,  $r$  changes to

$$\frac{ac}{|a||c|}r = \begin{cases} r & \text{if } a \text{ and } c \text{ have the same signs} \\ -r & \text{if } a \text{ and } c \text{ have opposite signs} \end{cases}$$

- (a) For a numerical verification of this property of  $r$ , consider the data of Exercise 3.18. Change the  $x$  and  $y$  measurements accordingly to

$$\begin{aligned} x' &= 2x - 3 \\ y' &= -y + 10 \end{aligned}$$

Calculate  $r$  from the  $(x', y')$  measurements and compare with the result of Exercise 3.18.

- (b) Suppose from a data set of height measurements in inches and weight measurements in pounds, the value of  $r$  is found to be .86. What would the value of  $r$  be if the heights were measured in centimeters and weights in kilograms?

3.31 The amount of municipal solid waste created has become a major problem. According to the Environmental Protection Agency, the yearly amounts (millions of tons) are:

- (a) Plot the amount of garbage (millions of tons) versus year.
- (b) Visually, does there appear to be a strong correlation? Explain.
- (c) Give one possible lurking variable.

3.32 Refer to the data on garbage in Exercise 3.31.

- (a) Plot the amount of garbage (millions of tons) versus population (millions).
- (b) Does there appear to be a strong correlation? Explain.
- (c) How does your interpretation of the association differ from that in Exercise 3.31, parts (b) and (c)?

3.33 Refer to the data on garbage in Exercises 3.31.

- (a) Replace year by (year—1960). Calculate the correlation coefficient between (year—1960) and amount of garbage in millions of tons.
- (b) Based on your calculation in part (a), what is the correlation between the year itself and the amount of garbage? Explain.

3.34 Refer to the data on garbage in Exercises 3.31.

- (a) Calculate the correlation coefficient between the amount of garbage in millions of tons and the population size in millions.
- (b) Based on your calculation in part (a), give the correlation coefficient between the amount of garbage in pounds and population size in number of persons. Explain your answer. [*Hint:* Recall that there are 2000 pounds in a ton so (number of pounds) = 2000 × (number of tons).]

## 6. PREDICTION OF ONE VARIABLE FROM ANOTHER (LINEAR REGRESSION)

An experimental study of the relation between two variables is often motivated by a need to predict one from the other. The administrator of a job training program may wish to study the relation between the duration of training and the score of the trainee on a subsequent skill test. A forester may wish to estimate the timber volume of a tree from the measurement of the trunk diameter a few feet above the ground. A medical technologist may be interested in predicting the blood alcohol measurement from the read-out of a newly devised breath analyzer.

In such contexts as these, the **predictor** or **input** variable is denoted by  $x$ , and the **response** or **output** variable is labeled  $y$ . The object is to find the nature of relation between  $x$  and  $y$  from experimental data and use the relation to predict the response variable  $y$  from the input variable  $x$ . Naturally, the first step in such a study is to plot and examine the scatter diagram. If a linear relation emerges, the calculation of the numerical value of  $r$  will confirm the strength of the linear relation. Its value indicates how effectively  $y$  can be predicted from  $x$  by fitting a straight line to the data.

A line is determined by two constants as illustrated in Figure 9: its height above the origin (**intercept**) and the amount that  $y$  increases whenever  $x$  is increased by one unit (**slope**).

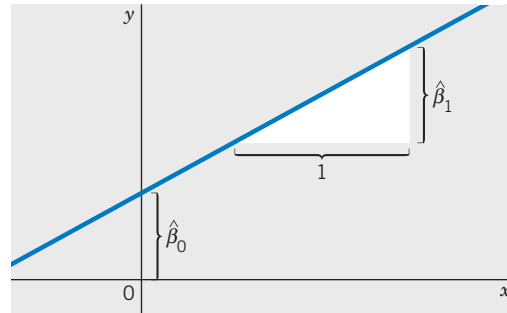


Figure 9 The line  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x$ .

Chapter 11 explains an objective method of best fitting a straight line, called the **method of least squares**. This best fitting line, or least squares line, is close to the points graphed in the scatter plot in terms of minimizing the average amount of vertical distance.

### Equation of the Line Fitted by Least Squares

where

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x$$

$$\text{Slope } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$\text{Intercept } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Besides the sample mean  $\bar{x}$  and  $\bar{y}$ , the **fitted line** involves the sum of the squared deviations of the  $x$  observations,  $S_{xx}$ , and the sum of the cross products of the  $x$  observations and the  $y$  deviations,  $S_{xy}$ . The formulas will be examined in more detail in Chapter 11.

### Example 7 Calculation of the Line Fitted by Least Squares

A chemist wishes to study the relation between the drying time of a paint and the concentration of a base solvent that facilitates a smooth application. The data of concentration (grams)  $x$  and the observed drying times (minutes)  $y$  are recorded in the first two columns of Table 11. Plot the data, calculate  $r$ , and obtain the fitted line.

**TABLE 11** Data of Concentration  $x$  and Drying Time  $y$  (in minutes) and the Basic Calculations

Concentration $x$ (g)	Drying Time $y$ (min)	$x^2$	$y^2$	$xy$
0	1	0	1	0
1	5	1	25	5
2	3	4	9	6
3	9	9	81	27
4	7	16	49	28
Total 10	25	30	165	66

**SOLUTION** The scatter diagram in Figure 10 gives the appearance of a linear relation. To calculate  $r$  and determine the equation of the fitted line, we first calculate the basic quantities  $\bar{x}$ ,  $\bar{y}$ ,  $S_{xx}$ ,  $S_{yy}$ , and  $S_{xy}$  using the totals in Table 11.

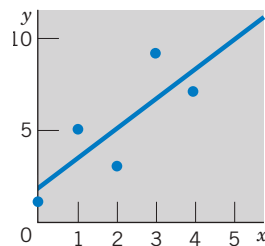


Figure 10 Scatter diagram.

$$\bar{x} = \frac{10}{5} = 2 \quad \bar{y} = \frac{25}{5} = 5$$

$$S_{xx} = 30 - \frac{(10)^2}{5} = 10$$

$$S_{yy} = 165 - \frac{(25)^2}{5} = 40$$

$$S_{xy} = 66 - \frac{10 \times 25}{5} = 16$$

$$r = \frac{16}{\sqrt{40 \times 10}} = \frac{16}{20} = .8$$

$$\hat{\beta}_1 = \frac{16}{10} = 1.6$$

$$\hat{\beta}_0 = 5 - (1.6)2 = 1.8$$

The equation of the fitted line is

$$\hat{y} = 1.8 + 1.6x$$

The estimated slope 1.6 tells us that one additional gram of solvent results in an increase of 1.6 minutes in average drying time. The fitted line is shown on the scatter diagram in Figure 10.

If we are to predict the drying time  $y$  corresponding to the concentration 2.5, we substitute  $x = 2.5$  in our prediction equation and get the result.

At  $x = 2.5$ , the predicted drying time =  $1.8 + 1.6(2.5) = 5.8$  minutes.

Graphically, this amounts to reading the ordinate of the fitted line at  $x = 2.5$ .

Software programs greatly simplify the calculation and plotting of the fitted line. The MINITAB calculations for Example 7 are shown in Figure 11. Column 1 is named  $x$  and column 2,  $y$ .

**Data:**

C1: 0 1 2 3 4

C2: 1 5 3 9 7

**Stat > Regression > Fitted Line Plot.**

Type C2 in **Response**. Type C1 in Predictors.

Under **Type of Regression Model** choose **Linear**. Click **OK**.

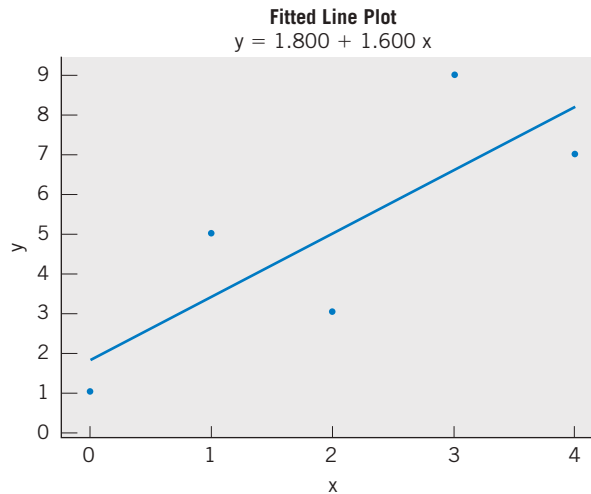


Figure 11 MINITAB output for fitted line in Example 7.

The sample correlation  $r$  was introduced as a measure of association between two variables. When  $r$  is near 1 or  $-1$ , points in the scatter plot are closely clustered about a straight line and the association is high. In these circumstances, the value of one variable can be accurately predicted from the value of the other. Said another way, when the value of  $r^2$  is near 1, we can predict the value of  $y$  from its corresponding  $x$  value. In all cases, the slope of the least squares line  $\hat{\beta}_1$  and the sample correlation  $r$  are related since  $\hat{\beta}_1 = r\sqrt{S_{yy}}/\sqrt{S_{xx}}$ . If the sample correlation is positive, then the slope of the least squares line is positive. Otherwise, both are negative or both zero.

Here we have only outlined the basic ideas concerning the prediction of one variable from another in the context of a linear relation. Chapter 11 expands on these ideas and treats statistical inferences associated with the prediction equation.

## Exercises

- 3.35 Plot the line  $y = 2 + 3x$  on graph paper by locating points for  $x = 1$  and  $x = 4$ . What is its intercept? Its slope?
- 3.36 Plot the line  $y = 6 - 2x$  on graph paper by locating the points for  $x = 0$  and  $x = 3$ . What is its intercept? Its slope?
- 3.37 A store manager has determined that the monthly profit  $y$  realized from selling a particular brand of car battery is given by

$$y = 10x - 155$$

where  $x$  denotes the number of these batteries sold in a month.

- (a) If 41 batteries were sold in a month, what was the profit?
- (b) At least how many batteries must be sold in a month in order to make a profit?

3.38 Identify the predictor variable  $x$  and the response variable  $y$  in each of the following situations.

- (a) A training director wishes to study the relationship between the duration of training for new recruits and their performance in a skilled job.
- (b) The aim of a study is to relate the carbon monoxide level in blood samples from smokers with the average number of cigarettes they smoke per day.
- (c) An agronomist wishes to investigate the growth rate of a fungus in relation to the level of humidity in the environment.
- (d) A market analyst wishes to relate the expenditures incurred in promoting a product in test markets and the subsequent amount of product sales.

3.39 Given these five pairs of  $(x, y)$  values

$x$	1	2	3	4	5
$y$	1	2.2	2.6	3.4	3.9

- (a) Plot the points on graph paper.
- (b) From a visual inspection, draw a straight line that appears to fit the data well.
- (c) Compute the least squares estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$  and draw the fitted line.

3.40 For the data set

$x$	1	2	7	4	6
$y$	5	4	1	3	2

- (a) Construct a scatter diagram.
- (b) From a visual inspection, draw a straight line that appears to fit the data well.
- (c) Calculate the least squares estimates and draw the least squares fitted line on your plot.

3.41 In an experiment to study the relation between the time waiting in line,  $y$  (minutes), to get to the head of the checkout line at her favorite grocery store and the number of persons ahead in line,  $x$ , a student collected the following statistics:

$$\begin{aligned} n &= 9 & \Sigma x &= 19 & \Sigma y &= 39.9 \\ S_{xx} &= 9.4 & S_{yy} &= 17.8 & S_{xy} &= 10.2 \end{aligned}$$

- (a) Find the equation of the least squares fitted line.
- (b) Using the fitted line, predict the time waiting in line when 3 persons are already in line.

3.42 Wolves used to range over much of Michigan, Minnesota, and Wisconsin. They were reintroduced several years ago, but counts over the winter showed that the populations are expanding rapidly over the past few years.

Total Number of Wolves in Wisconsin

Year	1998	1999	2000	2001	2002
No. wolves	178	204	248	257	327
Year	2003	2004	2005	2006	2007
No. wolves	325	373	436	467	546

- (a) Plot the number of wolves versus the year the count was taken.
- (b) Fit a least squares line to summarize the growth. To simplify the calculation, code 1998 as  $x = 1$ , 1999 as  $x = 2$ , and so on.
- (c) Does your fitted straight line summarize the growth in the wolf population over this period of time? If so, what numerical value summarizes the change in population size from one winter to the next?

3.43 The amount of municipal solid waste created has become a major problem. According to the Environmental Protection Agency, the yearly amount (millions of tons) are:

Year	1960	1970	1980	1990	2000	2007
Garbage (million tons)	88	121	152	205	232	254
Population (millions)	179	203	227	249	282	302

- (a) Plot the amount of garbage (millions of tons) versus population (millions).
- (b) Fit a straight line using  $x =$  population in millions.
- (c) According to the fitted line, how much garbage is created by a typical person?

## USING STATISTICS WISELY

1. To study the association between two variables, you need to collect the pair of values obtained from each unit in the sample. There is no information about association in the summaries of the observations on individual variables.
2. To study association when both variables are categorical, cross-tabulate the frequencies in a two-way table. Calculate relative frequencies based on the total number.
3. To look for association between any pair of variables whose values are numerical, create a scatter diagram and look for a pattern of association.
4. Never confuse a strong association with a causal relationship. The relation may be due to a lurking variable.
5. Remember that the correlation coefficient measures the clustering of points about a straight line. It is not appropriate for a relationship on a curve or disjoint groups of points.
6. Before using a fitted line to predict one variable from another, create a scatter plot and inspect the pattern to see if a straight-line relationship is appropriate.

## KEY IDEAS AND FORMULAS

Comparative trials often have a **placebo**, or inactive treatment, which serves as a control. This eliminates from the comparison a **placebo effect** where some subjects in the control group responded positively to an ineffective treatment because their expectation to improve is so strong. An experiment has **double blind trials** when neither the subject nor the person administering the treatments knows which treatment is given.

A **random assignment** of treatments helps prevent uncontrolled sources of variation from systematically influencing the responses.

Data on two traits can be summarized in a two-way table of frequencies where the categories for one trait are in the left margin and categories for the other trait along the upper margin. These are said to be **cross-classified** or **cross-tabulated data** and the summary tables of frequencies are called **contingency tables**.

The combining of two contingency tables that pertain to the same two traits, but are based on very different populations, can lead to very misleading conclusions if the two tables are combined. This is called **Simpson's paradox** when there is a third variable that strongly influences the other two.



A **scatter plot** or **scatter diagram** shows all the values  $(x_i, y_i)$  of a pair of variables as points in two dimensions. This plot can be visually inspected for the strength of association between the two variables.

The **correlation coefficient**  $r$  measures how closely the scatter approximates a straight-line pattern.

A positive value of correlation indicates a tendency of large values of  $x$  to occur with large values of  $y$ , and also for small values of both to occur together.

A negative value of correlation indicates a tendency of large values of  $x$  to occur with small values of  $y$  and vice versa.

A high correlation does not necessarily imply a causal relation.

In fact, a high value of correlation between two variables may be **spurious**. That is, the two variables may not be connected but the apparent correlation is caused by a third **lurking variable** that strongly influences both of the original two variables.

A least squares fit of a straight line helps describe the relation of the **response** or **output** variable  $y$  to the **predictor** or **input** variable  $x$ .

A  $y$  value may be predicted for a known  $x$  value by reading from the fitted line  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ .

For pairs of measurements  $(x, y)$

$$\text{Sample correlation } r = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}}$$

where

$$S_{xx} = \sum (x - \bar{x})^2, S_{yy} = \sum (y - \bar{y})^2, \text{ and } S_{xy} = \sum (x - \bar{x})(y - \bar{y}).$$

$$\text{Fitted line } \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where

$$\text{Slope } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \text{Intercept } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

## TECHNOLOGY

### *Fitting a straight line and calculating the correlation coefficient*

#### MINTAB

#### *Fitting a straight line*

Begin with the values for the predictor variable  $x$  in *C1* and the response variable  $y$  in *C2*.

Stat > Regression > Regression.

Type *C2* in Response. Type *C1* in Predictors.

Click OK.

To calculate the correlation coefficient, start as above with data in  $C1$  and  $C2$ .

**Stat > Basic Statistics > Correlation.**  
Type  $C1$   $C2$  in **Variables**. Click **OK**.

## EXCEL

### *Fitting a straight line*

Begin with the values of the predictor variable in column A and the values of the response variable in column B.

Highlight the data and go to **Insert**, then **Chart**.  
Select **XY (Scatter)** and click **Finish**.  
Go to **Chart**, then **Add Trendline**.  
Click on the **Options** tab and check **Display equation on chart**.  
Click **OK**.

To calculate the correlation coefficient, begin with the predictor variable in column A and the response variable in column B.

Click on a blank cell. Select **Insert** and then **Function** (or click on the  $f_x$  icon).  
Select **Statistical** and then **CORREL**. Click **OK**.  
**Highlight** the data in Column A for **Array1** and **Highlight** the data in Column B for **Array2**. Then, click **OK**.

## TI-84/83 PLUS

### *Fitting a straight line*

Press **STAT** then **1: Edit**.  
Enter the values of the predictor variable in **L1** and those of the response variable in **L2**.

Select **STAT** then **Calc** and then **4: LinReg (ax+b)**.  
With **LinReg** on the Home screen press **Enter**.

The calculator will return the intercept  $a$ , slope  $b$ , correlation coefficient  $r$ .  
(If  $r$  is not shown, go to the 2nd **O: CATALOG** and select **DiagnosticON**. Press **ENTER** twice. Then go back to **LinReg**.)

## 7. REVIEW EXERCISES

- 3.44 Applicants for welfare are allowed an appeals process when they feel they have been unfairly treated. At the hearing, the applicant may choose self-representation or representation by an attorney. The appeal may result in an increase, decrease, or no change in benefit recommendation. Court records of 320 appeals cases provided the data at the top of the next page. Calculate the relative frequencies for each row and compare the patterns of the appeals decisions between the two types of representation.

Type of Representation	Amount of Aid			Total
	Increased	Unchanged	Decreased	
Self	59	108	17	
Attorney	70	63	3	
Total				

3.45 Sugar content (g) and carbohydrate content (g) are obtained from the package of the breakfast cereals referred to in Exercise 3.2.

General Mills		Kellogg's		Quaker	
Sugar	Carb.	Sugar	Carb.	Sugar	Carb.
13	12	13	15	9	14
1	18	4	18	6	17
13	11	14	14	10	12
13	11	12	15	0	12
12	12	3	21	14	29
5	16	18	21	13	23
19	19	15	8	9	31
16	22	16	13	12	23
14	26	15	31	16	15
16	28	4	20	13	15

- (a) Calculate the sample mean carbohydrates for all 30 cereals.
- (b) Construct a table like the one in Exercise 3.2 but using carbohydrates rather than sugar.
- (c) Calculate the relative frequencies separately for each row. Comment on any pattern.

3.46 Refer to Exercise 3.45.

- (a) Make a scatter plot for the cereals made by General Mills.
- (b) Calculate  $r$  for the cereals made by General Mills. Do sugar content and carbohydrate content seem to be associated or unrelated?

3.47 A dealer's recent records of 80 truck sales provided the following frequency information on size of truck and type of drive.

Truck Size	2-Wheel Drive	4-Wheel Drive
Small	12	23
Full	20	25

- (a) Determine the marginal totals.
- (b) Obtain the table of relative frequencies.
- (c) Calculate the relative frequencies separately for each row.
- (d) Does there appear to be a difference in the choice of drive for purchasers of small- and full-size trucks?

3.48 A high-risk group of 1083 male volunteers was included in a major clinical trial for testing a new vaccine for type B hepatitis. The vaccine was given to 549 persons randomly selected from the group, and the others were injected with a neutral substance (placebo). Eleven of the vaccinated people and 70 of the nonvaccinated ones later got the disease.

- (a) Present these data in the following two-way frequency table.
- (b) Compare the rates of incidence of hepatitis between the two subgroups.

	Hepatitis	No Hepatitis	Total
Vaccinated			
Not vaccinated			
Total			

3.49 Would you expect a positive, negative, or nearly zero correlation for each of the following? Give reasons for your answers.

- (a) The time a student spends playing computer games each week and the time they spend talking with friends in a group.
- (b) The number of finals taken by undergraduates and their number of hours of sleep during finals week.
- (c) A person's height and the number of movies he or she watched last month.

- (d) The temperature at a baseball game and beer sales.
- 3.50 Examine each of the following situations and state whether you would expect to find a high correlation between the variables. Give reasons why an observed correlation cannot be interpreted as a direct relationship between the variables and indicate at least one possible lurking variable.
- The correlation between the number of Internet users and truck sales in cities of varying sizes.
  - The correlation between yearly sales of satellite TV receivers and portable MP3 players over the past 10 years.
  - The correlation between yearly sales of cell phones and number of new automated teller machines over the past 10 years.
  - Correlation between the concentration  $x$  of air pollutants and the number of riders  $y$  on public transportation facilities when the data are collected from several cities that vary greatly in size.
  - Correlation between the wholesale price index  $x$  and the average speed  $y$  of winning cars in the Indianapolis 500 during the last 10 years.
- 3.51 The tar yield of cigarettes is often assayed by the following method: A motorized smoking machine takes a two-second puff once every minute until a fixed butt length remains. The total tar yield is determined by laboratory analysis of the pool of smoke taken by the machine. Of course, the process is repeated on several cigarettes of a brand to determine the average tar yield. Given here are the data of average tar yield and the average number of puffs for six brands of filter cigarettes.

Average tar (milligrams)	12.2	14.3	15.7	12.6	13.5	14.0
Average no. of puffs	8.5	9.9	10.7	9.0	9.3	9.5

- Plot the scatter diagram.
- Calculate  $r$ .

**Remark:** Fewer puffs taken by the smoking machine mean a faster burn time. The amount of tar inhaled by a human smoker depends largely on how often the smoker puffs.

- 3.52 As part of a study of the psychobiological correlates of success in athletes, the following measurements (courtesy of W. Morgan) are obtained from members of the U.S. Olympic wrestling team.

Anger $x$	6	7	5	21	13	5	13	14
Vigor $y$	28	23	29	22	20	19	28	19

- Plot the scatter diagram.
  - Calculate  $r$ .
  - Obtain the least squares line.
  - Predict the vigor score  $y$  when the anger score is  $x = 8$ .
- 3.53 Refer to Exercise 3.45.
- Make a scatter plot for the cereals made by Kellogg's.
  - Calculate  $r$  for the cereals made by Kellogg's. Do sugar content and carbohydrate content seem to be associated or unrelated?
- 3.54 Given the following  $(x, y)$  values

$x$	0	2	5	4	1	6
$y$	5	4	4	2	7	2

- Make a scatter plot.
  - Calculate  $r$ .
- 3.55 Given these five pairs of values

$x$	0	3	5	8	9
$y$	1	2	4	3	5

- Plot the scatter diagram.
  - From a visual inspection, draw a straight line that appears to fit the data well.
  - Compute the least squares estimates  $\hat{\beta}_0$ ,  $\hat{\beta}_1$  and draw the fitted line.
- 3.56 For samples collected concerning the following pairs of variables, decide whether you should make a scatter plot or create a contingency table.
- The amount earned and the number of weeks worked during the last summer vacation.

- (b) Home ownership (own/rent) and having own bedroom (no/yes) during the freshman year of high school.
  - (c) Number of days that groceries were purchased and the number of days gas was purchased, in the past week.
- 3.57 Identify the predictor variable  $x$  and the response variable  $y$  in each of the following situations.
- (a) The state highway department wants to study the relationship between road roughness and a car's gas consumption.
  - (b) A concession salesperson at football games wants to relate total fall sales to the number of games the home team wins.
  - (c) A sociologist wants to investigate the number of weekends a college student goes home in relation to the trip distance.
- 3.60 Use MINITAB or some other computer package to obtain the scatter diagram, correlation coefficient, and regression line of:
- (a) The final on the initial times to row given in Table D.4 in the Data Bank.
  - (b) Drop one unusual pair and repeat part(a). Comment on any major differences.
- 3.61 A director of student counseling is interested in the relationship between the numerical score  $x$  and the social science score  $y$  on college qualification tests. The following data (courtesy of R. W. Johnson) are recorded.

$x$	41	39	53	67	61	67
$y$	29	19	30	27	28	27

$x$	46	50	55	72	63	59
$y$	22	29	24	33	25	20

$x$	53	62	65	48	32	64
$y$	28	22	27	22	27	28

$x$	59	54	52	64	51	62
$y$	30	29	21	36	20	29

$x$	56	38	52	40	65	61
$y$	34	21	25	24	32	29

$x$	64	64	53	51	58	65
$y$	27	26	24	25	34	28

- (a) Plot the scatter diagram.
- (b) Calculate  $r$ .

**The Following Exercises Require a Computer**

3.58 In Figure 11, we have illustrated the output from MINITAB commands for fitting a straight line. To create the scatter plot, without the fitted line, choose:

**Graph > Scatter plot.** Choose **Simple**. Click **OK**. Type **C2** in **Y variables** and **C1** in **X variables**. Click **OK**.

Use MINITAB (or another package program) to obtain the scatter diagram, correlation coefficient, and regression line for:

- (a) The GPA and GMAT scores data of Table 7 in Example 3.
  - (b) The hydrogen  $x$  and carbon  $y$  data in Exercise 3.25.
- 3.59 For fitting body length to weight for all wolves given in Table D.9 in the Data Bank, use MINITAB or some other computer package to obtain:
- (a) The scatter diagram.
  - (b) The correlation coefficient.
  - (c) The regression line.

# 4

## Probability

1. Introduction
2. Probability of an Event
3. Methods of Assigning Probability
4. Event Relations and Two Laws of Probability
5. Conditional Probability and Independence
6. Bayes' Theorem
7. Random Sampling from a Finite Population
8. Review Exercises

---

---

## *Uncertainty of Weather Forecasts*

Today's forecast: Increasing cloudiness with a 25% chance of snow.



Probabilities express the chance of events that cannot be predicted with certainty. Even unlikely events sometimes occur. © age fotostock/Superstock.

---

---



## 1. INTRODUCTION

---

In Chapter 1, we introduced the notions of *sample* and *statistical population* in the context of investigations where the outcomes exhibit variation. Although complete knowledge of the statistical population remains the target of an investigation, we typically have available only the partial information contained in a sample. Chapter 2 focused on some methods for describing the salient features of a data set by graphical presentations and calculation of the mean, standard deviation, and other summary statistics. When the data set represents a sample from a statistical population, its description is only a preliminary part of a statistical analysis. Our major goal is to make generalizations or inferences about the target population on the basis of information obtained from the sample data. An acquaintance with the subject of probability is essential for understanding the reasoning that leads to such generalizations.

In everyday conversations, we all use expressions of the kind:

“Most likely our team will win this Saturday.”

“It is unlikely that the weekend will be cold.”

“I have a 50–50 chance of getting a summer job at the camp.”

The phrases “most likely,” “probable,” “quite likely,” and so on are used qualitatively to indicate the chance that an event will occur. Probability, as a subject, provides a means of quantifying uncertainty. In general terms, the probability of an event is a numerical value that gauges how likely it is that the event will occur. We assign probability on a scale from 0 to 1 with a very low value indicating extremely unlikely, a value close to 1 indicating very likely, and the intermediate values interpreted accordingly. A full appreciation for the concept of a numerical measure of uncertainty and its role in statistical inference can be gained only after the concept has been pursued to a reasonable extent. We can, however, preview the role of probability in one kind of statistical reasoning.

*Suppose it has been observed that in 50% of the cases a certain type of muscular pain goes away by itself. A hypnotist claims that her method is effective in relieving the pain. For experimental evidence, she hypnotizes 15 patients and 12 get relief from the pain. Does this demonstrate that hypnotism is effective in stopping the pain?*

*Let us scrutinize the claim from a statistical point of view. If indeed the method had nothing to offer, there could still be a 50–50 chance that a patient is cured. Observing 12 cures out of 15 amounts to obtaining 12 heads in 15 tosses of a coin. We will see later that the probability of at least 12 heads in 15 tosses of a fair coin is .018, indicating that the event is not likely to happen. Thus, if we tentatively assume the model (or hypothesis) that the method is ineffective, 12 or more cures are very unlikely. Rather than agree that an unlikely*



*event has occurred, we conclude that the experimental evidence strongly supports the hypnotist's claim.*

This kind of reasoning, called *testing a statistical hypothesis*, will be explored in greater detail later. For now, we will be concerned with introducing the ideas that lead to assigned values for probabilities.

## 2. PROBABILITY OF AN EVENT

The probability of an event is viewed as a numerical measure of the chance that the event will occur. The idea is naturally relevant to situations where the outcome of an experiment or observation exhibits variation.

Although we have already used the terms “experiment” and “event,” a more specific explanation is now in order. In the present context, the term experiment is not limited to the studies conducted in a laboratory. Rather, it is used in a broad sense to include any operation of data collection or observation where the outcomes are subject to variation. Rolling a die, drawing a card from a shuffled deck, sampling a number of customers for an opinion survey, and quality inspection of items from a production line are just a few examples.

An **experiment** is the process of observing a phenomenon that has variation in its outcomes.

Before attempting to assign probabilities, it is essential to consider all the eventualities of the experiment. Pertinent to their description, we introduce the following terminologies and explain them through examples.

The **sample space** associated with an experiment is the collection of all possible distinct outcomes of the experiment.

Each outcome is called an **elementary outcome**, a **simple event**, or an **element of the sample space**.

An **event** is the set of elementary outcomes possessing a designated feature.

The elementary outcomes, which together comprise the sample space, constitute the ultimate breakdown of the potential results of an experiment. For instance, in rolling a die, the elementary outcomes are the points 1, 2, 3, 4, 5, and 6, which together constitute the sample space. The outcome of a football game

would be either a win, loss, or tie for the home team. Each time the experiment is performed, one and only one elementary outcome can occur. A sample space can be specified by either listing all the elementary outcomes, using convenient symbols to identify them, or making a descriptive statement that characterizes the entire collection. For general discussion, we denote:

The sample space by  $\mathcal{S}$

The elementary outcomes by  $e_1, e_2, e_3, \dots$

Events by  $A, B$ , and so on.

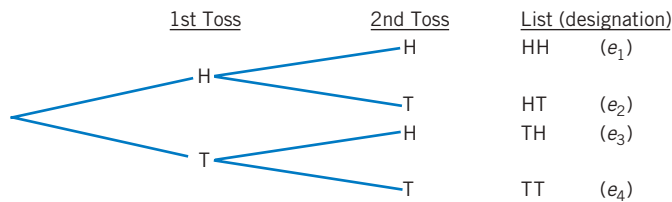
In specific applications, the elementary outcomes may be given other labels that provide a more vivid identification.

An event  $A$  occurs when any one of the elementary outcomes in  $A$  occurs.

### Example 1 A Tree Diagram and Events for Coin Tossing

Toss a coin twice and record the outcome head (H) or tail (T) for each toss. Let  $A$  denote the event of getting exactly one head and  $B$  the event of getting no heads at all. List the sample space and give the compositions of  $A$  and  $B$ .

**SOLUTION** For two tosses of a coin, the elementary outcomes can be conveniently identified by means of a **tree diagram**.



The sample space can then be listed as  $\mathcal{S} = \{HH, HT, TH, TT\}$ . With the designation given above, we can also write

$$\mathcal{S} = \{e_1, e_2, e_3, e_4\}$$

The order in which the elements of  $\mathcal{S}$  are listed is inconsequential. It is the collection that matters.

Consider the event  $A$  of getting exactly one head. Scanning the above list, we see that only the elements  $HT$  ( $e_2$ ) and  $TH$  ( $e_3$ ) satisfy this requirement. Therefore, the event  $A$  has the composition

$$A = \{e_2, e_3\}$$

which is, of course, a subset of  $S$ . The event  $B$  of getting no heads at all consists of the single element  $e_4$  so  $B = \{e_4\}$ . That is,  $B$  is a simple event as well as an event. The term “event” is a general term that includes simple events.

### Example 2 A Sample Space and an Event Based on a Count

On a Saturday afternoon, 135 customers will be observed during check-out and the number paying by card, credit or debit, will be recorded. Identify (a) the sample space and (b) the event that more than 50% of purchases are made with a card.

#### SOLUTION

- (a) Since the number of customers who purchase with a card could be any of the numbers 0, 1, 2, . . . , 135, the sample space can be listed simply as

$$S = \{0, 1, 2, \dots, 135\}$$

Using the notation  $e$  for elementary outcome, one can also describe this sample space as  $S = \{e_0, e_1, e_2, \dots, e_{135}\}$

- (b) Let  $A$  stand for the event that more than 50% of the customers purchase with a card. Calculating  $.5 \times 135 = 67.5$ , we identify

$$A = \{68, 69, \dots, 135\}$$

Both Examples 1 and 2 illustrate sample spaces that have a finite number of elements. There are also sample spaces with infinitely many elements. For instance, suppose a gambler at a casino will continue pulling the handle of a slot machine until he hits the first jackpot. The conceivable number of attempts does not have a natural upper limit so the list never terminates. That is,  $S = \{1, 2, 3, \dots\}$  has an infinite number of elements. However, we notice that the elements could be arranged one after another in a sequence. An infinite sample space where such an arrangement is possible is called “countably infinite.” Either of these two types of sample spaces is called a **discrete sample space**.

Another type of infinite sample space is also important. Suppose a car with a full tank of gasoline is driven until its fuel runs out and the distance traveled recorded. Since distance is measured on a continuous scale, any nonnegative number is a possible outcome. Denoting the distance traveled by  $d$ , we can describe this sample space as  $S = \{d; d \geq 0\}$ , that is, the set of all real numbers greater than or equal to zero. Here the elements of  $S$  form a continuum and cannot be arranged in a sequence. Any  $S$  that is an interval is called a **continuous sample space**.

To avoid unnecessary complications, we will develop the basic principles of probability in the context of finite sample spaces. We first elaborate on the notion of the probability of an event as a numerical measure of the chance that it will occur. The most intuitive interpretation of this quantification is to consider the fraction of times the event would occur in many repeated trials of the experiment.

The **probability of an event** is a numerical value that represents the proportion of times the event is expected to occur when the experiment is repeated many times under identical conditions.

The probability of event  $A$  is denoted by  $P(A)$ .

Since a proportion must lie between 0 and 1, the probability of an event is a number between 0 and 1. To explore a few other important properties of probability, let us refer to the experiment in Example 1 of tossing a coin twice. The event  $A$  of getting exactly one head consists of the elementary outcomes HT ( $e_2$ ) and TH ( $e_3$ ). Consequently,  $A$  occurs if either of these occurs. Because

$$\left[ \begin{array}{c} \text{Proportion of times} \\ A \text{ occurs} \end{array} \right] = \left[ \begin{array}{c} \text{Proportion of times} \\ \text{HT occurs} \end{array} \right] + \left[ \begin{array}{c} \text{Proportion of times} \\ \text{TH occurs} \end{array} \right]$$

the number that we assign as  $P(A)$  must be the sum of the two numbers  $P(\text{HT})$  and  $P(\text{TH})$ . Guided by this example, we state some general properties of probability.

**The probability of an event is the sum of the probabilities assigned to all the elementary outcomes contained in the event.**

Next, since the sample space  $S$  includes all conceivable outcomes, in every trial of the experiment some element of  $S$  must occur. Viewed as an event,  $S$  is certain to occur, and therefore its probability is 1.

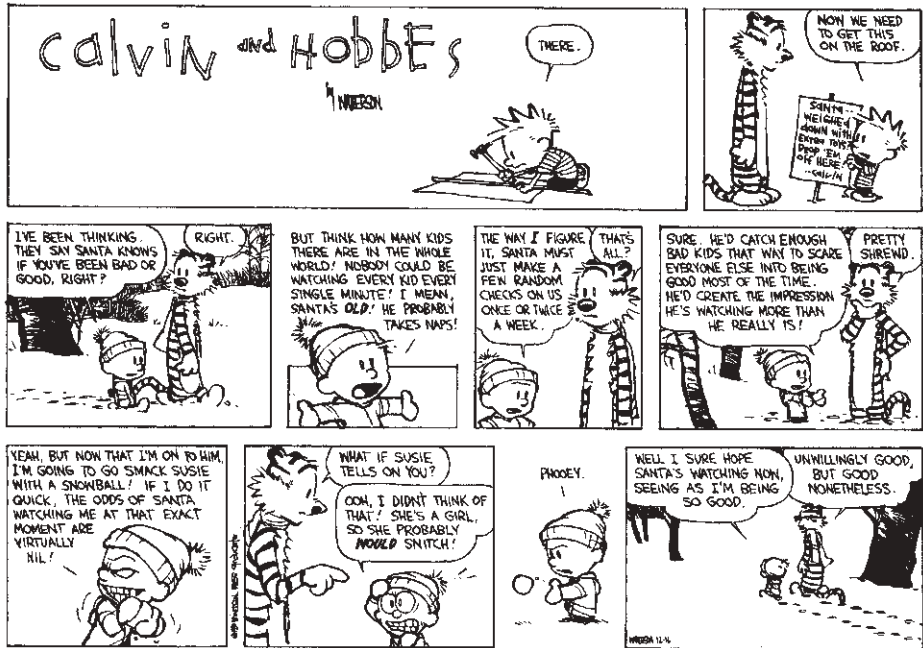
**The sum of the probabilities of all the elements of  $S$  must be 1.**

In summary:

**Probability** must satisfy:

1.  $0 \leq P(A) \leq 1$  for all events  $A$
2.  $P(A) = \sum_{\text{all } e \text{ in } A} P(e)$
3.  $P(S) = \sum_{\text{all } e \text{ in } S} P(e) = 1$

We have deduced these basic properties of probability by reasoning from the definition that the probability of an event is the proportion of times the event is expected to occur in many repeated trials of the experiment.



An assessment of the probabilities of events and their consequences can help to guide decisions. Calvin and Hobbes © 1990 Universal Press Syndicate. Reprinted with permission. All rights reserved.

### Exercises

4.1 Match the proposed probability of  $A$  with the appropriate verbal description. (More than one description may apply.)

Probability	Verbal Description
(a) .03	(i) No chance of happening
(b) .96	(ii) Very likely to happen
(c) 2.0	(iii) As much chance of occurring as not
(d) $-.1$	(iv) Very little chance of happening
(e) .3	(v) May occur but by no means certain
(f) 0	(vi) An incorrect assignment
(g) .5	

- (a) 1.2 (b)  $\frac{1}{1.2}$  (c)  $\frac{1}{2}$  (d)  $\frac{43}{47}$   
 (e)  $\frac{1}{79}$  (f) 1.0

Verbal statements: (i) cannot be a probability, (ii) the event is very unlikely to happen, (iii) 50–50 chance of happening, (iv) sure to happen, (v) more likely to happen than not.

4.2 For each numerical value assigned to the probability of an event, identify the verbal statements that are appropriate.

- 4.3 Identify the statement that best describes each  $P(A)$ .  
 (a)  $P(A) = .04$  (i)  $P(A)$  is incorrect.  
 (b)  $P(A) = .33$  (ii)  $A$  rarely occurs.  
 (c)  $P(A) = 1.4$  (iii)  $A$  occurs moderately often.

4.4 Construct a sample space for each of the following experiments.

- (a) Someone claims to be able to taste the difference between the same brand of bottled, tap, and canned draft beer. A glass of each is poured and given to the subject in an unknown order. The subject is asked to identify the contents of each glass. The number of correct identifications will be recorded.
- (b) Record the number of traffic fatalities in a state next year.
- (c) Observe the length of time a new digital video recorder will continue to work satisfactorily without service.

Which of these sample spaces are discrete and which are continuous?

- 4.5 Identify these events in Exercise 4.4.
- (a) Not more than one correct identification.
  - (b) Less accidents than last year.  
(Note: If you don't know last year's value, use 345.)
  - (c) Longer than the 90-day warranty but less than 425.4 days.
- 4.6 When bidding on two projects, the president and vice president of a construction company make the following probability assessments for winning the contracts.

President	Vice President
$P(\text{win none}) = .1$	$P(\text{win none}) = .1$
$P(\text{win only one}) = .5$	$P(\text{win Project 1}) = .4$
$P(\text{win both}) = .4$	$P(\text{win Project 2}) = .2$
	$P(\text{win both}) = .3$

For both cases, examine whether or not the probability assignment is permissible.

- 4.7 Bob, John, Linda, and Sue are the finalists in the campus bowling tournament. The winner and the first runner-up will be sent to a statewide competition.
- (a) List the sample space concerning the outcomes of the local tournament.
  - (b) Give the composition of each of the following events.  
 $A =$  Linda wins the local tournament  
 $B =$  Bob does not go to the state tournament

- 4.8 Consider the following experiment: A coin will be tossed twice. If both tosses show heads, the experiment will stop. If one head is obtained in the two tosses, the coin will be tossed one more time, and in the case of both tails in the two tosses, the coin will be tossed two more times.

- (a) Make a tree diagram and list the sample space.
- (b) Give the composition of the following events.

$$A = [\text{Two heads}] \quad B = [\text{Two tails}]$$

- 4.9 There are four elementary outcomes in a sample space. If  $P(e_1) = .3$ ,  $P(e_2) = .4$ , and  $P(e_3) = .2$ , what is the probability of  $e_4$ ?
- 4.10 Suppose  $S = \{e_1, e_2, e_3\}$ . If the simple events  $e_1, e_2$ , and  $e_3$  are all equally likely, what are the numerical values  $P(e_1), P(e_2)$ , and  $P(e_3)$ ?
- 4.11 The sample space for the response of a single person's attitude toward a political issue consists of the three elementary outcomes  $e_1 = \{\text{Unfavorable}\}$ ,  $e_2 = \{\text{Favorable}\}$ , and  $e_3 = \{\text{Undecided}\}$ . Are the following assignments of probabilities permissible?
- (a)  $P(e_1) = .8, P(e_2) = .1, P(e_3) = .1$
  - (b)  $P(e_1) = .3, P(e_2) = .3, P(e_3) = .3$
  - (c)  $P(e_1) = .5, P(e_2) = .5, P(e_3) = .0$
- 4.12 A campus organization will select one day of the week for an end-of-year picnic. Assume that the weekdays, Monday through Friday, are equally likely and that each weekend day, Saturday and Sunday, is twice as likely as a weekday to be selected.
- (a) Assign probabilities to the seven outcomes.
  - (b) Find the probability a weekday will be selected.
- 4.13 The month in which the year's highest temperature occurs in a city has probabilities in the ratio 1:3:6:10 for May, June, July, and August, respectively. Find the probability that the highest temperature occurs in either May or June.
- 4.14 **Probability and odds.** The probability of an event is often expressed in terms of odds. Specifically, when we say that the odds are  $k$  to  $m$  that

an event will occur, we mean that the probability of the event is  $k/(k + m)$ . For instance, “the odds are 4 to 1 that candidate Jones will win” means that  $P(\text{Jones will win}) = \frac{4}{5} = .8$ . Express the following statements in terms of probability.

- (a) The odds are 3 to 1 that there will be good weather tomorrow.
- (b) The odds are 7 to 3 that the city council will delay the funding of a new sports arena.

### 3. METHODS OF ASSIGNING PROBABILITY

An assignment of probabilities to all the events in a sample space determines a probability model. In order to be a valid probability model, the probability assignment must satisfy the properties 1, 2, and 3 stated in the previous section. Any assignment of numbers  $P(e_i)$  to the elementary outcomes will satisfy the three conditions of probability provided these numbers are nonnegative and their sum over all the outcomes  $e_i$  in  $S$  is 1. However, to be of any practical import, the probability assigned to an event must also be in agreement with the concept of probability as the proportion of times the event is expected to occur. Here we discuss the implementation of this concept in two important situations.

#### 3.1. EQUALLY LIKELY ELEMENTARY OUTCOMES— THE UNIFORM PROBABILITY MODEL

Often, the description of an experiment ensures that each elementary outcome is as likely to occur as any other. For example, consider the experiment of rolling a fair die and recording the top face. The sample space can be listed as

$$S = \{e_1, e_2, e_3, e_4, e_5, e_6\}$$

where  $e_1$  stands for the elementary outcome of getting the face 1, and similarly,  $e_2, \dots, e_6$ . Without actually rolling a die, we can deduce the probabilities. Because a fair die is a symmetric cube, each of its six faces is as likely to appear as any other. In other words, each face is expected to occur one-sixth of the time. The probability assignments should therefore be

$$P(e_1) = P(e_2) = \dots = P(e_6) = \frac{1}{6}$$

and any other assignment would contradict the statement that the die is fair. We say that rolling a fair die conforms to a **uniform probability model** because the total probability 1 is evenly apportioned to all the elementary outcomes.

What is the probability of getting a number higher than 4? Letting  $A$  denote this event, we have the composition  $A = \{e_5, e_6\}$ , so

$$P(A) = P(e_5) + P(e_6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

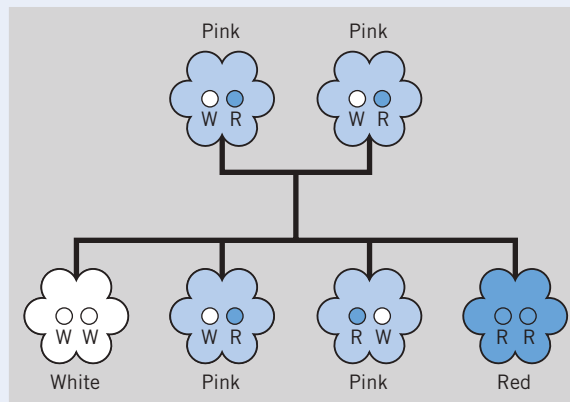
When the elementary outcomes are modeled as equally likely, we have a uniform probability model. If there are  $k$  elementary outcomes in  $\mathcal{S}$ , each is assigned the probability of  $1/k$ .

An event  $A$  consisting of  $m$  elementary outcomes is then assigned

$$P(A) = \frac{m}{k} = \frac{\text{No. of elementary outcomes in } A}{\text{No. of elementary outcomes in } \mathcal{S}}$$

Gregor Mendel, pioneer geneticist, perceived a pattern in the characteristics of generations of pea plants and conceived a theory of heredity to explain them. According to Mendel, inherited characteristics are transmitted from one generation to another by genes. Genes occur in pairs and the offspring obtain their pair by taking one gene from each parent. A simple uniform probability model lies at the heart of Mendel's explanation of the selection mechanism.

One experiment that illustrated Mendel's theory consists of cross fertilizing a pure strain of red flowers with a pure strain of white flowers. This produces hybrids having one gene of each type that are pink-flowered. Crossing these hybrids leads to one of four possible gene pairs. Under Mendel's laws, these four are equally likely. Consequently,  $P[\text{Pink}] = \frac{1}{2}$  and  $P[\text{White}] = P[\text{Red}] = \frac{1}{4}$ . (Compare with the experiment of tossing two coins.)



An experiment carried out by Correns, one of Mendel's followers, resulted in the frequencies 141, 291, and 132 for the white, pink, and red flowers, respectively. These numbers are nearly in the ratio 1 : 2 : 1. (Source: W. Johannsen, *Elements of the Precise Theory of Heredity*, Jena: G. Fischer, 1909.)



**Example 3** The Uniform Probability Model for Tossing a Fair Coin

Find the probability of getting exactly one head in two tosses of a fair coin.

**SOLUTION** As listed in Example 1, there are four elementary outcomes in the sample space:  $S = \{HH, HT, TH, TT\}$ . The very concept of a fair coin implies that the four elementary outcomes in  $S$  are equally likely. We therefore assign the probability  $\frac{1}{4}$  to each of them. The event  $A = [\text{One head}]$  has two elementary outcomes—namely, HT and TH. Hence,  $P(A) = \frac{2}{4} = .5$ .

**Example 4** Random Selection and the Uniform Probability Model

Suppose that among 50 students in a class, 42 are right-handed and 8 left-handed. If one student is randomly selected from the class, what is the probability that the selected student is left-handed?

**SOLUTION** The intuitive notion of random selection is that each student is as likely to be selected as any other. If we view the selection of each individual student as an elementary outcome, the sample space consists of 50  $e$ 's of which 8 are in the event "left-handed." Consequently,  $P[\text{Left-handed}] = \frac{8}{50} = .16$ .

**Note:** Considering that the selected student will be either left-handed ( $L$ ) or right-handed ( $R$ ), we can write the sample space as  $S = \{L, R\}$ , but we should be aware that the two elements  $L$  and  $R$  are not equally likely.

**3.2. PROBABILITY AS THE LONG-RUN RELATIVE FREQUENCY**

In many situations, it is not possible to construct a sample space where the elementary outcomes are equally likely. If one corner of a die is cut off, it would be unreasonable to assume that the faces remain equally likely and the assignments of probability to various faces can no longer be made by deductive reasoning. When speaking of the probability (or risk) that a man will die in his thirties, one may choose to identify the occurrence of death at each decade or even each year of age as an elementary outcome. However, no sound reasoning can be provided in favor of a uniform probability model. In fact, from extensive mortality studies, demographers have found considerable disparity in the risk of death for different age groups.

When the assumption of equally likely elementary outcomes is not tenable, how do we assess the probability of an event? The only recourse is to repeat the experiment many times and observe the proportion of times the

event occurs. Letting  $N$  denote the number of repetitions (or trials) of an experiment, we set

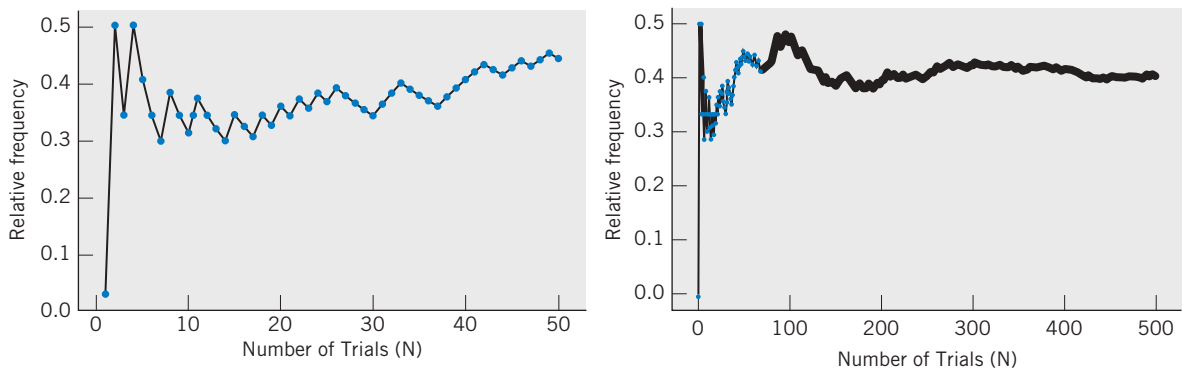
$$\text{Relative frequency of event } A \text{ in } N \text{ trials} = \frac{\text{No. of times } A \text{ occurs in } N \text{ trials}}{N}$$

For instance, let  $A$  be the event of getting a 6 when rolling a die. If the die is rolled 100 times and 6 comes up 23 times, the observed relative frequency of  $A$  would be  $\frac{23}{100} = .23$ . In the next 100 tosses, 6 may come up 18 times. Collecting these two sets together, we have  $N = 200$  trials with the observed relative frequency

$$\frac{23 + 18}{200} = \frac{41}{200} = .205$$

Imagine that this process is continued by recording the results from more and more tosses of the die and updating the calculations of relative frequency.

Figure 1 shows a typical plot of the relative frequency of an event  $A$  versus the number  $N$  of trials of the experiment. We see that the relative frequencies fluctuate as  $N$  changes, but the fluctuations become damped with increasing  $N$ . Two persons separately performing the same experiment  $N$  times are not going to get exactly the same graph. However, the numerical value at which the relative frequency stabilizes, in the long run, will be the same. This concept, called **long-run stability of relative frequency**, is illustrated in Figure 1*b*.



(a) Relative frequency versus number of trials. First 1–50. (b) Relative frequency versus number of trials. First 500 trials.

Figure 1 Stabilization of relative frequency.

Figure 1*a* graphically displays the considerable fluctuations present in the relative frequency as the number of trials increases from 1 to 50. Figure 1*b* displays the relative frequencies for the first 500 trials. In Figure 1*b*, the stabilization of relative frequency is evident, although the results for the first 50 trials are a little hard to discern in this view.

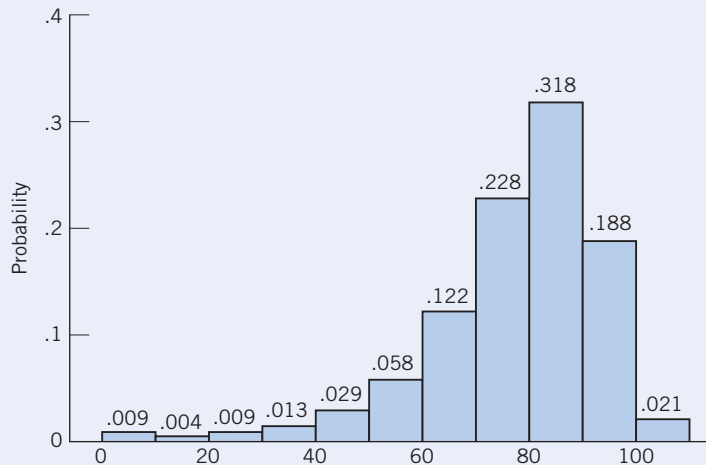
### Probability as Long-Run Relative Frequency

We define  $P(A)$ , the probability of an event  $A$ , as the value to which the relative frequency stabilizes with increasing number of trials.

Although we will never know  $P(A)$  exactly, it can be estimated accurately by repeating the experiment many times.

The property of the long-run stabilization of relative frequencies is based on the findings of experimenters in many fields who have undertaken the strain of studying the behavior of the relative frequencies under prolonged repetitions of their experiments. French gamblers, who provided much of the early impetus for the study of probability, performed experiments tossing dice and coins, drawing cards, and playing other games of chance thousands and thousands of times. They observed the stabilization property of relative frequency and applied this knowledge to achieve an understanding of the uncertainty involved in these games. Demographers have compiled and studied volumes of mortality data to examine the relative frequency of the occurrence of such events as death in particular age groups. In each

### How Long Will a Baby Live?



The probabilities for life length of a baby born in the United States. (Obtained from the *National Vital Statistics Reports* 54 [2006]).

context, the relative frequencies were found to stabilize at specific numerical values as the number of cases studied increased. Life and accident insurance companies actually depend on the stability property of relative frequencies.

As another example of an idealized model, consider the assignment of probabilities to the day of the week a child will be born. We may tentatively assume the simple model that all seven days of the week are equally likely. Each day is then assigned the probability  $\frac{1}{7}$ . If  $A$  denotes the event of a birth on the weekend (Saturday or Sunday), our model leads to the probability  $P(A) = \frac{2}{7}$ . The plausibility of the uniform probability model can only be ascertained from an extensive set of birth records.

Each newborn in the United States can be considered as a trial of the experiment where the day of birth determines whether or not the event  $A$  occurs. One year<sup>1</sup>, the outcomes for 4138.3 thousand newborns constitute a very large number of replications.

The resulting proportion of babies born on either Saturday or Sunday is  $\frac{830.4}{4138.3} = .201$ . This is quite far from the value  $\frac{2}{7} = .286$  predicted by the uniform model. The difference  $.285 - .201 = .084$  is much larger than would ordinarily occur by chance. A reasonable explanation is the increasing prevalence of elective induction of labor which is mostly performed on weekdays.

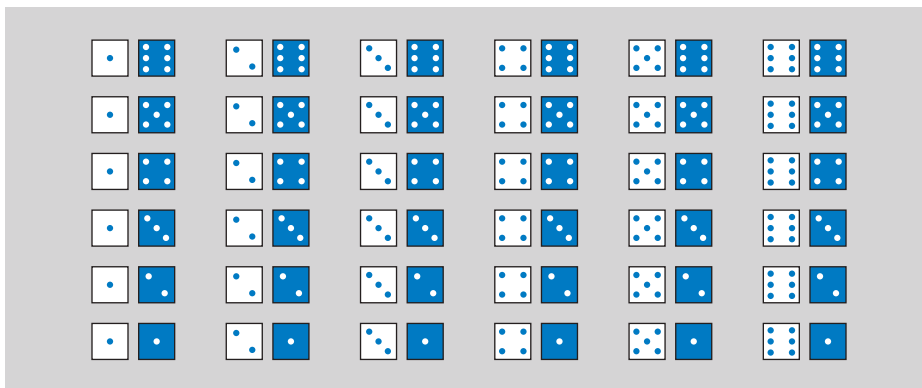
## Exercises

---

- 4.15 Refer to the day of birth data in the preceding text above. Assuming conditions are the same today, estimate the probability that a baby will be born during a weekday. That is, not on Saturday or Sunday.
- 4.16 Among 41,131 turkey permit holders for a recent hunting season in Wisconsin, 8845 harvested a bird. Assuming conditions are the same today, estimate the probability that a turkey will be harvested for a single permit.
- 4.17 Consider the experiment of tossing a coin three times.
- List the sample space by drawing a tree diagram.
  - Assign probabilities to the elementary outcomes.
  - Find the probability of getting exactly one head.

<sup>1</sup>National Vital Statistics Reports 56 (6) (December 5 2007).

- 4.18 A letter is chosen at random from the word "TEAM." What is the probability that it is a vowel?
- 4.19 A stack contains eight tickets numbered 1, 1, 2, 2, 2, 3, 3, 3. One ticket will be drawn at random and its number will be noted.
- List the sample space and assign probabilities to the elementary outcomes.
  - What is the probability of drawing an odd-numbered ticket?
- 4.20 Suppose you are eating at a pizza parlor with two friends. You have agreed to the following rule to decide who will pay the bill. Each person will toss a coin. The person who gets a result that is different from the other two will pay the bill. If all three tosses yield the same result, the bill will be shared by all. Find the probability that:
- Only you will have to pay.
  - All three will share.
- 4.21 A white and a colored die are tossed. The possible outcomes are shown in the illustration below.
- Identify the events  $A = [\text{Sum} = 6]$ ,  $B = [\text{Sum} = 7]$ ,  $C = [\text{Sum is even}]$ ,  $D = [\text{Same number on each die}]$ .
  - If both die are "fair," assign probability to each elementary outcome.
  - Obtain  $P(A)$ ,  $P(B)$ ,  $P(C)$ ,  $P(D)$ .
- 4.22 A roulette wheel has 34 slots, 2 of which are green, 16 are red, and 16 are black. A successful bet on black or red doubles the money, whereas one on green fetches 30 times as much. If you play the game once by betting \$5 on the black, what is the probability that:
- You will lose your \$5?
  - You will win \$5?
- \*4.23 One part of a quiz consists of two multiple-choice questions with the suggested answers: True (T), False (F), or Insufficient Data to Answer (I). An unprepared student randomly marks one of the three answers to each question.
- Make a tree diagram to list the sample space, that is, all possible pairs of answers the student might mark.
  - What is the probability of exactly one correct answer?
- 4.24 Based on the data of the Center for Health Statistics, the 2005 birth rates in 50 states are grouped in the following frequency table.



Birth rate (per thousand)	10–12	12–14	14–16
No. of states	7	23	16
Birth rate (per thousand)	16–18	18 and over	Total
No. of states	3	1	50

(Endpoint convention: Lower point is included, upper is not.)

If one state is selected at random, what is the probability that the birth rate there is:

- (a) Under 16?
- (b) Under 18 but not under 14?
- (c) 16 or over?

4.25 Fifteen persons reporting to a Red Cross center one day are typed for blood, and the following counts are found:

Blood group	O	A	B	AB	Total
No. of persons	3	5	6	1	15

If one person is randomly selected, what is the probability that this person's blood group is:

- (a) AB?
- (b) Either A or B?
- (c) Not O?

4.26 Friends will be called, one after another, and asked to go on a weekend trip with you. You will call until one agrees to go (A) or four friends are asked, whichever occurs first. List the sample space for this experiment.

4.27 Campers arriving at a summer camp will be asked one after another whether they have protection against Lyme disease (Y) or not (N). The inspection will continue until one camper is found to be not protected or until five campers are checked, whichever occurs first. List the sample space for this experiment.

- 4.28 (a) Consider the simplistic model that human births are evenly distributed over the 12 calendar months. If a person is randomly selected, say, from a phone directory, what is the probability that his or her birthday would be in November or December?
- (b) The following record shows a classification of births (thousands) in the United States. Calculate the relative frequency of births for each month and comment on the plausibility of the uniform probability model.

Jan.	331.5	July	357.1
Feb.	309.6	Aug.	369.3
March	349.3	Sept.	363.4
April	332.5	Oct.	344.6
May	346.3	Nov.	335.7
June	350.9	Dec.	348.3
		Total	<u>4,188.5</u>

4.29 A government agency will randomly select one of the 14 paper mills in a state to investigate its compliance with federal safety standards. Suppose, unknown to the agency, 9 of these mills are in compliance, 3 are borderline cases, and 2 are in gross violation.

- (a) Formulate the sample space in such a way that a uniform probability model holds.
- (b) Find the probability that a gross violator will be detected.

4.30 A plant geneticist crosses two parent strains, each with gene pairs of type  $aA$ . An offspring receives one gene from each parent.

- (a) Construct the sample space for the genetic type of the offspring.
- (b) Assign probabilities assuming that the selection of genes is random.
- (c) If  $A$  is dominant and the  $aa$  offspring are short while all the others are tall, find  $P$ [short offspring].

4.31 Explain why the long-run relative frequency interpretation of probability does not apply to the following situations.

- (a) The proportion of days when the home loan rate at your bank is above its value at the start of the year.
- (b) The proportion of cars that do not meet emission standards if the data are collected from service stations where the mechanics have been asked to check emissions while attending to other requested services.
- 4.32 A local bookstore intended to award three gift certificates in the amounts \$100, \$50, and \$25 to the first, second, and third customer to identify a mystery author. Unfortunately, a careless clerk in charge of mailing forgot the order and just randomly placed the gift certificates in the already addressed envelopes.
- (a) List the sample space using  $F$ ,  $S$ , and  $T$  for the three persons.
- (b) State the compositions of the events
- $$A = \text{[exactly one certificate is sent to the correct person]}$$
- $$B = \text{[all of the certificates are sent to incorrect persons]}$$
- 4.33 Refer to Exercise 4.32.
- (a) Assign probabilities to the elementary outcomes.
- (b) Find  $P(A)$  and  $P(B)$ .
- 4.34 Refer to Exercise 4.28. Using relative frequencies to estimate probabilities, find which 3 consecutive months have the lowest probability of a new birth.

## 4. EVENT RELATIONS AND TWO LAWS OF PROBABILITY

Later, when making probability calculations to support generalizations from the actual sample to the complete population, we will need to calculate probabilities of combined events, such as whether the count of no shows for a flight is either large or low.

Recall that the probability of an event  $A$  is the sum of the probabilities of all the elementary outcomes that are in  $A$ . It often turns out, however, that the event of interest has a complex structure that requires tedious enumeration of its elementary outcomes. On the other hand, this event may be related to other events that can be handled more easily. The purpose of this section is to first introduce the three most basic event relations: **complement**, **union**, and **intersection**. These event relations will then motivate some laws of probability.

The event operations are conveniently described in graphical terms. We first represent the sample space as a collection of points in a diagram, each identified with a specific elementary outcome. The geometric pattern of the plotted points is irrelevant. What is important is that each point is clearly tagged to indicate which elementary outcome it represents and to watch that no elementary outcome is missed or duplicated in the diagram. To represent an event  $A$ , identify the points that correspond to the elementary outcomes in  $A$ , enclose them in a boundary line, and attach the tag  $A$ . This representation, called a **Venn diagram**, is illustrated in Figure 2.

**Example 5** Venn Diagram for Coin Tossing

Make a Venn diagram for the experiment of tossing a coin twice and indicate the following events.

$A$ : Tail at the second toss

$B$ : At least one head

**SOLUTION** Here the sample space is  $\mathcal{S} = \{HH, HT, TH, TT\}$ , and the two events have the compositions  $A = \{HT, TT\}$ ,  $B = \{HH, HT, TH\}$ . Figure 2 shows the Venn diagram.

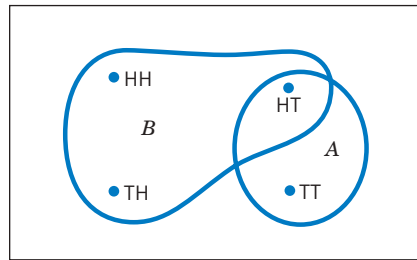


Figure 2 Venn diagram of the events in Example 5.

**Example 6** A Venn Diagram for the Selection of Puppies

Four young lab puppies from different litters are available for a new method of training.

Dog	Sex	Age (weeks)
1	M	10
2	M	15
3	F	10
4	F	10

Two dogs will be selected by lottery to receive the training. Considering all possible choices of two puppies, make a Venn diagram and show the following events.

$A$ : The selected dogs are of the same sex.

$B$ : The selected dogs are of the same age.



**SOLUTION** Here the elementary outcomes are the possible choices of a pair of numbers from  $\{1, 2, 3, 4\}$ . These pairs are listed and labeled as  $e_1, e_2, e_3, e_4, e_5, e_6$  for ease of reference.

$\{1, 2\}$	$(e_1)$	$\{2, 3\}$	$(e_4)$
$\{1, 3\}$	$(e_2)$	$\{2, 4\}$	$(e_5)$
$\{1, 4\}$	$(e_3)$	$\{3, 4\}$	$(e_6)$

The pair  $\{1, 2\}$  has both puppies of the same sex, and so does the pair  $\{3, 4\}$ . Consequently,  $A = \{e_1, e_6\}$ . Those with the same ages are  $\{1, 3\}, \{1, 4\}$ , and  $\{3, 4\}$ , so  $B = \{e_2, e_3, e_6\}$ . Figure 3 shows the Venn diagram.

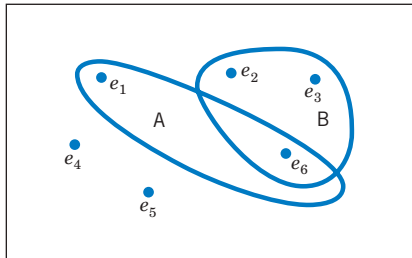


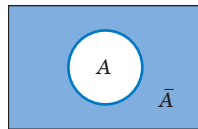
Figure 3 Venn diagram of the events in Example 6.

We now proceed to define the three basic event operations and introduce the corresponding symbols.

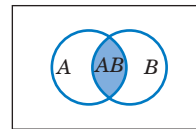
The **complement** of an event  $A$ , denoted by  $\bar{A}$ , is the set of all elementary outcomes that are not in  $A$ . The occurrence of  $\bar{A}$  means that  $A$  *does not occur*.

The **union** of two events  $A$  and  $B$ , denoted by  $A \cup B$ , is the set of all elementary outcomes that are in  $A$ ,  $B$ , or both. The occurrence of  $A \cup B$  means that *either  $A$  or  $B$  or both occur*.

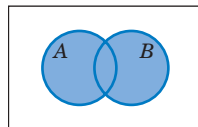
The **intersection** of two events  $A$  and  $B$ , denoted by  $AB$ , is the set of all elementary outcomes that are in  $A$  and  $B$ . The occurrence of  $AB$  means that *both  $A$  and  $B$  occur*.



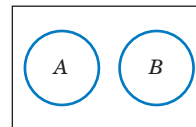
Complement  $\bar{A}$



Intersection  $AB$



Union  $A \cup B$



Incompatible events

Note that  $A \cup B$  is a larger set containing  $A$  as well as  $B$ , whereas  $AB$  is the common part of the sets  $A$  and  $B$ . Also it is evident from the definitions that  $A \cup B$  and  $B \cup A$  represent the same event, while  $AB$  and  $BA$  are both expressions for the intersection of  $A$  and  $B$ . The operations of union and intersection can be extended to more than two events. For instance,  $A \cup B \cup C$  stands for the set of all points that are in *at least one* of  $A$ ,  $B$ , and  $C$ , whereas  $ABC$  represents the *simultaneous occurrence* of all three events.

Two events  $A$  and  $B$  are called **incompatible** or **mutually exclusive** if their intersection  $AB$  is empty. Because incompatible events have no elementary outcomes in common, they cannot occur simultaneously.

### Example 7 Determining the Composition of Events Defined by Complement, Union, or Intersection

Refer to the experiment in Example 6 of selecting two puppies out of four. Let  $A = [\text{Same sex}]$ ,  $B = [\text{Same age}]$ , and  $C = [\text{Different sexes}]$ . Give the compositions of the events

$$C, \bar{A}, A \cup B, AB, BC$$

#### SOLUTION

The pairs consisting of different sexes are  $\{1, 3\}$ ,  $\{1, 4\}$ ,  $\{2, 3\}$ , and  $\{2, 4\}$ , so  $C = \{e_2, e_3, e_4, e_5\}$ . The event  $\bar{A}$  is the same as the event  $C$ . Employing the definitions of union and intersection, we obtain

$$\begin{aligned} A \cup B &= \{e_1, e_2, e_3, e_6\} \\ AB &= \{e_6\} \\ BC &= \{e_2, e_3\} \end{aligned}$$

Let us now examine how probabilities behave as the operations of complementation, union, and intersection are applied to events. It would be worthwhile for the reader to review the properties of probability listed in Section 2. In particular, recall that  $P(A)$  is the sum of probabilities of the elementary outcomes that are in  $A$ , and  $P(S) = 1$ .

First, let us examine how  $P(\bar{A})$  is related to  $P(A)$ . The sum  $P(A) + P(\bar{A})$  is the sum of the probabilities of all elementary outcomes that are in  $A$  plus the sum of the probabilities of elementary outcomes not in  $A$ . Together, these two sets comprise  $S$  and we must have  $P(S) = 1$ . Consequently,  $P(A) + P(\bar{A}) = 1$ , and we arrive at the following law.

#### Law of Complement

$$P(A) = 1 - P(\bar{A})$$

This law or formula is useful in calculating  $P(A)$  when  $\bar{A}$  is of a simpler form than  $A$  so that  $P(\bar{A})$  is easier to calculate.

Turning to the operation of union, recall that  $A \cup B$  is composed of points (or elementary outcomes) that are in  $A$ , in  $B$ , or in both  $A$  and  $B$ . Consequently,  $P(A \cup B)$  is the sum of the probabilities assigned to these elementary outcomes, each probability taken *just once*. Now, the sum  $P(A) + P(B)$  includes contributions from all these points, but it double counts those in the region  $AB$  (see the figure of  $A \cup B$ ). To adjust for this double counting, we must therefore subtract  $P(AB)$  from  $P(A) + P(B)$ . This results in the following law.

**Addition Law**

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

If the events  $A$  and  $B$  are incompatible, their intersection  $AB$  is empty, so  $P(AB) = 0$ , and we obtain

**Special Addition Law for Incompatible Events**

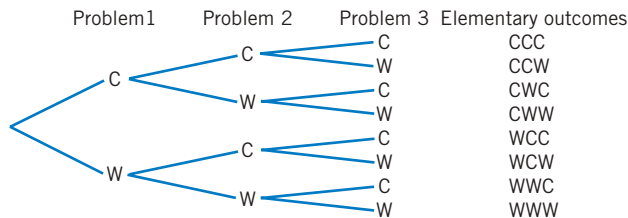
$$P(A \cup B) = P(A) + P(B)$$

The addition law expresses the probability of a larger event  $A \cup B$  in terms of the probabilities of the smaller events  $A$ ,  $B$ , and  $AB$ . Some applications of these two laws are given in the following examples.

**Example 8** Using the Law of Complement for Probability

A child is presented with three word-association problems. With each problem, two answers are suggested—one is correct and the other wrong. If the child has no understanding of the words whatsoever and answers the problems by guessing, what is the probability of getting at least one correct answer?

**SOLUTION** Let us denote a correct answer by  $C$  and a wrong answer by  $W$ . The elementary outcomes can be conveniently enumerated by means of a tree diagram.



There are 8 elementary outcomes in the sample space and, because they are equally likely, each has the probability  $\frac{1}{8}$ . Let  $A$  denote the event of getting at least one correct answer. Scanning our list, we see that  $A$  contains 7 elementary outcomes, all except WWW. Our direct calculation yields  $P(A) = \frac{7}{8}$ .

Now let us see how this probability calculation could be considerably simplified. First, making a complete list of the sample space is not necessary. Since the elementary outcomes are equally likely, we need only determine that there are a total of 8 elements in  $S$ . How can we obtain this count without making a list? Note that an outcome is represented by three letters. There are 2 choices for each letter—namely, C or W. We then have  $2 \times 2 \times 2 = 8$  ways of filling the three slots. The tree diagram explains this multiplication rule of counting. Evidently, the event  $A$  contains many elementary outcomes. On the other hand,  $\bar{A}$  is the event of getting all answers wrong. It consists of the single elementary outcome WWW, so  $P(\bar{A}) = \frac{1}{8}$ . According to the law of complement,

$$\begin{aligned} P(A) &= 1 - P(\bar{A}) \\ &= 1 - \frac{1}{8} = \frac{7}{8} \end{aligned}$$

### Example 9 Using the Addition Law for Probability

Refer to Example 6 where two puppies are selected from four by lottery. What is the probability that the selected puppies are either of the same sex or the same age?

**SOLUTION** In Example 6, we already enumerated the six elementary outcomes that comprise the sample space. The lottery selection makes all choices equally likely and the uniform probability model applies. The two events of interest are

$$\begin{aligned} A &= [\text{Same sex}] = \{e_1, e_6\} \\ B &= [\text{Same age}] = \{e_2, e_3, e_6\} \end{aligned}$$

Because  $A$  consists of two elementary outcomes and  $B$  consists of three,

$$P(A) = \frac{2}{6} \quad \text{and} \quad P(B) = \frac{3}{6}$$

Here we are to calculate  $P(A \cup B)$ . To employ the addition law, we also need to calculate  $P(AB)$ . In Figure 3, we see  $AB = \{e_6\}$ , so  $P(AB) = \frac{1}{6}$ . Therefore,

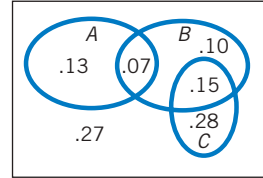
$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(AB) \\ &= \frac{2}{6} + \frac{3}{6} - \frac{1}{6} = \frac{4}{6} = \frac{2}{3} \end{aligned}$$

which is confirmed by the observation that  $A \cup B = \{e_1, e_2, e_3, e_6\}$  indeed has four outcomes.

**Example 10** Determining Probabilities from Those Given in a Venn Diagram

The accompanying Venn diagram shows three events  $A$ ,  $B$ , and  $C$  and also the probabilities of the various intersections. [For instance,  $P(AB) = .07$ ,  $P(A\bar{B}) = .13$ .] Determine:

- $P(A)$
- $P(B\bar{C})$
- $P(A \cup B)$



**SOLUTION** To calculate a probability, first identify the set in the Venn diagram. Then add the probabilities of those intersections that together comprise the stated event. We obtain

- $P(A) = .13 + .07 = .20$
- $P(B\bar{C}) = .10 + .07 = .17$
- $P(A \cup B) = .13 + .07 + .10 + .15 = .45$

**Example 11** Expressing Relations between Events in Set Notation

Refer to Example 10. Express the following events in set notation and find their probabilities.

- Both  $B$  and  $C$  occur.
- $C$  occurs and  $B$  does not.
- Exactly one of the three events occurs.

**SOLUTION** The stated events and their probabilities are

- $BC$   $P(BC) = .15$
- $\bar{B}C$   $P(\bar{B}C) = .28$
- $(A\bar{B}\bar{C}) \cup (\bar{A}B\bar{C}) \cup (\bar{A}\bar{B}C)$

$$\text{The probability} = .13 + .10 + .28 = .51$$

**Exercises**

- 4.35 A day of the week will be selected to hold an all-day club picnic. The sample space has seven elementary outcomes  $e_1, e_2, \dots, e_7$  where  $e_1$  represents Sunday,  $e_2$  Monday, and so on. Two events are given as  $A = \{e_4, e_5, e_6, e_7\}$  and  $B = \{e_1, e_6, e_7\}$ .
- Draw a Venn diagram and show the events  $A$  and  $B$ .
  - Determine the composition of the following events: (i)  $AB$  (ii)  $\bar{B}$  (iii)  $A\bar{B}$  (iv)  $A \cup B$ .
- 4.36 A sample space consists of 8 elementary outcomes with the following probabilities.
- $$P(e_1) = .08 \quad P(e_2) = P(e_3) = P(e_4) = .12$$
- $$P(e_5) = P(e_6) = P(e_7) = P(e_8) = .14$$

Three events are given as

$$A = \{e_1, e_2, e_5, e_6, e_7\},$$

$$B = \{e_2, e_3, e_6, e_7\}, \text{ and } C = \{e_6, e_8\}.$$

- (a) Draw a Venn diagram and show these events.
- (b) Give the composition and determine the probability of (i)  $\bar{B}$  (ii)  $BC$  (iii)  $A \cup C$  (iv)  $\bar{A} \cup C$ .
- 4.37 Refer to Exercise 4.36. Corresponding to each verbal description given here, write the event in set notation, give its composition, and find its probability.
- (a)  $C$  does not occur.
- (b) Both  $A$  and  $B$  occur.
- (c)  $A$  occurs and  $B$  does not occur.
- (d) Neither  $A$  nor  $C$  occurs.
- 4.38 Suppose you have had interviews for summer jobs at a grocery store, a discount store, and a movie theater. Let  $G$ ,  $D$ , and  $M$  denote the events of your getting an offer from the grocery store, the discount store, and the movie theater, respectively. Express the following events in set notation.
- (a) You get offers from the discount store and the movie theater.
- (b) You get offers from the discount store and the movie theater but fail to get an offer from the grocery store.
- (c) You do not get offers from the grocery store and the movie theater.
- 4.39 Four applicants will be interviewed for an administrative position with an environmental lobby. They have the following characteristics.
1. Psychology major, male, GPA 3.5
  2. Chemistry major, female, GPA 3.3
  3. Journalism major, female, GPA 3.7
  4. Mathematics major, male, GPA 3.8
- One of the candidates will be hired.
- (a) Draw a Venn diagram and exhibit these events:
- $A$ : A social science major is hired.  
 $B$ : The GPA of the selected candidate is higher than 3.6.  
 $C$ : A male candidate is hired.
- (b) Give the composition of the events  $A \cup B$  and  $AB$ .
- 4.40 For the experiment of Exercise 4.39, give a verbal description of each of the following events and also state the composition of the event.
- (a)  $\bar{C}$   
 (b)  $C\bar{A}$   
 (c)  $A \cup \bar{C}$
- 4.41 A sample space consists of 9 elementary outcomes  $e_1, e_2, \dots, e_9$  whose probabilities are
- $$P(e_1) = P(e_2) = .04 \quad P(e_3) = P(e_4) = P(e_5) = .2$$
- $$P(e_6) = P(e_7) = .1 \quad P(e_8) = P(e_9) = .06$$
- Suppose  $A = \{e_1, e_5, e_8\}$ ,  $B = \{e_2, e_5, e_8, e_9\}$ .
- (a) Calculate  $P(A)$ ,  $P(B)$ , and  $P(AB)$ .
- (b) Using the addition law of probability, calculate  $P(A \cup B)$ .
- (c) List the composition of the event  $A \cup B$  and calculate  $P(A \cup B)$  by adding the probabilities of the elementary outcomes.
- (d) Calculate  $P(\bar{B})$  from  $P(B)$  and also by listing the composition of  $\bar{B}$ .
- 4.42 Refer to Exercise 4.35. Suppose the elementary outcomes are assigned these probabilities.
- $$P(e_1) = P(e_2) = P(e_3) = .15 \quad P(e_4) = P(e_5) = .06$$
- $$P(e_6) = .2 \quad P(e_7) = .23$$
- (a) Find  $P(A)$ ,  $P(B)$ , and  $P(AB)$ .
- (b) Employing the laws of probability and the results of part (a), calculate  $P(\bar{A})$  and  $P(A \cup B)$ .
- (c) Verify your answers to part (b) by adding the probabilities of the elementary outcomes in each of  $\bar{A}$  and  $A \cup B$ .
- 4.43 Consider the two events.
- $$A = [\text{Obese}] \quad B = [\text{Male}]$$
- for persons in the age group 20–39 years old. A survey taken in early 2008 by the National Center for Health Statistics, suggests the probabilities

$P(A) = .25$      $P(B) = .50$      $P(AB) = .12$     4.47  
for a randomly selected person.

- (a) Enter these probabilities in the following table.
- (b) Determine the probabilities of  $A\bar{B}$ ,  $\bar{A}B$ , and  $\bar{A}\bar{B}$  and fill in the table.

	$B$	$\bar{B}$
$A$		
$\bar{A}$		

- 4.44 Refer to Exercise 4.43. Express the following events in set notation and find their probabilities.
- (a)  $B$  occurs and  $A$  does not occur.
  - (b) Neither  $A$  nor  $B$  occurs.
  - (c) Either  $A$  occurs or  $B$  does not occur.
- 4.45 Consider the two events.

$A =$  [5 or more alcoholic drinks in one day last year]

$B =$  [Female]

for persons in the age group 18–24 years old. A survey taken in early 2008 by the National Center for Health Statistics, suggests the probabilities

$P(B) = .50$      $P(\bar{A}\bar{B}) = .23$      $P(AB) = .14$

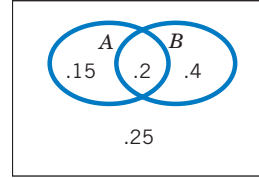
for a randomly selected person. The following table shows the probabilities concerning  $A$  and  $B$ .

	$B$	$\bar{B}$
$A$	.14	.23
$\bar{A}$		
	.50	

- (a) Determine the missing entries.
  - (b) What is the probability that  $A$  does not occur and  $B$  does occur?
  - (c) Find the probability that either  $A$  or  $B$  occurs.
  - (d) Find the probability that one of these events occurs and the other does not.
- 4.46 If  $P(A) = .2$  and  $P(B) = .9$ , can  $A$  and  $B$  be mutually exclusive? Why or why not?

From the probabilities shown in this Venn diagram, determine the probabilities of the following events.

- (a)  $A$  does not occur.
- (b)  $A$  occurs and  $B$  does not occur.
- (c) Exactly one of the events  $A$  and  $B$  occurs.



- 4.48 In a class of 32 seniors and graduate students, 20 are men and 12 are graduate students of whom 8 are women. If a student is randomly selected from this class, what is the probability that the selected student is (a) a senior? (b) a male graduate student?
- 4.49 Of 18 fast food restaurants in a city, 7 are in violation of sanitary standards, 8 are in violation of safety standards, and 4 are in violation of both. If a fast food restaurant is chosen at random, what is the probability that it is in compliance with both safety and sanitary standards?
- 4.50 Given that the probability that  $A$  occurs is .3, the probability that  $B$  does not occur is .6, and the probability that either  $A$  or  $B$  occurs is .5, find:
- (a) The probability that  $A$  does not occur.
  - (b) The probability that both  $A$  and  $B$  occur.
  - (c) The probability that  $A$  occurs and  $B$  does not occur.
- 4.51 The medical records of the male diabetic patients reporting to a clinic during one year provide the following percentages.

Age of Patient	Light Case		Serious Case	
	Diabetes in Parents		Diabetes in Parents	
	Yes	No	Yes	No
Below 40	15	10	8	2
Above 40	15	20	20	10

Suppose a patient is chosen at random from this group, and the events  $A$ ,  $B$ , and  $C$  are defined as follows.

- A: He has a serious case.  
 B: He is below 40.  
 C: His parents are diabetic.

- (a) Find the probabilities  $P(A)$ ,  $P(B)$ ,  $P(BC)$ ,  $P(ABC)$ .  
 (b) Describe the following events verbally and find their probabilities: (i)  $\overline{AB}$  (ii)  $A \cup \overline{C}$  (iii)  $\overline{ABC}$ .

- 4.52 The following frequency table shows the classification of 58 landfills in a state according to their concentration of the three hazardous chemicals arsenic, barium, and mercury.

	Barium			
	High		Low	
Arsenic	Mercury		Mercury	
	High	Low	High	Low
High	1	3	5	9
Low	4	8	10	18

If a landfill is selected at random, find the probability that it has:

- (a) A high concentration of barium.  
 (b) A high concentration of mercury and low concentrations of both arsenic and barium.  
 (c) High concentrations of any two of the chemicals and low concentration of the third.  
 (d) A high concentration of any one of the chemicals and low concentrations of the other two.

- 4.53 A bank rewards its employees by giving awards to any employee who is cited by a customer for giving special service. Each award consists of two gift certificates contained in a sealed envelope. Each envelope contains certificates for one of the five following combinations of items.

- Dinner and box of candy.
- Round of golf and flowers.
- Lunch and flowers.
- Box of candy and lunch.
- Music CD and lunch.

- (a) An employee, cited twice for service, first selects one envelope from a collection of five and then the second from the full collection of five choices. List the sample space and assign probabilities to the simple events.

- (b) State the compositions of the events

$$A = \{\text{The employee gets flowers}\}$$

$$B = \{\text{The employee gets lunch}\}$$

$$AB = \{\text{The employee gets flowers and lunch}\}$$

and give their probabilities.

- 4.54 Refer to Exercise 4.53. Let  $C$  denote the event that the employee gets either lunch or flowers or both.

- (a) Relate  $C$  to the events  $A$  and  $B$ , and calculate  $P(C)$  using a law of probability.

- (b) State the composition of  $C$  and calculate its probability by adding the probabilities of the simple events.

## 5. CONDITIONAL PROBABILITY AND INDEPENDENCE

The probability of an event  $A$  must often be modified after information is obtained as to whether or not a related event  $B$  has taken place. Information about some aspect of the experimental results may therefore necessitate a revision of the probability of an event concerning some other aspect of the results. The revised probability of  $A$  when it is known that  $B$  has occurred is called the **conditional probability of  $A$  given  $B$**  and is denoted by  $P(A|B)$ . To illustrate how such modification is made, we consider an example that will lead us to the formula for conditional probability.



### Example 12 Conditional Probability of Using Alternative Medicine Given Body Weight

Complementary alternative medicine (CAM), including acupuncture, yoga, and massage has become more popular. By combining information in two tables,<sup>2</sup> we obtain information concerning use of CAM in the past year and weight class based on body mass index. The proportions in the various categories appear in Table 1.

**TABLE 1** Body Weight and Complementary and Alternative Medicine

	Under-weight	Healthy weight	Over-weight	Obese	Total
CAM	.01	.13	.12	.12	.32
No CAM	.02	.19	.21	.20	.62
Total	.03	.32	.33	.32	1.00

- What is the probability that a person selected at random from this population will have used complementary and alternative medicine in the past year?
- A person selected at random is found to be overweight. What is the probability that this person used complementary and alternative medicine in the past year?

**SOLUTION** Let  $A$  denote the event that a person used CAM, and let  $B$  denote the event that a person is overweight.

- Because 32% of the people used CAM and the individual is selected at random, we conclude that  $P(A) = .32$ . This is the unconditional probability of  $A$ .
- When we are given the information that the selected person is overweight, the categories in the first, second, and fourth columns of Table 1 are not relevant to this person. The third column shows that among the subgroup of overweight persons, the proportion using CAM is  $.12/.33$ . Therefore, given the information that the person is in this subgroup, the probability that he or she used CAM

$$P(A|B) = \frac{.12}{.33} = .364$$

Noting that  $P(AB) = .12$  and  $P(B) = .33$ , we have derived  $P(A|B)$  by taking the ratio  $P(AB)/P(B)$ . In other words,  $P(A|B)$  is the proportion of the population having the characteristic  $A$  among all those having the characteristic  $B$ .

<sup>2</sup>Statistical Abstract of the United States, Table 203 (2009) and National Health Statistics Reports, Table 7 12 (December 10, 2008).

The **conditional probability** of  $A$  given  $B$  is denoted by  $P(A|B)$  and defined by the formula

$$P(A|B) = \frac{P(AB)}{P(B)}$$

Equivalently, this formula can be written

$$P(AB) = P(B)P(A|B)$$

This latter version is called the **multiplication law of probability**.

Similarly, the conditional probability of  $B$  given  $A$  can be expressed

$$P(B|A) = \frac{P(AB)}{P(A)}$$

which gives the relation  $P(AB) = P(A)P(B|A)$ . Thus, the multiplication law of probability states that the conditional probability of an event multiplied by the probability of the conditioning event gives the probability of the intersection.

The multiplication law can be used in one of two ways, depending on convenience. When it is easy to compute  $P(A)$  and  $P(AB)$  directly, these values can be used to compute  $P(A|B)$ , as in Example 12. On the other hand, if it is easy to calculate  $P(B)$  and  $P(A|B)$  directly, these values can be used to compute  $P(AB)$ .

### Example 13 Conditional Probability of Survival

Refer to the box “How Long Will a Baby Live?” in Section 4.3. It shows the probabilities of death within 10-year age groups.

- What is the probability that a newborn child will survive beyond age 90?
- What is the probability that a person who has just turned 80 will survive beyond age 90?

#### SOLUTION

- Let  $A$  denote the event “Survive beyond 90.” Adding the probabilities of death in the age groups 90–100 and beyond, we find

$$P(A) = .188 + .021 = .209$$

- Letting  $B$  denote the event “Survive beyond 80,” we see that the required probability is the conditional probability  $P(A|B)$ . Because  $AB = A$ ,  $P(A) = .209$ , and

$$P(B) = .318 + .188 + .021 = .527$$

we obtain

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{.209}{.527} = .397$$

**Example 14** Using the Multiplication Law of Probability

There are 25 pens in a container on your desk. Among them, 20 will write well but 5 have defective ink cartridges. You will select 2 pens to take to a business appointment. Calculate the probability that:

- Both pens are defective.
- One pen is defective but the other will write well.

**SOLUTION** We will use the symbols  $D$  for “defective” and  $G$  for “writes well”, and attach subscripts to identify the order of the selection. For instance,  $G_1D_2$  will represent the event that the first pen checked will write well and the second is defective.

- Here the problem is to calculate  $P(D_1D_2)$ . Evidently,  $D_1D_2$  is the intersection of the two events  $D_1$  and  $D_2$ . Using the multiplication law, we write

$$P(D_1D_2) = P(D_1)P(D_2|D_1)$$

In order to calculate  $P(D_1)$ , we need only consider selecting one pen at random from 20 good and 5 defective pens. Clearly,  $P(D_1) = \frac{5}{25}$ . The next step is to calculate  $P(D_2|D_1)$ . Given that  $D_1$  has occurred, there will remain 20 good and 4 defective pens at the time the second selection is made. Therefore, the conditional probability of  $D_2$  given  $D_1$  is  $P(D_2|D_1) = \frac{4}{24}$ . Multiplying these two probabilities, we get

$$P(\text{both defective}) = P(D_1D_2) = \frac{5}{25} \times \frac{4}{24} = \frac{1}{30} = .033$$

- The event [exactly one defective] is the union of the two incompatible events  $G_1D_2$  and  $D_1G_2$ . The probability of each of these can be calculated by the multiplication law as in part (a). Specifically,

$$P(G_1D_2) = P(G_1)P(D_2|G_1) = \frac{20}{25} \times \frac{5}{24} = \frac{1}{6}$$

$$P(D_1G_2) = P(D_1)P(G_2|D_1) = \frac{5}{25} \times \frac{20}{24} = \frac{1}{6}$$

The required probability is  $P(G_1D_2) + P(D_1G_2) = \frac{2}{6} = .333$ .

**Remark:** In solving the problems of Example 14, we have avoided listing the sample space corresponding to the selection of two pens from a collection of 25. A judicious use of the multiplication law has made it possible to focus attention on one draw at a time, thus simplifying the probability calculations.

A situation that merits special attention occurs when the conditional probability  $P(A|B)$  turns out to be the same as the unconditional probability  $P(A)$ .

Information about the occurrence of  $B$  then has no bearing on the assessment of the probability of  $A$ . Therefore, when we have the equality  $P(A|B) = P(A)$ , we say the events  $A$  and  $B$  are independent.

Two events  $A$  and  $B$  are **independent** if

$$P(A|B) = P(A)$$

Equivalent conditions are

$$P(B|A) = P(B)$$

or

$$P(AB) = P(A)P(B)$$

The last form follows by recalling that  $P(A|B) = P(AB)/P(B)$ , so that the condition  $P(A|B) = P(A)$  is equivalent to

$$P(AB) = P(A)P(B)$$

which may be used as an alternative definition of independence. The other equivalent form is obtained from

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B)$$

The form  $P(AB) = P(A)P(B)$  shows that the definition of independence is symmetric in  $A$  and  $B$ .

### Example 15 Demonstrating Dependence between Use of CAM and Overweight

Are the two events  $A = [\text{used CAM}]$  and  $B = [\text{Overweight}]$  independent for the population in Example 12?

**SOLUTION** Referring to that example, we have

$$P(A|B) = \frac{P(A)}{P(AB)} = \frac{.32}{.12} = .33 \quad \text{and} \quad P(A) = .364$$

Because these two probabilities are different, the two events  $A$  and  $B$  are dependent.

**Caution:** Do not confuse the terms “incompatible events” and “independent events.” We say  $A$  and  $B$  are incompatible when their intersection  $AB$  is empty, so  $P(AB) = 0$ . On the other hand, if  $A$  and  $B$  are independent,  $P(AB) = P(A)P(B)$ . Both these properties cannot hold as long as  $A$  and  $B$  have nonzero probabilities.

We introduced the condition of independence in the context of checking a given assignment of probability to see if  $P(A|B) = P(A)$ . A second use of this condition is in the assignment of probability when the experiment consists of two physically unrelated parts. When events  $A$  and  $B$  refer to unrelated parts of an experiment,  $AB$  is assigned the probability  $P(AB) = P(A)P(B)$ .

### Example 16 Using Independence to Assign Probability

Engineers use the term “reliability” as an alternative name for the probability that a device does not fail. Suppose a mechanical system consists of two components that function independently. From extensive testing, it is known that component 1 has reliability .98 and component 2 has reliability .95. If the system can function only if both components function, what is the reliability of the system?

**SOLUTION** Consider the events

$A_1$ : Component 1 functions

$A_2$ : Component 2 functions

$S$ : System functions

Here we have the event relation  $S = A_1A_2$ . Given that the components operate independently, we take the events  $A_1$  and  $A_2$  to be independent. Consequently, the multiplication law assigns

$$P(S) = P(A_1)P(A_2) = .98 \times .95 = .931$$

and the system reliability is .931.

In this system, the components are said to be connected in series, and the system is called a series system. A two-battery flashlight is an example. The conventional diagram for a series system is shown in the illustration:



### Example 17 Independence and Assigning Probabilities When Sampling with Replacement

In the context of Example 14, suppose that a box contains 25 cards identifying the pens and their ability to write. One card is drawn at random. It is returned to the box and then another card is drawn at random. What is the probability that both draws produce pens that will not write?

**SOLUTION** As before, we will use the letter  $D$  for defective and  $G$  for a pen that will write. By returning the first card to the box, the contents of the box remain unchanged. Hence, with each draw,  $P(D) = \frac{5}{25}$ , and the results of the two draws are independent. Instead of working with conditional probability as we did in Example 11, we can use the property of independence to calculate

$$P(D_1D_2) = P(D_1)P(D_2) = \frac{5}{25} \times \frac{5}{25} = .04$$

**Remark 1:** Evidently, this method of probability calculation extends to any number of draws if after each draw the selected card is returned to the box. For instance, the probability that the first draw produces a  $D$  and the next two draws produce  $G$ 's is

$$P(D_1G_2G_3) = \frac{5}{25} \times \frac{20}{25} \times \frac{20}{25} = .128$$

**Remark 2:** Sampling with replacement is seldom used in practice, but it serves as a conceptual frame for simple probability calculations when a problem concerns sampling from a large population. For example, consider drawing 3 cards from a box containing 2500 cards, of which 2000 are  $G$ 's and 500  $D$ 's. Whether or not a selected card is returned to the box before the next draw makes little difference in the probabilities. The model of independence serves as a reasonable approximation.

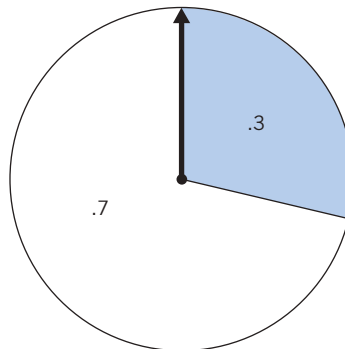
The connection between dependent trials and the size of the population merits further emphasis.

### Example 18 Dependence and Sampling without Replacement

If the outcome of a single trial of any experiment is restricted to just two possible outcomes, it can be modeled as drawing a single ball from an urn containing only red ( $R$ ) and white ( $W$ ) balls. In the previous example, these two possible outcomes were good and defective. Consider taking a sample of size 3, without replacement, from each of two populations:

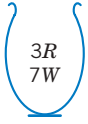
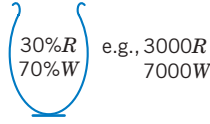
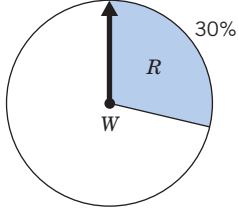
1. Small population where the urn contains 7  $W$  and 3  $R$ .
2. Large population where the urn contains 7000  $W$  and 3000  $R$ .

Compare with a sample of size 3 generated from a spinner having a probability .7 of white, where  $R$  or  $W$  is determined by a separate spin for each trial.



- (a) Calculate the probability assigned to each possible sample.
- (b) Let  $D =$  [At least one  $W$ ]. Calculate the probability of  $D$ .

**TABLE 2** A Comparison of Finite Populations and the Spinner Model

Draw 3 balls without replacement			
	Small population	Large population	Spinner
			
	$P(ABC) = P(A)P(B A)P(C AB)$	$P(ABC) \approx P(A)P(B)P(C)$	$P(ABC) = P(A)P(B)P(C)$
Outcome	Not independent	Approximately independent	Independent
<i>RRR</i>	$\frac{3}{10} \times \frac{2}{9} \times \frac{1}{8} = \frac{6}{720}$	$\frac{3000}{10,000} \times \frac{2999}{9999} \times \frac{2998}{9998} \approx (.3)(.3)(.3)$	$(.3)(.3)(.3)$
<i>RRW</i>	$\frac{3}{10} \times \frac{2}{9} \times \frac{7}{8} = \frac{42}{720}$	$\approx (.3)(.3)(.7)$	$(.3)^2(.7)$
<i>RWR</i>	$\frac{3}{10} \times \frac{7}{9} \times \frac{2}{8} = \frac{42}{720}$	$\approx (.3)(.7)(.3)$	$(.3)^2(.7)$
<i>WRR</i>	$\frac{7}{10} \times \frac{3}{9} \times \frac{2}{8} = \frac{42}{720}$	$\approx (.7)(.3)(.3)$	$(.3)^2(.7)$
<i>RWW</i>	$\frac{3}{10} \times \frac{7}{9} \times \frac{6}{8} = \frac{126}{720}$	$\approx (.3)(.7)(.7)$	$(.3)(.7)^2$
<i>WRW</i>	$\frac{7}{10} \times \frac{3}{9} \times \frac{6}{8} = \frac{126}{720}$	$\approx (.7)(.3)(.7)$	$(.3)(.7)^2$
<i>WWR</i>	$\frac{7}{10} \times \frac{6}{9} \times \frac{3}{8} = \frac{126}{720}$	$\approx (.7)(.7)(.3)$	$(.3)(.7)^2$
<i>WWW</i>	$\frac{7}{10} \times \frac{6}{9} \times \frac{5}{8} = \frac{210}{720}$	$\approx (.7)(.7)(.7)$	$(.7)^3$
$P(D) = 1 - P(\bar{D})$	$= 1 - \frac{6}{720}$	$\approx 1 - (.3)^3$	$= 1 - (.3)^3$
If <i>A</i> = 1 <sup>st</sup> is <i>R</i> , then	<i>B</i> = 2 <sup>nd</sup> is <i>R</i> $ABC = \{RRR\} = \bar{D}$	<i>C</i> = 3 <sup>rd</sup> is <i>R</i> $D = \{RRW, RWR, WRR, RWW, WRW, WWR, WWW\}$	<i>D</i> = at least one <i>W</i>

**SOLUTION**

(a) We will write *RWR* for the outcome where the first and third draws are *R* and the second is *W*. Applying the general multiplication rule  $P(ABC) = P(A)P(B|A)P(C|AB)$ , when sampling the small population, we get

$$P(RWR) = P(R)P(W|R)P(R|RW) = \frac{3}{10} \times \frac{7}{9} \times \frac{2}{8} = \frac{42}{720}$$

For the larger population,

$$P(RWR) = \frac{3000}{10,000} \times \frac{7000}{9999} \times \frac{2999}{9998} \approx (.3) \times (.7) \times (.3) = (.3)^2(.7)$$

When the population size is large, the assumption of independence produces a very good approximation.

Under the spinner model, the probability of  $R$  is .3 for the first trial and this probability is the same for all trials. A spinner is a classic representation of a device with no memory, so that the outcome of the current trial is independent of the outcomes of all the previous trials. According to the product rule for independence, we assign

$$P(RWR) = (.3) \times (.7) \times (.3)$$

Notice that the spinner model is equivalent to sampling with replacement from either of the two finite populations.

The results for all eight possible samples are shown in Table 2.

- (b) The event  $D$  is complicated, whereas  $\bar{D} = \{RRR\}$ , a single outcome. By the law of the complement,

$$P(D) = 1 - P(\bar{D}) = 1 - \frac{3}{10} \times \frac{2}{9} \times \frac{1}{8} = 1 - \frac{6}{720}$$

In the second case,  $P(D)$  is approximately  $1 - (.3) \times (.3) \times (.3)$  and this answer is exact for the spinner model.

Table 2 summarizes sampling from a small finite population, a large but finite population, and the spinner model. Dependence does matter when sampling without replacement from a small population.

## 6. BAYES' THEOREM

We first show how the multiplication rule of probability leads to a result called the rule of total probability. An event  $A$  can occur either when an event  $B$  occurs or when it does not occur. That is,  $A$  can be written as the disjoint union of  $AB$  and  $A\bar{B}$ . Consequently,

$$P(A) = P(AB) + P(A\bar{B})$$

Using the multiplication rule of probability, we obtain the **rule of total probability**.

### Rule of Total Probability

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$$



**Example 19** Rule of Total Probability and Uncertainty with Medical Tests

Let  $A$  be the event that a person tests positive for a serious virus and  $B$  be the event that the person actually has the virus. Suppose that the virus is present in 1.4% of the population. Because medical tests are sometimes incorrect, we model uncertainty by assigning probability. Suppose the conditional probability that the test is positive, given that the person has the virus, is  $.995 = P(A|B)$ . Also, suppose that  $.01 = P(A|\bar{B})$  is the conditional probability that a person not having the virus tests positive; a false positive.

Determine the probability that a person will test positive,  $P(A)$ .

**SOLUTION** We are given  $P(B) = .014$  so  $1 - P(B) = P(\bar{B}) = .986$ . Then

$$\begin{aligned} P(A) &= P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) \\ &= .995 \times .014 + .01 \times .986 = .024 \end{aligned}$$

The same reasoning prevails if there are three events  $B_1$ ,  $B_2$ , and  $B_3$  that are mutually exclusive and whose union is the whole sample space as illustrated in Figure 4. Then,  $A$  is the union of the mutually exclusive events  $AB_1$ ,  $AB_2$ , and  $AB_3$  and the rule of total probability becomes

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)$$

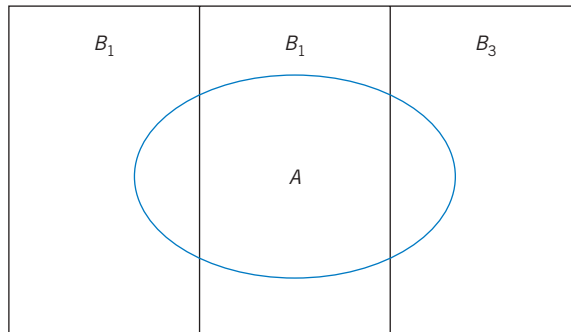


Figure 4 Event  $A$  and mutually exclusive events  $B_1$ ,  $B_2$ , and  $B_3$  with  $B_1 \cup B_2 \cup B_3 = S$ .

Suppose the two events  $A$  and  $B$  can occur together and, before observing either, we know the probability  $P(B)$  so  $P(\bar{B}) = 1 - P(B)$  is also known. We call these two probabilities the **prior probabilities** since they represent the probabilities associated with  $B$  and  $\bar{B}$  before we know the status of event  $A$  or any other event. When we also know the two conditional probabilities  $P(A|B)$  and  $P(A|\bar{B})$ , the probability of  $B$  can be updated when we observe the status of  $A$ .

Once we know  $A$  has occurred, the updated or **posterior probability** of  $B$  is given by the conditional probability

$$P(B|A) = \frac{P(AB)}{P(A)}$$

The numerator can be written as  $P(A|B)P(B)$  by the multiplication rule and the denominator  $P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$  by the rule of total probability. Substituting these two alternate expressions into the formula for conditional probability, we obtain **Bayes' Theorem**.

### Bayes' Theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})}$$

The posterior probability of  $\bar{B}$  is then  $P(\bar{B}|A) = 1 - P(B|A)$

### Example 20 Bayes' Theorem and the Uncertainty with Medical Tests

Refer to Example 19 where  $A$  is the event that a person tests positive for a serious virus and  $B$  is the event that the person actually has the virus.

Suppose a person tests positive. Use Bayes' Theorem to update the probability that the person has the virus. That is, determine the posterior probability  $P(B|A)$

**SOLUTION** From the previous example, we have  $P(A|B) = .995$ ,  $P(A|\bar{B}) = .01$  and  $P(B) = .014$  so  $P(\bar{B}) = .986$ . By Bayes' Theorem, the posterior probability of having the virus is

$$\begin{aligned} P(B|A) &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})} \\ &= \frac{.995 \times .014}{.995 \times .014 + .01 \times .986} = .586 \end{aligned}$$

The probability of having the virus has increased dramatically from .014 to .586 but it is still far below 1.

When there are three events  $B_1$ ,  $B_2$ , and  $B_3$  that are mutually exclusive and whose union is the whole sample space, as in Figure 4, Bayes' Theorem becomes

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)}$$

### Exercises

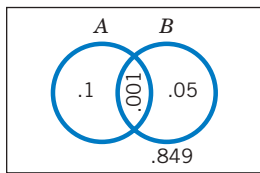
- 4.55 A person is randomly selected from persons working in your state. Consider the two events  
 $A =$  [Earned over \$60,000 last year]  
 $B =$  [College graduate]
- Given that the person is a college graduate, would you expect the probability of  $A$  to be larger, the same, or smaller than the unconditional probability  $P(A)$ ? Explain your answer. Are  $A$  and  $B$  independent according to your reasoning?
- 4.56 A person is randomly selected from persons working in your state. Consider the two events

$A =$  [Lawyer]  
 $B =$  [Driving a new luxury car]

Given that the person selected drives a new luxury car, would you expect the probability of  $A$  to be larger, the same, or smaller than the unconditional probability  $P(A)$ ? Explain your answer. Are  $A$  and  $B$  independent according to your reasoning?

- 4.57 Refer to Exercise 4.43. Find
- (a) the conditional probability that  $B$  occurs given that  $A$  occurs.
  - (b) The conditional probability that  $B$  does not occur given that  $A$  occurs.
  - (c) The conditional probability that  $B$  occurs given that  $A$  does not occur.
- 4.58 Refer to Exercise 4.45. Find
- (a) the conditional probability that  $A$  occurs given that  $B$  occurs.
  - (b) the conditional probability that  $B$  occurs given that  $A$  does not occur.
- 4.59 The following data relate to the proportions in a population of drivers.
- $A =$  Defensive driver training last year  
 $B =$  Accident in current year

The probabilities are given in the accompanying Venn diagram. Find  $P(B|A)$ . Are  $A$  and  $B$  independent?



- 4.60 Suppose  $P(A) = .55$ ,  $P(B) = .32$ , and  $P(\overline{A}B) = .20$ .
- (a) Determine all the probabilities needed to fill in the accompanying table.

	$B$	$\overline{B}$	
$A$			.55
$\overline{A}$	.20		
	.32		

- (b) Find the conditional probability of  $A$  given that  $B$  does not occur.

- 4.61 For two events  $A$  and  $B$ , the following probabilities are given.

$$P(A) = .4 \quad P(B) = .25 \quad P(A|B) = .7$$

Use the appropriate laws of probability to calculate

- (a)  $P(\overline{A})$
  - (b)  $P(AB)$
  - (c)  $P(A \cup B)$
- 4.62 Records of student patients at a dentist's office concerning fear of visiting the dentist suggest the following proportions.

	School		
	Elementary	Middle	High
Fear	.12	.08	.05
Do not fear	.28	.25	.22

For a student selected at random, consider the events

$A =$  [Fear]       $M =$  [Middle school]

- (a) Find the probabilities

$$P(A) \quad P(AM)$$

$$P(M) \quad P(A \cup M)$$

- (b) Are  $A$  and  $M$  independent?

- 4.63 An urn contains two green balls and three red balls. Suppose two balls will be drawn at random one after another and *without replacement* (i.e., the first ball drawn is *not* returned to the urn before the second one is drawn).

- (a) Find the probabilities of the events

$A =$  [Green ball appears in the first draw]  
 $B =$  [Green ball appears in the second draw]

- (b) Are the two events independent? Why or why not?

- 4.64 Refer to Exercise 4.63. Now suppose two balls will be drawn *with replacement* (i.e., the first ball drawn will be returned to the urn before the second draw). Repeat parts (a) and (b).
- 4.65 In a county, men constitute 60% of the labor force. The rates of unemployment are 5.1% and 4.3% among males and females, respectively.
- In the context of selecting a worker at random from the country labor force, state what probabilities the foregoing percentages represent. (Use symbols such as  $M$  for male,  $E$  for employed.)
  - What is the overall rate of unemployment in the county?
  - If a worker selected at random is found to be unemployed, what is the probability that the worker is a woman?
- 4.66 If the probability of running out of gas is .03 and the probability the electronic starting system will not work is .01,
- what is the probability that there will be enough gas and that the starting system will work? Assume the two events are independent.
  - When may independence be a poor assumption?
- 4.67 Suppose  $P(A) = .6$  and  $P(B) = .22$ .
- Determine  $P(A \cup B)$  if  $A$  and  $B$  are independent.
  - Determine  $P(A \cup B)$  if  $A$  and  $B$  are mutually exclusive.
  - Find  $P(A|\bar{B})$  if  $A$  and  $B$  are mutually exclusive.
- 4.68 Refer to Exercise 4.49.
- If a fast food restaurant selected at random is found to comply with safety standards, what is the probability that it violates sanitary standards?
  - If a restaurant selected at random is found to violate at least one of the two standards, what is the probability that it complies with safety standards?
- 4.69 In a shipment of 12 room air conditioners, there are 3 with defective thermostats. Two air conditioners will be selected at random and inspected one after another. Find the probability that:
- The first is defective.
  - The first is defective and the second good.
  - Both are defective.
  - The second air conditioner is defective.
  - Exactly one is defective.
- 4.70 Refer to Exercise 4.69. Now suppose 3 air conditioners will be selected at random and checked one after another. Find the probability that:
- All 3 are good.
  - The first 2 are good and the third defective.
  - Two are good and 1 defective.
- 4.71 Of 20 rats in a cage, 12 are males and 9 are infected with a virus that causes hemorrhagic fever. Of the 12 male rats, 7 are infected with the virus. One rat is randomly selected from the cage.
- If the selected rat is found to be infected, what is the probability that it is a female?
  - If the selected rat is found to be a male, what is the probability that it is infected?
  - Are the events "the selected rat is infected" and "the selected rat is male" independent? Why or why not?
- 4.72 A restaurant critic goes to a place twice. If she has an unsatisfactory experience during both visits, she will go once more. Otherwise she will make only the two visits. Assuming that the results for different visits are independent and that the probability of a satisfactory experience in any one visit is .8
- assign probabilities to each outcome.
  - Find the probability of at least two unsatisfactory visits.
  - Find the conditional probability of at least one satisfactory visit given at least one unsatisfactory visit.

- 4.73 Of three events,  $A$ ,  $B$ , and  $C$ , suppose events  $A$  and  $B$  are independent and events  $B$  and  $C$  are mutually exclusive. Their probabilities are  $P(A) = .7$ ,  $P(B) = .2$ , and  $P(C) = .3$ . Express the following events in set notation and calculate their probabilities.
- Both  $B$  and  $C$  occur.
  - At least one of  $A$  and  $B$  occurs.
  - $B$  does not occur.
  - All three events occur.
- 4.74 Approximately 40% of the Wisconsin population have type O blood. If 4 persons are selected at random to be donors, find  $P$ [at least one type O].
- 4.75 The primary cooling unit in a nuclear power plant has reliability .999. There is also a back-up cooling unit to substitute for the primary unit when it fails. The reliability of the back-up unit is .910. Find the reliability of the cooling system of the power plant. Assume independence.
- 4.76 An accountant screens large batches of bills according to the following sampling inspection plan. She inspects 4 bills chosen at random from each batch and passes the batch if, among the 4, none is irregular. Find the probability that a batch will be passed if, in fact:
- 5% of its bills are irregular.
  - 20% of its bills are irregular.
- 4.77 An electronic scanner is successful in detecting flaws in a material in 80% of the cases. Three material specimens containing flaws will be tested with the scanner. Assume that the tests are independent.
- List the sample space and assign probabilities to the simple events.
  - Find the probability that the scanner is successful in at least two of the three cases.
- 4.78 Refer to Exercise 4.52. Given that a landfill selected at random is found to have a high concentration of mercury, what is the probability that its concentration is:
- High in barium?
  - Low in both arsenic and barium?
  - High in either arsenic or barium?
- 4.79 Of the patients reporting to a clinic with the symptoms of sore throat and fever, 25% have strep throat, 40% have an allergy, and 10% have both.
- What is the probability that a patient selected at random has strep throat, an allergy, or both?
  - Are the events “strep throat” and “allergy” independent?
- \*4.80 Consider tossing two fair coins and the events
- $A$ : Head in the first toss  
 $B$ : Head in the second toss  
 $C$ : Both heads or both tails in the two tosses
- Verify that the property of independence holds for all event pairs.
  - Show that  $P(ABC)$  is different from the product  $P(A)P(B)P(C)$ . (This illustrates the fact that pairwise independence does not ensure complete independence.)
- 4.81 **Imperfect clinical test.** Repeat Example 20 but change  $P(A|B)$  to .96.
- 4.82 Carol and Karl both solve difficult computer problems that come to the student desk. Carol makes 60% of the repairs and Karl 40%. However, Carol’s repairs are incomplete 4% of the time and Karl’s are incomplete 6% of the time.
- Determine the probability that a repair is incomplete.
  - If a repair is found to be incomplete, what is the probability that the repair was made by Karl?

## 7. RANDOM SAMPLING FROM A FINITE POPULATION

In our earlier examples of probability calculations, we have used the phrase “randomly selected” to mean that all possible selections are equally likely. It usually is not difficult to enumerate all the elementary outcomes when both the population size and sample size are small numbers. With larger numbers, making a list of all the possible choices becomes a tedious job. However, a counting rule is available that enables us to solve many probability problems.

We begin with an example where the population size and the sample size are both small numbers so all possible samples can be conveniently listed.

### Example 21 Selecting a Random Sample of Size 2 from a Population of Size 5

There are five qualified applicants for two editorial positions on a college newspaper. Two of these applicants are men and three women. If the positions are filled by randomly selecting two of the five applicants, what is the probability that neither of the men is selected?

**SOLUTION** Suppose the three women applicants are identified as  $a$ ,  $b$ , and  $c$  and the two men as  $d$  and  $e$ . Two members are selected at random from the population:

$$\underbrace{\{a, b, c\}}_{\text{women}} \quad \underbrace{\{d, e\}}_{\text{men}}$$

The possible samples may be listed as

$$\begin{array}{cccc} \{a, b\} & \{b, c\} & \{c, d\} & \{d, e\} \\ \{a, c\} & \{b, d\} & \{c, e\} & \\ \{a, d\} & \{b, e\} & & \\ \{a, e\} & & & \end{array}$$

As the list shows, our sample space has 10 elementary outcomes. The notion of random selection entails that these are all equally likely, so each is assigned the probability  $\frac{1}{10}$ . Let  $A$  represent the event that two women are selected. Scanning our list, we see that  $A$  consists of the three elementary outcomes

$$\{a, b\} \quad \{a, c\} \quad \{b, c\}$$

Consequently,

$$P(A) = \frac{\text{No. of elements in } A}{\text{No. of elements in } S} = \frac{3}{10} = .3$$

Note that our probability calculation in Example 21 only requires knowledge of the two counts: the number of elements in  $S$  and the number of elements in  $A$ . Can we arrive at these counts without formally listing the sample space? An important counting rule comes to our aid. (See Appendix A.2)

### The Rule of Combinations

**Notation:** The number of possible choices of  $r$  objects from a group of  $N$  distinct objects is denoted by  $\binom{N}{r}$ , which reads as “ $N$  choose  $r$ .”

**Formula:**

$$\binom{N}{r} = \frac{N \times (N - 1) \times \cdots \times (N - r + 1)}{r \times (r - 1) \times \cdots \times 2 \times 1}$$

More specifically, the numerator of the formula  $\binom{N}{r}$  is the product of  $r$  consecutive integers starting with  $N$  and proceeding downward. The denominator is also the product of  $r$  consecutive integers, but starting with  $r$  and proceeding down to 1.

To motivate the formula, let us consider the number of possible choices (or collections) of three letters from the seven letters  $\{a, b, c, d, e, f, g\}$ . This count is denoted by  $\binom{7}{3}$ .

It is easier to arrive at a formula for the number of ordered selections. The first choice can be any of the 7 letters, the second can be any of the remaining 6, and the third can be any of the remaining 5. Thinking in terms of a tree diagram (without actually drawing one), we arrive at the following count.

***The number of ordered selections of 3 letters from 7 is given by the product  $7 \times 6 \times 5$ .***

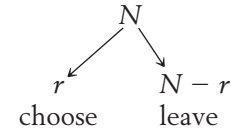
Next, note that a particular collection, say  $\{a, b, c\}$ , can produce  $3 \times 2 \times 1$  orderings, as one can also verify by a tree diagram. The  $\binom{7}{3}$  number of collections, each producing  $3 \times 2 \times 1$  orderings, generate a total of  $\binom{7}{3} \times 3 \times 2 \times 1$  orderings. Because this count must equal  $7 \times 6 \times 5$ , we get

$$\binom{7}{3} = \frac{7 \times 6 \times 5}{3 \times 2 \times 1}$$

This explains the formula of  $\binom{N}{r}$  for the case  $N = 7$  and  $r = 3$ .

Although not immediately apparent, there is a certain symmetry in the counts  $\binom{N}{r}$ . The process of selecting  $r$  objects is the same as choosing  $N - r$

objects to leave behind. Because every choice of  $r$  objects corresponds to a choice of  $N - r$  objects,

$$\binom{N}{r} = \binom{N}{N-r}$$


This relation often simplifies calculations. Since  $\binom{N}{N} = 1$ , we take  $\binom{N}{0} = 1$ .

### Example 22 Evaluating Some Combinations

Calculate the values of  $\binom{5}{2}$ ,  $\binom{15}{4}$ , and  $\binom{15}{11}$ .

**SOLUTION**

$$\binom{5}{2} = \frac{5 \times 4}{2 \times 1} = 10 \quad \binom{15}{4} = \frac{15 \times 14 \times 13 \times 12}{4 \times 3 \times 2 \times 1} = 1365$$

Using the relation  $\binom{N}{r} = \binom{N}{N-r}$ , we have

$$\binom{15}{11} = \binom{15}{4} = 1365$$

### Example 23 Calculating a Probability Using Combinations

Refer to Example 21 concerning a random selection of two persons from a group of two men and three women. Calculate the required probability without listing the sample space.

**SOLUTION** The number of ways two persons can be selected out of five is given by

$$\binom{5}{2} = \frac{5 \times 4}{2 \times 1} = 10$$

Random selection means that the 10 outcomes are equally likely. Next, we are to count the outcomes that are favorable to the event  $A$  that both selected persons are women. Two women can be selected out of three in

$$\binom{3}{2} = \frac{3 \times 2}{2 \times 1} = 3 \text{ ways}$$

Taking the ratio, we obtain the result

$$P(A) = \frac{3}{10} = .3$$



**Example 24** Probabilities of Being Selected under Random Selection

After some initial challenges, there remain 16 potential jurors of which 10 are male and 6 female. The defense attorney can dismiss 4 additional persons on the basis of answers to her questions.

- How many ways can the 4 additional jurors be selected for dismissal?
- How many selections are possible that result in 1 male and 3 females being dismissed?
- If the selection process were random, what is the probability that 1 male and 3 females would be dismissed?

**SOLUTION**

- According to the counting rule  $\binom{N}{r}$ , the number of ways 4 jurors can be selected out of 16 is

$$\binom{16}{4} = \frac{16 \times 15 \times 14 \times 13}{4 \times 3 \times 2 \times 1} = 1820$$

- One male can be chosen from 10 in  $\binom{10}{1} = 10$  ways. Also, 3 females can be chosen from 6 in

$$\binom{6}{3} = \frac{6 \times 5 \times 4}{3 \times 2 \times 1} = 20 \text{ ways}$$

Each of the 10 choices of a male can accompany each of the 20 choices of 3 females. Reasoning from the tree diagram, we conclude that the number of possible samples with the stated composition is

$$\binom{10}{1} \times \binom{6}{3} = 10 \times 20 = 200$$

- Random sampling requires that the 1820 possible samples are all equally likely. Of these, 200 are favorable to the event  $A = [1 \text{ male and } 3 \text{ females}]$ . Consequently,

$$P(A) = \frac{200}{1820} = .110$$

The notion of a random sample from a finite population is crucial to statistical inference. In order to generalize from a sample to the population, it is imperative that the sampling process be impartial. This criterion is evidently met if we allow the selection process to be such that all possible samples are given equal opportunity to be selected. This is precisely the idea behind the term **random sampling**, and a formal definition can be phrased as follows.

A sample of size  $n$  selected from a population of  $N$  distinct objects is said to be a **random sample** if each collection of size  $n$  has the same probability  $1 / \binom{N}{n}$  of being selected.

Note that this is a conceptual rather than an operational definition of a random sample. On the surface, it might seem that a haphazard selection by the experimenter would result in a random sample. Unfortunately, a seemingly haphazard selection may have hidden bias. For instance, when asked to name a random integer between 1 and 9, more persons respond with 7 than any other number. Also, odd numbers are more popular than even numbers. Therefore, the selection of objects must be entrusted to some device that cannot think; in other words, some sort of mechanization of the selection process is needed to make it truly haphazard!

To accomplish the goal of a random selection, one may make a card for each of the  $N$  members of the population, shuffle, and then draw  $n$  cards. This method is easy to understand but awkward to apply to large-size populations. It is best to use random numbers as described in Chapter 1. Random numbers are conveniently generated on a computer (see Chapter 4 Technology section).

At the beginning of this chapter, we stated that probability constitutes the major vehicle of statistical inference. In the context of random sampling from a population, the tools of probability enable us to gauge the likelihood of various potential outcomes of the sampling process. Ingrained in our probability calculations lies the artificial assumption that the composition of the population is *known*. The route of statistical inference is exactly in the opposite direction, as depicted in Figure 5. It is the composition of the population that is unknown

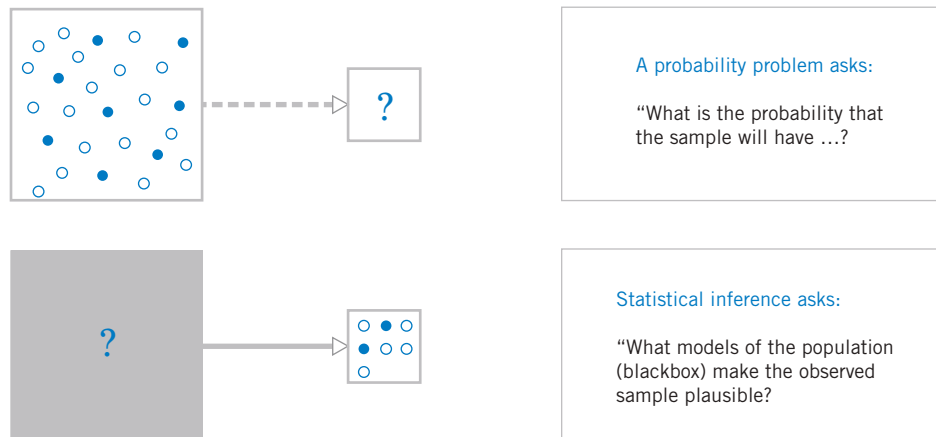


Figure 5 Probability versus statistical inference.

while we have at hand the observations (data) resulting from a random sample. Our object of inference is to ascertain what compositions (or models) of the population are compatible with the observed sample data. We view a model as plausible unless probability calculations based on this model make the sample outcome seem unlikely.

## Exercises

- 4.83 Evaluate:
- (a)  $\binom{6}{3}$     (b)  $\binom{10}{4}$     (c)  $\binom{22}{2}$   
 (d)  $\binom{22}{20}$     (e)  $\binom{30}{3}$     (f)  $\binom{30}{27}$
- 4.84 List all the samples from  $\{a, b, c, d, e\}$  when (a) 2 out of 5 are selected, (b) 3 out of 5 are selected. Count the number of samples in each case.
- 4.85 Of 10 available candidates for membership in a university committee, 6 are men and 4 are women. The committee is to consist of 4 persons.
- (a) How many different selections of the committee are possible?  
 (b) How many selections are possible if the committee must have 2 men and 2 women?
- 4.86 If a coin is tossed 11 times, the outcome can be recorded as an 11-character sequence of  $H$ 's and  $T$ 's according to the results of the successive tosses. In how many ways can there be 4  $H$ 's and 7  $T$ 's? (Put differently, in how many ways can one choose 4 positions out of 11 to put the letter  $H$ ?)
- 4.87 A psychologist will select 5 preschool children from a class of 11 students in order to try out new abuse awareness material.
- (a) How many different selections are possible?  
 (b) Suppose 4 of the 11 children are males. If the 5 selected children were to consist of 2 males and 3 females, how many different selections are possible?
- 4.88 Out of 12 people applying for an assembly job, 3 cannot do the work. Suppose two persons will be hired.
- (a) How many distinct pairs are possible?  
 (b) In how many of the pairs will 0 or 1 person not be able to do the work?  
 (c) If two persons are chosen in a random manner, what is the probability that neither will be able to do the job?
- 4.89 After a preliminary screening, the list of qualified jurors consists of 10 males and 7 females. The 5 jurors the judge selects from this list are all males. Did the selection process seem to discriminate against females? Answer this by computing the probability of having no female members in the jury if the selection is random.
- 4.90 Suppose you participate in a lottery conducted by a local store to give away four prizes. Each customer is allowed to place 2 cards in the barrel. Suppose the barrel contains 5000 cards from which the 4 winning cards will be chosen at random. What is the probability that at least one of your cards will be drawn?
- 4.91 A batch of 20 used automobile alternators contains 4 defectives. If 3 alternators are sampled at random, find the probability of the event
- (a)  $A$  = [None of the defectives appear]  
 (b)  $B$  = [Exactly two defectives appear]
- 4.92 **Ordered sampling versus unordered sampling.** Refer to Exercise 4.91. Suppose the sampling of 3 alternators is done by randomly choosing one after another and without replacement. The event  $A$  can then be described as  $G_1G_2G_3$ , where  $G$  denotes "good" and the suffixes refer to the order of the draws. Use the

method of Example 14 to calculate  $P(A)$  and  $P(B)$ . Verify that you get the same results as in Exercise 4.91.

This illustrates the following fact: To arrive at a random sample, we may randomly draw one object at a time without replacement and then disregard the order of the draws.

4.93 A college senior is selected at random from each state. Next, one senior is selected at random from the group of 50. Does this procedure produce a senior selected at random from those in the United States?

4.94 An instructor will choose 3 problems from a set of 7 containing 3 hard and 4 easy problems. If the selection is made at random, what is the probability that only the hard problems are chosen?

4.95 Nine agricultural plots for an experiment are laid out in a square grid as shown. Three plots are to be selected at random.

(a) Find the probability that all 3 are in the same row.

(b) Find the probability that all 3 are in different rows.

1	2	3
4	5	6
7	8	9

4.96 In one area of an orchard, there are 17 trees, of which 10 are bushy and 7 lean. If 4 trees are randomly selected for testing a new spray, what is the probability that exactly 2 bushy trees are selected?

\*4.97 Referring to Exercise 4.96, now suppose that the trees are located in two rows: Row  $A$  has 8 trees of which 4 are bushy, and row  $B$  has 9 trees of which 6 are bushy. Two trees are to be randomly selected from each row for testing the spray, and the selections are independent for the two rows.

(a) Find the probability that the trees selected in row  $A$  are both bushy and those selected in row  $B$  are both lean.

(b) Find the probability that of the total of 4 trees selected in the manner described above, exactly 2 are bushy.

4.98 Are the following methods of selection likely to produce a random sample of 5 students from your school? Explain.

(a) Pick 5 students throwing flying discs on the mall.

(b) Pick 5 students who are studying in the library on Friday night.

(c) Select 5 students sitting near you in your statistics course.

4.99 An advertisement seeking volunteers for a clinical research draws 11 respondents. Of these respondents, 5 are below age 30 and 6 are over 30. The researcher will randomly select 4 persons to assign to a particular treatment regimen.

(a) How many selections are possible?

(b) What is the probability exactly 3 of the selected persons are below age 30?

\*4.100 Refer to Exercise 4.99, and further suppose that the 5 respondents who are below 30 consist of 2 males and 3 females, whereas those above 30 consist of 4 males and 2 females. Now, the researcher wants to randomly select 2 males and 2 females to be assigned to the treatment regimen. (The random selections from the different sexes are, of course, independent.)

(a) How many selections are possible?

(b) What is the probability that both selected males are over 30 and both selected females are under 30?

4.101 A box of tulip bulbs contains six bulbs that produce yellow flowers and five bulbs that produce red flowers. Four bulbs are to be randomly selected without replacement. Find the probability that:

(a) Exactly two of the selected bulbs produce red flowers.

(b) At least two of the selected bulbs produce red flowers.

(c) All four selected bulbs produce flowers of an identical color.

4.102 A file cabinet has eight student folders arranged alphabetically according to last name. Three files are selected at random.

(a) How many different selections are possible?

(b) Find the probability that the selected folders are all adjacent.

(Hint: Enumerate the selections of adjacent folders.)

## USING STATISTICS WISELY

1. Begin by creating a sample space  $S$  which specifies all possible outcomes for the experiment.
2. Always assign probabilities to events that satisfy the axioms of probability. In the discrete case, the possible outcomes can be arranged in a sequence. The axioms are then automatically satisfied when probability  $P(e)$  is assigned to the elementary outcome  $e$ , where

$$0 \leq P(e) \quad \text{and} \quad \sum_{\text{all } e \text{ in } S} P(e) = 1$$

and then the probability of any event  $A$  is defined as

$$P(A) = \sum_{\text{all } e \text{ in } A} P(e)$$

3. Always use the rules of probability when combining the probabilities of events.
4. Do not confuse independent events with mutually exclusive events. When  $A$  and  $B$  are mutually exclusive, only one of them can occur. Their intersection is empty and so has probability 0.
5. Do not apply probability to  $AB$  according to the special product rule

$$P(AB) = P(A)P(B)$$

unless the conditions for independence hold. Independence may be plausible when the events  $A$  and  $B$  pertain to physically unrelated parts of a large system and there are no common causes that jointly affect the occurrence of both events.

## KEY IDEAS AND FORMULAS

An **experiment** is any process of observing a phenomenon that has variation in its outcomes. Each possible outcome is called an **elementary outcome**, a **simple event**, or an **element of the sample space**. The **sample space** is the collection of all of these outcomes. A **discrete sample space** has outcomes that can be arranged in a, possibly infinite, sequence. In contrast, a **continuous sample space** is an interval of possible outcomes.

A **tree diagram**, with separate sets of branches for each stage of an experiment, can help identify the elementary outcomes.

If an experiment is repeated a large number of times, experimentally we observe that the relative frequency of an event  $A$

$$\frac{\text{Number of times } A \text{ occurs}}{\text{Number of times experiment is performed}}$$

will stabilize at a numerical value. This **long-run stability of relative frequency** motivates us to assign a number  $P(A)$  between 0 and 1 as the probability of the event  $A$ . In the other direction, we can approximate the probability of any event by repeating an experiment many times.

When the **sample space** is **discrete**, **probability** is then expressed as any assignment of non-negative numbers to the elementary outcomes so that probability 1 is assigned to the whole sample space. The probability model of an experiment is described by:

1. The **sample space**, a list or statement of all possible distinct outcomes.
2. Assignment of probabilities to all the elementary outcomes.  $P(e) \geq 0$  and  $\sum P(e) = 1$ , where the sum extends over all  $e$  in  $\mathcal{S}$ .

The **probability of an event**  $A$  is the sum of the probabilities of all the elementary outcomes that are in  $A$ .

$$P(A) = \sum_{\text{all } e \text{ in } A} P(e)$$

A uniform probability model holds when all the elementary outcomes in  $\mathcal{S}$  are equiprobable. With a **uniform probability model**,

$$P(A) = \frac{\text{No. of } e \text{ in } A}{\text{No. of } e \text{ in } \mathcal{S}}$$

In all cases,  $P(A)$ , viewed as the long-run relative frequency of  $A$ , can be approximately determined by repeating the experiment a large number of times.

Elementary outcomes and events can be portrayed in a **Venn diagram**. The event operations **union**, **intersection**, and **complement** can be depicted as well as the result of combining several operations.

The three basic laws of **probability** are

**Law of complement**  $P(A) = 1 - P(\bar{A})$

**Addition law**  $P(A \cup B) = P(A) + P(B) - P(AB)$

**Multiplication law**  $P(AB) = (B)P(A|B)$

These are useful in probability calculations when events are formed with the operations of complement, union, and intersection.

Two events are **incompatible** or **mutually exclusive** if their intersection is empty. In that case we have the **special addition law for incompatible events**

$$P(A \cup B) = P(A) + P(B)$$

The concept of conditional probability is useful to determine how the probability of an event  $A$  must be revised when another event  $B$  has occurred. It forms the basis of the multiplication law of probability and the notion of **independence** of events.

**Conditional probability of  $A$  given  $B$**

$$P(A|B) = \frac{P(AB)}{P(B)}$$

Two events  $A$  and  $B$  are said to be **independent** if  $P(A|B) = P(A)$ . An equivalent condition for independence is that  $P(AB) = P(A)P(B)$ .

### Rule of Total Probability

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$$

### Bayes' Theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})}$$

The notion of **random sampling** is formalized by requiring that all possible samples are equally likely to be selected. The rule of combinations facilitates the calculation of probabilities in the context of random sampling from  $N$  distinct units.

### Rule of Combinations

$$\binom{N}{r} = \frac{N \times (N - 1) \times \dots \times (N - r + 1)}{r \times (r - 1) \times \dots \times 2 \times 1}$$

## TECHNOLOGY

### Generating random digits

#### Minitab

The following commands illustrate the generation of 5 random digits between 1 and 237 inclusive. As with random-digit tables, it is possible to get repeated values. It is prudent to generate a few more digits than you need in order to get enough unique numbers.

**Dialog box:**

**Calc > Random Data > Integer.** Type *CI* in **Store**.

Type 5 in **Generate**, 1 in **Minimum**, and 237 in **Maximum**.

Click **OK**.

#### EXCEL

The following commands illustrate the generation of 5 random digits between 1 and 237 inclusive. As with random-digit tables, it is possible to get repeated values.

Select **Tools**, then **Data Analysis**,<sup>3</sup> and then **Random Number Generation**.

Click **OK**. Type 1 in **Number of Variables**, 5 in **Number of Random Numbers**.

Select *Uniform* for **Distribution**, type 1 for **Between** and 238 after and (238 is 1 larger than the desired limit 237)

Type (any positive number) 743 in **Random Seed**.

Click **OK**.

<sup>3</sup>If **Data Analysis** is not on tools menu, see directions for adding in Chapter 2, Technology EXCEL.

The random numbers appear in the first column of the spreadsheet. You just ignore the decimal part of each entry to obtain random digits between 1 and 237 inclusive.

### TI-84/83 PLUS

The following commands show the generation of 5 random digits between 1 and 237 inclusive. As with random-digit tables, it is possible to get repeated values.

Enter any nonzero number on the Home screen.

Press the **STO** → button. Press the **MATH** button.

Select the **PRB** menu and then select **1: rand**.

From the Home screen press **ENTER**.

Press the **MATH** button. Select the **PRB** menu and then **5: randInt(**.

With **randInt(** on the Home screen, enter 1 and 237 so that the following appears  
`randInt (1,237, 5)`

Press **ENTER** to obtain the 5 random digits.

## 8. REVIEW EXERCISES

- 4.103 Describe the sample space for each of the following experiments.
- The number of different words used in a sentence containing 24 words.
  - The air pressure (psi) in the right front tire of a car.
  - In a survey, 50 students are asked to respond “yes” or “no” to the question “Do you hold at least a part-time job while attending school?” Only the number answering “yes” will be recorded.
  - The time a TV satellite remains in operation.
- 4.104 For the experiments in Exercise 4.103, which sample spaces are discrete and which are continuous?
- 4.105 Identify these events in the corresponding parts of Exercise 4.103.
- More than 22 words.
  - Air pressure less than or equal to 28 psi.
  - At most 25% hold jobs.
  - Less than 500.5 days.
- 4.106 Examine each of these probability assignments and state what makes it improper.
- Concerning tomorrow’s weather,
 
$$P(\text{Rain}) = .4$$

$$P(\text{Cloudy but no rain}) = .4$$

$$P(\text{Sunny}) = .3$$
  - Concerning your passing of the statistics course,
 
$$P(\text{Pass}) = 1.1 \quad P(\text{Fail}) = .1$$
  - Concerning your grades in statistics and economics courses,
 
$$P(A \text{ in statistics}) = .3$$

$$P(A \text{ in economics}) = .7$$

$$P(A\text{'s in both statistics and economics}) = .4$$
- 4.107 A driver is stopped for erratic driving, and the alcohol content of his blood is checked. Specify the sample space and the event  $A =$  [level exceeds legal limit] if the legal limit is .10%.
- 4.108 The Wimbledon men’s tennis championship ends when one player wins three sets.
- How many elementary outcomes end in three sets? In four?
  - \*If the players are evenly matched, what is the probability that the tennis match ends in four sets?



4.109 There are four tickets numbered 1, 2, 3, and 4. Suppose a two-digit number will be formed by first drawing one ticket at random and then drawing a second ticket at random from the remaining three. (For instance, if the first ticket drawn shows 3 and the second shows 1, the number recorded is 31.) List the sample space and determine the following probabilities.

- (a) An even number.
- (b) A number larger than 20.
- (c) A number between 22 and 30.

4.110 To compare two varieties of wheat, say,  $a$  and  $b$ , a field trial will be conducted on four square plots located in two rows and two columns. Each variety will be planted on two of these plots.

1	2
3	4

Plot arrangement

- (a) List all possible assignments for variety  $a$ .
- (b) If the assignments are made completely at random, find the probability that the plots receiving variety  $a$  are:
  - (i) In the same column.
  - (ii) In different rows and different columns.

4.111 Refer to Exercise 4.110. Instead of a completely random choice, suppose a plot is chosen at random from each row and assigned to variety  $a$ . Find the probability that the plots receiving  $a$  are in the same column.

4.112 Chevalier de Méré, a French nobleman of the seventeenth century, reasoned that in a single throw of a fair die,  $P(1) = \frac{1}{6}$ , so in two throws,  $P(1 \text{ appears at least once}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$ . What is wrong with the above reasoning? Use the sample space of Exercise 4.21 to obtain the correct answer.

4.113 A letter is chosen at random from the word "VOLUNTEER."

- (a) What is the probability that it is a vowel?
- (b) What is the probability that it is a  $T$  or  $V$ ?

4.114 Does the uniform model apply to the following observations? Explain.

- (a) Day of week on which the most persons depart by airplane from Chicago.
- (b) Day of week on which the monthly low temperature occurs.
- (c) Day of week on which the maximum amount of ozone is recorded.
- (d) Month of year when a department store has the maximum sales revenues.

4.115 A three-digit number is formed by arranging the digits 1, 2, and 5 in a random order.

- (a) List the sample space.
- (b) Find the probability of getting a number less than 400.
- (c) What is the probability that an even number is obtained?

4.116 A late shopper for Valentine's flowers calls by phone to have a flower wrapped. The store has only 5 roses, of which 3 will open by the next day, and 6 tulips, of which 2 will open by the next day.

- (a) Construct a Venn diagram and show the events  $A = [\text{Rose}]$ , and  $B = [\text{Will open next day}]$ .
- (b) If the store selects one flower at random, find the probability that it will not open by the next day.

4.117 In checking the conditions of a used car, let  $A$  denote the event that the car has a faulty transmission,  $B$  the event that it has faulty brakes, and  $C$  the event that it has a faulty exhaust system. Describe in words what the following events represent:

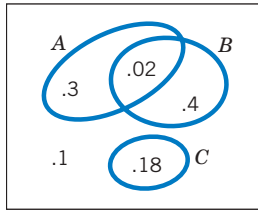
- (a)  $A \cup B$     (b)  $ABC$
- (c)  $\overline{A} \overline{B} \overline{C}$     (d)  $\overline{A} \cup \overline{B}$

4.118 Express the following statements in the notations of the event operations.

- (a)  $A$  occurs and  $B$  does not.
- (b) Neither  $A$  nor  $B$  occurs.
- (c) Exactly one of the events  $A$  and  $B$  occurs.

4.119 Suppose each of the numbers .1, .3, and .5 represents the probability of one of the events  $A$ ,  $AB$ , and  $A \cup B$ . Connect the probabilities to the appropriate events.

4.120 From the probabilities exhibited in this Venn diagram, find  $P(\bar{A})$ ,  $P(AB)$ ,  $P(B \cup C)$ , and  $P(BC)$ .



- 4.121 Using event relations, express the following events in terms of the three events  $A$ ,  $B$ , and  $C$ .
- (a) All three events occur.
  - (b) At least one of the three events occurs.
  - (c)  $A$  and  $B$  occur and  $C$  does not.
  - (d) Only  $B$  occurs.
- 4.122 Concerning three events  $A$ ,  $B$ , and  $C$ , the probabilities of the various intersections are given in the accompanying table. [For instance,  $P(ABC) = .10$ .]

	$B$		$\bar{B}$	
	$C$	$\bar{C}$	$C$	$\bar{C}$
$A$	.05	.10	.05	.17
$\bar{A}$	.20	.15	.18	.10

- (a) Draw a Venn diagram, identify the intersections, and mark the probabilities.

- (b) Determine the probabilities

$$P(AB) \quad P(A\bar{C}) \quad P(C)$$

- (c) Fill in the accompanying probability table concerning the events  $A$  and  $B$ .

	$B$	$\bar{B}$
$A$		
$\bar{A}$		

- 4.123 Referring to Exercise 4.122, calculate the probabilities of the following events.
- (a) Both  $B$  and  $C$  occur.
  - (b) Either  $B$  or  $C$  occurs.
  - (c)  $B$  occurs and  $C$  does not occur.
  - (d) Only one of the three events  $A$ ,  $B$ , and  $C$  occurs.
- 4.124 Concerning three events  $A$ ,  $B$ , and  $C$ , the following probabilities are specified.
- $$P(A) = .51 \quad P(AB) = .17 \quad P(ABC) = .12$$
- $$P(B) = .45 \quad P(BC) = .20$$
- $$P(C) = .50 \quad P(AC) = .33$$

Draw a Venn diagram and determine the probabilities of all the intersections that appear in the diagram. Also, make a probability table like the one given in Exercise 4.122.

- 4.125 Referring to Exercise 4.124, find the probability that:
- (a)  $B$  occurs and  $C$  does not occur.
  - (b) At least one of the events  $A$  and  $B$  occurs.
  - (c) Exactly two of the events  $A$ ,  $B$ , and  $C$  occur.
- 4.126 Suppose a fair die has its even-numbered faces painted red and the odd-numbered faces are white. Consider the experiment of rolling the die once and the events

$$A = [2 \text{ or } 3 \text{ shows up}]$$

$$B = [\text{Red face shows up}]$$

Find the following probabilities:

- (a)  $P(A)$  (b)  $P(B)$  (c)  $P(AB)$   
 (d)  $P(A|B)$  (e)  $P(A \cup B)$

4.127 Given  $P(AB) = .4$  and  $P(B) = .5$ , find  $P(A|B)$ . If, further,  $P(A) = .8$ , are  $A$  and  $B$  independent?

4.128 Suppose three events  $A$ ,  $B$ , and  $C$  are such that  $B$  and  $C$  are mutually exclusive and

$$P(A) = .6 \quad P(B) = .3 \quad P(C) = .25$$

$$P(A|B) = \frac{2}{3} \quad P(\bar{A}C) = .1$$

- (a) Show the events in a Venn diagram.  
 (b) Determine the probabilities of all the intersections and mark them in the Venn diagram.  
 (c) Find the probability that only one of the three events occurs.

4.129 Refer to Exercise 4.128. For each pair of events given below, determine whether or not the events are independent.

- (a)  $A$ ,  $C$   
 (b)  $A\bar{B}$ ,  $C$

4.130 Let  $A$  be the event that a person is a moderate or heavy drinker and  $B$  be the event that the person is female. For a person selected at random in the United States, the probabilities are<sup>4</sup>

$$P(B) = .50 \quad P(A|B) = .12 \quad P(A|\bar{B}) = .29$$

- (a) Express in words, in the context of this problem, the third probability statement.  
 (b) Determine the probability that the person selected is a moderate or heavy drinker.  
 (c) If the person selected is found to be a moderate or heavy drinker, what is the probability of being female?

4.131 Refer to the probability table given in Exercise 4.122 concerning three events  $A$ ,  $B$ , and  $C$ .

- (a) Find the conditional probability of  $A$  given that  $B$  does not occur.  
 (b) Find the conditional probability of  $B$  given that both  $A$  and  $C$  occur.  
 (c) Determine whether or not the events  $A$  and  $C$  are independent.

4.132 Mr. Hope, a character apprehended by Sherlock Holmes, was driven by revenge to commit two murders. He presented two seemingly identical pills, one containing a deadly poison, to an adversary who selected one while Mr. Hope took the other. The entire procedure was then to be repeated with the second victim. Mr. Hope felt that Providence would protect him, but what is the probability of the success of his endeavor?

4.133 A bowl contains 15 marbles, of which 10 are numbered 1 and 5 are numbered 2. Two marbles are to be randomly drawn from the bowl one after another and without replacement, and a two-digit number will be recorded according to the results. (For instance, if the first marble drawn shows 2 and the second shows 1, the number recorded is 21.)

- (a) List the sample space and determine the probability of each outcome.  
 (b) Find the probability of getting an even number.  
 (c) Find the probability that the number is larger than 15.

4.134 Three production lines contribute to the total pool of a company's product. Line 1 provides 20% to the pool and 10% of its products are defective; Line 2 provides 50% to the pool and 5% of its products are defective; Line 3 contributes 30% to the pool and 6% of its products are defective.

- (a) What percent of the items in the pool are defective?  
 (b) Suppose an item is randomly selected from the pool and found to be defective. What is the probability that it came from Line 1?

4.135 In an optical sensory experiment, a subject shows a fast response ( $F$ ), a delayed response ( $D$ ), or no response at all ( $N$ ). The experiment will be performed on two subjects.

- (a) Using a tree diagram, list the sample space.  
 (b) Suppose, for each subject,  $P(F) = .4$ ,  $P(D) = .3$ ,  $P(N) = .3$ , and the responses of different subjects are independent.

<sup>4</sup>National Center for Health Statistics, *Health Behavior of Adults, United States* (September 2006).

- (i) Assign probabilities to the elementary outcomes.
- (ii) Find the probability that at least one of the subjects shows a fast response.
- (iii) Find the probability that both of the subjects respond.

4.136 Four upper level undergraduate students are available to serve on a committee.

Student	Gender	Year in school
1	M	Junior
2	M	Senior
3	F	Junior
4	F	Senior

Two students will be selected at random to serve on the committee. Let

$A$ : The students selected are of the same gender.

$B$ : The students selected are the same year in school.

- (a) Make a Venn diagram showing the outcomes and the two events.
- (b) Find the probability of  $A \cup B$ .
- (c) Are  $A$  and  $B$  independent? Explain why or why not.
- (d) Find the probability of  $\overline{AB}$ .

4.137 A local moving company owns 11 trucks. Three are randomly selected for compliance with emission standards and all are found to be noncompliant. The company argues that these are the only three which do not meet the standards. Calculate the probability that, if only three are noncompliant, all three would be in the sample. Comment on the veracity of the company's claim.

4.138 In all of William Shakespeare's works, he used 884,647 different words. Of these, 14,376 appeared only once and 4343 appeared twice. If one word is randomly selected from a list of these 884,647 different words:

- (a) What is the probability that the selected word appears only once?

- (b) What is the probability that the selected word appears exactly twice?
- (c) What is the probability that the selected word appears more than twice?
- (d) Suppose that, instead of randomly selecting a word from the list of different words, you randomly select a word from a book of Shakespeare's complete works by selecting a page, line, and word number from a random-digit table. Is the probability of selecting a word that appears exactly once larger, smaller, or the same as your answer to part (a)?

4.139 An IRS agent receives a batch of 18 tax returns that were flagged by computer for possible tax evasions. Suppose, unknown to the agent, 7 of these returns have illegal deductions and the other 11 are in good standing. If the agent randomly selects 4 of these returns for audit, what is the probability that:

- (a) None of the returns that contain illegal deductions are selected?
- (b) At least 2 have illegal deductions?

\*4.140 *Polya's urn scheme.* An urn contains 4 red and 6 green balls. One ball is drawn at random and its color is observed. The ball is then returned to the urn, and 3 new balls of the same color are added to the urn. A second ball is then randomly drawn from the urn that now contains 13 balls.

- (a) List all outcomes of this experiment (use symbols such as  $R_1G_2$  to denote the outcome of the first ball red and the second green).
- (b) What is the probability that the first ball drawn is green?
- (c) What is the conditional probability of getting a red ball in the second draw given that a green ball appears in the first?
- (d) What is the (unconditional) probability of getting a green ball in the second draw?

\*4.141 *Birthdays.* It is somewhat surprising to learn the probability that 2 persons in a class share the same birthday. As an approximation,

assume that the 365 days are equally likely birthdays.

- (a) What is the probability that, among 3 persons, at least 2 have the same birthday? (*Hint:* The reasoning associated with a tree diagram shows that there are  $365 \times 365 \times 365$  possible birthday outcomes. Of these,  $365 \times 364 \times 363$  correspond to no common birthday.)
- (b) Generalize the above reasoning to  $N$  persons. Show that

$P[\text{No common birthday}] =$

$$\frac{365 \times 364 \times \cdots \times (365 - N + 1)}{(365)^N}$$

Some numerical values are

$N$	5	9	18	22	23
$P[\text{No common birthday}]$	.973	.905	.653	.524	.493

We see that with  $N = 23$  persons, the probability is greater than  $\frac{1}{2}$  that at least two share a common birthday.)